Jastin Pompeu Soares

Explorar diferentes estratégias de *data mining* aplicadas a dois problemas no pré-processamento de dados

Dissertação de Mestrado

Julho de 2017

· U C ·

Universidade de Coimbra

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Explorar diferentes estratégias de *data mining* aplicadas a dois problemas no pré-processamento de dados

Jastin Pompeu Soares

Coimbra, Julho 2017

Explorar diferentes estratégias de *data mining* aplicadas a dois problemas no pré-processamento de dados

Orientadores:

Professor Doutor Pedro Henriques Abreu

Professor Doutor Hélder de Jesus Araújo

Júri:

Professora Doutora Bernardete Martins Ribeiro

Professor Doutor Paulo José Monteiro Peixoto

Professor Doutor Pedro Henriques Abreu

Dissertação submetida ao Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra em cumprimento parcial dos requisitos para o grau de Mestre.

Coimbra, Julho 2017

Agradecimentos

É costume dizer-se que “Roma não foi construída num dia!” e eu neste momento passo a acrescentar “Muito menos sozinha”. Como tal gostaria de agradecer a todos os intervenientes que de alguma forma fizeram parte deste puzzle, a realização desta dissertação, e em especial:

Ao Professor Doutor Pedro Henriques Abreu, orientador da dissertação no Departamento de Engenharia Informática, agradeço o excelente desafio que me propôs, o apoio que solicitou em todas as etapas deste trabalho e o grande contributo prestado no enriquecimento da minha formação académica.

Ao Professor Doutor Hélder de Jesus Araújo, orientador da dissertação no Departamento de Engenharia Electrotécnica e de Computadores, agradeço o apoio e a oportunidade que me deu em realizar este trabalho noutro departamento.

À Miriam Seoane Santos, aluna de doutoramento, agradeço o seu contributo, a partilha de saber, o apoio e toda a persistência e perseverança necessária para tornar este trabalho possível.

Aos meus amigos e colegas de curso, que ouviram todos os meus dilemas e dramas ao longo dos últimos cincos anos, agradeço todos os momentos de cumplicidade e a ajuda que me proporcionaram.

Por fim, à mais importante peça do puzzle, a Família, o maior “Obrigado!”, pois mesmo tendo se tornada incompleta ao longo deste meu percurso académico, aumentou ainda mais a minha força e vontade em tornar o meu sonho e o dele real – “Os três obterem o seu diploma e profissão.”.

Resumo

Com o aumento de volumes de dados, melhorias tecnológicas, e a necessidade crescente em extrair conhecimento de dados, as técnicas de *Machine Learning* têm sido alvo de grande estudo, focando-se as principais contribuições no desenvolvimento e melhoria dos seus algoritmos. Nesse contexto, a qualidade dos dados é um ponto crucial na obtenção de bons resultados. Incluído na análise de dados, o pré-processamento é uma das etapas da extração de conhecimentos que possibilita a melhoria da qualidade dos dados. Esta dissertação visa contribuir em dois problemas que podem surgir na fase de pré-processamento: dados incompletos e dados não balanceados.

Para resolver o primeiro problema, os investigadores usam tipicamente estratégias *brute-force* que, para além do seu elevado custo computacional, não têm em consideração a natureza dos dados e, portanto, não possibilitam a sua generalização para diferentes contextos. Neste trabalho é explorada a relação entre o desempenho das técnicas de imputação estado-da-arte e a distribuição dos dados, procurando desenvolver uma heurística que permita escolher a técnica de imputação mais apropriada para cada variável incluída no estudo, evitando a necessidade de testar várias técnicas. Os resultados mostram que existe uma relação entre a distribuição das variáveis e o desempenho dos algoritmos. Este desempenho parece ser influenciado pela estratégia e taxa de geração dos dados em falta.

No segundo problema pretende-se medir o desempenho dos classificadores em contextos de dados não balanceados. A abordagem utilizada para proceder à validação cruzada (antes ou depois do pré-processamento) pode levar a desempenhos sobre-otimistas, aquando da aplicação de técnicas de sobre-amostragem para atenuar a diferença entre classes. Este trabalho visa mostrar qual a abordagem mais correta na validação cruzada e relacionar o motivo do sobre-otimismo com a complexidade dos *datasets*. Os resultados demostram que a abordagem de validação cruzada mais adequada é aquela onde a divisão do *dataset* é efetuada antes do pré-processamento, e o sobre-otimismo aparenta estar relacionado com a semelhança na complexidade dos conjuntos de treino e teste.

Palavras-Chave

Dados incompletos; Imputação; Distribuição de dados; Dados não balanceados; Sobre-amostragem; Complexidade.

Abstract

With increasing volumes of data, technological improvements, and the need to extract knowledge from data, Machine Learning techniques have been subjected to great study, where the main contributions are currently focused in the development and improvement of algorithms. In this context, data quality is a crucial point to achieve good results. Included in data analysis, preprocessing is one of the stages of knowledge-discovery in databases that enables the improvement of data quality. This dissertation aims to contribute to two problems that may arise in the preprocessing stage: Missing Data and Imbalanced Data.

To solve the first problem, researchers typically use brute-force strategies that, in addition to their high computational cost, do not take into account the nature of the data and therefore do not allow their generalization to different contexts. In this work, the relationship between the performance of the state-of-art imputation techniques and the data distribution is explored, by trying to develop a heuristic that allows choosing the most appropriate imputation technique for each feature included in the study, to avoid the need of testing several techniques. The results show that there is a relationship between the features' distributions and the imputation performance. This performance seems to be influenced by the strategy and rate of the missing data generation.

In the second problem, the intention is to measure the performance of classifiers in imbalanced data contexts. The approach used to perform cross-validation (before or after pre-processing) can lead to over-optimistic performances when applying oversampling techniques to attenuate the between-class imbalance. This work aims to show the most correct approach of cross-validation and to relate the over-optimistic performance with the datasets' complexity. The results show that the most appropriate cross-validation approach is the one where the dataset splitting is performed before the pre-processing stage, and over-optimistic performances seem to be related to the similarity of the complexity of training and test sets.

Keywords

Missing Data; Imputation; Data Distribution; Imbalance Data; Oversampling; Complexity.

Conteúdo

Lista de Acrónimos

ACC Exatidão da Classificação (*Classification Accuracy*)

ADASYN *Adaptive Synthetic Sampling Approach*

ADOMS *Adjusting the Direction Of the synthetic Minority clasS*

ADTree *Alternating Decision Tree*

AHC *Agglomerative Hierarchical Clustering*

ANN Redes Neuronais Artificiais (*Artificial Neural Networks*)

AUC Área abaixo da Curva ROC (*Area Under the ROC Curve*)

BAN *Boosted Augmented Naive Bayes*

BPCA *Bayesian PCA*

CART *Classification and Regression Trees*

CBO *Cluster-Based Oversampling*

CDF Função Distribuição Acumulada (*Cumulative Distribution Function*)

CE Erro de Classificação (*Classification Error*)

DAC Exatidão Distribucional (*Distributional Accuracy*)

DT Árvores de Decisão (*Decision Tree*)

ECM *Expectation Conditional Maximization*

EM *Expectation–Maximization*

ENN *Edited Nearest Neighbor*

FIMUS *Framework for Imputing Missing Values Using Co-Appearance, Correlation and Similarity Analysis*

FURIA *Fuzzy Unordered Rule Induction Algorithm*

GBN *General Bayes Network Classifiers*

GoF *Goodness of Fit*

HEOM Métrica Heterogénea de Sobreposição Euclidiana (*Heterogeneous Euclidean-Overlap Metric*)

ID3 *Iterative Dichotomiser 3*

IDE *Input Decimated Ensemble*

IR Taxa de Desequilíbrio (*Imbalance Ratio*)

k-NN *k-Nearest Neighbors*

KDD Exploração de Conhecimento (*Knowledge Discovery in Databases*)

KEEL *Knowledge Extraction based on Evolutionary Learning*

KMC *K-Means Clustering*

LLR *Locally Linear Regression*

LLS *Local Least Squares*

LR *Logistic Regression*

MAE Erro Absoluto Médio (*Mean Absolute Error*)

MAR Falta de Dados Aleatória (*Missing At Random*)

MCAR Falta de Dados Completamente Aleatória (*Missing Completely At Random*)

MI Imputação Múltipla (*Multiple Imputation*)

MLP *Multi-Layer Perceptron*

MM Média/Moda

MoG *Mixture of Gaussians*

MR Taxa de Valores em Falta (*Missing Rate*)

MSE Erro Quadrático Médio (*Mean Squared Error*)

MWMOTE *Majority Weighted Minority Over-sampling Technique*

NB *Naive Bayes*

NMAR Falta de Dados Não Aleatória (*Not Missing At Random*)

NRMSE Raiz Normalizada do Erro Médio Quadrático (*Normalized Root Mean Square Error*)

PAC Exatidão Preditiva (*Predictive Accuracy*)

PCA Análise de Componentes Principais (*Principal Component Analysis*)

PDF Função Densidade de Probabilidade (*Probability Density Function*)

RBF Função de Base Radial (*Radial Basis Function*)

RDR *Ripple-Down Rules*

REPTree *Reduced Error-Prunning Tree*

RF *RandomForest*

RMSE Raiz do Erro Médio Quadrático (*Root Mean Square Error*)

ROC Característica de Operação do Receptor (*Receiver Operating Characteristic*)

ROS *Random Over-Sampling*

RUS *Random Under-Sampling*

SEN Sensibilidade

SMOTE *Synthetic Minority Over-sampling Technique*

SOM Mapa Auto-Organizável (*Self-Organizing Map*)

SPEC Especificidade (*Specificity*)

SPECT Tomografia Computorizada por Emissão de Fotão Único (*Single Proton Emission Computed Tomography*)

SPIDER *Selective Preprocessing of Imbalanced Data*

SVD Decomposição em Valores Singulares (*Singular Value Decomposition*)

SVM Máquinas de Vetores de Suporte (*Support Vector Machine*)

TAN *Tree Augmented Naive Bayes*

TL *Tomek Links*

VPP Valor Preditivo Positivo

WEKA *Waikato Environment for Knowledge Analysis*

Lista de Figuras

Lista de Tabelas

Lista de Listagens

1 Introdução

Este trabalho foi elaborado no Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologias da Universidade de Coimbra. Neste capítulo é efetuada uma breve apresentação das temáticas, dos objetivos e da estrutura desta dissertação.

1.1 Contextualização

Antes do aparecimento das tecnologias e sistemas de gestão de dados, a informação era armazenada em papel e a extração de conhecimento de um conjunto de dados era efetuada de forma manual, o que tornava esse processo demorado, complexo e pouco eficiente. O começo do armazenamento de dados em formato digital e a necessidade de obter informações a partir deles, despertou um grande interesse na comunidade científica, na década de 70 [1]. No final dos anos 80, Gregory Piatetsky-Shapiro na *Internacional Joint Conference on Artificial Intelligence* introduziu pela primeira vez a expressão que atualmente delimita uma área das ciências da computação, a Exploração de Conhecimento (KDD)[2]. Fayyad, Piatetsky-Shapiro e Smyth [3] definem KDD como um processo não trivial de obtenção de nova informação, válido, compreensível e útil, que se decompõe em cinco etapas distintas (Figura 1.1):

1. Seleção;
2. Pré-processamento;
3. Transformação;
4. *Data Mining*;
5. Avaliação e Interpretação.

A primeira etapa consiste na identificação e definição do conhecimento a obter, e na contextualização e recolha dos dados necessários. O pré-processamento resume-se à definição de estratégias capazes de efetuar a limpeza dos dados (remoção de ruído, tratamento de dados incompletos, entre outras). Na terceira etapa realiza-se a redução da dimensionalidade do

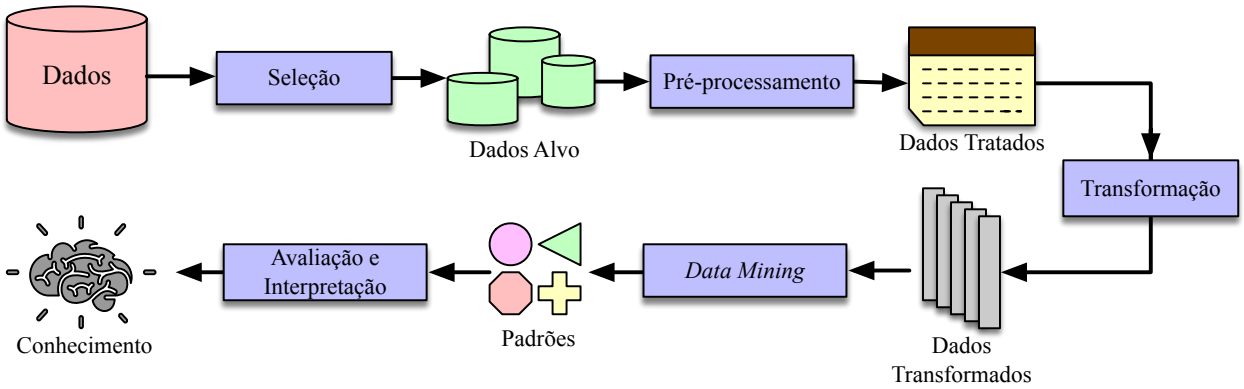


Figura 1.1: Esquema dos passos que compões o processo de KDD. Adaptado da fonte [3].

problema, quer através da selecção das características mais discriminativas, quer através da sua transformação (e.g. projecção). O *data mining* é a quarta etapa do processo, onde se efetua o reconhecimento de padrões, de modo a construir um modelo de classificação ou regressão. Por fim, a última etapa é onde se realiza a interpretação e avaliação do modelo obtido na etapa anterior, de maneira a atingir o conhecimento pretendido.

O foco deste trabalho incide sobre dois problemas que podem ocorrer na fase de pré-processamento: a existência de dados incompletos e dados não balanceados. Estes dois problemas comprometem o desempenho dos algoritmos de *Machine Learning* e a fiabilidade dos seus resultados.

No caso dos dados em falta, existem estratégias capazes de lidar com os dados incompletos na fase de pré-processamento. Algumas dessas estratégias fazem a remoção das instâncias incompletas, o que pode levar à exclusão de informação útil para a exploração de conhecimento. Outras estratégias fazem o preenchimento das instâncias incompletas com base nas restantes instâncias, que apesar de poderem ser estimativas plausíveis, são valores sintéticos, ou seja, não existe uma forma efectiva de saber se se assemelham aos reais que estão em falta.

No caso dos dados não balanceados, existem estratégias para lidar com o problema, tais como: a remoção de dados da classe majoritária (contudo o uso desta estratégia pode levar à exclusão de informação que eventualmente podia ser relevante); a introdução de réplicas de exemplos da classe minoritária (no entanto aumenta a probabilidade de ocorrer *overfitting*) e a criação de dados sintéticos na classe minoritária (porém esses dados podem não ter um significado real).

1.2 Objetivos e Questões de Investigação

Como referido anteriormente, as contribuições deste trabalho irão incidir sobre dois problemas que podem ocorrer no pré-processamento: como lidar com dados incompletos e dados não balanceados.

Assim, os objetivos principais do trabalho são:

- Estudar o efeito da distribuição dos dados no desempenho de várias técnicas de imputação, considerando vários cenários diferentes (e.g. percentagem de dados em falta, método de geração dos valores em falta);
- Avaliar o impacto de diferentes estratégias de validação cruzada no sobre-otimismo dos modelos de classificação, quando são usadas estratégias de sobre-amostragem;

De forma a atingir os objetivos identificados, várias questões de investigação necessitam de ser respondidas. Para o problema de dados incompletos, definiram-se as seguintes:

- Qual a influência da distribuição dos dados no desempenho dos algoritmos de imputação estado-da-arte?
- Quais os critérios para avaliação do processo de imputação como objetivo final?
- Como se comportam os algoritmos de imputação face a contextos diferentes (e.g. número de variáveis, percentagem de dados em falta, etc)?

Para o problema de dados não balanceados, definiram-se as seguintes:

- Qual a forma mais correta para avaliar os algoritmos de sobre-amostragem? Validação cruzada antes ou depois do pré-processamento?
- Como influenciam as diferentes técnicas de sobre-amostragem a complexidade dos *datasets*?
- Será possível inferir o desempenho da classificação através da complexidade dos *datasets*?

1.3 Contribuições Científicas

O trabalho desenvolvido nesta dissertação resultou nas seguintes publicações:

- Miriam Seoane Santos, Jastin Pompeu Soares, Pedro Henriques Abreu, Hélder Araújo e João Santos. "Influence of data distribution in missing data imputation". Em: *16th Conference on Artificial Intelligence in Medicine (AIME)*. 21 – 24 de Junho de 2017 (Publicado)

- Jastin Pompeu Soares, Miriam Seoane Santos, Pedro Henriques Abreu, Hélder Araújo e João Santos. "Exploring the Effects of Data Distribution in Missing Data Imputation". Em: *IEEE Computational Intelligence Magazine*. Submetido a 8 de Junho de 2017 (Em fase de revisão)

- Jastin Pompeu Soares, Miriam Seoane Santos, Pedro Henriques Abreu, Hélder Araújo e João Santos. "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches". (Em fase de submissão a revista científica)

1.4 Estrutura do Documento

O restante desta dissertação está organizada da seguinte forma: no Capítulo 2, são apresentados os fundamentos teóricos enquadrados nos dois problemas a tratar, no Capítulo 3, uma descrição detalhada sobre os trabalhos científicos publicados nas áreas de dados em falta e dados não balanceados é realizada, no Capítulo 4 é exposta a metodologia e a discussão dos resultados do estudo relacionado com a imputação dos dados. De forma análoga, no Capítulo 5 é descrita a metodologia utilizada e os resultados alcançados para o problema de dados não balanceados e por fim, no Capítulo 6 são apresentadas as conclusões finais e propostas para trabalhos futuros.

2 Fundamentos Teóricos

Neste capítulo estão descritos os fundamentos teóricos dos dois problemas tratados neste trabalho.

2.1 Tratamento de Dados Incompletos

Um conjunto de dados é considerado incompleto quando existem padrões cujos valores estão em falta (numa ou mais variáveis). Ao longo desta secção são descritas as noções teóricas inerentes ao problema de dados incompletos.

2.1.1 Mecanismos de Dados Incompletos

A falta de dados num *dataset* pode ser causada por uma grande variedade de razões: um médico por lapso não regista algumas observações, uma balança defeituosa não efetua o envio de algumas medidas; um cliente que não responde a algumas das questões colocadas num questionário; um jovem que participa num estudo não comparece numa das suas consultas. Apesar de todas estas situações parecerem muito semelhantes a olho nu, para cada um destes casos podem existir diferentes processos responsáveis pela geração de dados incompletos. Estes processos são denominados de mecanismos de dados incompletos, introduzidos por Rubin [4] no final dos anos 80, e são divididos em três categorias: Falta de Dados Completamente Aleatória (MCAR), Falta de Dados Aleatória (MAR) e Falta de Dados Não Aleatória (NMAR).

2.1.1.1 Falta de Dados Completamente Aleatória

Os dados incompletos são considerados MCAR, quando a probabilidade de ocorrência das faltas é independente dos dados observados. Por outras palavras, se um valor estiver em falta e a sua causa não se relaciona com os valores das restantes variáveis, então diz-se que é MCAR. Tomemos como exemplo um estudo levado a cabo por uma empresa, com o objetivo

de relacionar o salário mensal de um dado cliente com o consumo dos serviços da empresa. Se um cliente por distração não responder a uma pergunta de um questionário desta empresa ("qual é o seu salário?"), então os dados em falta são considerados MCAR, uma vez que se trata de uma falta que podia ter ocorrido a qualquer cliente, com alto ou baixo salário ou consumo de serviço.

2.1.1.2 Falta de Dados Aleatória

Os dados em falta são classificados como MAR, quando a probabilidade de ocorrência dessas faltas é dependente dos valores de outras variáveis, mas independente dos valores que estariam presentes nesses dados em falta caso eles tivessem sido observados. De forma a clarificar esta definição suponhamos o mesmo cenário dado para MCAR, mas com a diferença de que, desta vez, o cliente não responde à pergunta "qual é o seu salário?", porque o seu consumo de serviços é "baixo". Neste caso, o valor em falta produzido na variável "salário" está relacionada com a variável "consumo". Pressupondo que a probabilidade de um valor estar em falta na variável "salário" não está relacionada com os valores do "salário" em si (por exemplo, o salário médio dos clientes com consumo baixo é semelhante ao dos clientes com consumo alto), então estamos perante o mecanismo MAR, mas não NMAR.

2.1.1.3 Falta de Dados Não Aleatória

Os dados incompletos são denominados NMAR, quando a probabilidade da ocorrência dessas faltas é dependente dos dados observados (noutras variáveis no estudo) e dos valores observados na variável em falta. Reutilizando o exemplo anterior, e supondo que na maioria dos casos, clientes com consumos baixos apresentam salários baixos: um cliente responde que tem consumos de serviços "baixo", mas recusa-se a dizer o seu salário por ser "baixo". Neste exemplo, estamos perante uma situação de NMAR.

2.1.2 Técnicas de Imputação de Dados

A imputação consiste na substituição de dados incompletos por valores calculados através dos dados observados. Nesta secção é apresentada uma breve explicação das técnicas de imputação mais frequentemente utilizadas em trabalhos relacionados.

2.1.2.1 Medidas de Posição Estatística

A imputação baseada na Média/Moda (MM) consiste na atribuição do valor médio dos valores existentes, caso a variável seja contínua, e da moda, em caso contrário, nos dados incompletos [5]. O cálculo da média (ou moda) pode ser efetuado através do uso de todas as instâncias observadas ou de todas as instâncias agrupadas por classe (média/moda condicionada) [6].

2.1.2.2 Árvores de Decisão

As Árvores de Decisão (DT) são algoritmos de *data mining* que efetuam, de modo recursivo, a divisão dos dados em nós de uma árvore desde a sua raiz até às folhas [7]. Cada nó corresponde a um teste que se realiza a uma variável e cada ramo corresponde a um valor (ou intervalo de valores) que essa variável assume. As folhas contêm uma classe associada, são o elemento final da árvore e definem a decisão a tomar para uma dada instância. Uma árvore de decisão constrói-se a partir de um conjunto inicial de dados (treino ou obtenção do modelo). Após o treino, é possível realizar atribuição de classes a dados novos (de teste) utilizando o modelo previamente obtido. Na imputação de dados com Árvores de Decisão (DT), cada variável incompleta é definida como a decisão a tomar (*target* ou classe) e as restantes variáveis são usadas para construir o modelo [8, 9].

2.1.2.3 *k-Nearest Neighbors*

O *k-Nearest Neighbors* (k-NN) é um algoritmo que se baseia na informação dos k vizinhos mais próximos de um padrão para efetuar a sua classificação e pode ser adaptado para efetuar o preenchimento de dados incompletos [5]. O valor a preencher pode ser a moda, caso a variável seja discreta e média ou média ponderada, caso a variável seja contínua. O uso da média ponderada permite atribuir pesos a cada vizinho atendendo à sua distância do padrão incompleto. Esta técnica têm um tempo de execução elevado, causado pela necessidade de cálculo das distâncias entre padrões. Além disso, este algoritmo apresenta dois requisitos, o conhecimento do número de vizinhos a procurar (k) e a escolha da distância entre padrões mais apropriada.

Existem várias formas de efetuar o cálculo das distâncias mas nem todas têm a capacidade de lidar simultaneamente com variáveis contínuas, nominais e ainda valores em falta. A Métrica Heterogénea de Sobreposição Euclidiana (HEOM), introduzida por Wilson e Martinez [10] tem sido muito utilizada pela sua capacidade de lidar com todas estas condicionantes

[5, 6, 11, 12]. Considerando dois padrões $\mathbf{x}_A$ e $\mathbf{x}_B$, a distância HEOM é obtida por:

$$d(\mathbf{x}_A, \mathbf{x}_B) = \sqrt{\sum_{j=1}^{n} d_j(x_{Aj}, x_{Bj})^2} \tag{2.1}$$

onde $d_j(x_{Aj}, x_{Bj})$ é a distância entre dois padrões na j-ésima variável, e corresponde a:

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 1 & , \text{ se faltar } x_j \text{ em } \mathbf{x}_A \text{ ou } \mathbf{x}_B; \\ d_O(x_{Aj}, x_{Bj}) & , \text{ se } x_j \text{ for nominal;} \\ d_N(x_{Aj}, x_{Bj}) & , \text{ se } x_j \text{ for contínuo.} \end{cases} \tag{2.2}$$

Na equação 2.2, a distância varia entre $[0, 1]$. No caso de um dos valores da j-ésima variável (x_{Aj} ou x_{Bj}) ser desconhecido, a distância é 1. Se ambos os valores são conhecidos, utiliza-se a métrica de sobreposição (*Overlap*), d_O, para as variáveis nominais (Equação 2.3), e a distância euclidiana normalizada, d_N, para as variáveis contínuas (Equação 2.4).

$$d_O(x_{Aj}, x_{Bj}) = \begin{cases} 0 & , \text{ se } x_{Aj} = x_{Bj}; \\ 1 & , \text{ caso contrário.} \end{cases} \tag{2.3}$$

$$d_N(x_{Aj}, x_{Bj}) = \frac{|x_{Aj} - x_{Bj}|}{\max(x_j) - \min(x_j)} \tag{2.4}$$

2.1.2.4 Mapa Auto-Organizável

O Mapa Auto-Organizável (SOM) é um algoritmo baseado em redes neuronais, que faz o mapeamento de um conjunto de dados de dimensão superior (várias variáveis) num mapa (matriz) de dimensão inferior, normalmente bidimensional [13] e pode ser adaptada para imputar dados de um *dataset* incompleto [11, 14]. Os dados de entrada são mapeados numa grelha de neurónios, designados de nós, onde cada um é associado a um vetor de pesos com a mesma dimensão que o espaço de variáveis. O Mapa Auto-Organizável (SOM) tem duas fases:

- Treino – Nesta fase é construído o mapa com os dados de entrada completos, através da colocação de cada padrão ao lado do nó com o vetor de pesos mais semelhante, também conhecido como a *Best Matching Unit*. Os pesos dos nós adjacentes à *Best Matching Unit* são atualizados em cada iteração, cada vez que um novo padrão de treino é apresentado ao mapa.
- Mapeamento – Nesta fase é efetuada a imputação de dados incompletos, consoante o mapa construído na fase anterior, procurando o nó mais semelhante a cada padrão incompleto. De modo a realizar a imputação, as variáveis com valores em falta são

ignoradas [14]. Após isso, seleciona-se o nó com o vetor de pesos mais próximo das variáveis completas de cada padrão incompleto. De seguida, obtém-se o grupo de ativação composto pelo nó vencedor e seus vizinhos, e por fim, os dados imputados são obtidos através do cálculo dos pesos do grupo de ativação de nós na dimensão correspondente aos valores em falta (Figura 2.1).

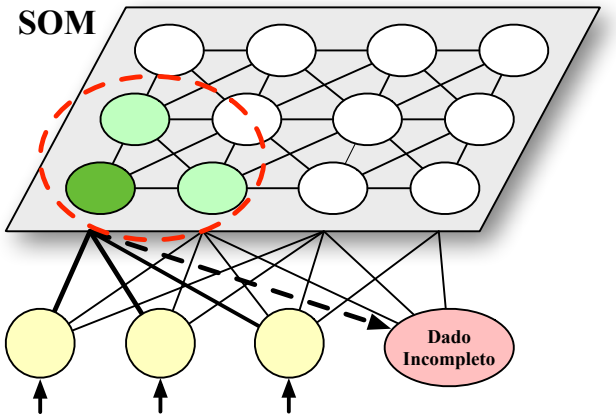


Figura 2.1: Modelo de SOM para imputação. Adaptado da fonte [14].

2.1.2.5 Máquinas de Vetores de Suporte

O conceito de Máquinas de Vetores de Suporte (SVM) foi introduzido nos anos 60 mas a implementação do algoritmo, como classificador não linear através da aplicação de funções *Kernel*, só surgiu no final do século XX [15–17]. Atualmente, no reconhecimento de padrões, esta técnica é o estado-da-arte, pois têm a capacidade de operar em problemas de classificação de elevado grau de complexidade [5, 18–20]. Esta técnica tenta encontrar o hiperplano ótimo que permita maximizar a separação de dados de classes distintas [15]. A margem de separação dos dados define a fronteira de decisão (hiperplano ótimo) e os padrões que estão mais próximos dela são os vetores de suporte. O SVM também pode ser utilizado como método de imputação, em que cada variável incompleta é definida como *target*, enquanto as restantes são consideradas para treino. Este método apresenta uma fase de treino, onde são utilizados os dados completos com o intuito de determinar os parâmetros ótimos do modelo. Por fim, os dados incompletos são imputados seguindo o modelo com os parâmetros obtidos.

2.1.3 Métricas de Avaliação

Em problemas de dados incompletos é frequente avaliar uma técnica de imputação com o Erro de Classificação (CE), ou seja, o melhor método de imputação é aquele que reduz o erro

de classificação. O uso do CE é, no entanto, controverso, no sentido em que a imputação que minimiza o erro pode produzir estimativas enviesadas que afetem a distribuição dos dados originais, especialmente se for aplicado o mesmo método em dados que seguem diferentes distribuições [21]. A comparação entre os dados originais e imputados é, portanto, uma forma mais adequada de avaliar a qualidade da imputação e pode ser efetuada utilizando várias métricas. Chambers [22] descreve cinco propriedades importantes para a avaliação do método de imputação. Devido à natureza dos dados, estas propriedades nem sempre são possíveis de analisar. Na primeira parte deste trabalho, as variáveis são contínuas e, como tal, os critérios relevantes são a Exatidão Preditiva (PAC) e a Exatidão Distribucional (DAC). A análise com estas abordagens é somente realizável aquando da existência dos dados originais completos. Estas duas abordagens serão expostas nas secções 2.1.3.1 e 2.1.3.2. Nas secções 2.1.3.3-2.1.3.5 são explicadas de forma breve as três métricas de avaliação das técnicas de imputação utilizadas neste trabalho.

2.1.3.1 Exatidão Preditiva

Aquando da utilização da técnica de imputação de dados pretende-se ter uma boa capacidade de reprodução dos valores **verdadeiros**. A propriedade que objetiva a preservação dos valores originais é denominada Exatidão Preditiva (PAC) [22]. A Figura 2.2 é um exemplo visual que demostra o que esta propriedade avalia: a distância entre os valores originais e os imputados.

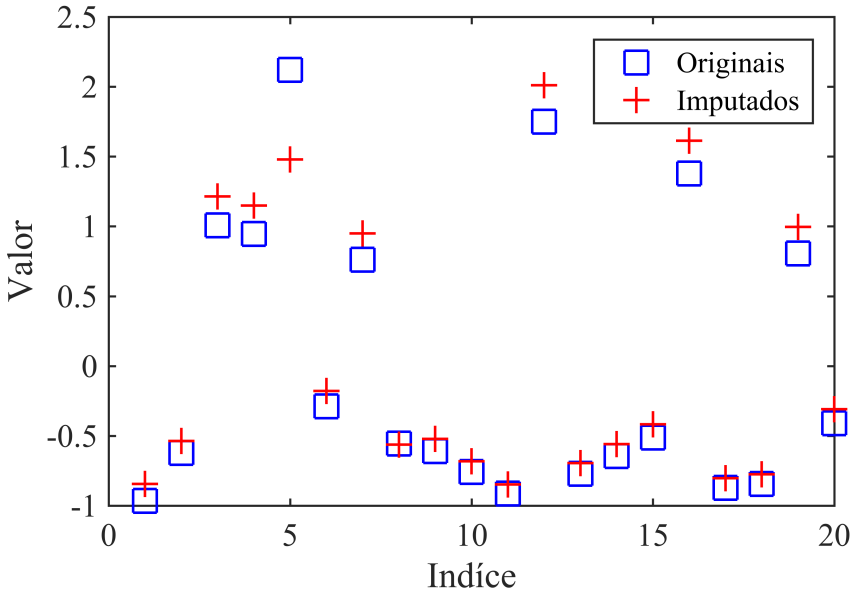


Figura 2.2: Exemplo de PAC

2.1.3.2 Exatidão Distribucional

Um bom método de imputação deve ter como objetivo preservar a distribuição dos dados originais. O critério de Exatidão Distribucional (DAC) visa avaliar a eficiência na preservação da distribuição dos dados originais [22]. A Figura 2.3 é um exemplo desta propriedade, onde está representada a Função Distribuição Acumulada (CDF) de um *dataset* original e da sua versão imputada.

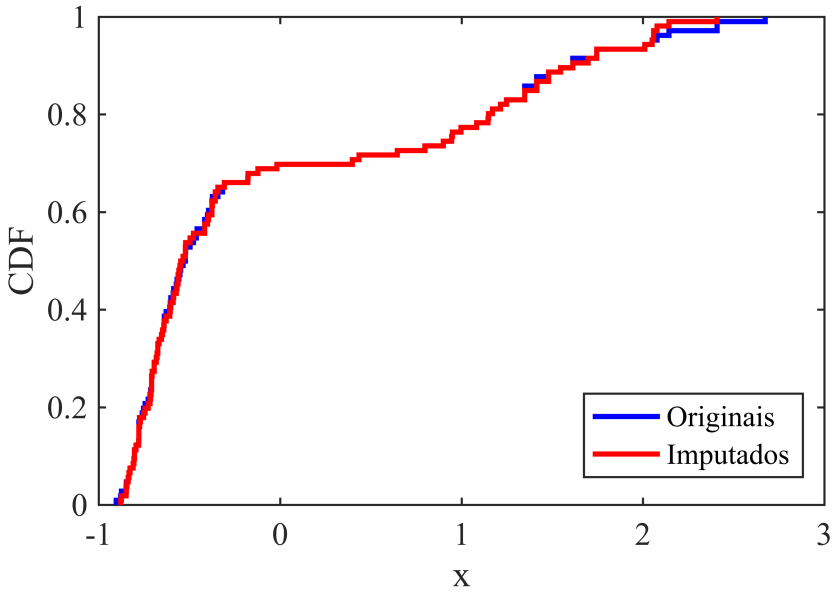


Figura 2.3: Exemplo de DAC

2.1.3.3 Coeficiente de Determinação

O coeficiente de determinação, r^2, é uma métrica que fornece a proporção das flutuações dos dados imputados a partir da comparação com os dados originais. O r^2 é equivalente ao quadrado do coeficiente de correlação de *Pearson*, e permite efetuar uma análise PAC [23]. A correlação de *Pearson* para uma variável de um *dataset* é dada por:

$$r = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\tilde{x}}_i)(\hat{x}_i - \bar{\hat{x}}_i)}{\sqrt{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\tilde{x}}_i)^2 \sum_{i=1}^{n} (\hat{x}_i - \bar{\hat{x}}_i)^2}} \tag{2.5}$$

onde $\tilde{x}$ é um vetor com os valores imputados, $\hat{x}$ os valores originais correspondentes, e n é o número de valores em falta. O valor de r^2 varia entre 0 e 1, e quanto maior for esse valor melhor é o desempenho da técnica de imputação.

2.1.3.4 Distância de *Kolmogorov-Smirnov*

A distância de *Kolmogorov-Smirnov*, D_{KS}, permite medir a diferença entre a distribuição de um *dataset* original e a distribuição da sua versão imputada, possibilitando uma análise DAC [24]. A D_{KS} para cada variável de um *dataset*, obtém-se da seguinte forma:

$$D_{KS} = \max(||F_{\hat{x}} - F_{\tilde{x}}||) \tag{2.6}$$

onde $F_{\hat{x}}$ e $F_{\tilde{x}}$ são as CDF empíricas dos valores imputados ($\tilde{x}$) e os originais correspondentes ($\hat{x}$), respetivamente. Valores de D_{KS} menores representam melhores resultados de imputação.

2.1.3.5 Erro Quadrático Médio

O Erro Quadrático Médio (MSE) pode ser usado para efetuar a medida do erro de imputação e analisar a robustez externa (robustez a *outlier*) de uma técnica de imputação [25]. Para cada variável de um *dataset*, todos os dados originais ($\hat{x}$) são comparados com os da versão imputada ($\tilde{x}$), e o cálculo do erro é efetuado através da expressão:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i) \tag{2.7}$$

onde n é o número de amostras dos dados $\hat{x}$ e $\tilde{x}$. Neste caso, quanto mais próximo de 0 for o valor de MSE, melhor é o desempenho do método de imputação.

2.2 Tratamento de Dados Não Balanceados

Um conjunto de dados considera-se não balanceado quando apresenta um número de instâncias díspar entre classes. Por exemplo, um *dataset* com uma classe negativa que tem 1000 instâncias e uma positiva que tem 200: neste caso, a classe negativa representa 80% dos casos ao passo que a positiva apenas 20%. Este desequilíbrio dos dados num *dataset* tem como consequência a construção de modelos de classificação enviesados que beneficiam a classe majoritária, produzindo por isso estimativas pouco fiáveis. Além do número díspar de instâncias entre classes (*between-class imbalance*), também existem casos em que ocorre desequilíbrios no número de instâncias nos subgrupos de cada classe, este problema é designado de *within-class imbalance* e *small disjuncts* [26–28].

Atualmente existem várias técnicas de amostragem capazes de contornar o problema dos conjuntos de dados não balanceados. De seguida são apresentadas algumas dessas técnicas e o seu respetivo funcionamento.

2.2.1 Técnicas de Sobre-amostragem

No estágio de pré-processamento de KDD, a utilização de técnicas de amostragem possibilitam a redução da disparidade no número de instâncias entre classes. Resumidamente, estas técnicas solucionam o desequilíbrio das classes através da remoção ou adição de instâncias. Nesta secção é apresentada uma breve explicação das várias técnicas de sobre-amostragem utilizadas neste trabalho. O motivo da escolha destas técnicas provém da sua presença marcada na literatura e por esta combinação efetuar a cobertura de várias estratégias existentes (replicação aleatória de dados da classe minoritária, inserção de dados sintéticos com ou sem o uso de *clusters*, remoção de instâncias ruidosas, etc.)

2.2.1.1 *Random Over-Sampling*

Random Over-Sampling (ROS) é uma técnica de amostragem que realiza a inserção de instâncias na classe minoritária. Essa inserção consiste na cópia de instâncias existentes, onde a seleção do exemplo da classe minoritária a replicar é aleatória.

Este método é simples de implementar e não exige um grande esforço computacional, relativamente a outras técnicas de sobre-amostragem. Porém, na literatura vários investigadores criticam o aumento da probabilidade de *overfitting* no ROS [29].

2.2.1.2 *Synthetic Minority Over-sampling Technique*

No sentido de diminuir a probabilidade de *overfitting*, métodos mais avançados foram propostos. Entre eles, o *Synthetic Minority Over-sampling Technique* (SMOTE) [30] que efetua a inserção de instâncias na classe minoritária, mas contrariamente ao ROS, as instâncias novas não são réplicas exatas.

O funcionamento desta técnica resume-se aos seguintes passos: para cada instância da classe minoritária (x_i na Figura 2.4) são procurados os k vizinhos mais próximos (x_{ik}, $k = \{1, 2, 3, 4\}$); seleciona-se aleatoriamente um dos seus vizinhos (e.g. x_{i2}); obtêm-se a diferença entre os dois padrões (x_i e x_{i2}), multiplica-se por um valor aleatório compreendido entre 0 e 1 e adiciona-se esse valor obtido ao padrão inicial (x_i) para originar um padrão sintético (r_2).

Contudo, a este algoritmo está associado o problema de sobre-generalização, relacionado com o modo de geração dos padrões sintéticos: o SMOTE ao computar as instâncias sintéticas não tem em consideração os exemplos próximos da classe majoritária, aumentando a ocorrência de sobreposição de classes [28, 31–34].

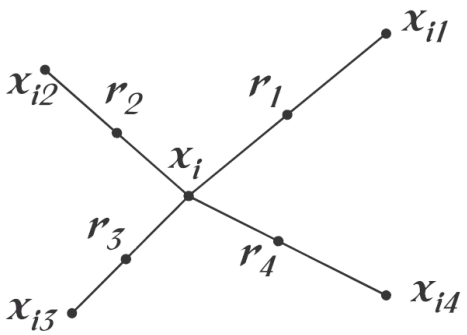


Figura 2.4: Exemplo de criação de dados sintéticos com SMOTE. Adaptado da fonte [28]

2.2.1.3 SMOTE+*Edited Nearest Neighbor*

Frequentemente, são observados conjuntos de dados que apesar de balanceados apresentam sobreposição de classes, o que dificulta o processo de *data mining* e deteriora o desempenho dos classificadores. O método de *Wilson's Edited Nearest Neighbor* (ENN), introduzido em 1972 [35], permite efetuar a remoção de padrões mal classificados, onde os dados mal classificado são aqueles que são incorretamente classificados com base na informação dos seus 3 vizinhos mais próximos.

O SMOTE + ENN consiste na combinação de uma técnica de geração de dados sintéticos e uma técnica de limpeza de dados ruidosos com o objetivo de evitar a sobreposição de classes [29]. De salientar, esta técnica efetua primeiramente a geração de dados sintéticos na classe minoritária seguida da limpeza das instâncias mal classificadas em ambas as classes.

2.2.1.4 SMOTE+*Tomek Links*

Outro método de limpeza conhecido é a aplicação de *Tomek Links* (TL), também chamado de *Tomek's modification of Condensed Nearest Neighbour* [36], e consiste na exclusão de instâncias que formem TL. Um par de pontos é considerado TL quando estes são os vizinhos mais próximos um do outro mas pertencem a classes diferentes. Se um par de instâncias é considerada uma TL então um dos exemplos é ruído ou ambos os exemplos fazem parte da fronteira.

A união do método descrito anterior com SMOTE, denominado de SMOTE+TL, tem uma sequência de execução e motivação semelhante ao da junção SMOTE+ENN. Todavia, o processo de limpeza de dados baseado em *Tomek Links* é menos intrusivo, ou seja, não efetua uma limpeza de ruído tão profunda como ENN [29].

2.2.1.5 *Borderline*-SMOTE

Os algoritmos de classificação focam-se em compreender e interpretar a zona de separação entre as classes com o objetivo de melhorar a previsão [37]. Assim sendo, as instâncias próximas das fronteiras são as mais relevantes e quanto melhor estiver delimitada a fronteira, mais precisa é a discriminação entre classes [28].

Han, Wang e Mao [37] propõem duas variações ao SMOTE denominadas *Borderline*-SMOTE1 e *Borderline*-SMOTE2. A principal característica que diferencia o *Borderline*-SMOTE do SMOTE é a inserção de novas instâncias sintéticas na classe minoritária ser exclusivamente efetuada nas proximidades das fronteiras das classes. As instâncias da classe minoritária são divididas em 3 grupos (**ruído**, **perigo** e **segura**) e a definição destas zonas é baseada no número de instâncias da classe majoritária (m) presentes nos k vizinhos mais próximos. Se $m = k$ então essa instância é etiquetada de **ruído**, caso $\frac{k}{2} < m < k$, a instância é considerada **perigo** e se $0 < m < \frac{k}{2}$ é considerada **segura**. No *Borderline*-SMOTE1, recorrendo ao mesmo conceito do SMOTE, geram-se dados sintéticos com os n vizinhos mais próximos da classe minoritária de cada exemplo do grupo **perigo**. Na versão 2, para além dos n vizinhos minoritários, considera-se também o vizinho mais próximos da classe majoritária de cada exemplo do conjunto **perigo**.

2.2.1.6 *Safe-Level*-SMOTE

Outra abordagem de SMOTE, introduzida por Bunkhumpornpat, Sinapiromsaran e Lursinsap [38], é o *Safe-Level*-SMOTE e destaca-se dos restantes métodos por só gerar dados sintéticos com base em instâncias consideradas **seguras**. Os exemplos da classe minoritária são nomeados **seguros** ou **ruído** consoante o valor de um coeficiente chamado de *safe level* (sl), que é obtido para cada instância minoritária e representa o número de instâncias minoritárias presentes nos k vizinhos mais próximos. Uma instância que tenha um sl próximo de 0 é considerado **ruído**, caso seja próximo de k então está numa região **segura**. Uma instância (p) que pertence à classe minoritária é selecionada, e após a obtenção dos seus k vizinhos, um destes vizinhos (n) pertencente à classe minoritária é selecionado. Para o p e n é calculado o *safe level*, $sl(p)$ e $sl(n)$, respetivamente, de modo a obter o rácio de *safe level* (slr) que é igual a $\frac{sl(p)}{sl(n)}$. Conforme o valores obtidos têm-se as seguintes regras [38]:

- $slr = \infty \bigwedge sl(p) = 0$ então p e n são tidos com **ruído** e não são utilizados para gerar dados sintéticos;

- $slr = \infty \bigwedge sl(p) > 0$ significa que n é considerado **ruído** e como tal é somente utilizado p para gerar a instância sintética, que será uma réplica de p;
- $slr = 1$ logo p e n apresentam o mesmos valor de sl (p é tão **seguro** quanto n) e a geração do dado sintético estará contido na linha que une os dois exemplos;
- $slr > 1$ significa que p é mais **seguro** que n e devido a tal a nova instância é gerada mais perto de p no intervalo $[0, \frac{1}{slr}]$;
- $slr < 1$ significa que p é menos **seguro** que n e devido a tal a nova instância é gerada mais perto de n no intervalo $[1 - slr, 1]$.

2.2.1.7 *Majority Weighted Minority Over-sampling Technique*

Motivados pelo aumento da dificuldade de aprendizagem dos algoritmos de *Machine Learning* em *datasets* sujeitos a algumas técnicas de amostragem que efetuam a inserção de dados sintéticos em regiões menos apropriadas, Barua et al. [39] propõem outra abordagem, *Majority Weighted Minority Over-sampling Technique* (MWMOTE). Nesta técnica é primeiramente efetuada a remoção dos padrões pertencentes à classe minoritária que tenham todos os seus k vizinhos da classe majoritária (consideram-se ruído) e procuram-se os dados minoritários responsáveis pelo aumento da dificuldade de treino. De seguida são atribuídos pesos aos dados minoritários com base nas seguintes diretivas: os exemplos mais próximos das fronteiras são mais importantes; os dados minoritários que fazem parte de um *cluster* mais esparso são mais importantes; os exemplos minoritários próximos de um *cluster* denso de dados majoritários são mais relevantes do que os exemplos próximos de um *cluster* esparso. Além disso, são identificados os *clusters* da classe minoritária, com o intuito dos novos dados inseridos estarem contidos nestes *clusters*. Por fim, MWMOTE realiza a geração dos dados sintéticos com base na informação dos pesos e *clusters*, de forma similar ao SMOTE.

2.2.1.8 *Adaptive Synthetic Sampling Approach*

He et al. [40] desenvolveram o *Adaptive Synthetic Sampling Approach* (ADASYN), onde a ideia é produzir de uma forma adaptada instâncias sintéticas na classe minoritária de acordo com a sua distribuição, por outras palavras, são gerados em maior quantidade dados sintéticos com base nos padrões mais difíceis de aprender, como o intuito de estes tornarem-se mais fáceis de aprender. ADASYN procura os k vizinhos mais próximos de cada dado minoritário (i) e atribui um rácio (r_i), que consiste no número de vizinhos majoritários sobre k, para efetuar o cálculo de quantos padrões sintéticos deverá produzir (g_i), com base

no i-éssimo padrão da classe minoritária. Após a obtenção de todos os rácios, estes são normalizados ($\hat{r}_i$) e efetua-se o computação de g_i através da equação 2.8 [40].

$$g_i = \hat{r}_i \times G \tag{2.8}$$

Onde G é o número total de exemplos sintéticos que são pretendidos produzir. Os dados sintéticos são gerados de acordo com a abordagem utilizada pelo SMOTE.

Ao contrário dos métodos de geração de dados sintéticos explicitados anteriormente, que ignoram os exemplos minoritários com todos os k vizinhos pertencentes à classe majoritária (consideram-se ruído), ADASYN faz uso desses exemplos e atribui-lhes a máxima importância.

2.2.1.9 *Adjusting the Direction Of the synthetic Minority clasS*

Tang e Chen [41] propõem um algoritmo que efetua o equilíbrio de classes através da inserção de instâncias sintéticas com base na Análise de Componentes Principais (PCA) denominada *Adjusting the Direction Of the synthetic Minority clasS* (ADOMS). Esta técnica pretende gerar os dados numa zona mais bem distribuída e para tal utiliza a informação dos k vizinhos mais próximos de uma instância da classe minoritária para efetuar o cálculo das PCA. O modo de proceder de ADOMS consiste na seleção aleatória de um exemplo da classe minoritária (x_i), procura dos k vizinhos mais próximos da classe minoritária de x_i, cálculo da PCA do x_i e seus k vizinhos e determinação da primeira componente principal, escolha aleatória de um dos vizinhos (e.g. x_{i1}), obtenção da distância (D) entre x_i e x_{i1} e geração de um padrão sintético através da projeção de $D \times rand(0, 1)$ no eixo da primeira componente principal.

Quando é utilizado $k = 1$ no ADOMS, o eixo da primeira componente, obtido pela PCA é o segmento de recta que une essas instâncias, o que torna esta abordagem equivalente à de SMOTE.

2.2.1.10 *Selective Preprocessing of Imbalanced Data*

Selective Preprocessing of Imbalanced Data (SPIDER) foi introduzido por Stefanowski e Wilk [42] e combina a replicação local de instâncias minoritárias com a filtragem de exemplos da classe majoritária difíceis de aprender. Na primeira etapa todas as instâncias são classificadas com um classificador k-NN e caso a sua predição seja correta, esse exemplo é etiquetado como **seguro**, e caso contrário como **ruído**. A etapa seguinte depende do parâmetro de entrada, o qual pode ser: amplificação fraca, amplificação fraca com reetiquetação

e amplificação forte. Na amplificação fraca os exemplos minoritários com etiqueta **ruído** (x_r) são replicados n vezes, onde n é o número de exemplos majoritários com etiqueta **seguro** presentes nos k vizinhos de x_r. A amplificação fraca com reetiquetação é uma extensão do parâmetro anterior com a modificação da classe dos exemplos majoritários com etiqueta **ruído** (y_r) que estão presentes na vizinhança k dos exemplos x_r. Por fim, na amplificação forte, os exemplos minoritários com etiqueta **seguro** são replicados n vezes, os exemplos x_r são classificados com um k maior (k_m) e se obtiverem uma classificação correta então são replicados n vezes, caso contrário são replicados n_m vezes, onde n_m é o número de exemplos majoritários com etiqueta **seguro** presentes nos k_m vizinhos de x_r. No final, os exemplos y_r restantes (não reetiquetados) são removidos.

O SPIDER, após a identificação das instâncias, efetua o processamento dos dados de ambas as classes em simultâneo. Porém, esse procedimento pode resultar numa deterioração excessiva na classe majoritária. Com o objetivo de evitar esse fenómeno, Napierała, Stefanowski e Wilk [43] apresentam o SPIDER2, onde primeiro os exemplos y_r são removidos ou reetiquetados (dependendo dos parâmetros de entrada), para que depois seja efetuada a amplificação dos dados minoritários.

2.2.1.11 *Cluster-Based Oversampling*

No artigo "Class Imbalances versus Small Disjuncts" [26], Jo e Japkowicz efetuam uma experiência com o objetivo de entender se a perda de desempenho dos classificadores é causada de forma direta por *between-class imbalance* ou se é motivada pelos *small disjuncts*. Ao longo da sua experiência, os investigadores sugerem uma técnica que permite corrigir ambos os problemas (*small disjuncts* e *between-class imbalance*) – *Cluster-Based Oversampling* (CBO).

O primeiro passo do CBO é identificar os *clusters* de cada classe, onde alguns autores adotam o *K-Means Clustering* (KMC) [26, 27]. O funcionamento do KMC resume-se a [44]: escolha aleatória de k instâncias e definição dessas como os centroides iniciais (m_k); cálculo da distância de cada exemplo aos centroides m_k e atribuição de um *cluster* a cada exemplo com base no centroide mais próximo; atualização dos centroides através do cálculo da média dos exemplos que lhes pertencem; repetição dos passos anteriores até não ocorrerem alterações nos *clusters*, ou até o máximo de iterações ser atingido.

Após a obtenção dos *clusters*, é efetuada a sobre-amostragem. A técnica de geração de dados mais comum é o ROS [26], mas existem outras variantes para adicionar instâncias

sintéticas (SMOTE) [27]. Na classe majoritária, todos os *clusters*, excepto o maior, são preenchidos com novos exemplos de forma a apresentarem o mesmo número de instâncias que o maior *cluster*. Na classe minoritária, cada *cluster* é povoado com novas instâncias até obter a dimensão t_{maj}/k_{min}, onde t_{maj} é o número total de exemplos da classe majoritária e k_{min} é o número de *clusters* da classe minoritária.

2.2.1.12 *Agglomerative Hierarchical Clustering*

Numa linha de raciocínio semelhante à anterior, Cohen et al. [45] propõem, entre várias técnicas, uma que consiste no uso de *Agglomerative Hierarchical Clustering* (AHC), onde os centroides dos *clusters* obtidos são utilizados para gerar dados sintéticos na classe minoritária. O algoritmo de identificação de *clusters*, AHC, tem um funcionamento distinto do *k-means* [46]: todas as instâncias começam por ser *clusters* individuais, e vão-se juntando os dois *clusters* mais próximos, até todos os padrões se encontrarem no mesmo *cluster*. Existem várias regras de cálculo da distância entre os *clusters* (*linkage rules*), por exemplo: *single linkage*, onde a distância entre dois *clusters* $D(c_1, c_2)$ corresponde à distância dos exemplos de cada *cluster* mais próximos; *complete linkage*, onde $D(c_1, c_2)$ corresponde à distância dos exemplos de cada *cluster* mais afastados; *average linkage between groups*, onde $D(c_1, c_2)$ corresponde à média das distâncias entre todos os pares de exemplos pertencente a clusters distintos. O processo de sobre-amostragem baseada em AHC é composto por três passos [47]: construir dendrogramas com AHC utilizando *single linkage* e *complete linkage*; observar os *clusters* de todos os níveis dos dendrogramas e cálculo dos seus centroides; adicionar os centroides, como novas instâncias sintéticas, aos exemplos originais e repetir até obter o equilíbrio entre classes.

2.2.2 Técnicas de Classificação

A classificação é um processo que pertence ao estágio de *data mining* de KDD, que tem como objetivo construir modelos capazes de tomar decisões futuras. A qualidade de um processo de extração de conhecimento é dependente do desempenho do modelo construído.

Ao longo desta secção são descritas as técnicas de classificação utilizadas neste trabalho. Estas técnicas têm princípios de funcionamento distintos (*tree-based, distance-based, statistical, kernel-based*).

2.2.2.1 Árvores de Decisão

O princípio base do funcionamento das Árvores de Decisão já foi mencionado na secção 2.1.2.2. Como tal, nesta secção pretende-se esclarecer e delimitar as diferenças entre algumas das implementações mais populares, como o C4.5 [48], o sucessor de *Iterative Dichotomiser 3* (ID3) [49], e as *Classification and Regression Trees* (CART) [8]. Estas técnicas utilizam métricas baseadas na entropia como critério de divisão, durante o processo de construção da árvore. A entropia caracteriza a impureza de um conjunto de exemplos arbitrários. Considerando um conjunto de instâncias D, que contêm c classes, a entropia desse conjunto $E(D)$ é obtida pela equação 2.9, onde p_i é a proporção de instâncias de D pertencentes à classe i [7]. Se $E(D)$ for 0 então todos os exemplos de D pertencem à mesma classe. Através da definição de entropia, pode-se medir a eficácia de uma variável A para discriminar as várias classes, pelo cálculo da entropia de D após ser utilizada a variável A para dividir D em v partições, $E(D, A)$ (equação 2.10). A métrica de divisão chamada ganho de informação (*Information Gain*), utilizada como critério na implementação de ID3, permite quantificar a redução da entropia causada pela divisão de exemplos de acordo com uma variável (equação 2.11).

$$E(D) = \sum_{i=1}^{c} -p_i \times log_2(p_i) \tag{2.9}$$

$$E(D, A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times E(D_j) \tag{2.10}$$

$$Gain(D, A) = E(D) - E(D, A) \tag{2.11}$$

O C4.5 aplica um critério denominado de rácio de ganho (*gain ratio*), que é uma normalização do ganho de informação (equações 2.12 e 2.13) [49, 50].

$$S(D, A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \frac{|D_j|}{|D|} \tag{2.12}$$

$$GainRatio(D, A) = \frac{Gain(D, A)}{S(D, A)} \tag{2.13}$$

O algoritmo CART utiliza outros critérios de divisão, o mais frequente é o *Gini impurity* (equações 2.14 e 2.15) [50, 51]. Neste caso, de forma semelhante ao ID3 e C4.5 pretende-se reduzir a impureza no conjunto D, e por isso a variável escolhida para efetuar o teste num nó é aquela que maximize essa redução (equação 2.16).

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{2.14}$$

$$Gini(D, A) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Gini(D_j) \tag{2.15}$$

$$\triangle Gini(D, A) = Gini(D) - Gini(D, A) \tag{2.16}$$

Quando uma DT é construída, alguns ramos podem refletir anomalias presentes nos dados de treino, devido a ruído ou *outliers*. Assim, alguns algoritmos de DT recorrem a uma técnica de poda (*prunning*) que efetua a remoção de ramos que levem a um modelo sobre-ajustado (*overfitting*). O método de poda através do custo de complexidade, utilizado pelo CART, tem em consideração uma função com o número de folhas da DT e a taxa de erro. De modo resumido, é efetuado, para cada nó, o cálculo do custo de complexidade da árvore com e sem esse nó, e caso o valor de complexidade sem esse nó seja menor, o nó em questão é removido. O C4.5 faz uso de uma técnica de poda "pessimista" baseada na estimativa da taxa de erro, mas não requer um conjunto de árvores podadas como no método usado pelo CART. Em vez disso, faz somente uso do conjunto de dados de treino para estimar a taxa de erro. Por fim, CART é caracterizado pelo facto de construir árvores binárias, ou seja, cada nó faz um teste que só pode ter duas respostas (dois ramos).

2.2.2.2 *k-Nearest Neighbors*

k-Nearest Neighbors (k-NN) é um algoritmo de classificação no qual k vizinhos mais próximos de um padrão são escolhidos, encontrados pela minimização de uma medida de distância (e.g. Euclidiana, Manhattan e Minkowski) [52]. Para determinar a classe de um padrão o k-NN efetua o cálculo das distâncias desse exemplo aos restantes exemplos, e identifica os seus k vizinhos e as respetivas classes. Esse padrão é então classificado através da classe que tiver maior número de ocorrências (sistema de votos) ou pela classificação que obtiver um peso maior, onde o peso das classes é atribuído consoante a distância ao exemplo a classificar (quanto menor distância maior o peso). A maior desvantagem desta técnica está relacionado com o facto de ser um "algoritmo preguiçoso" (*lazy learner*). Esse título advém da necessidade de computar um novo modelo para atribuir uma nova classificação. Por outras palavras, o k-NN não efetua a construção de um modelo generalizado com os dados de treino: sempre que k-NN efetua a classificação de um novo padrão (de teste), necessita de realizar o cálculo das distâncias para todo o *dataset*, o que é especialmente problemático em *datasets* de grandes dimensões. Outro problema é a escolha do melhor valor de k e da métrica de distância a aplicar, como descrito na secção 2.1.2.3.

2.2.2.3 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVM), como mencionado anteriormente (secção 2.1.2.5), é um algoritmo que permite realizar a classificação através de hiperplanos ótimos. As SVM podem fazer uso de funções *kernel*, que são capazes de operar em espaços dimensionais elevados e permitem projetar (mapear) os dados para um espaço com dimensão superior. A principal vantagem do uso dessas funções é a possibilidade de generalizar fronteiras de decisão não lineares, o que facilita a classificação de padrões [53, 54]. As SVM equilibram a complexidade do modelo com o seu sucesso no ajuste dos dados, o que se traduz num compromisso bem sucedido entre a flexibilidade do modelo e o erro no treino dos dados [53]. No entanto, para efetuar o treino é necessário definir alguns parâmetros de entrada com base na função *kernel* utilizada.

2.2.2.4 *Naive Bayes*

O classificador *Naive Bayes* (NB) tem em consideração a distribuição de probabilidades dos dados em cada classe para tomar uma decisão, assumindo que existe uma relação probabilística entre as variáveis e a classe [55]. O classificador Bayesiano determina a probabilidade de uma dada instância x pertencer à classe c_i, $P(c_i|x)$, que é denominada de probabilidade *a posteriori* (ou condicionada). Para um problema de classificação binário (c_1 e c_2), onde existem duas probabilidades *a posteriori*, a regra de decisão de NB atribui a classe c_1 à instância x se $P(c_1|x) > P(c_2|x)$, a classe c_2 se $P(c_1|x) < P(c_2|x)$, e caso $P(c_1|x) = P(c_2|x)$ atribui aleatoriamente uma das classes.

As probabilidades *a posteriori* são calculadas de acordo com a fórmula de *Bayes* (equação 2.17)

$$P(c_i|x) = P(x|c_i) \times \frac{P(c_i)}{P(x)} \tag{2.17}$$

onde $P(c_i)$ é a probabilidade *a priori* da classe c_i, que consiste na probabilidade estimada de ocorrer classe c_i, $P(x|c_i)$ é a verosimilhança de x, que pode ser obtida pela Função Densidade de Probabilidade (PDF) de x, e $P(x)$ é a probabilidade total de x, a qual é obtida pela equação 2.18:

$$P(x) = \sum_{i=1}^{c} P(x|c_i) \times P(c_i) \tag{2.18}$$

Esta técnica é capaz de operar com variáveis contínuas e discretas, apesar do processo para as variáveis contínuas ser mais complexo. Além disso, as variáveis são consideradas independentes, o que na realidade pode não ser verdade.

2.2.3 Validação Cruzada

A validação cruzada *k-folds* é uma estratégia comum de validação do desempenho de uma dada técnica de classificação, onde o *dataset* completo é inicialmente dividido em *k-folds*, as $k-1$ divisões são usadas para efetuar o treino do modelo de classificação enquanto a divisão sobrante serve de *dataset* de teste do modelo. Este processo é repetido iterativamente de modo a que todas as divisões sejam usadas para efetuar o treino e teste do modelo. No final é calculada a média das métricas de avaliação da classificação obtidas por cada divisão de teste. O objetivo da validação cruzada é assegurar que k conjuntos de validação independentes são usados para testar o modelo, simulando dados ainda não vistos pelo modelo de classificação: os dados do conjunto de teste nunca são vistos durante a fase de treino do modelo, para precaver o risco de *overfitting*. Contudo, num problema de dados não balanceados em que se efetua sobre-amostragem, assume-se a inserção de novas instâncias minoritárias que ou são réplicas das instâncias originais ou novas instâncias sintéticas que imitam as originais. Neste cenário, o uso conjunto da sobre-amostragem e validação cruzada tem que ser efetuada de forma cuidadosa, já que um uso menos apropriado da validação cruzada pode levar a estimações sobre-otimistas no desempenho do modelo. Caso a sobre-amostragem seja realizada antes de efetuar as k divisões do *dataset* necessárias para a validação cruzada, as cópias da mesmas instâncias (ou semelhantes) podem aparecer em ambos os conjuntos, treino e teste, durante a validação cruzada, provocando estimativas enviesadas.

2.2.4 Métricas de Avaliação

Atualmente, um dos critérios de avaliação de classificação mais comum é a Exatidão da Classificação (ACC). Porém, em problemas de dados não balanceados, esse critério não é ideal, uma vez que a ACC não faz a distinção do número correto de classificações das instâncias de classes diferentes [28]. Por exemplo, para um dado *dataset* que tenha 970 exemplos da classe negativa e 30 da classe positiva e que o classificador obtenha uma ACC de 0.97, não é conclusivo se é um bom ou mau resultado pois não se sabe se o classificador atribuiu a todas as instâncias a classe negativa. Assim sendo, a escolha das métricas de avaliação utilizadas neste trabalho é mais minuciosa, visto que é necessário ter em consideração as proporções de exemplos presentes nas classes. Nas secções seguintes são descritas de modo breve algumas métricas para contextualizar os critérios de avaliação implementados neste trabalho.

2.2.4.1 Matriz de Confusão, Sensibilidade, Especificidade e Precisão

A matriz de confusão possibilita a visualização do desempenho de um dado classificador. A matriz de confusão é uma matriz quadrada com o número de colunas e linhas igual ao número de classes, onde cada coluna contém as instâncias classificadas com uma classe especifica, enquanto cada linha tem as instâncias com a sua classe original (Tabela 2.1).

Tabela 2.1: Exemplo de uma matriz de confusão para um problema binário.

		Previsão	
		Positivo	**Negativo**
Atual	**Positivo**	Verdadeiro positivo (VP)	Falso positivo (FP)
	Negativo	Verdadeiro negativo (VN)	Falso negativo (FN)

A taxa de verdadeiros positivos (equação 2.19), também conhecida como Sensibilidade (SEN) ou *Recall*, indica a proporção de instâncias positivas corretamente identificadas (VP).

$$\text{SEN} = \frac{VP}{VP + FN} \tag{2.19}$$

A taxa de verdadeiros negativos (equação 2.20), frequentemente chamada Especificidade (SPEC), designa a proporção de instâncias negativas corretamente identificadas (VN).

$$\text{SPEC} = \frac{VN}{VN + FP} \tag{2.20}$$

O Valor Preditivo Positivo (VPP) – equação 2.21 – também chamado Precisão, enuncia a proporção de instâncias positivas corretamente identificadas no conjunto de instâncias previstas como positivas.

$$\text{VPP} = \frac{VP}{VP + FN} \tag{2.21}$$

2.2.4.2 *F-Measure* e *G-Mean*

A SEN e a VPP podem ter níveis de importância diferentes, que variam com as características dos conjuntos de dados e da classe de interesse [56]. A *F-Measure* é uma medida que possibilita o compromisso entre a sensibilidade e a precisão (equação 2.22). O compromisso é obtido através da média harmónica dessas duas métricas [57]:

$$\textit{F-Measure} = \frac{(1 + \beta^2) \times \text{VPP} \times \text{SEN}}{\beta^2 \times \text{VPP} + \text{SEN}} \tag{2.22}$$

onde β é a importância relativa dada à métrica de sensibilidade em relação à precisão. Quando $\beta > 1$, a sensibilidade é mais importante, caso $\beta < 1$, a precisão é mais relevante que a sensibilidade, mas se $\beta = 1$ (conhecida com *F-1*), o peso das métricas é igual.

Outra métrica que é independente da distribuição dos exemplos entre classes é a *G-Mean* (equação 2.23) [58]. Esta calcula a média geométrica da exatidão de cada classe, na tentativa de as maximizar enquanto visa manter um bom equilíbrio. Contudo, esta métrica baseada numa média geométrica não permite controlar o nível de importância da classe de interesse [28].

$$\textit{G-Mean} = \sqrt{\frac{VP}{VP+FN} \times \frac{VN}{VN+FP}} = \sqrt{\text{SEN} \times \text{SPEC}} \tag{2.23}$$

2.2.4.3 Área abaixo da Curva ROC

A Curva Característica de Operação do Receptor (ROC) é um critério de avaliação gráfico do desempenho da classificação, onde são representados os compromissos entre os benefícios (SEN) e os custos (taxa de FP) [59].

A AUC é uma forma de comparar diferentes modelos de classificação, através das suas curvas ROC. O significado estatístico da AUC é a probabilidade do classificador atribuir um *rank* superior a uma instância positiva escolhida aleatoriamente do que a uma instância negativa, e pode ser obtida pela seguinte equação [60]:

$$\text{AUC} = \frac{\sum r_i - n_+(n_+ + 1)/2}{n_+ n_-} \tag{2.24}$$

onde r_i é o *rank* da i-éssima instância positiva, n_+ é o número de instâncias positivas e n_- é o número de instâncias negativas. Quando maior for o valor da AUC melhor é o desempenho do classificador.

2.2.5 Métricas de Complexidade

Ho e Basu [61] estudaram métricas capazes de caracterizar a dificuldade na classificação de um problema, com base na complexidade da geometria das fronteiras entre classes. Os autores apresentaram um conjunto de métricas que estão sumariada na Tabela 2.2. De salientar que F1, F2 e F3 realizam a medida da sobreposição das variáveis entre classes diferentes, L3 e N4 avaliam a geometria e topologia, e N1, N2, N3, L1 e L2 identificam a separabilidade das classes.

Tabela 2.2: Descrição das métricas de complexidade. O ++ corresponde a valores elevados (ou seja, quanto maior forem os valores da métrica, mais complexo será o *dataset*) e −−, o contrário.

Métrica	**Descrição**	**Mais complexos**
F1	Rácio discriminativo de Fisher's máximo	−−
F2	Volume da região de sobreposição entre classes	++
F3	Eficiência máxima (individual) das variáveis	−−
L1	Erro mínimo de um classificador linear	++
L2	Taxa de erro de um classificador linear	++
L3	Não-Linearidade do classificador linear (usando SVM de *kernel* linear)	++
N1	Fração de pontos nas fronteiras	++
N2	Rácio médio da distância intra-vizinhos e inter-vizinhos	++
N4	Não-Linearidade de um classificador k-NN com $k = 1$	++

3 Trabalhos Relacionados

Ao longo deste capítulo serão apresentados trabalhos relacionados com os dois problemas tratados nesta dissertação: Dados Incompletos (secção 3.1) e Dados Não Balanceados (secção 3.2). As características principais de cada artigo descrito em cada secção estão sumariadas nas Tabelas 3.1 e 3.2, respetivamente.

3.1 Problema de Dados Incompletos

No artigo de Nanni, Lumini e Brahnam [62] o objetivo foi identificar os melhores algoritmos de imputação de dados incompletos, através do desempenho da classificação. Os investigadores utilizaram cinco *datasets* completos na área da saúde, com $[8, 32]$ variáveis, $[155, 768]$ instâncias, e geraram dados em falta com Taxa de Valores em Falta (MR) entre $[10, 50]\%$. As técnicas de imputação implementadas foram a Média, Redes Neuronais Artificiais (ANN), *InPaint*, *Bayesian PCA* (BPCA), k-NN, *Expectation–Maximization* (EM), *Dissimilarity* e *Learn*$^{++}$*MF*. O *Input Decimated Ensemble* (IDE) e as SVM foram os classificadores implementados nesse artigo. A análise do desempenho foi efetuada através das métricas AUC e *Rank*. Os investigadores concluíram que as suas técnicas de imputação, baseadas em *clustering* e sub-espaços aleatórios, apresentam melhor comportamento que todas as outras, obtendo um desempenho satisfatório para MR superiores a 30%.

Kang [63] investigou o desempenho de algoritmos de classificação e regressão em *datasets* com valores em falta imputados recorrendo à técnica de *Locally Linear Regression* (LLR). O modo de execução consistiu na criação de valores em falta aleatórios, com MR entre $[1, 50]\%$, em 13 e 9 *datasets* utilizados para classificação e regressão, respetivamente. Os conjuntos de dados nessa pesquisa provêm de diferentes áreas (e.g. médica, industrial e económica). Além disso, o número de variáveis varia entre $[4, 60]$ e o número de instâncias entre $[150, 14429]$ nos *datasets* para classificação, e para os de regressão $[194, 10000]$ e $[8, 32]$. As técnicas de imputação avaliadas foram as seguintes: Média, *Hot-deck*, k-NN, *Expectation Conditional*

Maximization (ECM), KMC, *Mixture of Gaussians* (MoG) e LLR. Os métodos de regressão e classificação aplicados foram k-NN, LLR, ANN, *Logistic Regression* (LR) e CART. As métricas ACC e Raiz do Erro Médio Quadrático (RMSE) foram utilizadas para avaliar os resultados das classificações e das regressões, respetivamente. Os resultados mostraram que todos os algoritmos de imputação ajudam a aumentar a ACC, a imputação de LLR proposta melhorou o desempenho dos modelos comparativamente aos restantes métodos (independentemente da MR e da técnica de aprendizagem utilizada), e o LLR, em termos de ACC, para MR relativamente elevadas, teve resultados semelhantes aos obtidos com os *datasets* originais (completos).

O estudo de Aisha, Adam e Shohaimi [19] teve com objetivo compreender o efeito do tratamento de dados incompletos na ACC de quatro classificadores de redes Bayesianos. O *dataset* utilizado não é completo, apresentava uma MR de $48,39\%$, insere-se na área da saúde (hepatite crónica aguda), contém 19 variáveis e 155 instâncias e não é balanceado pois a classe negativa só tem 32 instâncias. As técnicas de imputação aplicadas foram a média/moda, EM, SVM, 2 variações de k-NN e de KMC, Decomposição em Valores Singulares (SVD) e *Local Least Squares* (LLS). Os métodos de classificação usados foram NB, *Tree Augmented Naive Bayes* (TAN), *Boosted Augmented Naive Bayes* (BAN) e *General Bayes Network Classifiers* (GBN). A comparação foi efetuada através da ACC. As conclusões do estudo afirmam que os métodos de imputação de dados SVM e LLS foram os que melhoraram a ACC, comparativamente à classificação obtida quando os dados incompletos são ignorados.

No trabalho de García-Laencina et al. [12] foi utilizado um *dataset* real de 399 pacientes com cancro da mama, composto por 16 variáveis, e com dados em falta. A MR desse *dataset* varia entre 0% e 80.45% em cada variável, sendo a sua média de 17.99%. O objetivo destes estudo foi prever a sobrevivência global aos 5 anos de um paciente com cancro da mama. Das 399 instâncias, 282 pertencem à classe positiva (sobreviver), o que representa um desequilíbrio entre classes. Devido à falta de valores, os autores procuraram compreender qual o melhor método de imputação de forma a obter resultados de classificação mais exatos. A média/moda, EM e k-NN foram os algoritmos de imputação utilizados. As técnicas de classificação usadas foram k-NN, CART, LR e SVM. As métricas de avaliação aplicadas foram a ACC, a SEN, a SPEC e a AUC. Os resultados demostraram que o k-NN teve o melhor desempenho como classificador e técnica de imputação, para o cenário proposto, e a média/moda teve a pior performance.

No artigo de Jerez et al. [5], avaliou-se o efeito de diferentes técnicas de imputação na previsão da recurrência de cancro da mama. O *dataset* utilizado é real, composto por

8 variáveis e com informações sobre 3679 pacientes que tiveram cancro da mama. Esse conjunto de dados tem no global 5.61% de dados incompletos, mas a MR por variável varia entre $[0, 43]\%$. O algoritmo de classificação foi o ANN e as técnicas de imputação aplicadas foram a média/moda, *Hot-deck*, Imputação Múltipla (MI), *Multi-Layer Perceptron* (MLP), SOM e k-NN. A avaliação foi efetuada recorrendo à AUC. Os investigadores demostraram que todos os métodos de imputação, com a excepção do *Hot-deck*, melhoram a previsão. Além disso, as conclusões obtidas indicam que o k-NN teve os melhores resultados e que as técnicas de imputação baseadas em *Machine Learning*, em comparação com as baseadas em métodos estatísticos, tiveram um desempenho superior.

Na pesquisa de Mostafizur e Davis [18] foram exploradas várias técnicas de imputação num conjunto de dados incompletos de pacientes com doenças cardiovasculares. O *dataset* utilizado é composto por 832 registos e 26 variáveis, cuja MR varia entre $[0, 30]\%$, e apenas 120 instâncias pertencem à classe de sobrevivência. O KMC foi o algoritmo utilizado pelos autores para efetuar a classificação. Os algoritmos de imputação aplicados foram a Média/Moda, DT, SVM, *Fuzzy Unordered Rule Induction Algorithm* (FURIA) e *Ripple-Down Rules* (RDR). A ACC, SEN e SPEC foram as medidas de avaliação dos resultados utilizadas. Os resultados mostraram que todos os métodos de imputação baseados em *Machine Learning*, em termos de SEN e em alguns casos de ACC, apresentam melhores valores do que o método de média/moda, e o melhor entre as várias técnicas usadas foi o FURIA.

García-Laencina, Sancho-Gómez e Figueiras-Vidal [6] estudaram o efeito da imputação na ACC, utilizando *datasets* reais e sintéticos. Na primeira parte desse trabalho, os autores começaram por medir a qualidade da imputação usando as métricas PAC usando o coeficiente de *Pearson* (r) e MSE, e DAC através da distância de *Kolmogorov-Smirnov* (D_{KS}). Contudo, o uso dessas métricas de análise só foram aplicadas para a técnica de imputação k-NN, na primeira variável de um *dataset* sintético, composto por 1500 instâncias e 5 classes, e com valores em falta gerados aleatoriamente com MR entre $[5, 40]\%$. Essas métricas foram posteriormente excluídas em favor do CE, dado que o objetivo principal era a resolução do problema de classificação. Numa segunda parte do estudo foram aplicados o k-NN, MLP, SOM e EM como algoritmos de imputação. A classificação foi elaborada através de uma ANN. Nessa segunda parte, foram utilizados um *dataset* sintético (descrito anteriormente) e dois reais (um utilizado para identificar vogais e outro para diagnosticar problemas de tiróide). O primeiro *dataset* real é completo, contém 871 instâncias, 3 variáveis e 6 classes. Contudo, foram removidos valores aleatoriamente a uma das variáveis, com uma MR entre $[5, 40]\%$. O segundo conjunto de dados reais é constituído por 2800 casos com 28 variáveis,

mas apenas cinco variáveis contínuas são utilizadas, e contêm dados em falta com uma MR que varia entre $[6, 21]\%$. Os resultados da primeira parte mostraram que para valores de k baixos a distribuição tende a ser mantida, isto é, melhor valor de DAC, mas a taxa de valores verdadeiros é baixa, ou seja, pior valor de PAC, e para valores de k altos o fenómeno é contrário, quer dizer, maior número de valores verdadeiros mas a distribuição não é mantida. Na segunda parte, os resultados demonstraram que o melhor método de imputação, de modo a obter o menor CE, foi o método de EM, para o *dataset* sintético, e o k-NN, para os dois conjuntos de dados reais.

No trabalho de Rahman e Islam [64] foram implementadas várias técnicas de imputação e comparado o seu desempenho através de métricas PAC, como a correlação de *Pearson* (r), *Index of agreement* (d_2), RMSE e Erro Absoluto Médio (MAE). Porém, não foram efetuadas análises de desempenho com métricas DAC. De forma mais detalhada, os autores aplicaram o EM e uma variante de LLS para realizar a imputação dos dados, e além disso, desenvolveram duas técnicas de imputação baseadas em DT. Em termos de *datasets*, foram usados nove conjuntos de dados com diferentes áreas, número de variáveis $[2, 11]$ e número de instâncias $[871, 2800]$. Três destes *datasets* têm dados em falta originalmente. A esses *datasets* foram removidas as instâncias com valores em falta. No decorrer da experiência os autores removeram dados de forma aleatória dos nove conjuntos de dados, com uma MR de 1%, 3%, 5% e 10%. Os resultados demostraram que as duas variantes de DT implementadas apresentam sempre melhores valores nas métricas de avaliação em comparação com as outras duas técnicas utilizadas nesse trabalho.

O estudo de Abreu et al. [65] apresenta a comparação do desempenho entre três métodos de *ensemble* na classificação de um conjunto de dados incompleto. No âmbito desse estudo, com o objetivo de prever a sobrevivência de doentes com cancro da mama, foi construído um *dataset* com informações de 847 pacientes, onde cada um é caracterizado por 15 variáveis. Este conjunto de dados apresenta uma MR de 25%. A técnica de imputação utilizada foi k-NN e como classificadores foram usados *TreeBagger*, *LPBoost* e Sub-Espaços. As medidas de performance aplicadas para analisar a previsão de classificação foram a ACC e *rank*. As conclusões desse artigo apesar de serem mais direcionadas para análise de métodos de classificação, demonstram que k-NN para uma taxa relativamente elevada (25%) permite obter bons resultados.

Davis e Mostafizur [9] investigaram o efeito de diferentes métodos de imputação na classificação de um *dataset* clínico. Os autores efetuaram a combinação de dois conjuntos de dados de pacientes submetidos a uma cirurgia cardiovascular, e originaram um *dataset* com

a informação de 823 pacientes, composta por 22 variáveis, das quais 18 apresentam uma MR entre 1% e 30%. Os dados usados pretendem prever a morte ou ocorrência de um evento cardiovascular severo dentro dos 30 dias após a operação. A imputação dos dados incompletos foi efetuada com os seguintes algoritmos: Média/Moda, DT, k-NN, FURIA, SVM e RDR. Além disso, nesse artigo foram utilizadas as seguintes técnicas de classificação DT, k-NN, FURIA, ANN e KMC. Os resultados foram avaliados pela ACC, SEN e SPEC, e todos os métodos de classificação apresentaram, em termos de SEN, valores baixos, cerca de 20% (em média). Os investigadores justificam esses valores com falta de balanceamento dos dados. Apesar disso, concluíram que os melhores resultados foram obtidos com a técnica de imputação FURIA em junção com o classificador KMC.

Na pesquisa de Amiri e Jensen [66] foram implementadas três técnicas de imputação baseadas em *Fuzzy Rough Sets* com o objetivo de comparar os seus desempenhos com onze métodos de imputação considerados estado-da-arte. Neste artigo foram utilizados 27 conjuntos de dados reais, completos, de áreas diferentes e com características variadas. Concretamente, os *datasets* considerados têm $[106, 2201]$ instâncias, $[3, 60]$ variáveis e $[2, 11]$ classes. Amiri e Jensen [66] efetuaram a remoção de valores nos conjuntos de dados de acordo com um mecanismo MCAR com MR de 5%, 10%, 20% e 30%. Os onze métodos de imputação utilizados foram: duas variantes de KMC; duas técnicas baseadas em *MostCommon*; duas versões de k-NN; BPCA; SVD; SVM; EM; LLS. A métrica de avaliação aplicada foi o RMSE, o que permitiu efetuar uma análise PAC, mas os autores não efetuaram nenhuma análise DAC. As simulações mostraram que os melhores resultados foram obtidos para as duas versões de k-NN, SVM e as três versões implementados pelos investigadores baseadas em *Fuzzy Rough Sets*. Os investigadores concluem que a sua proposta é uma aposta igualmente válida para lidar com dados incompletos, mas necessitam de otimização com objetivo de reduzir o tempo e esforço computacional.

No trabalho de Deb e Liew [67] foi implementada uma técnica de imputação baseada em DT e a sua performance é comparada com outras técnicas. Foram utilizados quatro *datasets* com informações sobre acidentes rodoviários, o número de instâncias e variáveis varia entre $[13889, 89679]$ e $[11, 43]$, respetivamente. Os autores efetuaram a remoção de todas as instâncias incompletas nos conjuntos de dados e só depois efetuaram a remoção de dados com oito abordagens diferentes e quatro MR diferentes (2%, 4%, 8% e 10%). As técnicas de imputação aplicadas de modo a realizar a comparação com a proposta desse artigo foram: k-NN, DT e *Framework for Imputing Missing Values Using Co-Appearance, Correlation and Similarity Analysis* (FIMUS). Os autores analisaram o desempenho com

RMSE para as variáveis numéricas e ACC para as variáveis categóricas. Os investigadores concluíram que a sua proposta de imputação apresenta melhores resultados do que as outras técnicas implementadas. No entanto, os resultados mostraram também que o FIMUS e DT apresentam melhores resultados que o k-NN.

No artigo de Kumar et al. [68] foi proposta uma técnica de imputação baseada na SVD, *t-test* e *fold-change*, para ser aplicada em *datasets* com a finalidade de identificar biomarcadores metabólicos. Os autores efetuam a comparação da sua proposta utilizando um *dataset* real incompleto com 1388 instâncias e 57 variáveis, mas antes da sua utilização foram removidos os padrões incompletos. Além desse conjunto de dados reais, os investigadores simulam 100 *datasets* com 500 instâncias (não é descrito o número de variáveis). E todos os *datasets* são sujeitos à introdução de *outliers* (3% a 15%) e à remoção de 10%, 15% e 20% dos dados, onde metade dos valores são removidos de forma MCAR e a outra metade tirados de entre os valores menores (mínimo por ordem crescente). Os algoritmos de imputação implementados foram: k-NN, *Random Forest* e imputação *Zero* (valores em falta são substituídos por zeros). Em seguida, foi efetuada uma análise PAC com a métrica RMSE. No entanto, os investigadores também efetuam uma comparação de desempenho baseada na classificação, onde o classificador usado foi SVM e para a avaliação foram calculadas as seguintes medidas: ACC, SEN, SPEC, CE e AUC. Os resultados demostram que para os dados sintéticos sem *outliers* o k-NN apresenta os melhores resultados de RMSE, mas para os restantes casos a proposta dos autores foi sempre a melhor técnica de imputação. Além disso, os resultados obtidos mostram que k-NN e *Random Forest* têm comportamentos similares e sempre superiores aos resultados de *Zero*.

Em todos os trabalhos anteriormente citados, as técnicas de imputação são frequentemente avaliadas em termos de CE, e os efeitos que têm na distribuição de dados são ignorados. Além disso, todas as variáveis são imputadas com a mesma técnica, sem considerar a possibilidade de que algumas técnicas podem ter desempenhos diferentes para cada variável, no mesmo *dataset*. Alguns estudos utilizam *datasets* incompletos e apresentam MR muito diferentes em cada variável, e o efeito que esse facto pode originar não é analisado. Outro detalhe comum entre muitos dos artigos citados que utilizam conjuntos de dados completos é a geração de valores em falta de modo MCAR.

Neste trabalho, realizamos um estudo sobre a influência da distribuição de dados na imputação de dados em falta. O objetivo é avaliar como diferentes técnicas de imputação se comportam para diferentes distribuições de dados, o que, pelo nosso conhecimento, nunca foi realizado. A seleção dos diferentes métodos de imputação teve como base dois crité-

rios: algoritmos de imputação que seguem abordagens diferentes (*tree-based, distance-based, statistical, kernel-based, neural networks*); algoritmos mais frequentes na literatura e que tenham apresentado bons resultados. A Tabela 3.1 permite obter uma visão geral das informações principais da literatura citada anteriormente e perceber a escolha das técnicas de imputação utilizadas nas nossas simulações (DT, k-NN, Média/Moda, SVM e SOM).

Tabela 3.1: Sumário dos trabalhos relacionados com dados incompletos

	Algoritmos			***Datasets***		
Publicação	Imputação	Classificação	**Métricas**	Contexto	Variáveis	Instâncias
Nanni, Lumini e Brahnam [62]	*Dissimilarity*; EM; Média; ANN; k-NN; BPCA; *InPaint*; *Learn*$^{++}$*MF*	IDE; SVM	*Rank*; AUC	Saúde	[8, 32]	[155, 768]
Kang [63]	Média; *Hot-deck*; k-NN; ECM; KMC; MoG; LLR	k-NN; LLR; ANN; LR; CART	ACC; RMSE	Várias	[4, 60]	[150, 14429]
Aisha, Adam e Shohaimi [19]	Média/Moda; EM; SVM; k-NN; KMC; SVD; LLS	NB; TAN; BAN; GBN	ACC	Saúde	19	155
García-Laencina et al. [12]	Média/Moda; EM; k-NN	k-NN; CART; LR; SVM	ACC; SEN; SPEC; AUC	Saúde	16	399
Jerez et al. [5]	Média/Moda; MLP; MI; SOM; k-NN; *Hot-deck*	ANN	AUC	Saúde	8	3679
Mostafizur e Davis [18]	Média/Moda; RDR; DT; SVM; FURIA	KMC	ACC; SEN; SPEC	Saúde	26	832
García-Laencina, Sancho-Gómez e Figueiras-Vidal [6]	k-NN	Não aplicável	r; D_{KS}; MSE	Sintético	[a]	1500
	MLP; SOM; k-NN; EM	ANN	CE	Várias	[3, 28]	[871, 2800]
Rahman e Islam [64]	DT; EM; LLS;	Não aplicável	d_2; MAE; r; RMSE	Várias	[2, 11]	[398, 32561]
Abreu et al. [65]	k-NN	*TreeBagger*; *LPBoost*; Sub-Espaços	*Rank*; ACC	Saúde	15	847
Davis e Mostafizur [9]	Média/Moda; DT; RDR; SVM; k-NN; FURIA;	DT; KMC; k-NN; ANN; FURIA	ACC; SEN; SPEC	Saude	22	823

Continua na página seguinte...

[a] Desconhecido.

Tabela 3.1: Continuação da página anterior.

	Algoritmos			***Datasets***		
Publicação	Imputação	Classificação	**Métricas**	Contexto	Variáveis	Instâncias
Amiri e Jensen [66]	*Fuzzy-Rough*; EM; *MostCommon*; KMC; k-NN; BPCA; SVD; SVM; LLS	Não aplicável	RMSE	Várias	[3, 60]	[106, 2201]
Deb e Liew [67]	DT; k-NN; FIMUS;	Não aplicável	RMSE; ACC	Automóvel	[11, 43]	[13889, 89679]
Kumar et al. [68]	k-NN;*RandomForest*; *Zero*; Proposta	SVM	ACC; SEN; AUC; SPEC; CE; RMSE	Metabolismo	57	[500, 1388]

3.2 Problema de Dados Não Balanceados

Van Hulse, Khoshgoftaar e Napolitano [69] elaboraram uma análise detalhada do desempenho de várias técnicas de classificação e de amostragem em conjuntos de dados não balanceados. Para esse efeito, fizeram uso de 35 *datasets* reais com $[214, 20000]$ instâncias, $[4, 65]$ variáveis e $[1.86, 74.19]$ em termos de Taxa de Desequilíbrio (IR). Os autores, para realizar a validação cruzada 5-*folds*, procederam à divisão dos *datasets* antes de aplicar qualquer técnica de amostragem. Das sete técnicas de pré-processamento implementadas, quatro eram técnicas de sobre-amostragem: SMOTE, *Borderline*-SMOTE, CBO e ROS. As classificações foram realizadas com C4.5, SVM, NB, k-NN, MLP, *RandomForest* (RF), LR e Função de Base Radial (RBF). Os desempenhos foram medidos com ACC, AUC, SEN, *G-Mean*, *F-Measure* e *Rank*. Os investigadores concluíram que o melhor método de amostragem depende do classificador usado, por exemplo para C4.5 o melhor método foi uma técnica de sub-amostragem e para LR o melhor método foi ROS. Além disso, nos classificadores k-NN, LR e NB, em termos de AUC os métodos de pré-processamento não revelaram melhorias muito significativas.

A fim de estudar os problemas de dados não balanceados e dados ruidosos, Van Hulse e Khoshgoftaar [70] realizaram a análise do desempenho da classificação em *datasets* (nos quais inseriram ruído) para vários classificadores e técnicas de pré-processamento. Os sete

datasets reais utilizados nesta experiência continham $[302, 12964]$ instâncias, $[8, 64]$ variáveis e IR a variar entre $[4.03, 38.53]$. Os investigadores antes de utilizarem os *datasets* recolhidos na experiência, procederam à sua divisão de forma a possibilitar a validação cruzada 10-*folds*. Os algoritmos de sobre-amostragem selecionados foram ROS, SMOTE, CBO e *Borderline*-SMOTE e as técnicas de classificação usadas foram k-NN, C4.5, NB, MLP, LR, SVM e RF. A métrica de avaliação da classificação aplicada foi a AUC. Os resultados mostraram que, dentro das técnicas de sobre-amostragem: nos conjuntos de dados sem ruído, *Borderline*-SMOTE apresenta os melhores resultados; para ruído inserido na classe minoritária, o SMOTE teve o melhor desempenho; para ruído inserido em ambas as classes, o *Borderline*-SMOTE mostrou-se o mais capaz. Em termos de classificadores, os resultados mostraram que o NB foi o menos sensível a ruído e mais estável no seu desempenho.

No trabalho de Verbiest et al. [71] foi apresentada uma nova abordagem de amostragem para solucionar o problema de dados não balanceados e o seu desempenho foi comparado com métodos estado-da-arte. Os autores resolveram utilizar dados sintéticos com dimensão reduzida, duas variáveis, e onde a classe minoritária assume três estruturas topológicas diferentes. Deste modo, foram construídos 30 *datasets*, contendo 600 ou 800 instâncias, 5 ou 7 IR e 5 diferentes percentagens de ruído. Antes de ter sido feito o pré-processamento, os investigadores procederam à divisão dos dados em 5-*folds*, de forma a posteriormente ser realizada a validação cruzada. SMOTE, SMOTE+TL, SMOTE+ENN, *Borderline*-SMOTE (1 e 2), *Safe-Level*-SMOTE, SPIDER e SPIDER2 foram utilizados na avaliação e comparação de desempenho dos classificadores. A classificação dos *datasets* foi elaborada com k-NN, através da AUC e *Rank*. Os resultados mostraram que o SMOTE+TL e SMOTE+ENN apresentam uma melhoria do desempenho de classificação superior à do SMOTE. De salientar que nesta experiência, o SMOTE+TL é a técnica que obteve melhores resultados dentro das técnicas de sobre-amostragem utilizadas.

García et al. [72] elaboraram uma configuração experimental para avaliar o desempenho da classificação de três variações de SMOTE que propuseram e de sete técnicas de sobre-amostragem que consideram estado-da-arte. Os métodos de sobre-amostragem implementados foram AHC, SMOTE, ADASYN, ADOMS, ROS, *Borderline-SMOTE* e *Safe-Level-SMOTE*. Foram utilizados 39 *datasets* reais que continham uma IR entre $[1.82, 39.11]$, $[4, 19]$ variáveis e $[150, 5472]$ instâncias. Esses conjuntos de dados antes de serem balanceados foram divididos em 5-*folds*, com o objetivo de realizar a validação cruzada. A classificação foi realizada através de k-NN, C4.5 e MLP, e a sua avaliação feita com a AUC e *Rank*. Excluindo os resultados obtidos nas três propostas dos autores (que foram as melhores):

SMOTE obteve o melhor valor médio de AUC (0.830) e *Rank* no classificador k-NN; no classificador C4.5 a melhor técnica foi ADASYN com a AUC de 0.834; AHC e ADASYN obtiveram os melhores valores de AUC 0.845 e 0.844, respetivamente, nas classificações de MLP.

No artigo de Soufan et al. [73] analisou-se o desempenho de vários classificadores em contextos de dados não balanceados, pré-processados com técnicas estado-da-arte e uma abordagem proposta. Os nove *datasets* reais utilizados continham informações que permitem identificar atividade biológica em inúmeros compostos químicos e apresentavam uma IR entre $[2, 377]$, $[206, 184541]$ instâncias e 2940 variáveis. As técnicas de sobre-amostragem implementadas foram SMOTE e MWMOTE. Os autores realizaram antecipadamente a divisão dos *datasets* em 5-*folds*, de maneira a realizar a validação cruzada. SVM, k-NN, NB e RF foram os classificadores implementados neste trabalho e as métricas de avaliação do desempenho foram a SEN, precisão (VPP), SPEC, AUC, *F-1*, *F-0.5* e *Rank*. Das várias conclusões e resultados desse artigo, destacou-se o desempenho superior do SMOTE em relação ao MWMOTE.

Na pesquisa de Loyola-González et al. [74] estudou-se o impacto do uso de técnicas de amostragem no desempenho de um classificador baseado em *contrast pattern*. Nas simulações deste artigo foram utilizados 95 *datasets* reais com $[3, 34]$ variáveis, $[101, 4174]$ instâncias e IR entre $[1.82, 129.44]$. Antes de realizar o pré-processamento, os investigadores efetuaram a divisão dos conjuntos de dados de modo a implementar a validação cruzada 5-*fold*. No conjunto de técnicas de sobre-amostragem usadas neste artigo foram incluídas as seguintes: AHC; ADASYN; SMOTE; ADOMS; ROS; *Borderline*; SMOTE+ENN; SMOTE+TL; *Safe-Level*; SPIDER. O desempenho da classificação foi avaliado com ACC, AUC e *Rank*. Nos resultados, em termos de *Rank* para ambas as métricas (AUC e ACC), SMOTE+TL foi a melhor técnica de sobre-amostragem. Em relação ao número de vitórias no desempenho de classificação, ROS foi a melhor técnica de sobre-amostragem, tendo apresentado no final, 15 melhores valores de AUC. Por fim, a maior ocorrência de valores de AUC superiores a 0.9 nas técnicas de sobre-amostragem foi obtida pelo SMOTE+TL com 40 casos.

Ah-Pine e Soriano-Morales [75] exploraram o uso de sobre-amostragem com dados sintéticos em conjuntos de dados não balanceados. Os três *datasets* utilizados eram reais e continham $[1906, 4519]$ instâncias, $[1569, 3918]$ variáveis e apresentavam *IR* entre $[1.68, 3.14]$. Os autores efetuaram o balanceamento dos dados da classe minoritária em diferentes proporções, usando as seguintes técnicas: SMOTE, ADASYN e *Borderline*-SMOTE. A classificação foi efetuada com CART e LR, e os resultados foram avaliados em termos de *G-Mean*, *F-1* e

ACC. Os investigadores indicam a utilização da validação cruzada 5-*folds*, mas a fase na qual a realizaram não é explicada. De forma geral, as técnicas de sobre-amostragem aplicadas obtiveram melhores resultados de *G-Mean* e *F-1*, no entanto obtiveram menores ACC. No entanto, as simulações não permitiram identificar a melhor técnica no global.

Alejo et al. [76] desenvolveram uma técnica de amostragem para lidar com dados não balanceados e avaliaram o desempenho da classificação para diferentes técnicas de amostragens estado-da-arte. Os investigadores inicialmente selecionaram cinco *datasets* reais *multiclass* e transformaram-nos em *datasets* binários. No final desse procedimento, obtiveram 35 *datasets* com $[1470, 10944]$ instâncias, $[4, 38]$ variáveis e IR entre $[1.05, 46.75]$. Antes de elaborarem o pré-processamento dos conjuntos, estes foram divididos para posteriormente ser realizada a validação cruzada 10-*folds*. ADASYN, SMOTE, ADOMS, ROS, *Borderline*-SMOTE, SMOTE+ENN, SMOTE+TL, *Safe-Level*-SMOTE, SPIDER e SPIDER2 foram as técnicas de sobre-amostragens utilizadas. O classificador implementado foi ANN e as métricas de avaliação foram AUC e *Rank*. Todos os métodos de sobre-amostragem obtiveram melhores *Ranks* médios do que os *datasets* originais, e ROS obteve o melhor *Rank* médio. Além disso, segundo os investigadores o IR não está relacionado com o desempenho dos algoritmos.

Na pesquisa de Zhu et al. [77] analisou-se o desempenho de várias técnicas de amostragem em *datasets* não balanceados. Os 11 *datasets* reais utilizados continham entre $[2019, 100426]$ instâncias, $[9, 231]$ variáveis e a IR variava entre $[5.90, 54.56]$. As técnicas de sobre-amostragem implementadas foram ADASYN, SMOTE, *Borderline*-SMOTE, SMOTE+ENN, SMOTE+TL e MWMOTE. Antes de realizarem o pré-processamento, os investigadores realizaram a divisão dos conjuntos de dados necessária para a validação cruzada 2-*folds*, que foi implementada para avaliar o desempenho da classificação. Os classificadores usados nesta experiência foram C4.5, SVM, LR e RF, e a avaliação dos resultados foi feita usando a AUC e os *Ranks*. Os resultados revelaram que a eficácia das técnicas de pré-processamento depende do classificador, por exemplo para C4.5 o *Borderline*-SMOTE apresenta o melhor desempenho, contudo para SVM os melhores foram ROS e SMOTE+ENN.

Uralde et al. [78] propuseram um sistema automático de extração de conhecimento de nano-estruturas e no seu desenvolvimento elaboram a comparação do desempenho da classificação com e sem o uso da técnica SMOTE no conjunto de dados não balanceados obtido. Esse *dataset*, que foi obtido de imagens a partir de microscópio, contém 26 variáveis, 266 instâncias e 4 classes, compostas por 9, 51, 86 e 120 instâncias, resultando numa IR compreendida entre $[1.40, 13.33]$. Após o balanceamento com SMOTE, os investigadores realizaram

a validação cruzada 10-*folds* para os classificadores SVM, NB, k-NN e C4.5, nos conjuntos originais e balanceados. As métricas de avaliação implementadas foram AUC e ACC, e os resultados obtidos demonstraram que os dados balanceados obtiveram sempre melhores desempenhos.

Al-Bahrani, Agrawal e Choudhary [79] desenvolveram um processo para aumentar a ACC num *dataset* médico não balanceado, onde foi utilizado o SMOTE. O *dataset* inicial multi-classe com 105133 instâncias e 65 variáveis, foi transformado em três *datasets* com o mesmo número de instâncias e variáveis (13), classes binárias e com IR entre $[1.38, 3.66]$. Os investigadores realizam a validação cruzada 10-*folds* dos vários classificadores com as métricas de avaliação ACC e AUC, após o pré-processamento. Os classificadores utilizados pelos investigadores foram C4.5, RF, LR, *Alternating Decision Tree* (ADTree) e *Reduced Error-Prunning Tree* (REPTree). Em termos de AUC, os *datasets* balanceados obtiveram os melhores resultados para todos os classificadores (o que também se verificou em termos de ACC).

Na pesquisa de Naseriparsa e Kashani [80] foi proposta a combinação de uma técnica de redução de dimensionalidade (PCA) com uma técnica de sobre-amostragem (SMOTE) num conjunto de dados não balanceados e testado o desempenho de um classificador NB em termos de ACC, VPP, SEN e SPEC. O *dataset* real utilizado contém 56 variáveis e 32 instâncias, e é composto por 3 classes com 9, 13 e 10 instâncias (IR varia entre $[1.30, 1.44]$). Após a aplicação do algoritmos de PCA, o *dataset* passou a conter apenas 18 variáveis (componentes). O SMOTE foi realizado em 3 execuções, sendo que em cada execução foram inseridos novos exemplos sintéticos na classe minoritária (que varia em cada execução). Os autores não referenciam o modo como foi elaborada a validação do desempenho da classificação. Os resultados mostraram que após a redução de dimensionalidade (sem aplicar a sobre-amostragem) o desempenho da classificação piora para todas as métricas em relação ao *dataset* original. Aquando do balanceamento dos dados reduzidos com SMOTE, o desempenho melhora e registam-se melhores resultados do que no *dataset* original.

Na pesquisa de Fergus et al. [81] avaliou-se a influência da aplicação de SMOTE no desempenho da classificação. Neste artigo são utilizados dois *datasets* reais com a IR entre $[6.89, 7.89]$, $[4, 15]$ variáveis e $[169, 300]$ instâncias. As técnicas de classificação utilizadas foram: SVM, LR, k-NN e DT. Após o balanceamento dos *datasets*, os investigadores realizaram duas configurações de validação: na primeira os conjuntos de dados foram divididos num conjunto de treino e de teste, com 80% e 20% dos dados, respetivamente; na segunda realizou-se a validação cruzada 5-*folds*. Em ambas as configurações, as métricas de avaliação

do desempenho de classificação utilizadas foram o CE, SEN, SPEC e AUC. Os resultados da classificação dos datasets não balanceados foram baixos, não havendo nenhum classificador que atingisse mais de $0,61$ de AUC e a maioria apresentou SEN igual a 0. No caso dos dados balanceados os resultados foram melhores: SEN entre $80-90\%$ e AUC superiores a $0,70$ na maioria dos classificadores.

Helal et al. [82] analisaram algumas técnicas de *data mining* e propuseram dois processos de avaliação para dados não balanceados. O *dataset* real utilizado contém 4 classes com IR entre $[3.15, 18.61]$, 6 variáveis e 1728 instâncias. No decorrer das experiência é utilizado este *dataset* com um menor número de instâncias, de modo a perceber a influência no desempenho do classificador. A classificação foi realizada com DT, SVM, k-NN e RF. Os autores utilizaram dois procedimentos diferentes para realizar a avaliação do desempenho dos classificadores: o primeiro consiste numa divisão do *dataset* num conjunto de treino com 70% dos exemplos e de teste com os restantes exemplos, e na avaliação do desempenho da classificação com a métrica *G-Mean*; o segundo consiste numa validação cruzada 10-*folds* e na utilização da ACC como métrica de avaliação. A sobre-amostragem foi executada antes de realizar esses procedimentos. Nas simulações mostrou-se que o primeiro procedimento obtinha pior resultados do que o segundo procedimento. Além disso, realizou-se a comparação dos resultados de treino e validação cruzada para o conjunto de dados original e balanceado, com número de instâncias diferentes. Esta comparação mostra que o número de instâncias necessárias para obter bons resultados de classificação é menor para o *dataset* balanceado.

Rani, Ramadevi e Lavanya [83] investigaram a influência da técnica de sobre-amostragem SMOTE (com várias percentagens) e da técnica de redução de dimensão PCA no desempenho de classificação em problemas de dados não balanceados. Neste artigo foram usados quatro *datasets* reais com IR entre $[1.60, 3.21]$, número de instâncias $[198, 699]$ e 10 variáveis. Após o pré-processamento dos dados, efetuaram a validação cruzada 10-*folds* com ACC para os seguintes classificadores: C4.5, SVM, k-NN, LR e RF. Os autores apenas apresentaram os resultados dos conjuntos pré-processados, como tal, não é possível concluir-se se o SMOTE teve melhor desempenho do que a versão não balanceada. Contudo, foi conclusivo que em 60% dos casos a ACC dos conjuntos pré-processados com SMOTE foi superior à ACC obtida para os conjuntos transformados com PCA. Outra conclusão deste trabalho foi que para os classificadores k-NN, C4.5 e RF os dados balanceados por SMOTE tiveram melhores resultados comparativamente à abordagem de PCA.

No trabalho de Oppedal et al. [84] foram estudadas técnicas de extração de conhecimento a partir de imagens médicas. No decorrer do processo experimental obtiveram-se

conjuntos de dados não balanceados e de modo a contornar este problema, foram testadas duas configurações, mas apenas uma delas invoca uma técnica de amostragem na fase de pré-processamento, o SMOTE. Do *dataset* original que contém três classes com 36, 58 e 16 instâncias, foram criados vários subconjuntos com várias combinações entre as classes, criando *datasets* para análise com o número de instâncias a variar entre $[52, 110]$ e a IR $[1.61, 4, 27]$. O número total de variáveis que compõem esses conjuntos de dados não é especificado, apenas mencionado como um valor elevado. A classificação destes conjuntos foi realizada recorrendo à validação cruzada 10-*folds* com RF após o pré-processamento do dados, e o seu desempenho foi avaliado em termos de ACC, VPP e SEN. Os resultados mostraram que SMOTE melhorou o desempenho da classificação, principalmente nos subconjuntos com *IR* mais altas.

Luengo et al. [85] avaliaram a utilidade das medidas de complexidade $F1$, $N4$ e $L3$ na avaliação de dois algoritmos de sobre-amostragem em dados não balanceados: SMOTE e SMOTE+ENN. Os autores começaram por selecionar 44 *datasets* reais com IR entre $[1.82, 128.87]$ (apenas um tem IR superior a 40), com número diferente de instâncias $[150, 5472]$ e de variáveis $[3, 19]$, e mediram a complexidade de cada *dataset*. De seguida, realizaram a divisão dos conjuntos de dados em *5-folds*, de modo a implementar a validação cruzada, e efetuaram amostragem nos conjuntos de treino obtidos para balancear as suas classes. Em seguida, foi realizada a classificação com C4.5 e PART e avaliado o desempenho dos classificadores pelo valor médio da AUC para os conjuntos de treino e de teste, individualmente. Os resultados das classificações foram utilizados como referências para a obtenção de intervalos de complexidade onde os classificadores mostraram bons e maus desempenhos: um bom desempenho foi definido como uma AUC média no teste de pelo menos 0.8 e a diferença entre os resultados de treino e teste menor que 0.1 (os autores consideram esse fenómeno ausência de *overffiting*); um mau desempenho foi atribuido à presença de *overfitting* ou baixos valores médios da AUC no teste. Os investigadores foram capazes de extrair regras para identificar bons e maus conjuntos de dados para cada uma das técnicas de amostragem. Em particular, a $F1$ provou ser uma métrica muito discriminativa, onde valores acima de 1.469 e abaixo de 0.366 sugerem um desempenho bom e mau, respetivamente, para ambos os algoritmos C4.5 e PART.

A importância da execução da validação cruzada de maneira apropriada em problemas com dados não balanceados foi pela primeira vez salientada no artigo de Blagus e Lusa [86]: neste estudo foram analisados os efeitos da utilização indevida da validação cruzada quando combinada com técnicas de amostragem. Os autores utilizaram numa primeira abordagem

datasets sintéticos com IR de 9, número de instâncias $[100, 1000]$ e de variáveis $[10, 100]$, e na segunda abordagem usaram 16 *datasets* reais (apenas 10 eram públicos) com IR a variar entre $[1.14, 33.73]$, o número de instâncias $[32, 17307]$ e variáveis $[5, 22283]$. A amostragem foi elaborada antes da validação cruzada e durante a validação (após efetuar a divisão em *k-folds*) com ROS, *Random Under-Sampling* (RUS) e SMOTE. O classificador implementado foi o CART e o desempenho das classificações foi medido com as seguintes métricas: ACC, SEN, SPEC, *F-1* e *G-Mean*. Os resultados mostram que a validação cruzada após o pré-processamento proporciona estimativas sobre-otimistas para ROS e SMOTE, enquanto o RUS mostrou um desempenho invariante para ambos os procedimentos. Apesar deste artigo apresentar uma abordagem interessante ao problema do sobre-otimismo, algumas questões ficam por responder na configuração experimental. A primeira é que apesar de serem fornecidos os resultados dos dois procedimentos de validação cruzada e ser mostrado o desempenho sobre-otimista para a validação após a aplicação da técnica de sobre-amostragem, não existe uma análise relativa à presença de *overfitting*, uma vez que só é realizada a comparação dos resultados entre as duas abordagens, ao invés de uma análise completa do desempenho nos conjuntos de treino e teste em cada procedimento. Além disso, alguns tópicos discutidos na parte da avaliação não são concisos. Não existe grande variabilidade e distribuição uniforme em termos de IR nos *datasets* públicos usados pois oito têm a taxa entre $[1.14, 3.84]$ e os dois restantes têm uma taxa de 24 e 33.73, e como tal, apesar dos investigadores terem observado estimativas enviesadas para *datasets* pequenos, não é claro que os resultados estejam relacionados com o número de instâncias de cada conjunto ou em combinação com a Taxa de Desequilíbrio. Ademais, não é possível elaborar uma análise mais detalhada devido à inexistência dos resultados individuais de cada *dataset*. Por fim, os autores também relacionam o desempenho obtido na validação cruzada com a dificuldade da tarefa de classificação, afirmando que quando a tarefa de classificação é "fácil" a diferença entre as abordagens de validação cruzada é pequena, contudo não são apresentadas quaisquer métricas de complexidade para suportar essa afirmação.

Esta dissertação pretende melhorar o trabalho inicial de Blagus e Lusa de modo a considerar mais conjuntos de dados com uma maior gama representativa de IR, um maior número de técnicas de sobre-amostragem e algoritmos de classificação, incluir uma análise sobre o risco de sobre-otimismo, distinguindo-o do risco de *overfitting* e analisar a influência das técnicas de sobre-amostragem na complexidade dos *datasets*, relacionando-a com o desempenho dos classificadores.

Tabela 3.2: Sumário dos trabalhos relacionados com dados não balanceados

	Algoritmos			***Datasets***		
Publicação	Sobre-amostragem	Classificação	**Métricas**	Variáveis	Instâncias	IR
Van Hulse, Khoshgoftaar e Napolitano [69]	SMOTE; ROS; CBO; *Borderline*;	k-NN; C4.5; NB; MLP; LR; SVM; RF; RBF	AUC; *F-1*; *G-Mean*; ACC; *Rank*; SEN	[4, 65]	[214, 20000]	[1.86, 74.19]
Van Hulse e Khoshgoftaar [70]	SMOTE; ROS; CBO; *Borderline*;	k-NN; C4.5; NB; MLP; LR; SVM; RF; RBF	AUC	[8, 64]	[302, 12964]	[4.03, 38.53]
Verbiest et al. [71]	SMOTE; *Borderline*; SMOTE+ENN; SMOTE+TL; *Safe-Level*; SPIDER	k-NN	AUC; *Rank*	2	[600, 800]	[5, 7]
García et al. [72]	AHC; SMOTE; ADASYN; ADOMS; ROS; *Borderline*; *Safe-Level*	k-NN; MLP; C4.5	AUC; *Rank*	[4, 19]	[150, 5472]	[1.82, 39.11]
Soufan et al. [73]	SMOTE; MWMOTE	SVM; k-NN; NB; RF	AUC; *Rank*; VPP; *F-0.5*; *F-1*; SEN; SPEC	2940	[206, 184641]	[2, 377]
Loyola-González et al. [74]	AHC; ADASYN; SMOTE; ADOMS; ROS; *Borderline*; SMOTE+ENN; SMOTE+TL; *Safe-Level*; SPIDER	Baseado em *Contrast Pattern*	ACC; AUC; *Rank*	[3, 34]	[101, 4174]	[1.82, 129.44]
Ah-Pine e Soriano-Morales [75]	ADASYN; SMOTE; *Borderline*	LR; CART	*G-Mean*; *F-1*; ACC	[1569, 3918]	[1906, 4519]	[1.68, 3.14]

Continua na página seguinte...

Tabela 3.2: Continuação da página anterior.

Publicação	**Algoritmos**		**Métricas**	***Datasets***		
	Sobre-amostragem	Classificação		Variáveis	Instâncias	IR
Alejo et al. [76]	ADASYN; SMOTE; ADOMS; ROS; *Borderline*; SMOTE+ENN; SMOTE+TL; *Safe-Level*; SPIDER	ANN	AUC; *Rank*	[4, 38]	[1470, 10944]	[1.05, 46.75]
Zhu et al. [77]	ADASYN; SMOTE; *Borderline*; SMOTE+ENN; SMOTE+TL; MWMOTE	LR; SVM; C4.5; RF	AUC; *Rank*	[9, 231]	[2019, 100462]	[5.90, 54.56]
Uralde et al. [78]	SMOTE	NB; k-NN; SVM; C4.5	ACC; AUC	26	266	[1.40, 13.33]
Naseriparsa e Kashani [80]	SMOTE	NB	VPP; ACC; SEN; SPEC	52	32	[1.3, 1.44]
Al-Bahrani, Agrawal e Choudhary [79]	SMOTE	LR; ADTree; REPTree; RF; C4.5	ACC; AUC	13	105133	[1.38, 3.66]
Fergus et al. [81]	SMOTE	k-NN; LR; SVM; DT	AUC; CE; SEN; SPEC	[4, 15]	[169, 300]	[6.89, 7.89]
Helal et al. [82]	SMOTE	DT; SVM; k-NN; RF	*G-Mean*; ACC	6	1728	[3.15, 18.61]
Rani, Ramadevi e Lavanya [83]	SMOTE	C4.5; SVM; k-NN; RF; LR	ACC	10	[198, 699]	[1.60, 3.21]
Oppedal et al. [84]	SMOTE	RF	VPP; ACC; SEN	[a]	[52, 110]	[1.61, 4.27]
Luengo et al. [85]	SMOTE; SMOTE+ENN	C4.5; PART	AUC	[3, 19]	[150, 5472]	[1.82, 128.87]
Blagus e Lusa [86]	ROS; SMOTE	CART	AUC; ACC; *F-1*; *G-Mean*; SEN; SPEC	[5, 22283]	[32, 17307]	[1.14, 33.73]

[a] Desconhecido

4 Problema de Dados Incompletos

Na literatura, vários investigadores tentam resolver o problema de dados incompletos apresentando soluções que analisam o desempenho de vários algoritmos de *data mining*, recorrendo a estratégias *brute-force*: é efetuada uma procura exaustiva e sistemática do melhor algoritmo para realizar a imputação em cada conjunto de dados incompletos (novos *datasets* passarão pelo mesmo processo). Contudo, este tipo de abordagem está longe de ser ideal, não só porque apresenta um custo computacional elevado mas principalmente porque não é generalizável para novos contextos, sendo necessário efetuar novas simulações.

Como discutido nos trabalhos relacionados, a comunidade científica utiliza frequentemente o Erro de Classificação (CE) como métrica de avaliação do desempenho das técnicas de imputação, o que pode causar conclusões enviesadas: a melhor técnica de imputação no desempenho dos classificadores, não é imperativamente a melhor técnica a preencher os dados em falta com os verdadeiros valores e a manter a distribuição dos dados originais. Além disso, todas as variáveis são imputadas com a mesma técnica, não sendo considerada a influência da distribuição de dados que cada variável segue no desempenho das técnicas de imputação.

Deste modo, neste capítulo apresentamos uma abordagem que, atendendo à distribuição que os dados seguem, pretende identificar a melhor técnica de imputação a utilizar para lidar com o problema. Para isso, foi criada uma arquitetura que será descrita de seguida.

4.1 Arquitectura da Solução Desenvolvida

A metodologia utilizada é constituída por cinco etapas fundamentais: Recolha de Dados, Distribuição de Melhor Ajuste, Geração de Dados em Falta, Imputação de Dados e Avaliação (Figura 4.1).

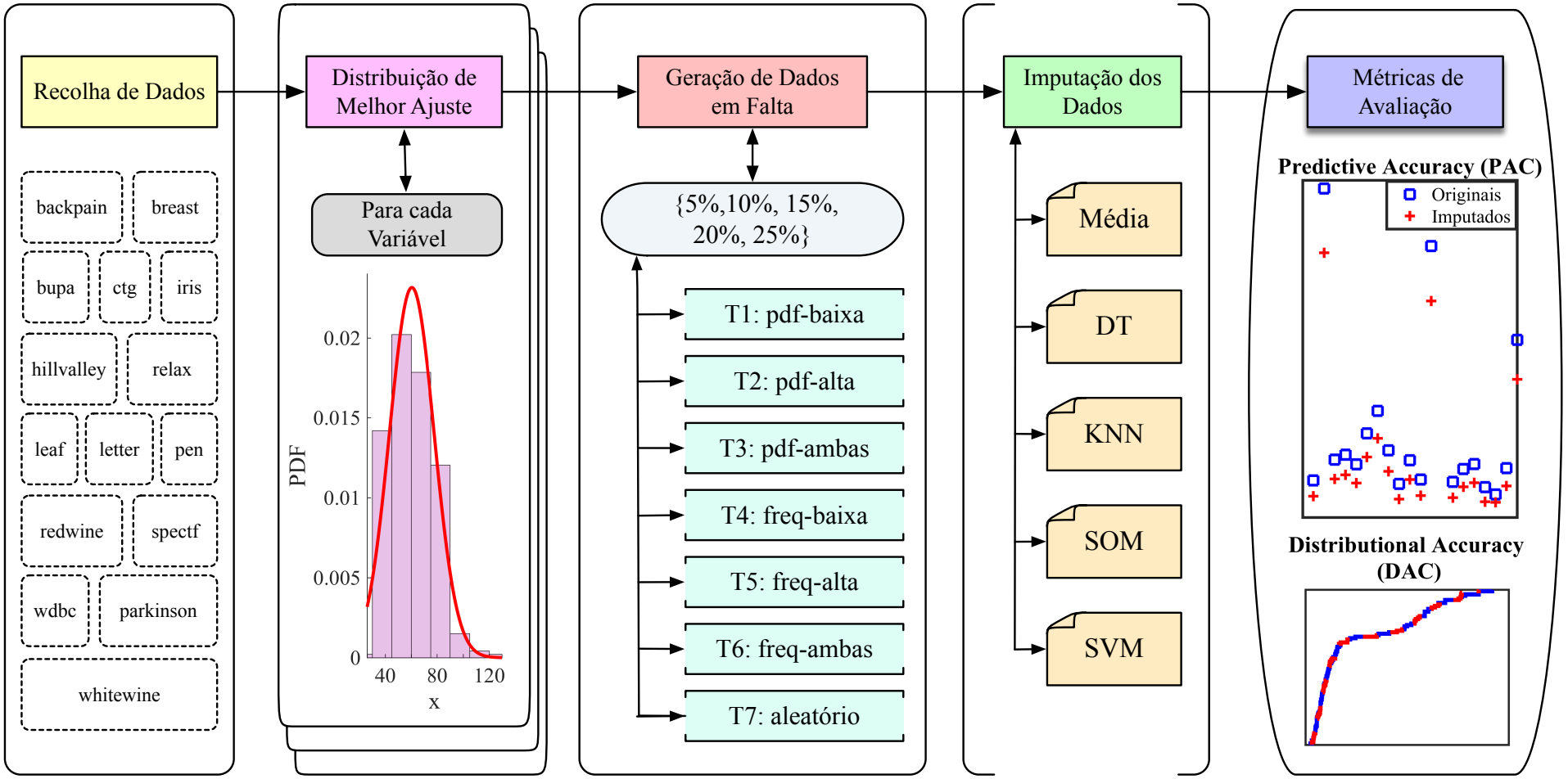


Figura 4.1: Metodologia utilizada para o problema de dados incompletos.

4.1.1 Recolha de Dados

O critério utilizado na escolha dos dados, consistiu na possibilidade de replicação das experiências por outros investigadores. Nesse sentido, a escolha de *datasets* existentes em repositórios livres *online* foram uma prioridade (esta estratégia também foi seguida no segundo problema, dados não balanceados, descrito no Capítulo 5). Para além disso, foram ainda considerados os seguintes critérios: *datasets* completos, com apenas variáveis contínuas, englobando diferentes contextos, número de instâncias e variáveis.

Os 15 *datasets* utilizados neste trabalho estão disponíveis no *UCI Machine Learning Repository* [88] e no *Kaggle* [87] e as suas principais características estão sumariadas na Tabela 4.1.

O **breast** é o *dataset* com menor número de instâncias (106) e o **letterhalf** com o maior (5000). Em termos de variáveis, o **iris** e o **hillvalley** têm o menor e maior número, com 4 e 100 variáveis, respetivamente.

Tabela 4.1: Sumário das propriedades dos *datasets* escolhidos

Datasets	**Problema**	**Instâncias**	**Variáveis**
backpain [87]	Detetar dor de costas anormal	310	12
breast [88]	Identificar carcinoma da mama	106	9
bupa [88]	Detetar problemas de alcoolismo	345	6
ctg [88]	Detetar patologias cardíacas em cardiotocografias	2126	21
hillvalley [88]	Detetar colinas e vales	1212	100
iris [88]	Distinguir entre diferentes espécies de íris (planta)	150	4
leaf [88]	Distinguir entre diferentes espécies de folhas	340	14
letterhalf[1] [88]	Identificar letras do alfabeto	5000	16
parkinson [88]	Diagnosticar casos de parkinsonismo	195	22
pen [88]	Identificar dígitos manuscritos	3498	16
redwine [88]	Classificar a qualidade de vinho tinto	1599	11
relax [88]	Distinguir entre o estado relaxado e agitado	182	12
spectf [88]	Detetar anomalias em imagens SPECT	267	44
wdbc [88]	Diagnosticar casos de cancro da mama	569	30
whitewine [88]	Classificar a qualidade de vinho branco	4898	11

[1]Subconjunto, no qual os dados seguem as mesmas distribuições que no *dataset* original (Tabela A.1)

4.1.2 Distribuição de Melhor Ajuste

Uma técnica de imputação de dados pode ter um desempenho distinto nas diferentes variáveis presentes num *dataset*. Nesta abordagem, pretende-se identificar a relação da distribuição com o desempenho dos algoritmos de imputação. Assim, o primeiro passo consistiu em identificar qual a distribuição seguida por cada variável em cada *dataset*.

A obtenção dos melhores parâmetros de ajuste de uma dada distribuição a um conjunto de dados é um assunto que já se encontra bem estudado [89, 90]. No entanto, para a avaliação da melhor distribuição que se ajusta a um vector de dados, num conjunto de distribuições, não existe um padrão definido. Como tal, foi implementado, em *MatLab*, um algoritmo que deteta qual a distribuição que melhor se ajusta a uma determinada variável, utilizando como critério de avaliação o *Goodness of Fit* (GoF). Este valor de avaliação representa a Raiz Normalizada do Erro Médio Quadrático (NRMSE), obtido pela comparação da CDF empírica com a CDF calculada da distribuição, e varia entre $[-\infty, 1]$ (pior e melhor valor). A nossa implementação, para cada variável, efetua o cálculo dos melhores parâmetros, da CDF e do GoF para as seguintes distribuições:

- Log-Normal;
- Exponencial;
- Log-Logística;
- Valor Extremo;
- t Location-scale;
- Gaussiana Inversa;
- Birnbaum-Saunders;
- Pareto Generalizada;
- Valor Extremo Generalizado;
- Beta;
- Gama;
- Normal;
- Rician;
- Weibull;
- Logística;
- Rayleigh;
- Nakagami.

As razões que nos levaram à escolha das distribuições anteriores foram estas serem todas distribuições contínuas e apresentarem CDF com características muito variadas, o que permite uma maior cobertura.

De forma a melhor ilustrar a implementação do ajuste, utilizemos como exemplo o *dataset* **backpain**. Aplicando a metodologia explicada, a primeira variável deste *dataset* segue uma

distribuição Gama com 0.9129 de GoF (Tabela A.2). Outras distribuições para a mesma variável apresentam resultados inferiores como é o caso da distribuição Exponencial com GoF de 0.2369 (Figura 4.2).

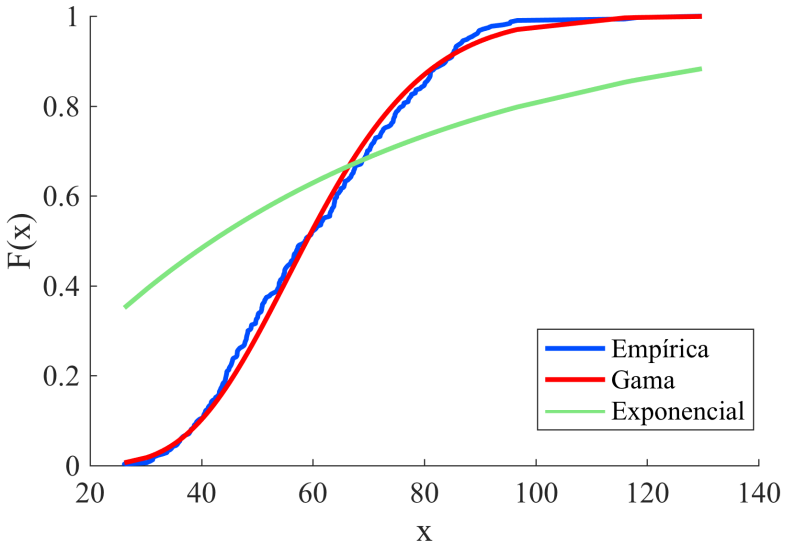


Figura 4.2: Exemplo de melhor ajuste para a primeira variável do *dataset* **backpain**.

4.1.3 Geração de Dados em Falta

A estratégia seguida consistiu na criação de dados em falta de forma artificial utilizando *datasets* completos, com o intuito de avaliar o desempenho das técnicas de imputação de dados na presença de diferentes distribuições e cenários (explicados em seguida).

Foram utilizadas três abordagens diferentes para a criação de dados em falta, empregando cinco MR diferentes (5%, 10%, 15%, 20% e 25%). Esta taxa é aplicada ao número de amostras de cada variável e não ao *dataset* como um todo, ou seja, garante-se que cada variável tem o mesmo número de valores em falta. Por exemplo, para o *dataset* **backpain** quando se aplica a taxa de 15% significa que removemos $\lceil 0.15 \times 310 \rceil = 47$ valores de cada uma das suas variáveis. A remoção é efetuada sempre para todas as variáveis do conjunto de dados.

As diferentes abordagens de criação de dados em falta implementadas foram baseadas nos seguintes critérios:

- Falta de Dados Completamente Aleatória (MCAR);
- Função Densidade de Probabilidade (PDF);
- Histograma/Frequência.

De seguida cada uma destas formas é explicada com maior detalhe.

Falta de Dados Completamente Aleatória (MCAR)

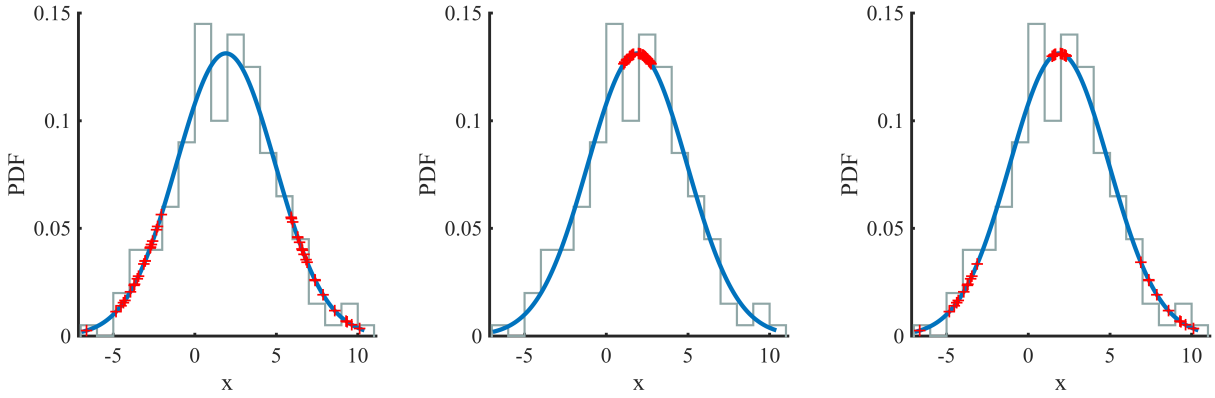
A perda de dados quando não é causada por um processo sistemático mas por obra do acaso, sem relação aos dados, é denominada por MCAR. De modo a executar esta abordagem, foi implementado um código que gera x índices de um vetor de dados, de forma aleatória, onde x é o número de amostras a remover por variável. A este método denominamos de *T7*. Devido à sua natureza aleatória são geradas trinta versões para cada MR e *dataset*, com o principal objetivo de evitar resultados enviesados. Esta decisão evita análises erradas do desempenho dos algoritmos de imputação de dados. Supondo que são efetuadas 3 versões onde os dados removidos aleatoriamente são sempre da primeira metade do *dataset*, os desempenhos finais até podem ser conformes mas não serão conclusivos em relação ao efeito da aleatoriedade dos dados em falta no dataset em global.

Função Densidade de Probabilidade (PDF)

Qualquer valor num determinado conjunto de dados tem uma determinada predisposição a ocorrer, que é quantificável através da PDF. A remoção de dados baseada na PDF é o conceito base desta abordagem. O conhecimento da distribuição que melhor se ajusta aos dados é um requisito para a obtenção da PDF. Nesta abordagem temos três métodos de seleção de valores *T1*, *T2* e *T3*, dados com baixos, elevados e ambos (baixos e elevados) valores de PDF, respetivamente (Figuras 4.3 a, b e c).

A ideia é remover valores aleatoriamente de uma zona específica, mas esta ideia gera um conflito de escolha. A seleção de uma zona grande de valores candidatos a remover aumenta a aleatoriedade mas diminui a localidade, e a escolha de zonas pequenas o contrário. Como tal, decidimos que o número de valores candidatos a serem eliminados é o dobro dos que se pretendem remover, isto é, se se pretender remover 100 valores então são pelo menos selecionados 200 candidatos, um compromisso entre a localidade e a aleatoriedade. O processo de remoção dos dados candidatos é então executado de modo aleatório.

No método *T3* o número de dados procurados, em cada zona, é igual ao número total de dados que se pretende remover. Enquanto o número de dados removidos, em cada zona, é metade do número de dados procurados. Esta decisão tenciona efetuar uma remoção de dados equilibrada nas duas zonas.



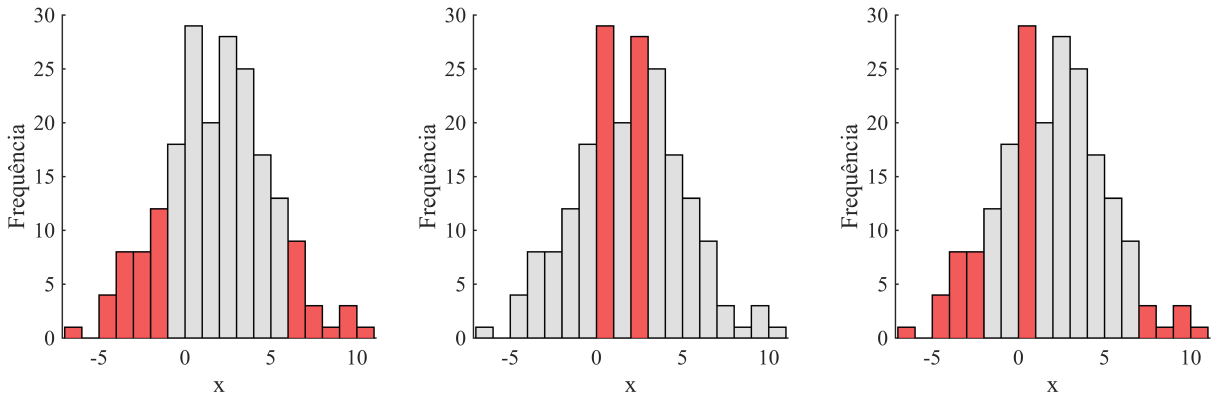
(a) *T1: pdf-baixa.* (b) *T2: pdf-alta.* (c) *T3: pdf-ambas.*

Figura 4.3: Exemplo do método de remoção de dados pela PDF, onde os pontos vermelhos são os dados candidatos (>20%) a serem removidos (taxa de remoção, 10%).

Histograma/Frequência

Esta abordagem é muito semelhante à anterior, com exceção do critério de escolha dos dados candidatos a remover, que neste caso é baseada no histograma dos dados. Foram definidos três métodos de seleção de valores, *T4*, *T5* e *T6*, que selecionam dados pouco, muito e ambos (pouco e muito) frequentes, respetivamente (Figuras 4.4 a, b e c). O cálculo de um histograma requer a escolha do número de barras, e como tal, para este trabalho decidimos usar a regra da raiz quadrada (equação 4.1), que é uma das regras amplamente usada em programas de análise de dados (e.g. Excel [91]).

$$k = \lceil \sqrt{n} \rceil, \; n \text{ amostras.} \tag{4.1}$$



(a) *T4: freq-baixa.* (b) *T5: freq-alta.* (c) *T6: freq-ambas.*

Figura 4.4: Exemplo do método de remoção de dados pela Histograma/Frequência, onde as barras vermelhas são os dados candidatos (> 20%) a serem removidos (taxa de remoção, 10%).

As abordagens *T1* a *T6* não seguem todos os pressupostos do NMAR, o que traria uma complexidade adicional às configurações experimentais, mas aproximam-se de um mecanismo NMAR, uma vez que são criadas limitações ao fenómeno de geração de faltas: os valores são retirados em função dos seus valores observados, de modo que a probabilidade de um valor estar em falta está relacionada com o valor observado nessa variável.

Antes de efetuar a remoção de qualquer valor de uma variável é efetuada uma normalização de dados. A elaboração deste passo é causada pelo requisito de dados normalizados em algumas técnicas de imputação. A normalização poderia ser efetuada somente para as técnicas que necessitam mas desse modo não teríamos os algoritmos no mesmo pé de igualdade. Assim, foi escolhido o método de normalização *Z-score* (equação 4.2) por ser uma normalização que não altera a distribuição de dados [92, 93]. O *z-score* de um valor, x, é obtido através da seguinte equação:

$$z = \frac{x - \mu}{\sigma} \tag{4.2}$$

onde μ e σ são a média e o desvio padrão da variável que se pretende normalizar, respetivamente. A justificação para a aplicação do *Z-score* antes da remoção de dados é porque este depende da média e do desvio padrão e, como tal, após a remoção estes parâmetros variam.

4.1.4 Imputação de Dados

Na imputação dos dados foram consideradas as seguintes técnicas: Média, DT, k-NN, SOM e SVM. Na secção 2.1.2 encontra-se um resumo teórico das técnicas utilizadas. De seguida é detalhado o modo de execução de cada um destes métodos.

Média/Moda

Nesta abordagem só lidamos com dados contínuos, como tal, é efetuado apenas o cálculo da média. São efetuadas os dois tipos de imputação com a média: média total e média por classes (média condicionada).

Árvores de Decisão

Foi implementado CART, que é um algoritmo que usa árvores de classificação e regressão [8], mas devido à natureza dos nossos dados, são somente usadas árvores de regressão.

k-Nearest Neighbors

Dado a natureza do nosso problema (apenas dados contínuos) foi efetuado o cálculo dos valores em falta com base na média ponderada. O uso da média ponderada permite que a informação relativa à distância de cada vizinho mais próximo seja tomada em conta. Além

disso, neste trabalho consideramos o valor de k entre $[1, 20]$ e a HEOM como métrica de distância entre os dados [10]. A razão que nos levou a efetuar testes para diferentes valores de k foi a tentativa de garantir o k ótimo. A escolha da HEOM como métrica advém pela sua capacidade de operar com valores em falta, e portanto inclui essa informação aquando da computação das distâncias.

Mapa Auto-Organizável

Na nossa implementação foram executadas várias configurações – mapas com diferentes números de nós, entre 10 e 100 (com incrementos de 5). A escolha destes valores pretende abranger o melhor cenário para cada variável.

Máquinas de Vetores de Suporte

Neste trabalho, foi implementado SVM de regressão para realizar a imputação dos dados, em particular, usou-se o SVM linear e a RBF. A execução do SVM RBF foi realizada várias vezes com diferentes valores de C e γ (ambos para os valores de 1×10^{-5} a 1×10^{5}, com incrementos de um fator de 10), com o objetivo de garantir a combinação de parâmetros ótima (*grid search*).

4.1.5 Avaliação

Dois dos aspectos analisados nesta experiência, são a Exatidão Preditiva (PAC) e a Exatidão Distribucional (DAC) [22]. A motivação pela qual se usou estes critérios é evitar uma avaliação da eficiência baseada no CE, que está focada nas tarefas de classificação, vendo apenas a imputação como um passo de pré-processamento necessário para atingir classificações satisfatórias, ao invés de avaliar a imputação *per se*.

Recorremos ao coeficiente de determinação (r^2) [23] para a análise da PAC. O desempenho da técnica de imputação é melhor quanto maior for o valor de r^2.

A análise da DAC, foi obtida através da implementação da distância de *Kolmogorov-Smirnov* (D_{KS}) [24]. Valores de D_{KS} menores representam melhores resultados de imputação. Além dos critérios anteriores, que se baseiam na análise dos valores verdadeiros, pretende-se também, analisar a robustez externa (robustez a *outlier*) e avaliar de forma diferente a eficiência de uma técnica de imputação [25]. Como tal, para esse efeito usou-se o MSE. Relativamente a este valor, quanto mais próximo estiver do 0 melhor é o desempenho do método de imputação.

A popularidade e familiaridade destas métricas, em contextos semelhantes, por parte da comunidade científica leva-nos ao seu uso nesta experiência [22–25, 94] .

4.2 Resultados Experimentais

Como mencionado no ponto 4.1.2, para cada variável dos conjuntos de dados utilizados, foi efetuada a procura da melhor distribuição de ajuste e os resultados obtidos estão sumariados na Tabela 4.2. Esta tabela contém também uma coluna com o rácio de variáveis por número de distribuições distintas que os dados seguem nesse conjunto de dados:

$$\text{Rácio} = \frac{\text{n.º de variáveis}}{(\text{n.º de distribuições distintas})^2} \tag{4.3}$$

Na equação 4.3, o denominador aparece ao quadrado porque se pretende atribuir maior importância ao número de distribuições distintas.

Dos resultados obtidos na Tabela 4.2 temos que: o **ctg** tem o maior número de distribuições distintas (10) e o menor rácio (0.21); o **hillvalley** apresenta o maior número de variáveis (100) e o menor número de distribuições (2), como tal têm o maior rácio (25); só o **hillvalley** e **spectf** têm o rácio superior a 1; a distribuição mais frequente é a *Birnbaum-Saunders* com 102 ocorrências; a distribuição que abrange mais *datasets* é o Valor Extremo Generalizado; não existem variáveis nos *datasets* que sigam uma distribuição de *Rician*; o conjunto com os piores valores de GoF, em média, é o **ctg**, pois apresenta cinco ajustes com GoF inferior a 0.5; o pior valor de GoF é 0.21 e o corre numa das variáveis de **pen** (Tabela A.2).

Apesar de haver oito variáveis que apresentam um GoF inferior a 0.5 e três inferiores a 0.3, todas são superiores a 0.2 e encontram-se distribuídas por diferentes distribuições. O efeito que essas variáveis podem realizar nas conclusões a tomar futuramente é inexistente pois os métodos de imputação competem através de um sistema de votos (explicado de forma mais detalhada de seguida), e estas estão sempre em minoria (Tabela A.3).

Após a obtenção dos *datasets* imputados, foi implementado um sistema de votação. Este consiste na atribuição de pontos nas técnicas de imputação que apresentam o melhor valor da métrica em análise, para cada variável em cada *dataset*, com a separação por MR a que o *dataset* imputado foi sujeito. Depois da obtenção dessas votações é efetuada a junção das variáveis desse *dataset* que seguem a mesma distribuição e, além disso, são realizadas junções de MR das votações:

- $5 - 10\%$;
- $15 - 20\%$;
- 25%;
- $5 - 10 - 15 - 20 - 25\%$ (Total).

De seguida, são selecionados os métodos vencedores em cada conjunto de votações, para cada variável, de todos os *datasets* e agrupados com os melhores métodos por distribuição. Por fim, estes métodos vencedores de cada distribuição competem numa votação. Este sistema é implementado para cada estratégia de geração de dados em falta e originam as Tabelas A.4 e A.5 com o valores das votações e as Tabelas A.6 e A.7 com a média, obtidos em cada métricas, distribuição e estratégia.

A análise do desempenho das técnicas de imputação, devido à densidade de informação, será dividida e exposta nas seguintes 4 subsecções: análise geral dos resultados; análise da performance em cada estratégia de geração de valores em falta (*T1*, *T2*, *T3*, *T4*, *T5*, *T6* e *T7*); análise dos desempenhos das técnicas de imputação para cada distribuição (Gama, Beta, Normal, Exponencial, etc.); e a construção de um modelo heurístico.

Tabela 4.2: Distribuição de melhor ajuste e respetivo GoF, obtidos para cada *dataset*.

Datasets	**Variáveis**	**Rácio**	**Distribuição**		**GoF**
			Tipo	**N.º**	$\mu \pm \sigma$
backpain	12	0.33	Beta	1	0.947
			Gama	2	0933±0029
			Pareto Generalizada	5	0916±0053
			Normal	1	0.859
			Nakagami	1	0.947
			t Location-scale	2	0936±0023
. breast	9	0.56	Birnbaum-Saunders	2	0909±0008
			Valor Extremo Generalizado	4	0.887±0.035
			Pareto Generalizada	2	0.92±0.019
			Log-Normal	1	0.921
bupa	6	0.24	Birnbaum-Saunders	1	0.934
			Exponencial	1	0.805
			Valor Extremo Generalizado	1	0.953
			Gaussiana Inversa	1	0.964
			Log-Logistica	2	0.945±0.005

Continua na página seguinte...

Tabela 4.2: Continuação da página anterior.

Datasets	**Variáveis**	**Rácio**	**Distribuição**		**GoF**
			Tipo	**N.º**	$\mu \pm \sigma$
ctg	21	0.21	Birnbaum-Saunders	1	0.947
			Gama	4	0.688±0.193
			Valor Extremo Generalizado	3	0.757±0.312
			Pareto Generalizada	2	0.77±0.205
			Gaussiana Inversa	1	0.959
			Logistica	2	0.69±0.376
			Normal	3	0.711±0.397
			Nakagami	1	0.91
			t Location-scale	2	0.602±0.496
			Weibull	2	0.901±0.099
hillvalley	100	25.00	Birnbaum-Saunders	94	0.875±0.001
			Valor Extremo Generalizado	6	0.877±0.002
iris	4	0.44	Valor Extremo	1	0.797
			Valor Extremo Generalizado	2	0.841±0.114
			Gaussiana Inversa	1	0.922
leaf	14	0.29	Beta	3	0.714±0.343
			Birnbaum-Saunders	1	0.93
			Valor Extremo Generalizado	2	0.842±0.084
			Pareto Generalizada	5	0.942±0.012
			Nakagami	1	0.928
			Log-Normal	1	0.916
			Rayleigh	1	0.963
letterhalf	16	0.64	Exponencial	1	0.859
			Gama	9	0.885±0.009
			Pareto Generalizada	2	0.88±0.005
			Normal	2	0.9±0.005
			Rayleigh	2	0.905±0.023
parkinson	22	0.45	Beta	1	0.892
			Gama	1	0.949
			Valor Extremo Generalizado	14	0.932±0.029
			Pareto Generalizada	2	0.936±0.00004
			Gaussiana Inversa	2	0.928±0.039
			Log-Logistica	1	0.861
			Weibull	1	0.92

Continua na página seguinte...

Tabela 4.2: Continuação da página anterior.

Datasets	**Variáveis**	**Rácio**	**Distribuição**		**GoF**
			Tipo	**N.º**	$\mu \pm \sigma$
pen	16	0.64	Valor Extremo	1	0.923
			Gama	2	0.68±0.361
			Valor Extremo Generalizado	4	0.816±0.118
			Pareto Generalizada	1	0.883
			Logistica	8	0.702±0.223
redwine	11	0.31	Birnbaum-Saunders	2	0.953±0.011
			Valor Extremo Generalizado	4	0.932±0.035
			Logistica	1	0.883
			Log-Logistica	1	0.98
			Nakagami	1	0.96
			t Location-scale	2	0.956±0.029
relax	12	0.75	Valor Extremo Generalizado	1	0.946
			Logistica	3	0.953±0.004
			Normal	1	0.944
			t Location-scale	7	0.948±0.009
spectf	44	4.89	Valor Extremo	30	0.925±0.028
			Logistica	3	0.932±0.006
			Weibull	11	0.95±0.008
wdbc	30	0.47	Birnbaum-Saunders	1	0.962
			Gama	5	0.932±0.036
			Valor Extremo Generalizado	17	0.951±0.02
			Pareto Generalizada	1	0.937
			Gaussiana Inversa	1	0.971
			Log-Logistica	2	0.968±0.001
			Log-Normal	2	0.978±0.003
			t Location-scale	1	0.97
whitewine	11	0.44	Valor Extremo Generalizado	4	0.944±0.041
			Pareto Generalizada	1	0.849
			Log-Logistica	3	0.968±0.01
			Nakagami	2	0.977±0.006
			t Location-scale	1	0.95

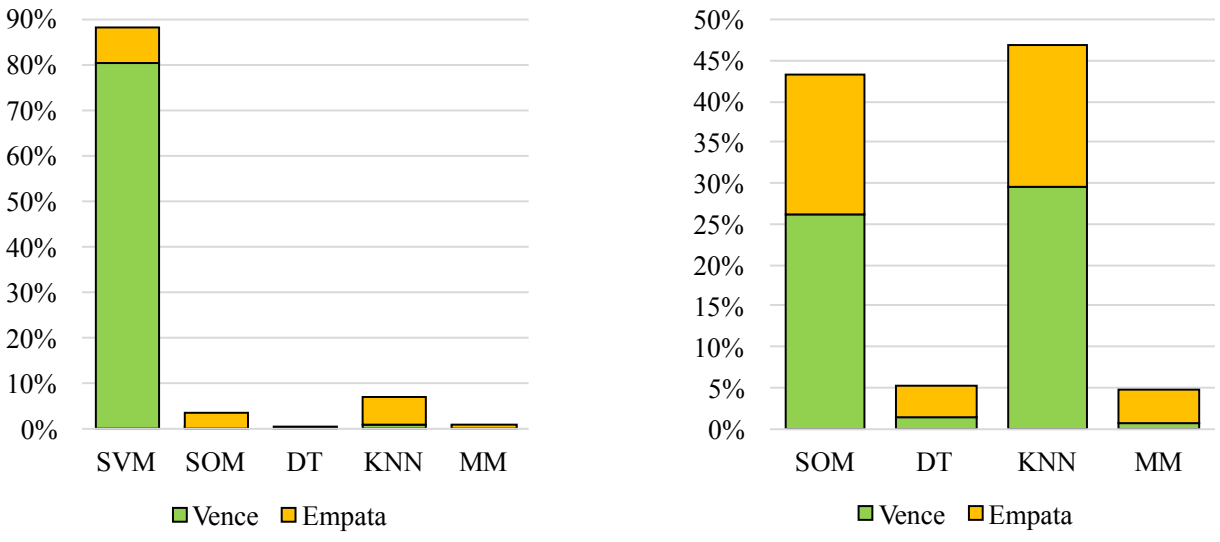
4.2.1 Análise Geral

A análise geral consiste na observação dos resultados de modo a identificar a técnica de imputação que teve o melhor desempenho no global, por outras palavras, contabilizar o número de vitórias e empates das técnicas de imputação para todas as configurações e *datasets* propostos.

A Figura 4.5 apresenta a percentagem de vitórias e de empates para cada algoritmo, considerando todas as métricas. Conclui-se que SVM é a melhor técnica de imputação, apresentando uma taxa total de vitórias superior a 80% (Figura 4.5a). Considerando todas as distribuições, o SVM obtém a média mais elevada para r^2 – 0.765 *versus* 0.723 obtidos com os restantes métodos – e a média mais baixa para MSE e D_{KS} – 0.015 e 0.106 *versus* 0.019 e 0.136 para os restantes métodos, respetivamente. Segundo os resultados das simulações, o SVM é uma técnica que aparenta não ser afetada pela distribuição que os dados seguem.

No entanto, o mesmo não se verifica para os restantes métodos, que apresentam comportamentos diferentes para distribuições e configurações distintas. Os resultados são, portanto, mais heterogéneos. Deste modo, para estudar esses comportamentos nos outros métodos, procedeu-se à análise dos resultados sem o SVM.

Na Figura 4.5b (retirando SVM da análise) é claro que os métodos k-NN e SOM são responsáveis pelas maiores percentagens de vitórias: 29.5% e 26.3%, respetivamente.

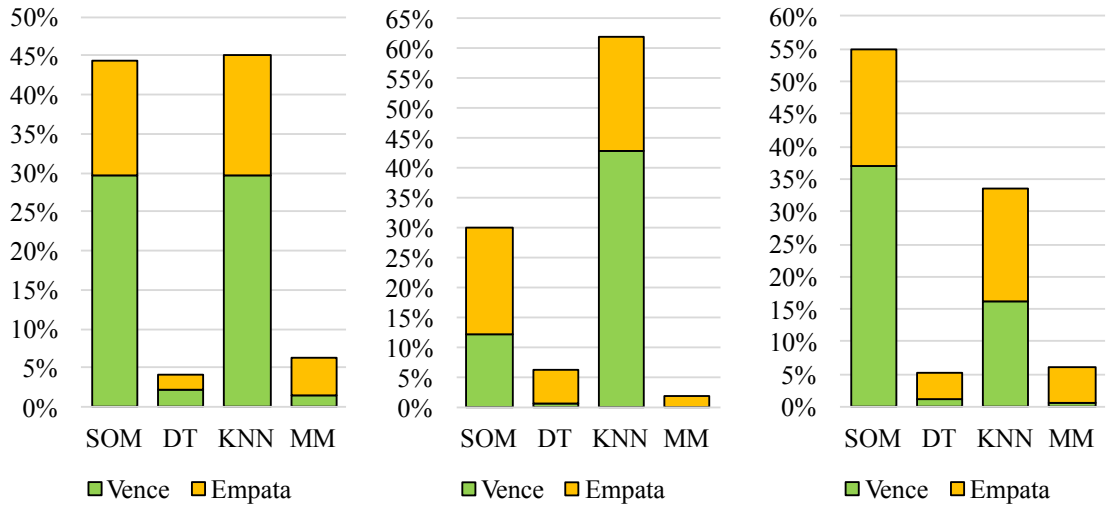


(a) Com SVM

(b) Sem SVM

Figura 4.5: Percentagem de vitórias e empates globais das técnicas de imputação.

Na análise global dos desempenhos das técnicas de imputação por métricas, a tendência enunciada anteriormente é mantida para a métrica r^2 (Figura 4.6a). Contudo, para D_{KS} e MSE os comportamentos são diferentes: k-NN é mais apropriado para manter a distribuição dos dados, enquanto SOM é responsável pelos melhores valores de MSE (Figuras 4.6b e 4.6c).



(a) r^2 (b) D_{KS} (c) MSE

Figura 4.6: Percentagem de vitórias e empates globais das técnicas de imputação por métrica.

A Figura 4.7 apresenta as vitórias e empates, em conjunto, para cada intervalo de MR considerado ($5-10\%$, $15-20\%$ e 25%). A média e SOM mostram um comportamento similar, sendo o seu desempenho maior no contexto onde a percentagem de dados em falta é maior. Contrariamente, DT e k-NN tendem a piorar o seu desempenho nesse mesmo contexto.

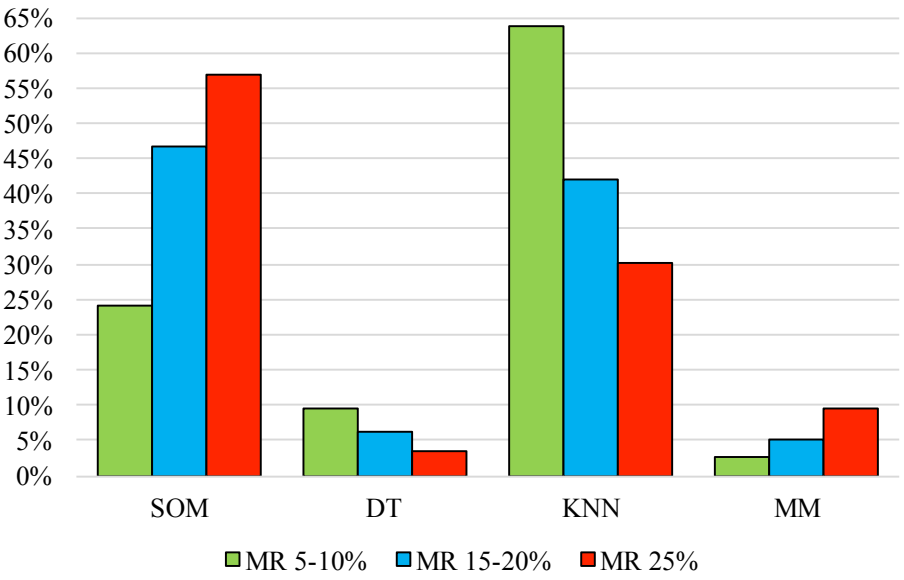


Figura 4.7: Percentagem de vitórias e empates globais das técnicas de imputação por MR.

Para um estudo mais detalhado deste comportamento, a Figura 4.8 e a Tabela 4.3 enunciam as percentagens de vitórias e empates de cada método, considerando cada métrica específica (MSE, r^2 e D_{KS}). O método de imputação k-NN para MR baixas ($5-10\%$) supera todos os outros métodos, sendo o vencedor mais frequente em todas as métricas (50%, 75.2% e 68.6% para MSE, r^2 D_{KS}, respetivamente). Quando a MR aumenta ($15-20\%$) k-NN perde o sua posição de vencedor mais frequente para SOM em termos de PAC e MSE, mas não em termos de DAC, onde aparece como vencedor 57.2% das vezes. Na MR de 25% o comportamento anterior é mantido, embora com as diferenças entre k-NN e SOM mais acentuadas. A superioridade de SOM torna-se clara, em termos de MSE e r^2, 66.9% e 59.6%, respetivamente. Apesar de k-NN continuar a ser o mais frequente vencedor, em termos de DAC, a sua dominância diminuiu para 49%.

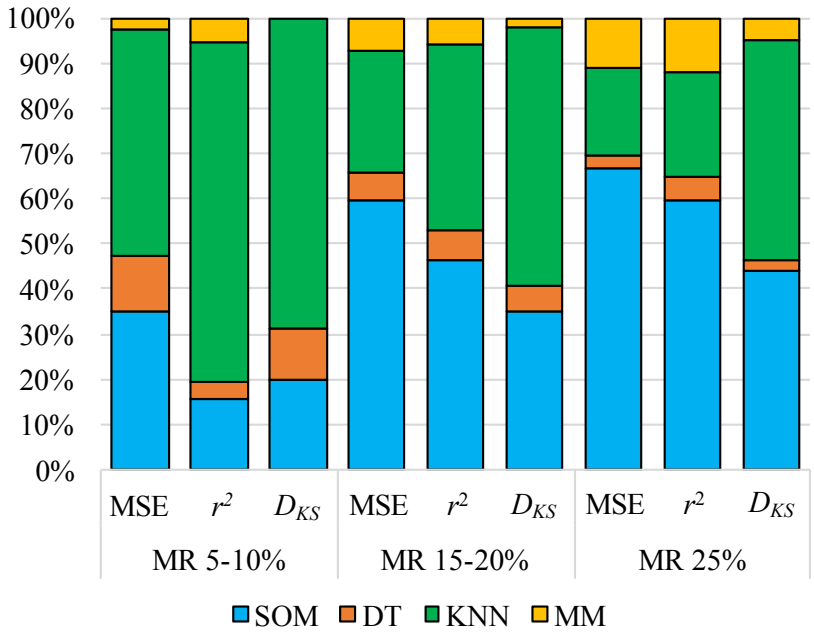


Figura 4.8: Percentagem de vitórias e empates dos métodos de imputação por MR e métrica.

Tabela 4.3: Percentagem de vitórias e empates das técnicas de imputação por MR e métrica.

MR	Métrica	SOM	DT	k-NN	MM
5-10%	MSE	34.9%	12.5%	**50.0%**	2.6%
	r^2	15.5%	3.9%	**75.2%**	5.4%
	D_{KS}	20.0%	11.4%	**68.6%**	0.0%
15-20%	MSE	**59.5%**	6.1%	27.0%	7.4%
	r^2	**46.3%**	6.7%	40.9%	6.0%
	D_{KS}	34.9%	5.9%	**57.2%**	2.0%
25%	MSE	**66.9%**	2.8%	19.0%	11.3%
	r^2	**59.6%**	5.1%	23.1%	12.2%
	D_{KS}	44.2%	2.0%	**49.0%**	4.8%

4.2.2 Análise por Estratégia de Geração de Dados em Falta

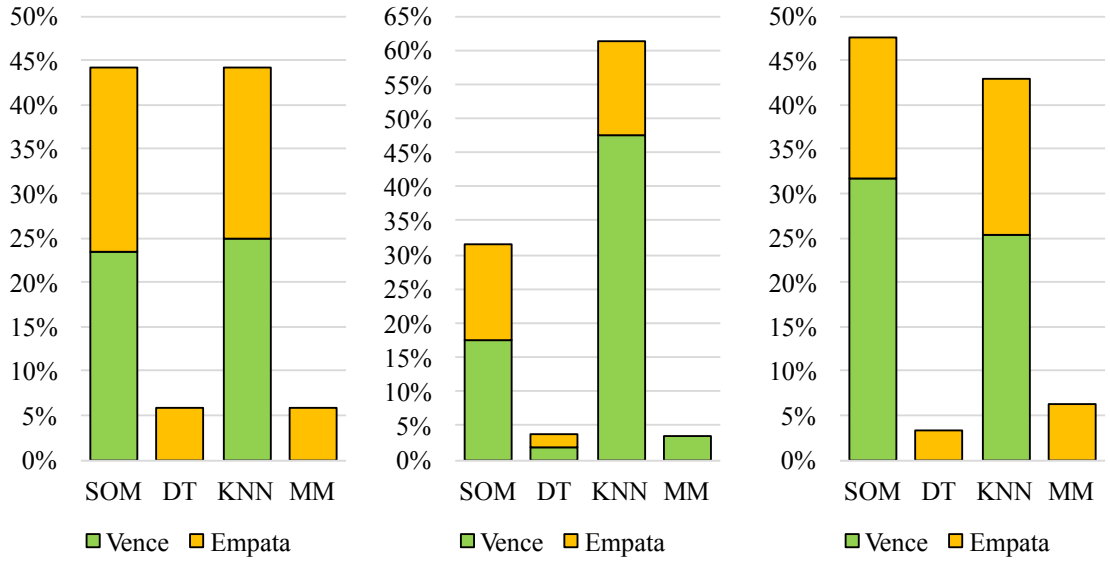
Nesta análise pretendeu-se destacar as técnicas de imputação com melhor desempenho para cada estratégia de geração de dados incompletos, enunciadas na secção 4.1.3.

De modo geral, SOM e k-NN apresentam percentagens de vitórias e empates similares na análise global de cada estratégia (Figura 4.9), exceto a estratégia $T2$ (pdf-alta) em que k-NN é vencedor em mais de 60% das vezes (Figura 4.9b).

Nas Tabelas 4.4 e 4.5 e nas Figuras 4.10–4.13 estão representadas o total de vitórias e empates para cada método, considerando intervalos de MR e estratégias de geração de dados em falta específicas. A tendência reportada para a Figura 4.8 é mantida nesta análise: o SOM obtém um desempenho superior à medida que a MR aumenta, enquanto que o k-NN apresenta o comportamento contrário: uma diminuição de vitórias com o aumento da MR. Em percentagens baixas de dados incompletos, k-NN é o vencedor para todos os cenários, ao passo que o SOM é o vencedor geral para taxas acima de 15%. Todavia, existe uma exceção para as estratégias $T2$ e $T5$, onde no intervalo $15 - 20\%$, k-NN é o vencedor mais frequente, contrariamente ao cenário anterior, onde SOM é vencedor na generalidade.

Tabela 4.4: Percentagem de vitórias e empates das técnicas de imputação por MR e estratégia.

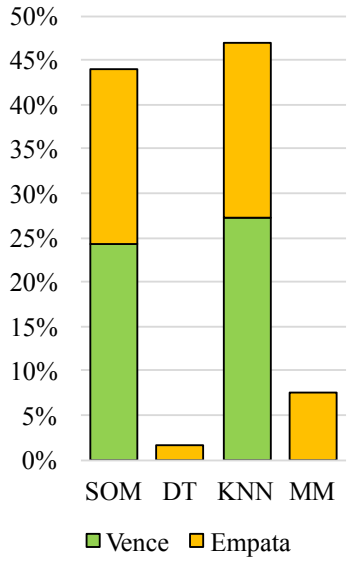
MR	Estratégia	SOM	DT	k-NN	MM
5-10%	T1	23.2%	5.4%	**71.4%**	0.0%
	T2	22.6%	8.1%	**66.1%**	3.2%
	T3	19.7%	9.8%	**68.9%**	1.6%
	T4	30.6%	3.2%	**62.9%**	3.2%
	T5	28.6%	9.5%	**52.4%**	9.5%
	T6	20.7%	10.3%	**69.0%**	0.0%
	T7	22.0%	20.3%	**57.6%**	0.0%
15-20%	T1	**58.5%**	6.2%	30.8%	4.6%
	T2	28.1%	7.8%	**57.8%**	6.3%
	T3	**50.8%**	6.3%	39.7%	3.2%
	T4	**52.9%**	5.7%	34.3%	7.1%
	T5	44.1%	1.7%	**52.5%**	1.7%
	T6	**49,2%**	3.2%	39.7%	7.9%
	T7	**43.1%**	12.3%	40.0%	4.6%
25%	T1	**65.0%**	0.0%	26.7%	8.3%
	T2	**45.5%**	4.5%	40.9%	9.1%
	T3	**59.0%**	4.9%	26.2%	9.8%
	T4	**61.9%**	4.8%	25.4%	7.9%
	T5	**47.7%**	4.6%	40.0%	7.7%
	T6	**63.1%**	3.1%	23.1%	10.8%
	T7	**56.9%**	1.5%	29.2%	12.3%



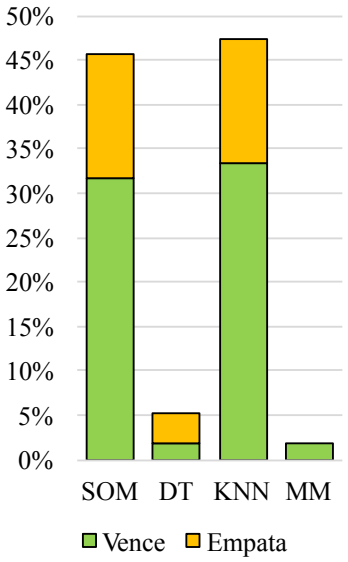
(a) $T1$

(b) $T2$

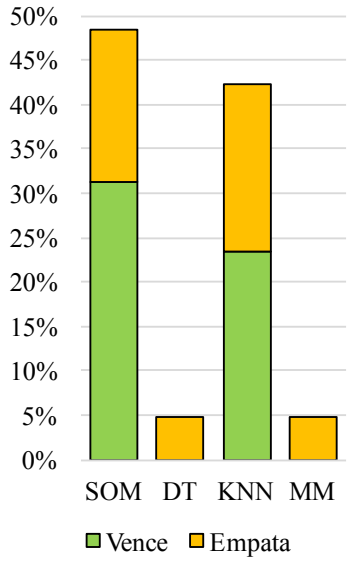
(c) $T3$



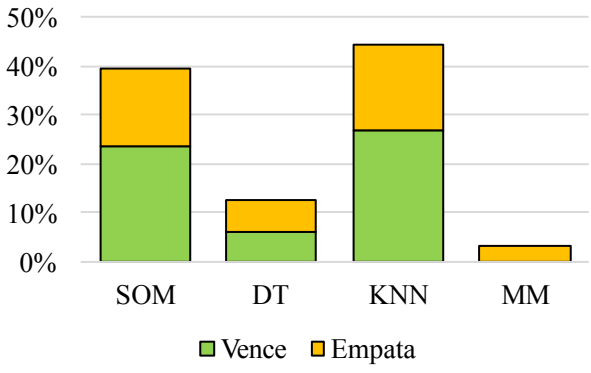
(d) $T4$



(e) $T5$



(f) $T6$



(g) $T7$

Figura 4.9: Percentagem de vitórias e empates globais das técnicas de imputação por estratégia.

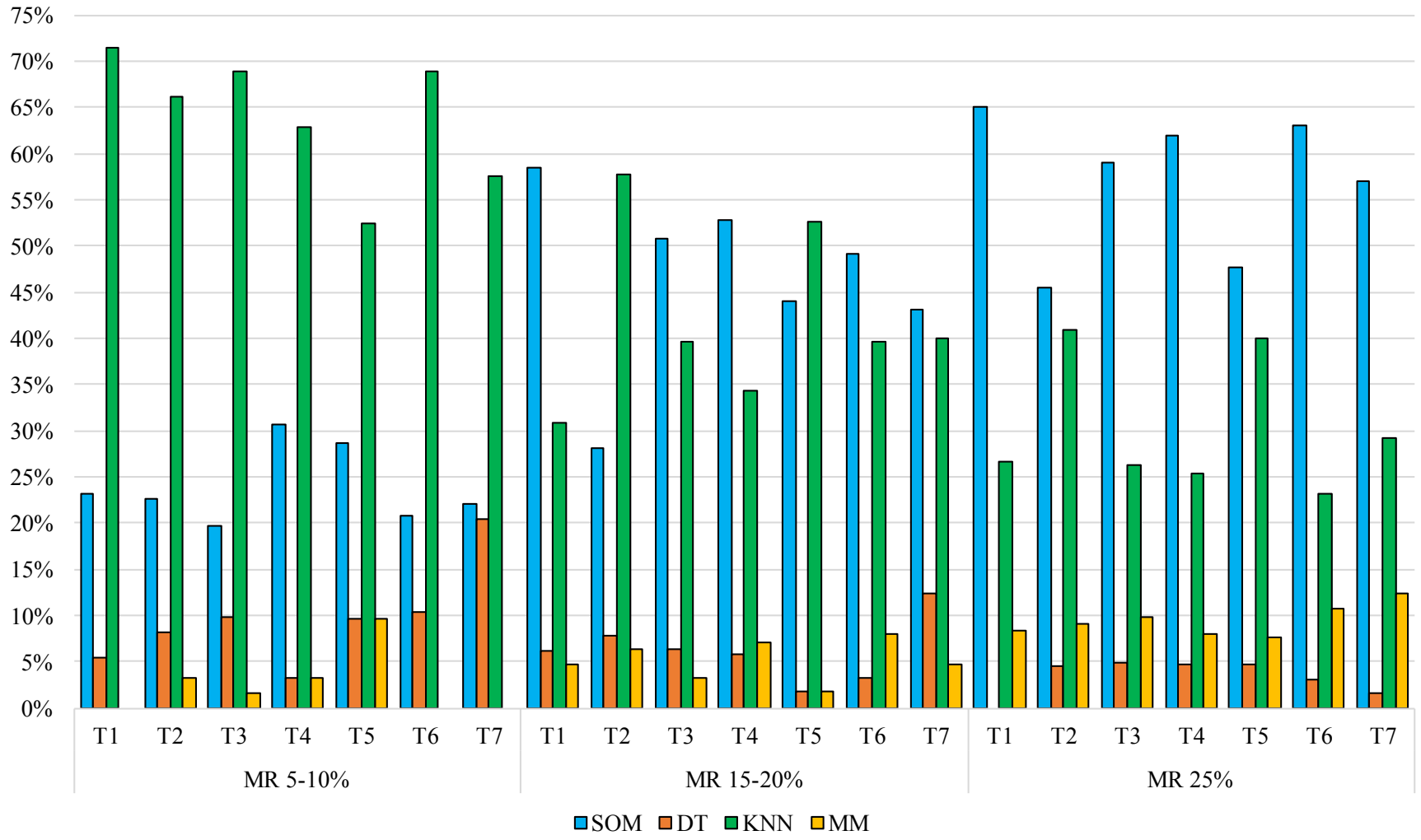


Figura 4.10: Percentagem de vitórias e empates das técnicas de imputação por MR e estratégia.

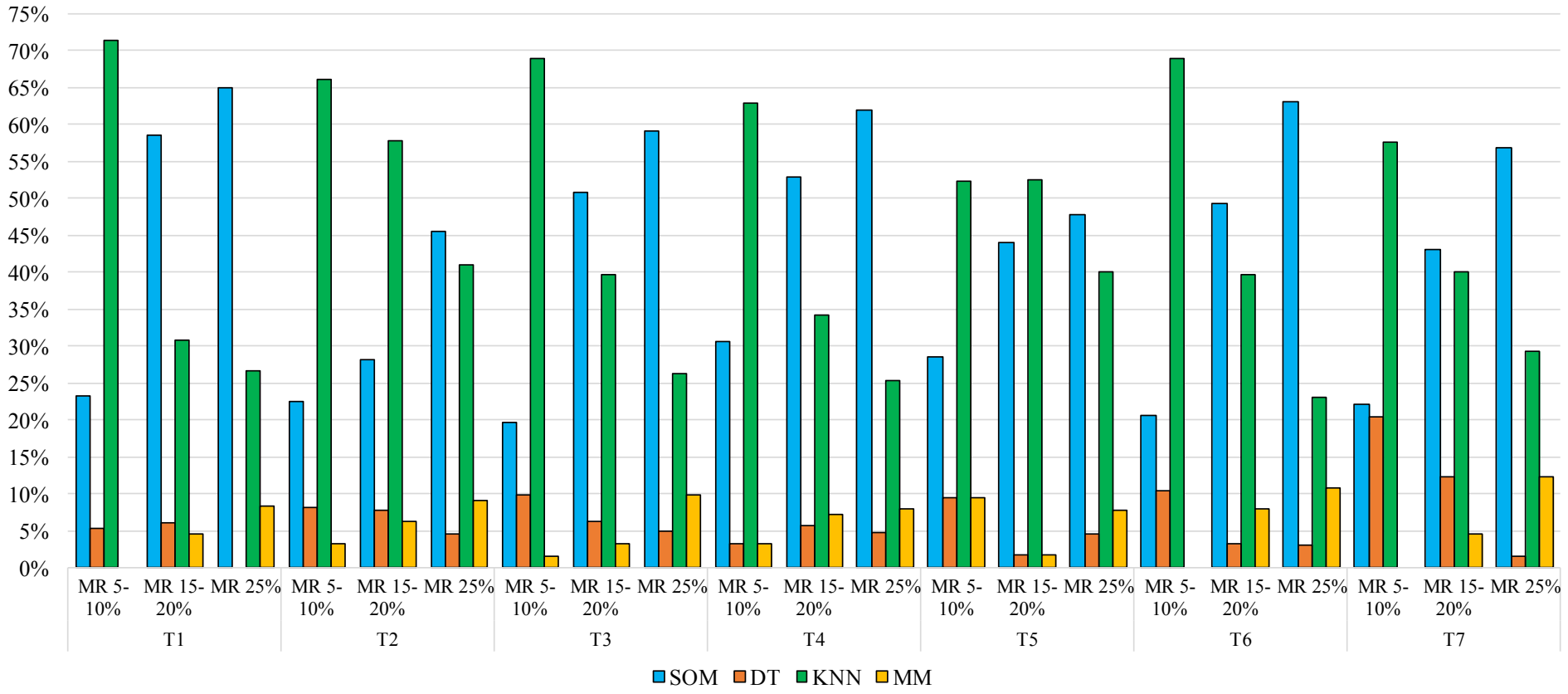


Figura 4.11: Percentagem de vitórias e empates das técnicas de imputação por estratégia e MR.

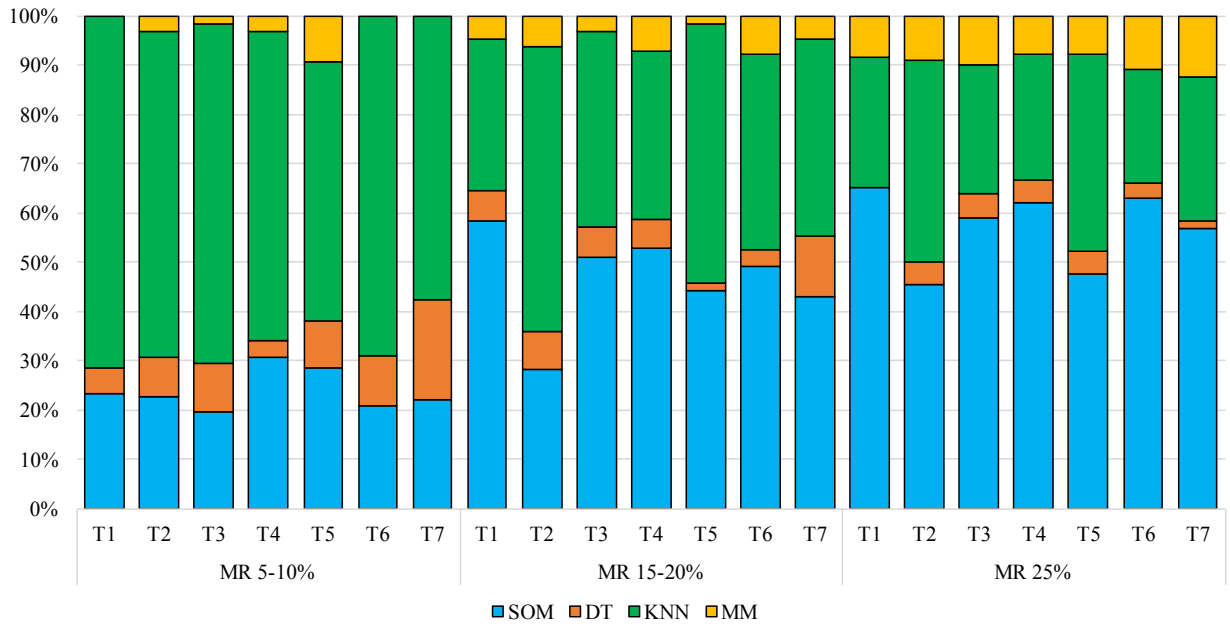


Figura 4.12: Percentagem de vitórias e empates das técnicas de imputação por MR e estratégia.

Tabela 4.5: Percentagem de vitórias e empates das técnicas de imputação por estratégia e MR.

Estratégia	MR	SOM	DT	k-NN	MM
T1	5-10%	23,2%	5,4%	**71,4%**	0,0%
	15-20%	**58,5%**	6,2%	30,8%	4,6%
	25%	**65,0%**	0,0%	26,7%	8,3%
T2	5-10%	22,6%	8,1%	**66,1%**	3,2%
	15-20%	28,1%	7,8%	**57,8%**	6,3%
	25%	**45,5%**	4,5%	40,9%	9,1%
T3	5-10%	19,7%	9,8%	**68,9%**	1,6%
	15-20%	**50,8%**	6,3%	39,7%	3,2%
	25%	**59,0%**	4,9%	26,2%	9,8%
T4	5-10%	30,6%	3,2%	**62,9%**	3,2%
	15-20%	**52,9%**	5,7%	34,3%	7,1%
	25%	**61,9%**	4,8%	25,4%	7,9%
T5	5-10%	28,6%	9,5%	**52,4%**	9,5%
	15-20%	44,1%	1,7%	**52,5%**	1,7%
	25%	**47,7%**	4,6%	40,0%	7,7%
T6	5-10%	20,7%	10,3%	**69,0%**	0,0%
	15-20%	**49,2%**	3,2%	39,7%	7,9%
	25%	**63,1%**	3,1%	23,1%	10,8%
T7	5-10%	22,0%	20,3%	**57,6%**	0,0%
	15-20%	**43,1%**	12,3%	40,0%	4,6%
	25%	**56,9%**	1,5%	29,2%	12,3%

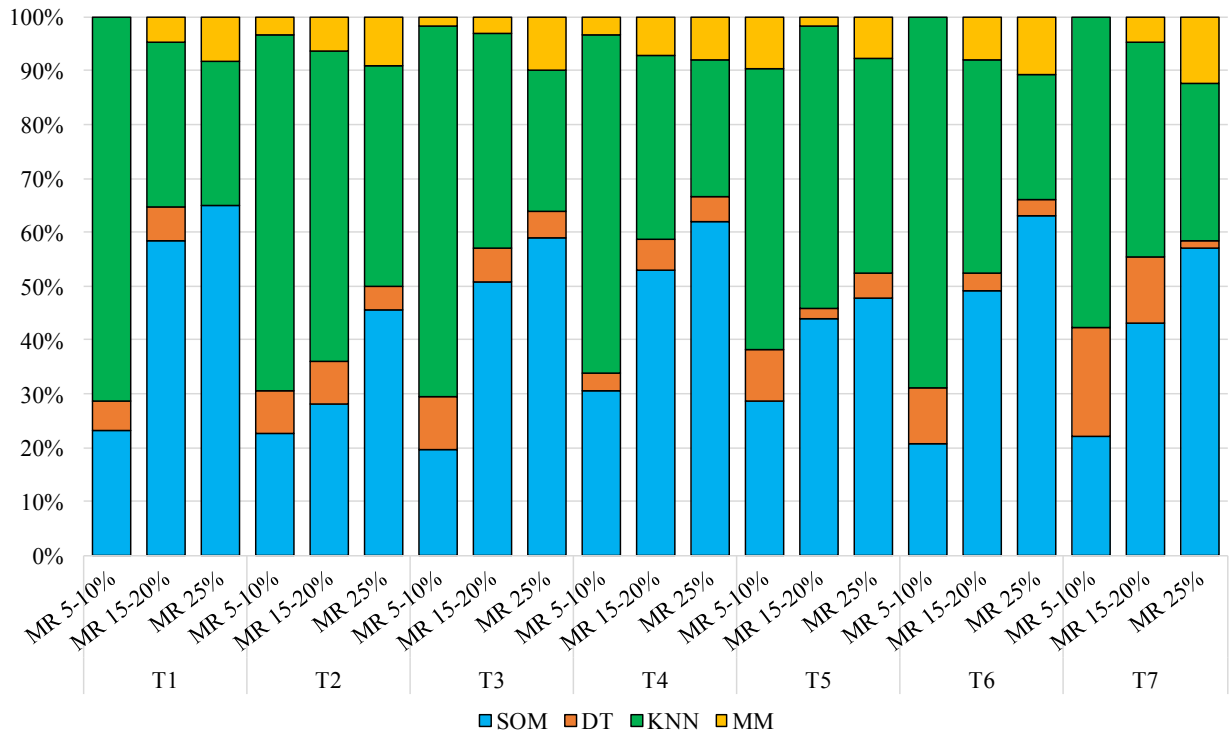
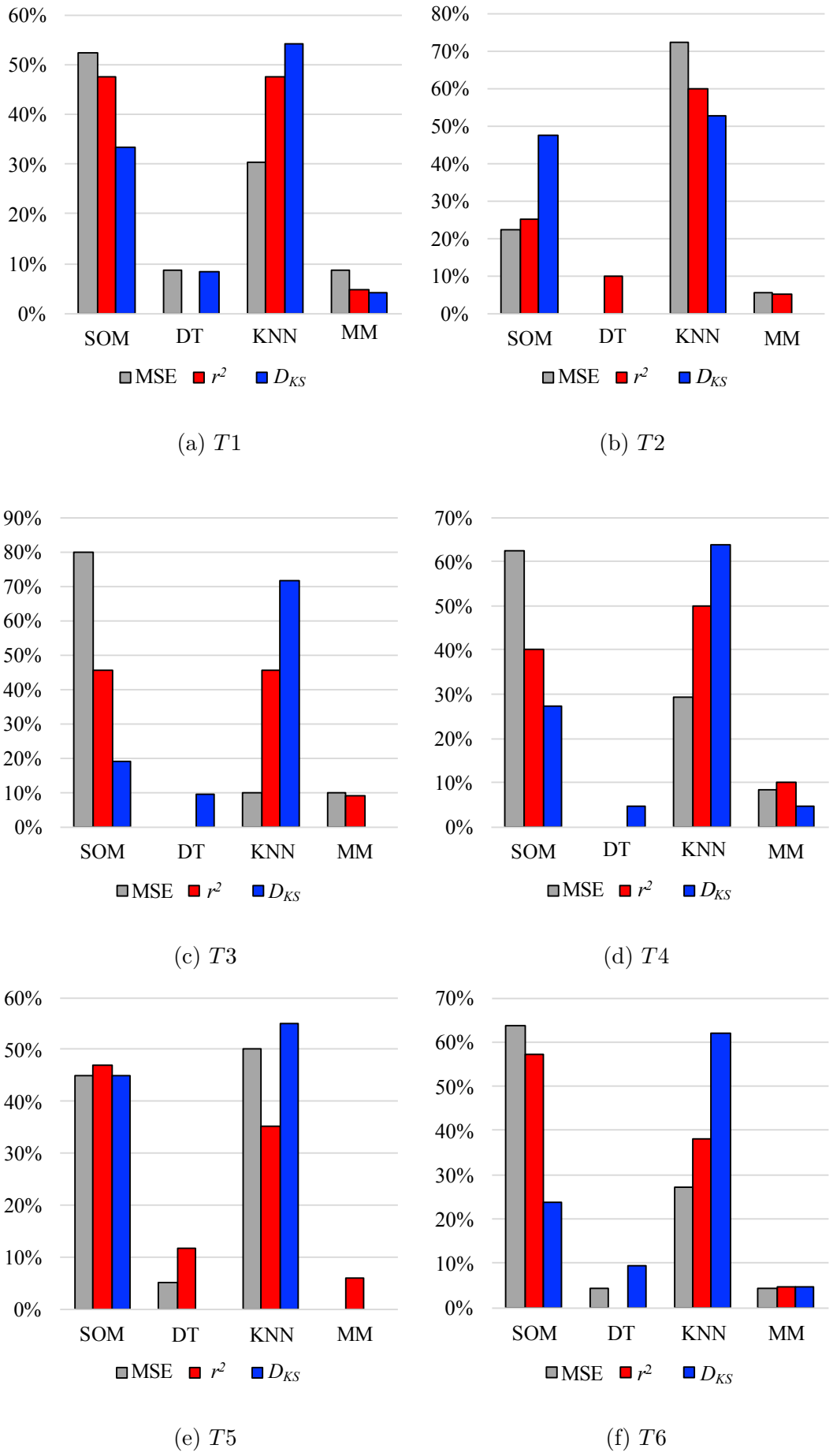


Figura 4.13: Percentagem de vitórias e empates das técnicas de imputação por estratégia e MR.

As Figuras 4.14 e 4.15 especificam o total de vitórias para cada técnica de imputação por métrica (MSE, r^2 e D_{KS}) e estratégia de geração de dados em falta ($T1$, $T2$, etc.). Da análise destas Figuras fica claro que k-NN é o vencedor mais frequente para todas as estratégias, em termos de DAC. Em termos de PAC e MSE, SOM aparenta ser o melhor método para todas as estratégias, exceto $T2$, onde a supremacia do k-NN é notável em ambos os termos, MSE e r^2. Para a estratégia análoga baseada em frequência ($T5$), o k-NN apresenta a maior percentagem de vitórias para a métrica MSE, mas o SOM apresenta uma taxa de vitórias próxima do k-NN.

As comparações dos pares análogos na geração de dados incompletos baseados na frequência e PDF estão representadas na Figura 4.16 e nas Tabelas 4.6–4.8. Não há diferenças relevantes a apontar, exceto o desequilíbrio nos resultados entre SOM e k-NN para a métrica PAC e MSE. A estratégia $T2$ é, na maioria das vezes, melhor imputada com k-NN em todas as métricas, com uma vantagem clara sobre o SOM. Em $T5$, esta diferença não é tão acentuada, como já foi dito anteriormente.



(a) $T1$ (b) $T2$

(c) $T3$ (d) $T4$

(e) $T5$ (f) $T6$

Figura 4.14: Percentagem de vitórias e empates das técnicas de imputação para as estratégias $T1$ a $T6$ agrupadas por métrica.

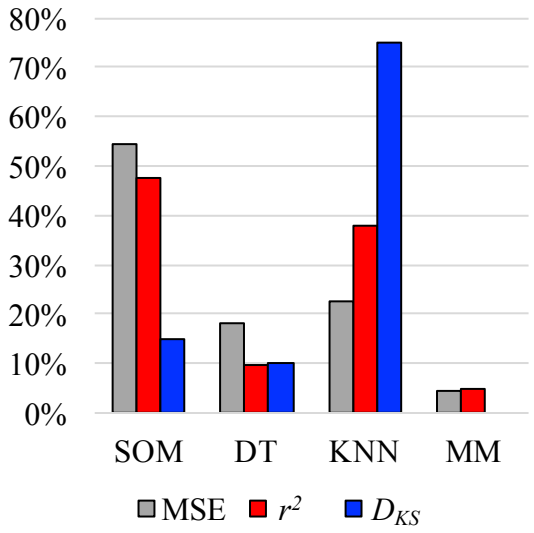
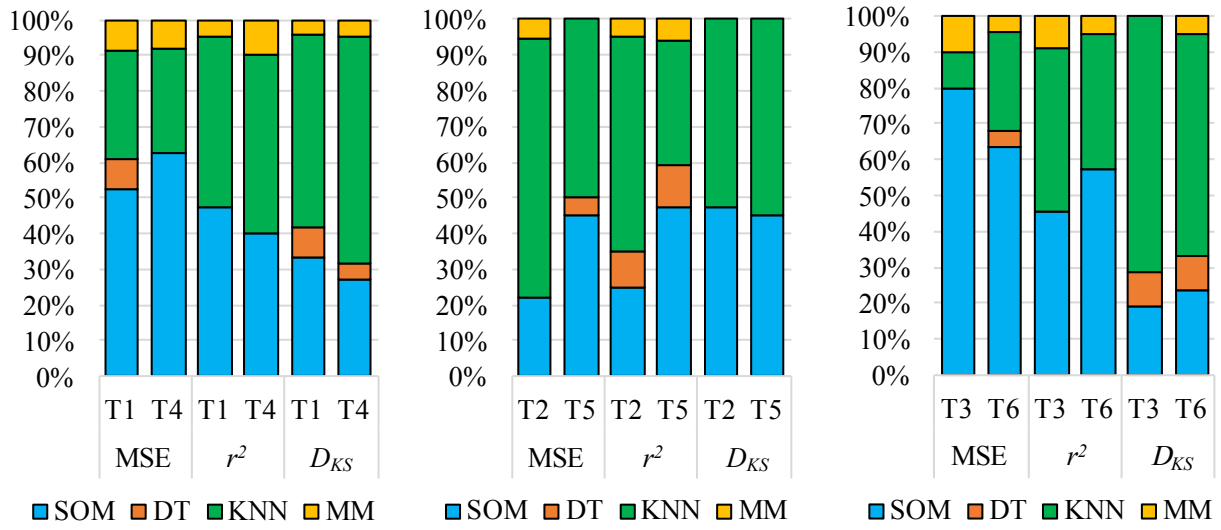


Figura 4.15: Percentagem de vitórias e empates das técnicas de imputação para a estratégia $T7$ agrupadas por métrica.



(a) $T1$ e $T4$. (b) $T2$ e $T5$. (c) $T3$ e $T6$.

Figura 4.16: Comparação do desempenho das técnicas de imputação (em percentagem total de vitórias) entre estratégias semelhantes por métricas.

Tabela 4.6: Comparação do desempenho das técnicas de imputação (em percentagem total de vitórias) entre $T1$ e $T4$ por métrica.

Métrica	Estratégia	SOM	DT	k-NN	MM
MSE	T1	**52.2%**	8.7%	30.4%	8.7%
	T4	**62.5%**	0.0%	29.2%	8.3%
r^2	T1	**47.6%**	0.0%	**47.6%**	4.8%
	T4	40.0%	0.0%	**50.0%**	10.0%
D_{KS}	T1	33.3%	8.3%	**54.2%**	4.2%
	T4	27.3%	4.5%	**63.6%**	4.5%

Tabela 4.7: Comparação do desempenho das técnicas de imputação (em percentagem total de vitórias) entre $T2$ e $T5$ por métrica.

Métrica	Estratégia	SOM	DT	k-NN	MM
MSE	T2	22.2%	0.0%	**72.2%**	5.6%
	T5	45.0%	5.0%	**50.0%**	0.0%
r^2	T2	25.0%	10.0%	**60.0%**	5.0%
	T5	**47.1%**	11.8%	35.3%	5.9%
D_{KS}	T2	47.4%	0.0%	**52.6%**	0.0%
	T5	45.0%	0.0%	**55.0%**	0.0%

Tabela 4.8: Comparação do desempenho das técnicas de imputação (em percentagem total de vitórias) entre $T3$ e $T6$ por métrica.

Métrica	Estratégia	SOM	DT	k-NN	MM
MSE	T3	**80,0%**	0,0%	10,0%	10,0%
	T6	**63,6%**	4,5%	27,3%	4,5%
r^2	T3	**45,5%**	0,0%	**45,5%**	9,1%
	T6	**57,1%**	0,0%	38,1%	4,8%
D_{KS}	T3	19,0%	9,5%	**71,4%**	0,0%
	T6	23,8%	9,5%	**61,9%**	4,8%

4.2.3 Análise por Distribuição

Nesta análise pretendeu-se identificar relações entre a distribuição que os dados seguem e o desempenho dos algoritmos de imputação para cada estratégia de geração de dados incompletos.

A Tabela 4.9 apresenta a junção da métrica PAC com MSE e a métrica DAC, considerando cada estratégia, as distribuições que têm 100% das vitórias num algoritmo de imputação específico. Em termos de $r^2 \cap$ MSE, SOM demostra ser a abordagem mais robusta, conquistando os melhores resultados em várias distribuições e estratégias. Em relação à DAC, k-NN é a melhor abordagem para maioria das configurações.

Focando nas distribuições, e considerando a junção das métricas todas, as distribuições *Weibull* e *Birnbaum-Saunders* são geralmente melhor imputadas com SOM na maioria das estratégias. Na Estratégia $T1$, o k-NN é a melhor abordagem para a distribuição *Pareto Generalizada*, enquanto que para $T2$, o k-NN é adequado para as distribuições *Exponencial* e *Valor Extremo Generalizado*. A estratégia $T3$ não é conclusiva, dado que diferentes métodos são mais apropriados para uma métrica específica. No entanto, para $T4$, o k-NN parece ser uma abordagem viável para *t Location-scale*, e para $T5$, também parece ser a melhor abordagem para as distribuições *Normal*, *Gaussiana Inversa* e *Logística*. Além disso, para $T5$, o SOM também é considerado o melhor método para a distribuição *Log-Normal*. Para a estratégia $T7$, não há um método que obtenha os melhores resultados para as mesmas distribuições em todas as métricas, exceto para DT, onde a distribuição *Log-Normal* obtém 100% das vitórias em todas as métricas.

É importante mencionar que a Tabela 4.9 inclui apenas as distribuições para as quais cada método é o único vencedor em cada métrica específica. No entanto, alguns métodos mostram um comportamento robusto, no sentido em que não são vencedores individuais mas aparecem frequentemente como vencedores. Neste caso a técnica de imputação DT, especialmente na estratégia $T7$, é frequentemente vencedora para as distribuições *Logística*, *Normal*, *Nakagami*, *Weibull* e *t Location-scale*.

A variação destes resultados quando a Taxa de Valores em Falta é considerada não é relevante, excepto para as distribuições *Valor Extremo Generalizado* e *Pareto Generalizada*, onde a tendência identificada anteriormente é novamente verificada: k-NN produz os melhores resultados PAC e DAC para MR baixas ($5-10\%$), mas perde o seu domínio para o SOM para taxas maiores.

Tabela 4.9: Distribuições com 100% de vitórias numa técnica de imputação para as estratégias $T1$ a $T7$ e as métricas $r^2 \cap$ MSE e D_{KS}. As distribuições comuns a ambas as métricas estão marcadas a vermelho.

Estratégia	**Imputação**	$\boldsymbol{r^2 \cap}$**MSE**	$\boldsymbol{D_{KS}}$
T1	k-NN	Pareto Generalizada	Normal
		Exponencial	Nakagami
			Beta
			Valor Extremo Generalizada
			Pareto Generalizada
			Logística
			Valor Extremo
	SOM	Weibull	Weibull
		Gaussiana Inversa	Birnbaum-Saunders
		Logística	
		Gama	
T2	k-NN	Normal	Exponencial
		Nakagami	Logística
		Valor Extremo Generalizada	t Location-scale
		Exponencial	Beta
		Logística	Valor Extremo Generalizada
		Gama	Pareto Generalizada
		Log-Logística	Birnbaum-Saunders
	SOM		Weibull
			Log-Normal
			Nakagami
			Gaussiana Inversa
			Valor Extremo
			Log-Logística
	MM	Log-Normal	
T3	k-NN		Gama
			Birnbaum-Saunders
			Pareto Generalizada
			Exponencial
			Valor Extremo
			Logística
			t Location-scale
			Normal
			Beta
			Valor Extremo Generalizada
	SOM	Weibull	Weibull
		Valor Extremo Generalizada	
		Birnbaum-Saunders	
		Logística	
		Gama	
		Log-Logística	

Continua na página seguinte...

Tabela 4.9: Continuação da página anterior.

Estratégia	Imputação	$r^2 \cap$MSE	D_{KS}
T4	k-NN	t Location-scale	Pareto Generalizada Gaussiana Inversa Log-Logística Valor Extremo Logística t Location-scale Normal Nakagami Beta Valor Extremo Generalizada
	SOM	Weibull Valor Extremo Generalizada Birnbaum-Saunders Logística Gama Log-Logística	Weibull
T5	k-NN	Normal Gaussiana Inversa Logística	Normal Valor Extremo Generalizada Pareto Generalizada Gaussiana Inversa Exponencial Logística Valor Extremo Log-Logística
	SOM	Weibull Valor Extremo Generalizada Log-Normal Pareto Generalizada Birnbaum-Saunders	Weibull Log-Normal Nakagami Birnbaum-Saunders
T6	k-NN		t Location-scale Normal Rayleigh Beta Valor Extremo Generalizada Pareto Generalizada Exponencial Gaussiana Inversa Logística Log-Logística
	SOM	Weibull Normal Log-Normal Valor Extremo Generalizada Birnbaum-Saunders Gaussiana Inversa Valor Extremo Logística	Weibull Birnbaum-Saunders

Continua na página seguinte...

Tabela 4.9: Continuação da página anterior.

Estratégia	Imputação	$r^2 \cap$MSE	D_{KS}
T7	k-NN		Valor Extremo Logística Log-Logística Valor Extremo Generalizada Pareto Generalizada Gaussiana Inversa t Location-scale Normal Nakagami Beta
	SOM	Valor Extremo Gama Weibull Valor Extremo Generalizada Birnbaum-Saunders Pareto Generalizada Gaussiana Inversa	
	DT	Log-Normal	Log-Normal

4.2.4 Modelo heurístico

Uma vez que este trabalho considera um conjunto vasto de configurações (várias distribuições, Taxa de Valores em Falta, estratégias e métricas), fornecer uma heurística clara não é um processo trivial. Para além disso, cada *dataset* contém várias informações (e.g. número de variáveis e de intâncias) que não são tidas em conta na análise anterior. Tal análise detalhada é demasiado complexa para ser realizada a olho nu e portanto, foi produzido um *dataset* com os resultados de cada variável de modo a analisar toda a informação disponível. Deste modo, os resultados não são avaliados em termos de votações onde as variáveis com a mesma distribuição do mesmo conjunto de dados só têm um voto, mas onde cada variável vai a votação individualmente. A construção deste *dataset* permite a inclusão de fatores importantes não considerados até então (e.g. GoF e número de instâncias). As informações e o formato considerado para estes conjuntos de dados está explícito na Listagem 4.1. O *dataset* criado incluí a informação do nome da distribuição (`Distribution_class`), Taxa de Valores em Falta (`MissingRate`), métrica (`Metric_class`), estratégia de geração de dados em falta (`GenType_class`), rácio obtido pela equação 4.3 (`FeatureRatio`), número de variáveis (`FeatureNo`), número de variáveis que seguem a mesma distribuição no *dataset* (`SameFeature`), número de instâncias (`SampleSize`), *Goodness of Fit* da distribuição atribuída à variável (`GoF`) e o melhor método de imputação (`bestMethod_class`).

```
@relation LowLevelInfoT1T2T3T4T5T6T7

@attribute Distribution_class {Beta,BirnbaumSaunders,Exponential,ExtremeValue,Gamma,
    GeneralizedExtremeValue,GeneralizedPareto,InverseGaussian,Logistic,Loglogistic,Lognormal,
    Nakagami,Normal,Rayleigh,Weibull,tLocationScale}
@attribute MissingRate {5,10,15,20,25}
@attribute Metric_class {ksdistance,mse,pearson}
@attribute GenType_class {T1,T2,T3,T4,T5,T6,T7}
@attribute FeatureRatio numeric
@attribute FeatureNo numeric
@attribute SameFeature numeric
@attribute SampleSize numeric
@attribute GoF numeric
@attribute bestMethod_class {DT,KNN,Mean/Mode,SOM}

@data
Gamma,5,mse,T1,0.33333,12,2,310,0.91288,SOM
Gamma,5,pearson,T1,0.33333,12,2,310,0.91288,SOM
Gamma,5,ksdistance,T1,0.33333,12,2,310,0.91288,DT
```

Listagem 4.1: Excerto do *dataset* produzido e utilizado no WEKA.

Com o novo conjunto de variáveis de interesse começou-se por efetuar o estudo das regras mais simples (*ZeroR* e *OneR*), que permitem obter uma ideia global da classificação dos dados. *ZeroR* sugere a classificação de todas as instâncias com SOM (AUC de 0.5) e *OneR* utiliza os valores de GoF para produzir um conjunto maior de regras de classificação (AUC de 0.608). Estes resultados indicam que o SOM é o vencedor global na maioria das configurações e sugerem que o GoF tem um poder discriminativo elevado. Motivados por estes resultados, foi também realizada a seleção das variáveis baseada no ganho de informação (*Information Gain*), o qual revelou que `GoF` (0.229), `SampleSize` (0.165) e `FeatureRatio` (0.158) são as três variáveis mais discriminativas. Consequentemente, como o intuito de determinar o subconjunto de variáveis que traduzem com maior exatidão o melhor método de imputação para cada variável realizámos uma seleção progressiva sequencial das variáveis (*forward selection*). Esta pesquisa retornou um subconjunto que inclui `GenType`, `SampleSize` e `GoF`, para o qual uma validação cruzada de *10-fold* de uma árvore de decisão C4.5 obteve uma AUC média de 0.725 (Tabela 4.10), o que representa um decréscimo de 0.027 relativamente aos resultados que incluem a informação toda (0.752).

No entanto, a árvore com toda a informação não é interpretável e a utilização do subconjunto `GenType`, `SampleSize` e `GoF` também não fornece uma árvore de decisão interpretável. Como tal, procurou-se um subconjunto de variáveis que possibilite simultaneamente uma interpretação clara, a fim de produzir regras relevantes, com a mínima perda de desempenho

(Tabela 4.10). O subconjunto de variáveis que favorece a árvore de decisão com melhor interpretabilidade é `Distribution_class`, `MissingRate`, `Metric_class` e `GenType_class`, com AUC média de 0.675 (decréscimo de 0.077 relativo à melhor AUC obtida). Apesar dessa diminuição de performance, este modelo providencia regras heurísticas gerais aos investigadores que conheçam a distribuição que os seus dados seguem e pretendam selecionar o melhor método de imputação. Um exemplo de uma árvore de decisão desse género está representada na Figura 4.17, onde é possível visualizar o melhor método de imputação para $T3$ (pdf-ambas). Neste caso o melhor método depende da distribuição dos dados (na maioria das distribuições o vencedor é SOM), e da MR (para distribuições especificas, o melhor método de imputação depende da MR).

Todavia, apesar desta árvore simples facilitar a visualização dos resultados, consideramos que é mais correto avaliar as regras geradas por todos os modelos presentes na Tabela 4.10, e selecionamos as regras mais comuns e precisas, as quais estão identificadas nas Listagens 4.2 e 4.3.

```
SampleSize <= 2126 and FeatureRatio > 4.88 and GenType = T1: SOM (1500,0)

SampleSize <= 2126 and FeatureRatio > 4.88 and GenType = T4 and MissingRate <=10: SOM (600,0)

SampleSize <= 2126 and FeatureRatio > 4.88 and GenType = T5 and
Metric_class = ksdistance: DT (500,0)
Metric_class = pearson: SOM (500,0)
Metric_class = mse: SOM (500,0)

Distribution_class = Exponential: KNN (470.0/195.0)

SampleSize <= 2126 and GenType = T1 and 0.86 <GoF < 0.88: SOM (1547, 11)
340 < SampleSize <= 569 and GenType = T1 and GoF < 0.96: SOM (390, 82)
SampleSize <= 2126 and GenType = T4 and 0.79 < GoF <=0.87: SOM (1660, 336)
SampleSize <= 2126 and GenType = T5 and 0.86 < GoF <=0.88: SOM (1545, 515)
SampleSize <= 2126 and GenType = T6 and 0.86 < GoF <=0.87: SOM (1535, 28)
2126 < SampleSize <= 3498: KNN (4111, 472)
SampleSize > 4898: KNN (4886, 1710)
```

Listagem 4.2: As melhores e mais comuns regras encontradas.

Tabela 4.10: Resultados para uma árvore de decisão C4.5, seguindo um esquema de validação cruzada de *10-fold* em diferentes subconjuntos de variáveis.

Variaveis	**AUC**
Todas	0.752

`Distribution_class`
`MissingRate`
`Metric_class`
`GenType_class`
`GoF`
`FeatureRatio`
`SampleSize`
	0.751

`Distribution_class`
`MissingRate`
`GenType_class`
`FeatureRatio`
`GoF`
	0.729

`GenType_class`
`GoF`
`SampleSize`
	0.725

`Distribution_class`
`MissingRate`
`Metric_class`
`GenType_class`
`GoF`
`FeatureRatio`
	0.721

`Distribution_class`
`MissingRate`
`Metric_class`
`GenType_class`
	0.675

`Distribution_class`
`Metric_class`
`GenType_class`
	0.665

`Distribution_class` `GenType_class`	0.655
`Distribution_class`	0.597
`GenType_class`	0.586

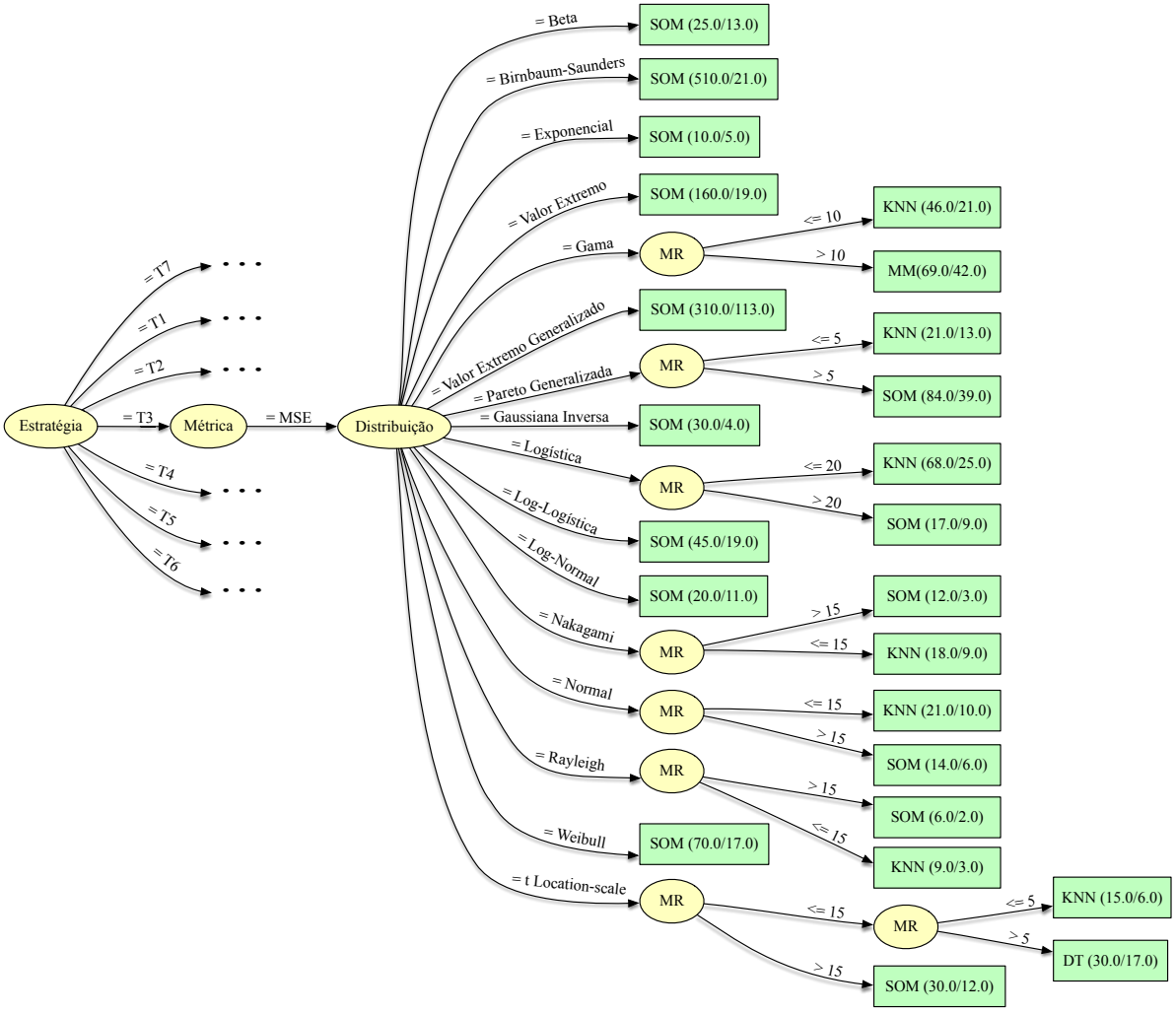


Figura 4.17: Fragmento de uma árvores de decisão exemplar do subconjunto de variáveis `Distribution_class`, `MissingRate`, `Metric_class` e `GenType_class`.

```
% GenType = T1
IF GenType = T1 and
Distribution_class = BirnbaumSaunders: SOM (1533.0/61.0)
Distribution_class = ExtremeValue: SOM (495.0/97.0)
Distribution_class = Weibull: SOM (230.0/76.0)
Distribution_class = Logistic: KNN (270.0/106.0)
Distribution_class = Loglogist: KNN (138.0/66.0)
Distribution_class = Normal: KNN (115.0/55.0)

% GenType = T2
IF GenType = T2 and
Distribution_class = BirnbaumSaunders and MissingRate > 20: SOM (306.0/15.0)
Distribution_class = BirnbaumSaunders and MissingRate <= 20: KNN (1974.0/861.0)
Distribution_class = ExtremeValue and Metric_class = ksdistance: SOM (166.0/35.0)
Distribution_class = ExtremeValue and Metric_class = mse: SOM (160.0/57.0)
Distribution_class = tLocationScale: KNN (227.0/94.0)
Distribution_class = ExtremeValue: SOM (496.0/198.0)
Distribution_class = Logistic: KNN (265.0/65.0)

% GenType = T3
IF GenType = T3 and Metric_class = ksdistance and
Distribution_class = Gamma: KNN (128.0/38.0)
Distribution_class = GeneralizedPareto: KNN (137.0/68.0)
Distribution_class = Logistic: KNN (111.0/42.0)
Distribution_class = tLocationScale: KNN (117.0/64.0)

IF GenType = T3 and Metric_class = mse
Distribution_class = BirnbaumSaunders: SOM (510.0/21.0)
Distribution_class = ExtremeValue: SOM (160.0/19.0)
Distribution_class = GeneralizedExtremeValue: SOM (310.0/113.0)
Distribution_class = Weibull: SOM (70.0/17.0)

IF GenType = T3  and Metric_class = pearson
Distribution_class = BirnbaumSaunders: SOM (510.0/21.0)
Distribution_class = ExtremeValue: SOM (160.0/29.0)
Distribution_class = GeneralizedExtremeValue and MissingRate <= 10: KNN (124.0/47.0)
Distribution_class = GeneralizedExtremeValue and MissingRate > 10: SOM (186.0/65.0)
% GenType = T4
IF GenType_class = T4 and Metric_class = ksdistance and
MissingRate <= 10: SOM (213.0/20.0)
MissingRate > 10 and MissingRate <= 20: KNN (208.0/20.0)
MissingRate > 10 and MissingRate > 20: SOM (103.0/48.0)

IF GenType_class = T4 and
Metric_class = mse: SOM (510.0/28.0)
Metric_class = pearson: SOM (510.0/42.0)

IF GenType_class = T4 and
Distribution_class = Logistic and MissingRate <= 20: KNN (213.0/87.0)
Distribution_class = Weibull: SOM (229.0/84.0)
Distribution_class = BirnbaumSaunders: SOM (1544.0/337.0)
```

```
Distribution_class = ExtremeValue: SOM (512.0/124.0)

% GenType = T5
IF GenType_class = T5 and
Metric_class = mse: SOM (510.0/20.0)
Metric_class = pearson: SOM (510.0/21.0)

IF GenType_class = T5 and Distribution_class = Gamma and
Metric_class = ksdistance: KNN (117.0/40.0)
Metric_class = mse: KNN (115.0/52.0)

GenType_class = T5 and
Distribution_class = Logistic: KNN (259.0/82.0)
Distribution_class = Loglogistic: KNN (137.0/51.0)
Distribution_class = BirnbaumSaunders: SOM (1536.0/541.0)

% GenType = T6
IF GenType_class = T6 and
Distribution_class = BirnbaumSaunders: SOM (1537.0/93.0)
Distribution_class = ExtremeValue: SOM (519.0/167.0)
Distribution_class = Weibull: SOM (235.0/83.0)
Distribution_class = Gamma and Metric_class = ksdistance: KNN (116.0/36.0)

% GenType = T7
IF GenType_class = T7 and
Distribution_class = GeneralizedExtremeValue: KNN (13797.0/9045)
Metric = ksdisnatce: KNN(3055.0/2166.0)
Metric = mse: DT(25320.0/15497.0)
Metric = pearson: KNN(39532.0/22486.0)
```

Listagem 4.3: Lista das melhores regras encontradas para uma estratégia especifica (T_1 a T_7).

4.3 Conclusões

Atendendo ao objetivo primordial deste capítulo, às simulações realizadas e aos resultados obtidos podem-se tirar as seguintes conclusões:

- No geral, o SVM apresenta uma taxa de vitórias superior a 80%, e não aparenta ser afetado pela distribuição de dados (Figura 4.5a). Quando o SVM não é considerado, o SOM e o k-NN apresentam taxas de vitórias e empates superiores a 40% (Figura 4.5b);
- O k-NN é o melhor método de imputação a manter a distribuição original dos dados, e o SOM é aquele que apresenta menores valores de MSE (Figura 4.6);
- Com o aumento da MR, os métodos SOM e MM sofrem um aumento nas taxas de vitórias e empates e as técnicas k-NN e DT sofrem uma diminuição dessas taxas (Figura 4.7 e 4.8 e Tabela 4.3);
- Todas as estratégias de geração de dados em falta apresentam no geral comportamentos semelhantes aos dois pontos anteriores (k-NN melhor para MR baixas e SOM melhor para MR altas) para os vários métodos de imputação (Figuras 4.9–4.16 e Tabelas 4.4–4.8). Com excepção da estratégia $T2$ (pdf-alta), onde o k-NN têm uma vantagem clara sobre os outros métodos (Figuras 4.9b, 4.11 e 4.14b);
- O SOM é a técnica de imputação vencedora na generalidade para as distribuições Weibull e Birnbaum-Saunders (Tabela 4.9);
- O modelo heurístico com melhor desempenho apresenta uma AUC de 0.752 e engloba todas as informações disponíveis (e.g. número de instâncias e de variáveis, GoF etc.) (Tabela 4.10);
- O rácio de variáveis por número de distribuições distintas que os dados seguem, o número de instâncias e o GoF são as três variáveis que têm o maior poder discriminativo no *dataset* utilizado na obtenção do modelo heurístico. O modelo que fez somente o uso desta variáveis obteve uma AUC de 0.725. Foi possível construir um modelo interpretável (DT) que permite a determinação de heurísticas para imputação a partir da análise de factores como a distribuição dos dados, métrica de avaliação, estratégia e taxa de geração de valores em falta. O modelo obteve uma AUC de 0.675, revelando apenas um decréscimo de 0.077 em relação ao melhor desempenho obtido (Tabela 4.10 e Figura 4.17).

5 Problema de Dados Não Balanceados

A utilização de conjuntos de dados com número de instâncias por classe desigual, pode levar a resultados de classificação enviesados. Para contornar este problema, vários investigadores implementam técnicas de amostragem para equilibrar o número de instâncias.

Como discutido nos trabalhos relacionados, alguns autores efetuam a divisão dos *datasets* em *k-folds* para a validação cruzada após o pré-processamento e outros fazem-no antes do pré-processamento (i.e. como parte da validação). Quando o pré-processamento consiste na utilização de algoritmos de sobre-amostragem, a forma como a validação cruzada é feita influencia os resultados da classificação: caso a validação cruzada seja feita após o processo de sobre-amostragem, os resultados da classificação tendem a ser sobrestimados. Esse estudo é realizado neste capítulo. Para além disso, as experiências pretendem ainda estudar o comportamento de um leque extenso de técnicas de sobre-amostragem, relacionando o seu desempenho com a complexidade que geram durante o pré-processamento dos *datasets*.

Em resumo, pretende-se identificar a abordagem mais correta na avaliação de diferentes técnicas de sobre-amostragem, e compreender a sua influência na complexidade dos conjuntos de dados. Para esse efeito, foi criada a arquitetura descrita de seguida.

5.1 Arquitectura da Solução Desenvolvida

A metodologia utilizada compreende cinco etapas: Recolha de Dados, Pré-processamento, Classificação, Avaliação da Classificação e Avaliação de Complexidade (Figura 5.1).

Com o objetivo de explicar o modo mais correto de realizar a validação cruzada num problema de dados não balanceado, são explicadas as duas abordagens de validação cruzada usadas neste trabalho, antes de efetuar a descrição das cinco etapas da metodologia utilizada.

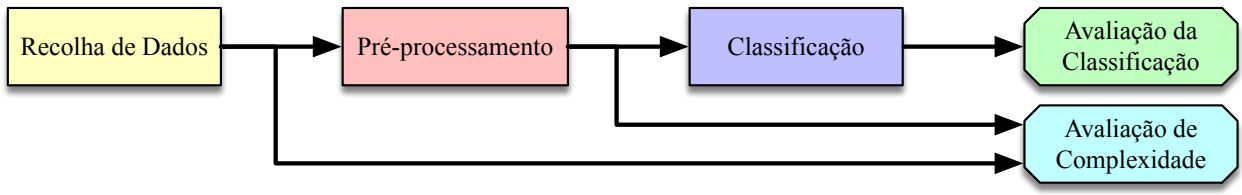


Figura 5.1: Metodologia utilizada para o problema de dados não balanceados.

5.1.1 Validação Cruzada

Antes de descrever as etapas da metodologia utilizada, é importante distinguir dois problemas que podem ocorrer em contextos de dados não balanceados: *overfitting* e sobre-otimismo.

- O *overfitting* ocorre quando o modelo de classificação se "ajusta demasiado" aos dados de treino e não consegue generalizar bem para o conjunto de teste. Neste caso, o desempenho da classificação é muito inferior no conjunto de teste comparativamente aos dados de treino, porque o modelo não tem uma boa generalização. O *overfitting* está associado a técnicas de sobre-amostragem que geram réplicas exactas de padrões já existentes no conjunto de treino, causando um sobre-ajuste do modelo na sua fase de aprendizagem.

- O sobre-otimismo ocorre quando se geram padrões iguais ou semelhantes em ambos os conjuntos de treino e teste. Neste caso o desempenho da classificação do conjunto de teste irá apresentar pouca diferença relativamente ao de treino, não porque o modelo seja capaz de generalizar eficientemente, mas porque contém padrões sensivelmente iguais em ambas as partições. O sobre-otimismo está associado à implementação incorrecta da validação cruzada, aquando do uso de estratégias de sobre-amostragem.

Por exemplo, imaginemos que se procedia à divisão de um *dataset* em 5-*folds*, retirando uma das divisões para teste. Se, nas 4 divisões de treino, se aplicar o ROS, geram-se réplicas exatas dos dados da classe minoritária, pelo que o modelo pode ficar tão ajustado aos dados de treino que vai classificar mal os dados de teste (isto é, ocorre *overfitting*). No entanto, se o ROS for aplicado no *dataset* completo antes da divisão, a probalidade de se obter o mesmo padrão em ambos os conjuntos (treino e teste) aumenta, e neste caso estamos perante uma situação de sobre-otimismo.

Para estudar a implementação mais apropriada de validação em problemas de dados não balanceados, neste capítulo é efetuada uma configuração experimental com dois estágios, ambos considerando um esquema de validação cruzada 5-*folds*. Na primeira fase, é realizada

uma configuração de validação cruzada propensa a sobre-otimismo: o *dataset* completo é primeiro pré-processado para balancear o número de instâncias entre classes e a validação cruzada é aplicada posteriormente. Neste caso, é possível haver réplicas da mesma instância em ambos os conjuntos de dados, teste e treino (Figura 5.2). Na segunda fase, o *dataset* é primeiro dividido em 5 partições estratificadas e somente os dados de treino (80% dos dados, que corresponde a 4 partições) são sobre-amostrados (Figura 5.2). Nesta situação, as instâncias no conjunto de teste nunca são consideradas na etapa de amostragem nem na construção do modelo, fornecendo assim estimativas mais fiáveis e permitindo uma avaliação adequada da capacidade do modelo para generalizar a partir dos dados de treino.

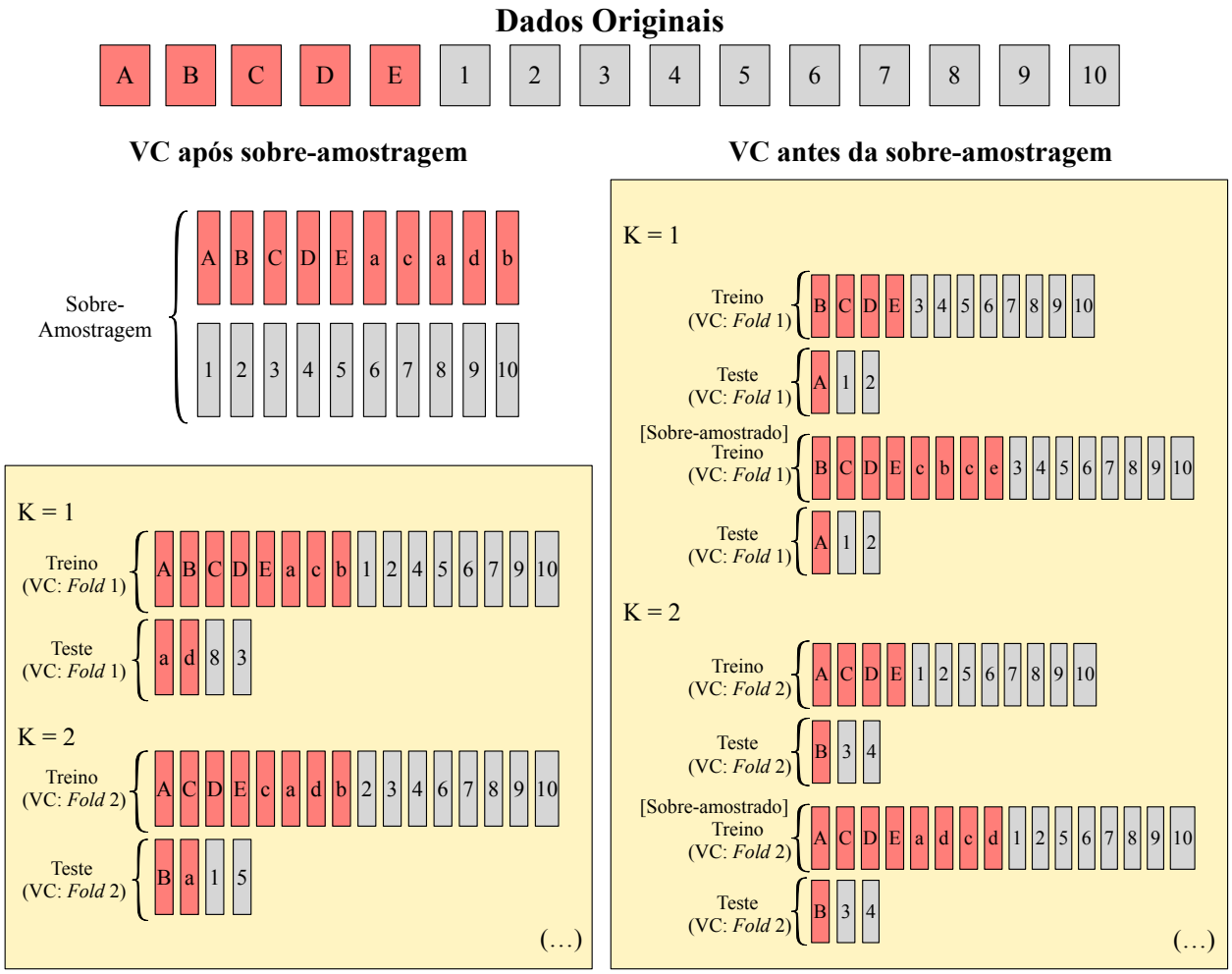


Figura 5.2: Os dois estágios da experiência: Validação Cruzada (VC) após a sobre-amostragem *versus* VC antes da sobre-amostragem.

5.1.2 Recolha de Dados

O critério utilizado na seleção dos dados para concretizar esta experiência é similar à experiência anterior (Capítulo 4) e teve como intuito possibilitar a reprodução do trabalho por parte da comunidade científica. Nesse sentido, a escolha de *datasets* existentes em repositórios livres *online* foi uma prioridade. Foram ainda considerados os seguintes critérios: *datasets* completos, com classes binárias, e diferentes IR (rácio do número de exemplos majoritários por exemplos minoritários), número de instâncias e variáveis. As decisões de procura de problemas só de classes binárias é motivada pelo uso de algumas de técnicas de sobre-amostragem somente serem capazes de lidar com esse tipo de problemas.

Os 86 *datasets* utilizados neste capítulo estão disponíveis no *UCI Machine Learning Repository* [88] e *Knowledge Extraction based on Evolutionary Learning* (KEEL) [95]. As principais características dos *datasets* selecionados estão sumariadas na Tabela 5.1.

O **glass-0-4vs5** tem o menor número de instâncias (92), e o **letter-Z** e o **letter-U** têm o maior número de instâncias (20000). Em termos do número de variáveis, o **haberman** é o *dataset* com menos (3) e o **dermatology-6** é o que tem mais (34). Por fim, o **bupa** e o **abalone-21vs8** são os *datasets* com menor e maior IR: 1.38 e 40.50, respetivamente.

Antes de utilizar os *datasets* é efetuada uma normalização dos dados. Tal como no Capítulo 4, o método de normalização escolhido foi o *Z-score* (equação 4.2).

Tabela 5.1: Sumário das propriedades dos *datasets* escolhidos.

Dataset	**Inst.**	**Variáveis**[1]	**IR**	***Dataset***	**Inst.**	**Variáveis**[1]	**IR**
bupa [88]	345	(6/1)	1.38	vowel0 [95]	988	(13/0)	9.98
page-blocks-1-3vs4 [95]	472	(10/0)	1.57	ecoli-0-6-7vs5 [95]	220	(7/0)	10.00
glass1 [95]	214	(9/0)	1.82	glass-0-1-6vs2 [95]	192	(9/0)	10.29
ecoli-0vs1 [95]	220	(7/0)	1.86	ecoli-0-1-4-7vs2-3-5-6 [95]	336	(7/0)	10.59
wisconsin [88, 95]	683	(9/0)	1.86	led7digit-0-2-4-5-6-7-8-9vs1 [95]	443	(7/0)	10.97
pima [88, 95]	768	(8/0)	1.87	ecoli-0-1vs5 [95]	240	(7/0)	11.00
cmc1vs2 [88]	961	(9/7)	1.89	glass-0-6vs5 [95]	108	(9/0)	11.00
iris0 [95]	150	(4/0)	2.00	glass-0-1-4-6vs2 [95]	205	(9/0)	11.06
glass0 [95]	214	(9/0)	2.06	glass2 [95]	214	(9/0)	11.59
german [88]	1000	(20/13)	2.33	ecoli-0-1-4-7vs5-6 [95]	332	(7/0)	12.28
yeast1 [95]	1484	(8/0)	2.46	cleveland-0vs4 [95]	173	(14/0)	12.31
haberman [88, 95]	306	(3/0)	2.78	ecoli-0-1-4-6vs5 [95]	280	(7/0)	13.00
vehicle2 [95]	846	(18/0)	2.88	shuttle-c0-vs-c4 [95]	1829	(9/0)	13.87
vehicle1 [95]	846	(18/0)	2.90	yeast-1vs7 [95]	459	(8/0)	14.30
vehicle3 [95]	846	(18/0)	2.99	glass4 [95]	214	(9/0)	15.46
glass-0-1-2-3vs4-5-6 [95]	214	(9/0)	3.20	ecoli4 [95]	336	(7/0)	15.8
transfusion [88]	748	(4/0)	3.20	abalone9-18 [95]	731	(8/1)	16.4
vehicle0 [95]	846	(18/0)	3.25	dermatology-6 [95]	358	(34/0)	16.9
ecoli1 [95]	336	(7/0)	3.36	thyroid-3vs2 [88]	703	(21/0)	18.00
newthyroid1 [95]	215	(5/0)	5.14	glass-0-1-6vs5 [95]	184	(9/0)	19.44
ecoli2 [95]	336	(7/0)	5.46	pageblocks-1vs3-4-5 [88]	5144	(10/0)	21.27
balance_scaleBvsR [88]	337	(4/0)	5.88	shuttle-6vs2-3 [95]	230	(9/0)	22.00
balance_scaleBvsL [88]	337	(4/0)	5.88	yeast-1-4-5-8vs7 [95]	693	(8/0)	22.10
segment0 [95]	2308	(19/0)	6.02	pageblocks-1-2vs3-4-5 [88]	5473	(10/0)	22.69
glass6 [95]	214	(9/0)	6.38	glass5 [95]	214	(9/0)	22.78
yeast3 [95]	1484	(8/0)	8.10	yeast-2vs8 [95]	482	(8/0)	23.10
ecoli3 [95]	336	(7/0)	8.60	letter-U [88]	20000	(16/0)	23.60
pageblocks0 [95]	5472	(10/0)	8.79	flare-F [95]	1066	(11/11)	23.79
ecoli-0-3-4vs5 [95]	200	(7/0)	9.00	car-good [95]	1728	(6/6)	24.04
yeast-2vs4 [95]	514	(8/0)	9.08	pageblocks-1vs4-5 [88]	5116	(10/0)	24.20
ecoli-0-6-7vs3-5 [95]	222	(7/0)	9.09	car-vgood [95]	1728	(6/6)	25.58
ecoli-0-2-3-4vs5 [95]	202	(7/0)	9.10	letter-Z [88]	20000	(16/0)	26.25
glass-0-1-5vs2 [95]	172	(9/0)	9.12	kr-vs-k-zero-onevsdraw [95]	2901	(6/6)	26.63
yeast-0-3-5-9vs7-8 [95]	506	(8/0)	9.12	yeast4 [95]	1484	(8/0)	28.10
yeast-0-2-5-6vs3-7-8-9 [95]	1004	(8/0)	9.14	winequality-red-4 [95]	1599	(11/0)	29.17
yeast-0-2-5-7-9vs3-6-8 [95]	1004	(8/0)	9.14	poker-9vs7 [95]	244	(10/0)	29.50
ecoli-0-4-6vs5 [95]	203	(7/0)	9.15	yeast-1-2-8-9vs7 [95]	947	(8/0)	30.57
ecoli-0-1vs2-3-5 [95]	244	(7/0)	9.17	abalone-3vs11 [95]	502	(8/1)	32.47
ecoli-0-2-6-7vs3-5 [95]	224	(7/0)	9.18	yeast5 [95]	1484	(8/0)	32.73
glass-0-4vs5 [95]	92	(9/0)	9.22	kr-vs-k-threevseleven [95]	2935	(6/6)	35.23
ecoli-0-3-4-6vs5 [95]	205	(7/0)	9.25	winequality-red-8vs6 [95]	656	(11/0)	35.44
ecoli-0-3-4-7vs5-6 [95]	257	(7/0)	9.28	abalone-17vs7-8-9-10 [95]	2338	(8/1)	39.31
yeast-0-5-6-7-9vs4 [95]	528	(8/0)	9.35	abalone-21vs8 [95]	581	(7/0)	40.50

[1](Número total de variáveis/Número de variáveis nominais).

5.1.3 Algoritmos de Sobre-amostragem

No balanceamento do conjunto de dados, as técnicas de sobre-amostragem utilizadas foram as seguintes:

- ADASYN;
- ADOMS;
- AHC;
- *Borderline*-SMOTE1;
- *Borderline*-SMOTE2;
- CBO+SMOTE;
- CBO+*Random*;
- MWMOTE;
- ROS;
- *Safe-Level*-SMOTE;
- SMOTE;
- SMOTE+ENN;
- SMOTE+TL;
- SPIDER;
- SPIDER2.

A escolha dos algoritmos de amostragem anteriormente citados teve com principal objetivo estudar as abordagens de sobre-amostragem consideradas estado-da-arte, incluindo abordagens focadas na replicação de instâncias, remoção de dados ruidosos, inserção de exemplos sintéticos (em alguns casos usando *clusters*) e na melhoria da fronteira entre classes e de zonas difíceis de aprender pelos classificadores.

A implementação do CBO+*Random*, CBO+SMOTE e MWMOTE foi realizada no *MatLab*, enquanto a implementação dos restantes métodos foi efetuada no KEEL.O resumo dos parâmetros de entrada utilizados em cada técnica encontra-se detalhado na Tabela B.1. Os parâmetros utilizados para os algoritmos de sobre-amostragem do KEEL seguem as recomendações do *software* [95], que correspondem aos parâmetros dados pelos autores de cada algoritmo nos seus artigos originais. A única modificação feita é respeitante ao valor de k a considerar para a vizinhança dos padrões minoritários, que foi definido com o mesmo valor para todos os algoritmos. Na secção 2.2.1 encontra-se um resumo teórico das técnicas escolhidas.

5.1.4 Classificação

A realização da classificação dos *datasets* originais e pré-processados foi realizada através dos seguintes métodos: CART, C4.5, k-NN, SVM (Linear e RBF) e NB. O apuramento deste conjunto de técnicas teve como base a preferência por algoritmos com princípios de funcionamentos diferentes, por exemplo, CART e C4.5 são modelos de Árvores de Decisão, o k-NN efetua o cálculo dos vizinhos para classificar uma instância (*instance-based learning*), o SVM usa os hiperplanos ótimos computados para classificar e NB utiliza distribuições de probabilidades dos dados para estimar a classificação. A explicação teórica destas técnicas encontra-se na secção 2.2.2.

Os métodos CART, k-NN, SVM Linear e SVM RBF foram implementados em *MatLab* e para os restantes métodos foram utilizadas as implementações existentes no *Waikato Environment for Knowledge Analysis* (WEKA). O algoritmo de k-NN foi executado para vários valores de k, $\{1, 2, 3, 4, 5\}$, e o SVM testado para diferentes valores de C (Linear e RBF) e γ (RBF) (ambos para valores de 1×10^{-3} a 1×10^{3}, com incrementos de um fator de 10), com o objetivo de garantir a combinação de parâmetros ótima.

5.1.5 Avaliação

As métricas de avaliação do desempenho da classificação utilizadas neste método foram: SEN; AUC; *G-Mean*; *F-1* (ou seja, *F-Measure* com $\beta = 1$). A razão pela qual foram escolhidas advém da necessidade de utilizar medidas capazes de avaliar o desempenho dos classificadores em problemas de dados não balanceados. De modo a possibilitar a análise da complexidade dos diferentes conjuntos de dados (i.e. geometria, topologia, sobreposição e separabilidade), selecionaram-se as seguintes métricas de complexidade: F1; F2; F3; L1; L2; L3; N1; N2; N3; N4. Inicialmente, todas as medidas de complexidade propostas por Ho e Basu [61] foram analisadas, mas é apenas reportado este subconjunto de medidas, uma vez que foram consideradas as mais informativas para os factores analisados neste trabalho. Durante a análise, com o objetivo de correlacionar as várias métricas de complexidade foi realizada a normalização *Z-score* (equação 4.2).

5.2 Resultados Experimentais

No decorrer desta secção estão descritas as duas análises realizadas: Comparação entre a Fase 1 e Fase 2 (divisão depois ou antes do pré-processamento) e Estudo da complexidade na Fase 2.

5.2.1 Avaliação do sobre-otimismo: Fase 1 *versus* Fase 2

A primeira análise a realizar nesta experiência é a comparação dos resultados entre a Fase 1, onde a validação cruzada 5-*folds* é realizada após o pré-processamento, e a Fase 2, onde a divisão dos *datasets* nos 5-*folds* é efetuada antes do pré-processamento, de modo a comprovar qual a abordagem mais correta de validação cruzada em problemas de dados não balanceados.

Os resultados confirmam a existência de resultados mais otimistas na Fase 1: os valores médios das várias métricas de avaliação (AUC, *G-Mean*, SEN e *F-1*) para os conjuntos de teste dos *datasets* utilizados são sempre superiores na Fase 1 (Figura 5.3). Nas análises seguintes, a métrica de avaliação do desempenho da classificação em foco é a AUC. Uma vez que o sobre-otimismo se verifica em geral para todas as métricas de classificação, foi selecionada a AUC para prosseguir com as análises seguintes, de modo a poder estabelecer-se uma comparação com os resultados obtidos em trabalhos relacionados, que usam maioritariamente a AUC (Tabela 3.2).

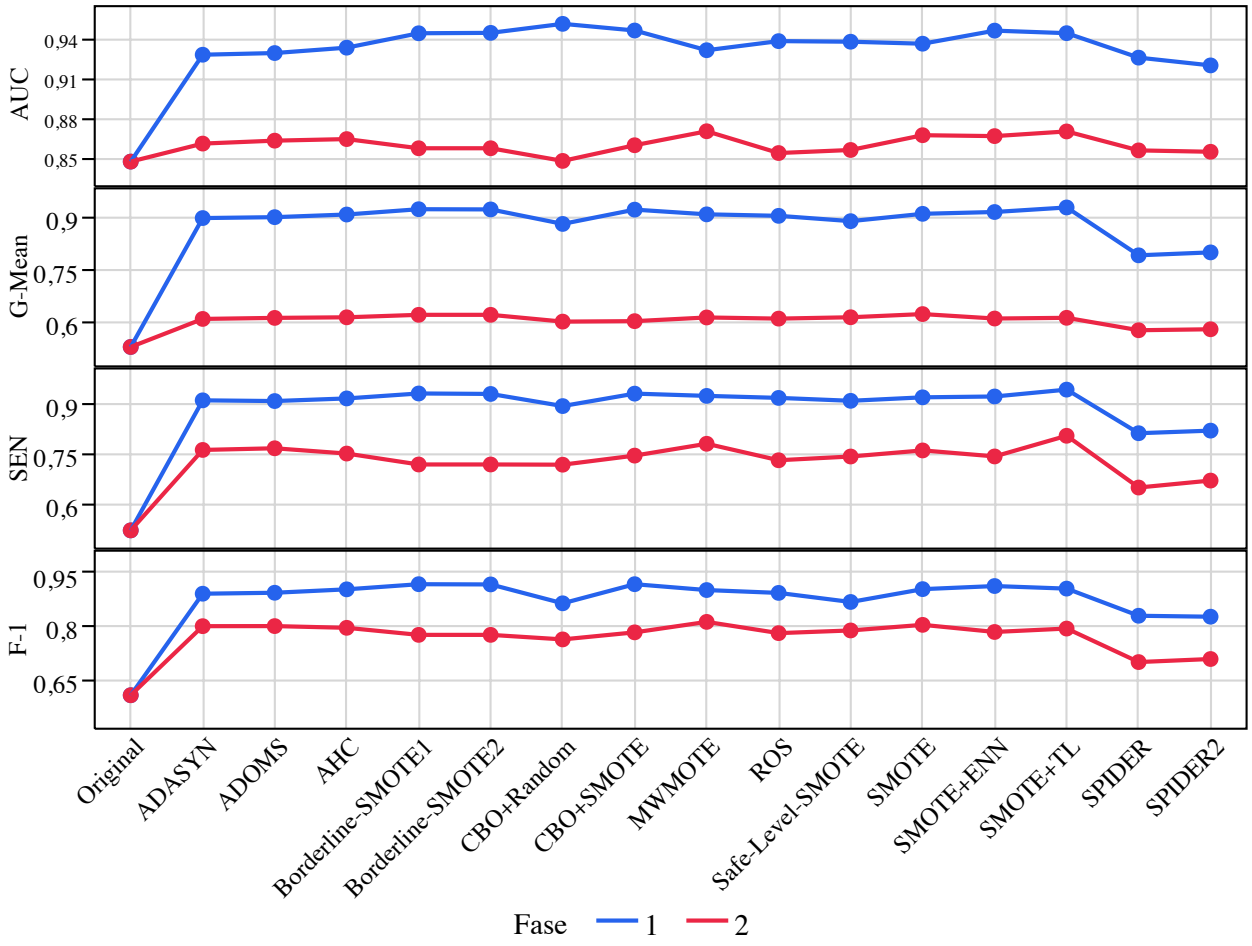


Figura 5.3: AUCs, *G-Mean*s, SENs e *F-1*s médias de teste em cada fase por técnica de sobre-amostragem.

A Figura 5.4 mostra a diferença dos desempenhos da classificação nos conjuntos de teste e treino em ambas as fases: os valores de treino não diferem entre as fases, indicando que o problema de sobre-otimismo está associado aos conjuntos de teste.

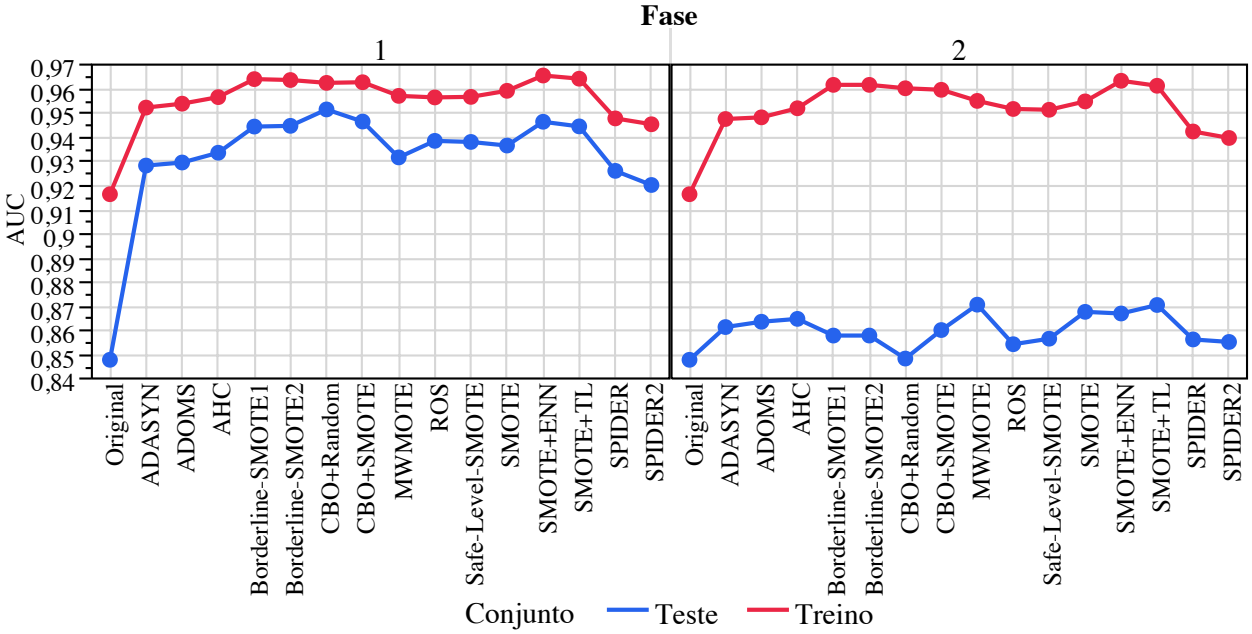


Figura 5.4: AUCs médias de teste e treino em cada técnica de sobre-amostragem por fase.

Os valores médios e o seus respetivos desvios padrão da AUC, nos diferentes classificadores e conjuntos (teste e treino), para cada fase estão sumariados na Tabela 5.2 (os valores das restantes métricas são apresentados nas Tabelas B.3, B.4 e B.5).

Na Figura 5.5, onde estão representados os valores médios da AUC de teste para cada classificador, o comportamento descrito anteriormente mantém-se (a Fase 1 mostra-se otimista em relação à Fase 2). Assim, uma vez que este comportamento é semelhante a todos os classificadores, podemos ainda inferir que o sobre-otimismo não advém do modelo usado para classificar os padrões, mas sim da metodologia usada para processar essa classificação.

Analisando as diferenças entre os resultados do treino e teste (Figura 5.4), é possível perceber que não está a ocorrer *overfitting* na Fase 1, mas os resultados dos conjuntos de teste são sobre-otimistas: os resultados de treino são semelhantes em ambas as fases, os classificadores comportam-se da mesma forma e como tal, a diferença entre as Fases 1 e 2 não reside nem nos dados de treino nem nos classificadores utilizados, mas sim nos dados usados para teste. Os conjuntos de teste da Fase 1 têm características idênticas aos conjuntos de treino (estão balanceados e podem ter réplicas ou padrões semelhantes aos originais), enquanto que na Fase 2 os conjuntos de teste seguem a mesma estrutura dos dados originais (têm a mesma distribuição entre classes que o *dataset* original e são constituídos por instâncias que não foram replicadas nem usadas de nenhuma forma para a sobre-amostragem). Assim sendo, na Fase 2 as técnicas de pré-processamento alteram os conjuntos de treino (balanceamento), mas os testes mantêm-se inalterados: a estrutura original do *dataset* é mantida nos conjuntos de teste, e altera-se a estrutura dos conjuntos de

treino (balanceamento), com o objetivo de melhorar as classificações num contexto de dados não balanceados (não sendo, por esse motivo, alterada a estrutura dos conjuntos de teste).

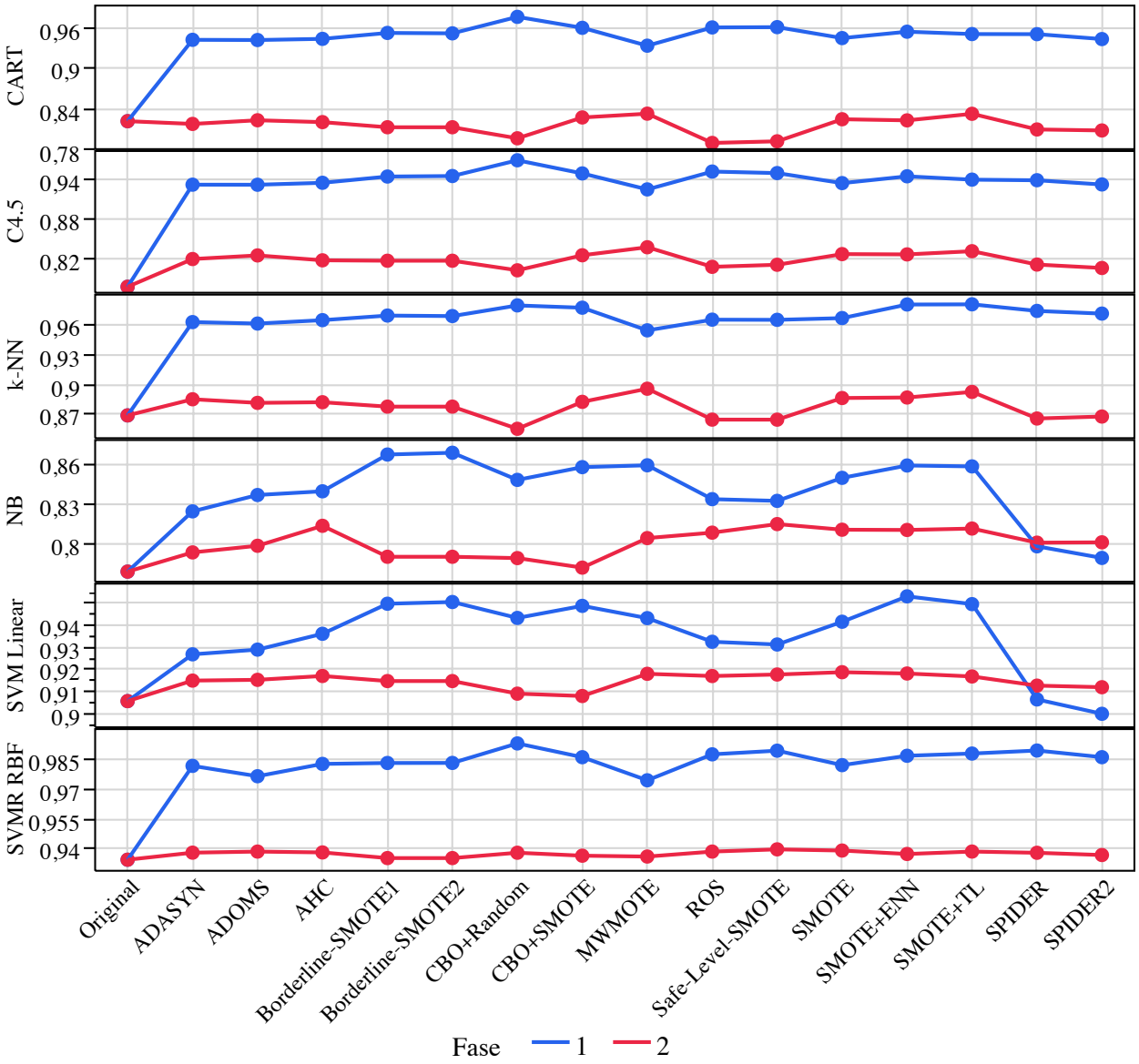


Figura 5.5: AUCs médias de teste para cada classificador por técnica de sobre-amostragem, divido por fases.

Pretende-se portanto, na Fase 2, que os classificadores aprendam a partir de um contexto balanceado, conseguindo generalizar para um contexto não balanceado, e obter melhores resultados do que nos datasets originais, em que esse balanceamento não é realizado. Essa melhoria ocorre, como se pode comprovar a partir da Figura 5.4, sem que os resultados do teste sejam sobre-estimados (como acontece na Fase 1). Existe no entanto, uma situação em que não se verifica a melhoria dos resultados originais para a Fase 2: quando é utilizado o CBO+*Random* para a sobre-amostragem. O segundo pior resultado é atribuído ao ROS, e ambos os métodos sofrem a desvantagem de gerar réplicas exactas de padrões existentes, estando portanto a causar *overfitting*.

Tabela 5.2: AUC de teste e treino média para todos os algoritmos de sobre-amostragem e métodos de classificação, considerando ambas as fases. Os melhores valores para cada coluna encontram-se a negrito.

		CART		**C4.5**		**k-NN**		**SVM Linear**		**SVM RBF**		**NB**	
	Método	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2
Teste	Original	0,8215±0,1297	0,8215±0,1297	0,7789±0,1602	0,7789±0,1602	0,8692±0,1148	0,8692±0,1148	0,9057±0,0999	0,9057±0,0999	0,9345±0,0839	0,9345±0,0839	0,7786±0,1437	0,7786±0,1437
	ADASYN	0,9423±0,0701	0,8174±0,1267	0,9323±0,08	0,8206±0,1329	0,9628±0,0609	0,8853±0,1049	0,9269±0,0835	0,915±0,0862	0,9817±0,0482	0,9381±0,0794	0,8246±0,1242	0,7932±0,1306
	ADOMS	0,942±0,0707	0,8229±0,1374	0,9322±0,0802	0,826±0,1389	0,9613±0,0633	0,8817±0,108	0,929±0,0803	0,9153±0,0879	0,9765±0,0543	0,9386±0,0801	0,8371±0,1197	0,7983±0,1364
	AHC	0,9437±0,0668	0,8201±0,128	0,9351±0,0763	0,8187±0,1384	0,9647±0,0583	0,8824±0,1071	0,9362±0,0753	0,9171±0,0869	0,9827±0,0434	0,9382±0,0803	0,8399±0,1147	**0,8135±0,1238**
	Borderline-SMOTE1	0,9525±0,0656	0,8125±0,1304	0,9445±0,0754	0,818±0,1378	0,9693±0,0528	0,878±0,1097	0,9498±0,0717	0,9148±0,0889	0,9831±0,0438	0,9354±0,0828	**0,868±0,116**	0,7899±0,1276
	Borderline-SMOTE2	0,9519±0,0642	0,8125±0,1304	0,9453±0,0721	0,818±0,1378	0,9688±0,0524	0,878±0,1097	**0,9506±0,0692**	0,9148±0,0889	0,9832±0,0423	0,9354±0,0828	**0,8693±0,1141**	0,7899±0,1276
	CBO+*Random*	**0,9764±0,0415**	0,7961±0,1374	**0,9689±0,0506**	0,8037±0,1383	0,9795±0,0416	0,8557±0,1163	0,9434±0,0733	0,9091±0,0952	**0,993±0,0247**	0,9381±0,0811	0,8486±0,1127	0,7889±0,1363
	CBO+SMOTE	0,9601±0,0549	0,8271±0,1259	0,9492±0,0656	0,8263±0,1274	0,9772±0,0443	0,8827±0,1075	0,9489±0,0678	0,908±0,0939	0,9861±0,0354	0,9366±0,0805	0,8583±0,1069	0,7816±0,1306
	MWMOTE	0,9336±0,0721	**0,8327±0,1248**	0,9252±0,0764	**0,8383±0,1282**	0,9545±0,0665	**0,8959±0,0976**	0,9433±0,0719	0,9181±0,0863	0,9745±0,0509	0,9361±0,0811	0,8598±0,1113	0,8042±0,1266
	ROS	0,9607±0,06	0,7892±0,1451	**0,952±0,0722**	0,8089±0,1405	0,9652±0,061	0,8649±0,1156	0,9326±0,077	0,9171±0,0872	0,9875±0,0392	0,9386±0,0798	0,8339±0,119	0,8083±0,1285
	Safe-Level-SMOTE	**0,9611±0,0606**	0,7916±0,1418	0,9497±0,0767	0,812±0,1373	0,965±0,0621	0,8649±0,1153	0,9313±0,0808	0,9177±0,0866	0,9893±0,0365	**0,9398±0,0792**	0,8326±0,1186	**0,8149±0,1288**
	SMOTE	0,9449±0,0658	0,8244±0,1245	0,9346±0,0771	0,828±0,1296	0,9668±0,0548	0,8866±0,1055	0,9417±0,076	**0,9188±0,0857**	0,9821±0,0462	**0,9392±0,0799**	0,8502±0,1205	0,8106±0,1255
	SMOTE+ENN	0,9544±0,053	0,8227±0,129	0,9449±0,0584	0,8275±0,1318	**0,9804±0,0314**	0,8871±0,1059	**0,9532±0,0578**	**0,9182±0,0874**	0,9868±0,0283	0,9374±0,0794	0,8596±0,114	0,8104±0,1239
	SMOTE+TL	0,9509±0,0558	**0,8325±0,1288**	0,9399±0,069	**0,8324±0,131**	**0,9806±0,0363**	**0,8927±0,1012**	0,9496±0,065	0,9169±0,0883	0,9879±0,0304	0,9386±0,0781	0,8589±0,115	0,8114±0,1246
	SPIDER	0,9507±0,0482	0,8092±0,1397	0,939±0,0626	0,8123±0,1409	0,974±0,0428	0,866±0,1157	0,9065±0,0954	0,9127±0,0912	**0,9894±0,033**	0,938±0,0798	0,7979±0,1255	0,8007±0,1282
	SPIDER2	0,9434±0,054	0,8077±0,1424	0,9326±0,0663	0,8071±0,1377	0,9713±0,0489	0,868±0,1153	0,8999±0,0998	0,912±0,0908	0,9861±0,0366	0,9369±0,0796	0,789±0,1257	0,8009±0,1283
Treino	Original	0,9662±0,0347	0,9662±0,0347	0,8697±0,1368	0,8697±0,1368	0,9695±0,0512	0,9695±0,0512	0,9163±0,0965	0,9163±0,0965	0,9688±0,0585	0,9688±0,0585	0,809±0,136	0,809±0,136
	ADASYN	0,9936±0,0121	0,9942±0,011	0,9718±0,0433	0,9741±0,0403	0,9923±0,0216	0,9912±0,022	0,9319±0,0779	0,9301±0,078	0,9954±0,0221	0,9608±0,0697	0,8294±0,1206	0,8355±0,118
	ADOMS	0,993±0,0115	0,9936±0,0101	0,9707±0,0452	0,9731±0,0419	0,99±0,0219	0,9877±0,0282	0,9343±0,0734	0,9322±0,074	0,9933±0,0263	0,9675±0,0596	0,8428±0,1171	0,8362±0,121
	AHC	0,993±0,013	0,9938±0,0112	0,973±0,0431	0,9738±0,0405	0,9916±0,031	0,9882±0,0377	0,9413±0,0689	0,9384±0,0699	0,9971±0,0107	0,9689±0,0567	0,8442±0,1113	0,8495±0,1092
	Borderline-SMOTE1	0,9939±0,0108	0,9942±0,0108	0,975±0,0391	0,9763±0,0361	0,9929±0,0158	0,9916±0,0211	0,9542±0,0631	**0,9541±0,0612**	0,9974±0,0091	**0,9782±0,0455**	**0,8717±0,1113**	**0,8764±0,1086**
	Borderline-SMOTE2	0,9937±0,0115	0,9942±0,0108	0,975±0,0399	0,9763±0,0361	0,9915±0,021	0,9916±0,0211	0,9546±0,063	**0,9541±0,0612**	0,9965±0,0145	**0,9782±0,0455**	**0,8715±0,1116**	**0,8764±0,1086**
	CBO+*Random*	**0,9962±0,0091**	**0,9974±0,0069**	**0,9872±0,0254**	**0,9897±0,02**	**0,9959±0,0163**	0,9922±0,0289	0,9465±0,0677	0,9458±0,0635	0,9981±0,0083	0,975±0,0454	0,8519±0,1086	0,8623±0,0963
	CBO+SMOTE	0,9951±0,0111	0,9958±0,0094	0,9795±0,0358	0,9816±0,0314	0,9934±0,0187	0,9913±0,0272	0,952±0,0623	0,9486±0,062	0,9968±0,0102	0,9745±0,0449	0,8604±0,1043	0,8668±0,097
	MWMOTE	0,9924±0,0117	0,9927±0,0117	0,9687±0,0432	0,9692±0,0428	0,9848±0,0295	0,9862±0,0284	0,9483±0,0645	0,9462±0,0654	0,985±0,0359	0,9693±0,0567	0,8642±0,1075	0,8675±0,1069
	ROS	0,9946±0,0123	0,9958±0,0105	0,9807±0,0393	0,9832±0,0342	0,9921±0,0308	0,9869±0,0405	0,9371±0,0719	0,9346±0,0717	0,9963±0,0208	0,9671±0,0606	0,8388±0,1153	0,8434±0,1145
	Safe-Level-SMOTE	0,9943±0,0133	0,9958±0,0107	0,9799±0,043	0,9833±0,0346	0,993±0,0302	0,9854±0,0441	0,9371±0,0726	0,9355±0,0716	0,9979±0,0097	0,9662±0,0612	0,8388±0,1147	0,8429±0,1142
	SMOTE	0,9933±0,0119	0,994±0,011	0,9738±0,0417	0,9757±0,0394	0,992±0,0195	0,9915±0,0212	0,9466±0,0685	0,944±0,0705	0,9961±0,0177	0,9688±0,0567	0,8541±0,1163	0,8558±0,1132
	SMOTE+ENN	**0,9963±0,006**	**0,9966±0,0058**	**0,983±0,0219**	**0,9855±0,0188**	0,9956±0,009	**0,9965±0,0071**	**0,9572±0,0524**	**0,9558±0,0538**	**0,9982±0,0059**	**0,9786±0,0426**	0,8638±0,1104	0,868±0,1067
	SMOTE+TL	0,9946±0,0086	0,9956±0,0069	0,9792±0,0301	0,9813±0,027	0,9952±0,0116	**0,9955±0,0117**	**0,9553±0,0559**	0,9534±0,0575	**0,9984±0,0048**	0,9761±0,0454	0,8633±0,1105	0,8666±0,1055
	SPIDER	0,9924±0,0122	0,9947±0,0087	0,9785±0,0402	0,9831±0,0312	**0,9965±0,016**	0,9925±0,0255	0,914±0,0914	0,9143±0,0887	0,998±0,01	0,9571±0,0735	0,8082±0,1239	0,8134±0,1222
	SPIDER2	0,9927±0,0113	0,9945±0,0083	0,9764±0,0365	0,9795±0,031	0,9956±0,0159	0,9926±0,0262	0,9099±0,0943	0,9121±0,0905	0,9975±0,0107	0,9518±0,0808	0,801±0,1252	0,8084±0,1243

De modo a relacionar os resultados da classificação com a complexidade dos conjuntos de treino e teste de ambas as fases, as Figuras 5.6, 5.7 e 5.8 expõem os valores médios da diferença (em módulo) entre as métricas de complexidade nos conjuntos de treino e teste, em cada fase (obtidos pela Tabela B.6), e com a informação adicional do valor da AUC média do respetivo teste. Numa observação global temos que a diferença (em módulo) das métricas é sempre superior na Fase 2 em comparação à Fase 1, o que é plausível com o comportamento observado anteriormente, já que na Fase 1 os conjuntos de teste e treino apresentam características semelhantes.

Os métodos SPIDER e SPIDER2 mostram por vezes um comportamento antagónico aos restantes métodos (Figuras 5.5 a 5.8), o que pode ser explicado pelos seus processos de geração de novos dados, que diferem dos restantes métodos. Na implementação utilizada neste trabalho, o SPIDER utiliza uma estratégia de amplificação fraca, em que os exemplos da classe minoritária são replicados mediante a existência de exemplos majoritários classificados como **seguro** entre os seus 5 vizinhos mais próximos. Caso o *dataset* seja complexo e não existam muitos exemplos "fáceis" de classificar (exemplos **seguros**), os exemplos minoritários não chegam a ser replicados. O SPIDER2 usa uma estratégia de amplificação forte, em que a amplificação dos exemplos minoritários é estendida a 7 vizinhos (k = 5 + 2) e procede-se ainda à reetiquetação dos exemplos majoritários classificados como **ruído**: as classes dos pontos originais são modificadas directamente. Adicionalmente, o SPIDER e o SPIDER2 são os únicos métodos que não garantem uma representação igual de classes (não garantem que o *dataset* resultante passa a ser balanceado). Estas diferenças relativamente aos outros métodos podem estar na origem do seu comportamento errático quer nas métricas de classificação quer nas métricas de complexidade: em particular para o *SPIDER*, verificámos que (i) gera *datasets* que não continham grandes modificações relativamente ao *dataset* original (IR e número de instâncias) e (ii) é o método que gera conjuntos de treino com o rácio discriminativo F1 mais baixo, dando origem às maiores diferenças entre treino e teste (Figura 5.6).

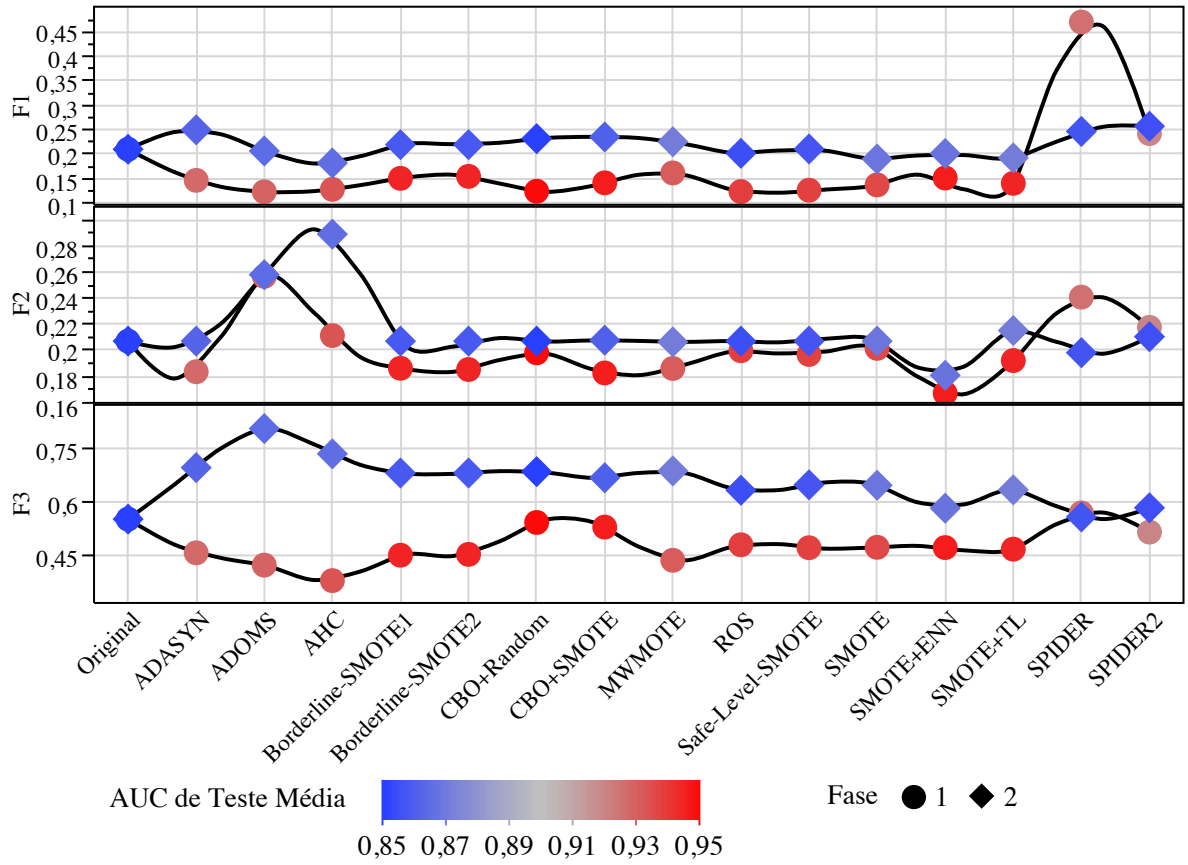


Figura 5.6: Diferença entre o módulo de treino e teste das métricas F1, F2 e F3 em cada fase por técnica de sobre-amostragem, onde a cor nos pontos representa a média da AUC de teste.

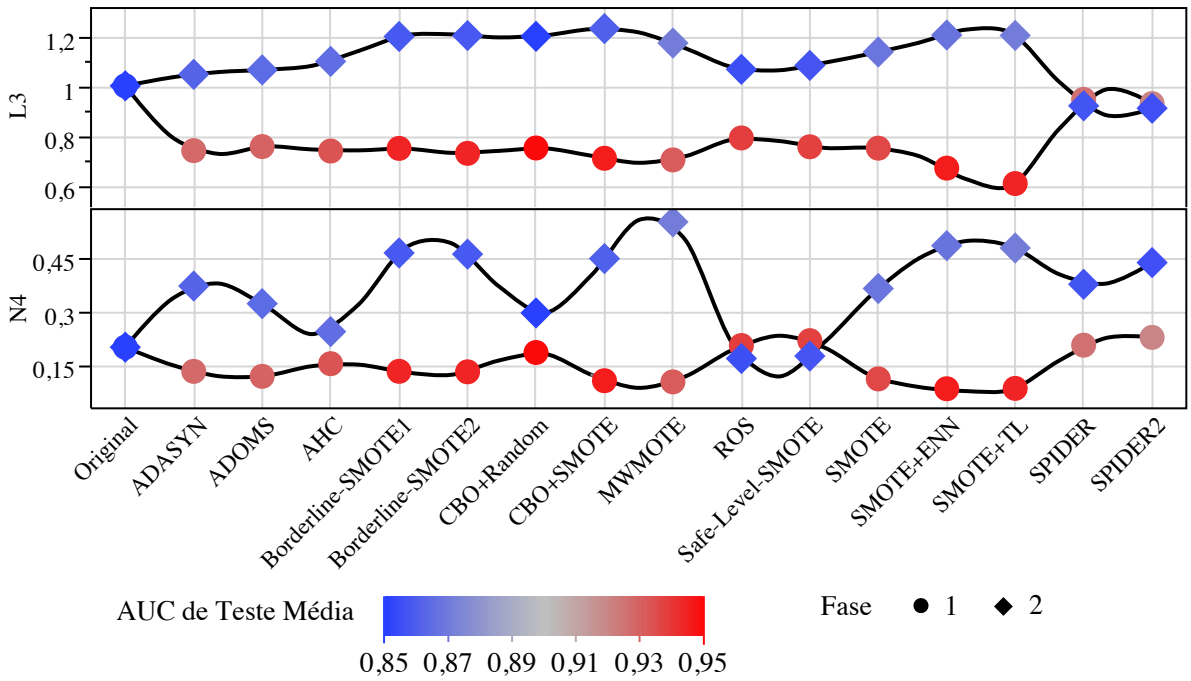


Figura 5.7: Diferença entre o módulo de treino e teste das métricas L3 e N4 em cada fase por técnica de sobre-amostragem, onde a cor nos pontos representa a média da AUC de teste.

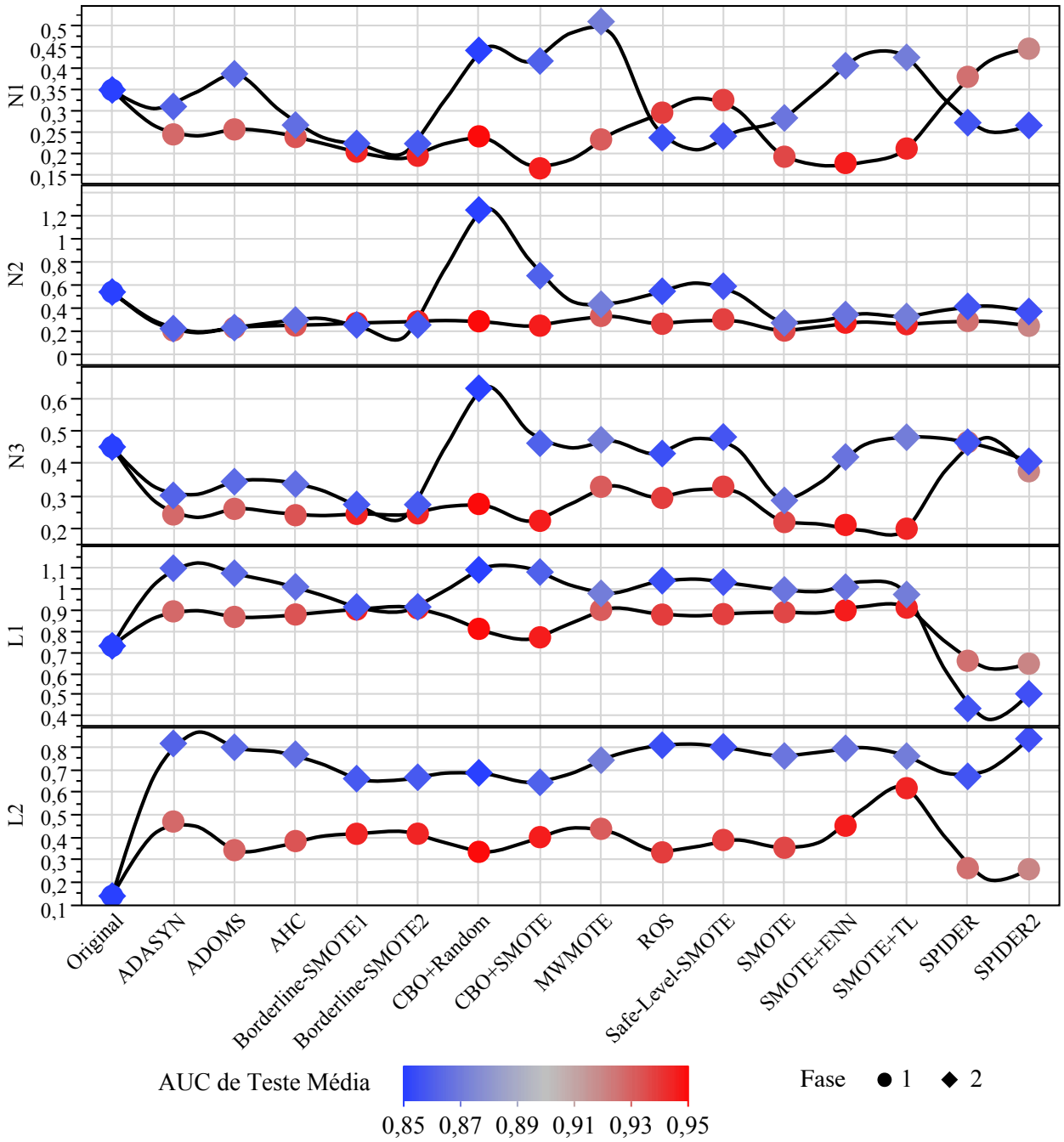


Figura 5.8: Diferença entre o módulo de treino e teste das métricas N1, N2, N3, L1 e L2 em cada fase por técnica de sobre-amostragem, onde a cor nos pontos representa a média da AUC de teste.

5.2.2 Exploração da Fase 2: Estudo da complexidade

Os resultados mostram que a Fase 2 é a abordagem mais correta de validação cruzada em dados não balanceados, e por isso para a restante análise foca-se somente nos resultados da Fase 2. No que diz respeito a identificação dos melhores métodos, foram aplicadas três estratégias de análise:

- A média total dos valores da AUC média dos conjuntos de teste para cada classificador (Estratégia 1);
- O *rank* médio dos valores da AUC (média de teste) por técnica de pré-processamento (onde a melhor recebe o valor 1 a seguinte 2 e assim sucessivamente) para cada classificador (Estratégia 2);
- A média do *rank* médio, para cada classificador, dos valores da AUC de teste em cada *dataset* por técnica de pré-processamento (Estratégia 3).

De salientar que os valores dos *ranks* para cada classificador utilizados para obter a média da última estratégia estão na Tabela B.2. Os valores obtidos nas três estratégias são detalhados na Tabela 5.3 e mostram que as quatro melhores técnicas de pré-processamento implementadas estão presentes em todas as três estratégicas: SMOTE+TL, MWMOTE, SMOTE e SMOTE+ENN. No entanto, a ordem do melhor para o pior varia na estratégia: na Tabela 5.4 essa informação está sumariada e permite afirmar que o melhor método é o SMOTE+TL seguido do MWMOTE.

Tabela 5.3: Resultados das três estratégias adoptadas para identificar as melhores técnicas de sobre-amostragem, onde os melhores valores estão negrito.

	Estratégia		
Método	**1**	**2**	**3**
Original	0,848±0,066	13,667±3,830	11,471±4,981
ADASYN	0,862±0,059	8,000±2,098	8,215±4,223
ADOMS	0,864±0,057	7,000±2,000	7,606±4,195
AHC	0,865±0,055	6,000±2,280	8,088±3,817
Borderline-SMOTE1	0,858±0,060	10,667±1,966	8,743±4,119
Borderline-SMOTE2	0,858±0,060	10,667±1,966	8,743±4,119
CBO+*Random*	0,849±0,063	13,500±2,811	9,821±4,419
CBO+SMOTE	0,860±0,058	9,333±5,317	8,745±4,475
MWMOTE	**0,871±0,053**	**4,333±4,844**	**7,199±4,332**
ROS	0,855±0,063	9,833±5,076	9,019±4,034
Safe-Level-SMOTE	0,857±0,061	8,000±6,753	8,411±4,075
SMOTE	**0,868±0,054**	**3,000±1,265**	7,222±3,460
SMOTE+ENN	0,867±0,054	5,000±2,828	**7,201±4,072**
SMOTE+TL	**0,871±0,052**	**3,167±1,941**	**6,535±4,094**
SPIDER	0,856±0,059	11,000±1,673	9,412±4,665
SPIDER2	0,855±0,059	11,833±2,137	9,569±4,764

Tabela 5.4: Os melhores métodos de sobre-amostragem e respetiva posição com base nas 3 estratégia.

	Estratégia		
Posição	**1**	**2**	**3**
1º	SMOTE+TL	SMOTE	SMOTE+TL
2º	MWMOTE	SMOTE+TL	MWMOTE
3º	SMOTE	MWMOTE	SMOTE+ENN
4º	SMOTE+ENN	SMOTE+ENN	SMOTE

Para compreender melhor a evolução das métricas de complexidade na Fase 2, apresentam-se as Figuras 5.9, 5.10 e 5.11, onde os valores das métricas no conjunto de treino estão organizadas e agrupadas pela AUC média dos conjuntos de testes. No panorama global, analisando estes resultados, parece existir uma relação quase linear entre as métricas de complexidade nos *datasets* de treino e o desempenho da classificação nos *datasets* de teste.

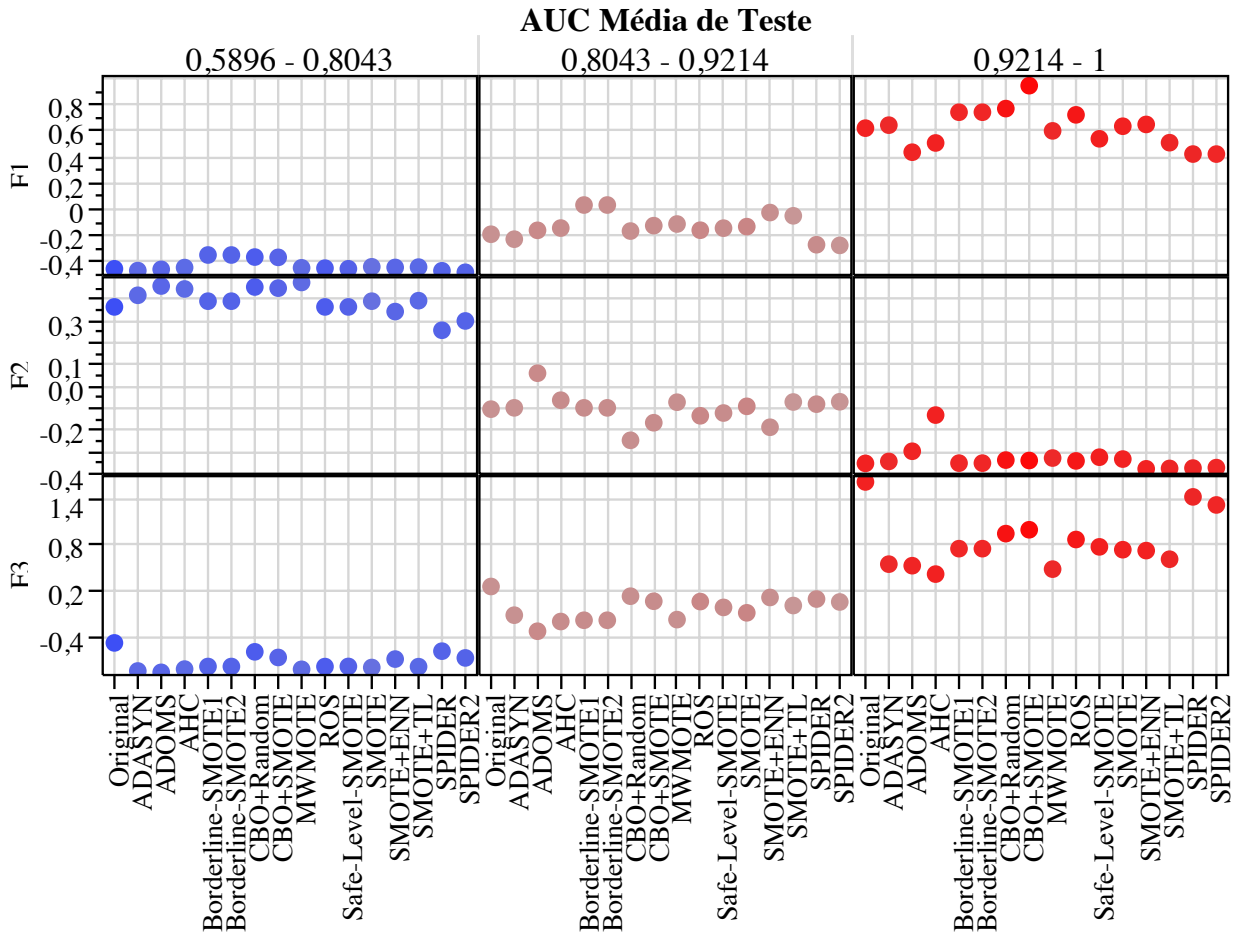


Figura 5.9: Evolução das métricas F1, F2 e F3 (treino) em termos de AUC média (teste) para cada técnica de sobre-amostragem, considerando a Fase 2.

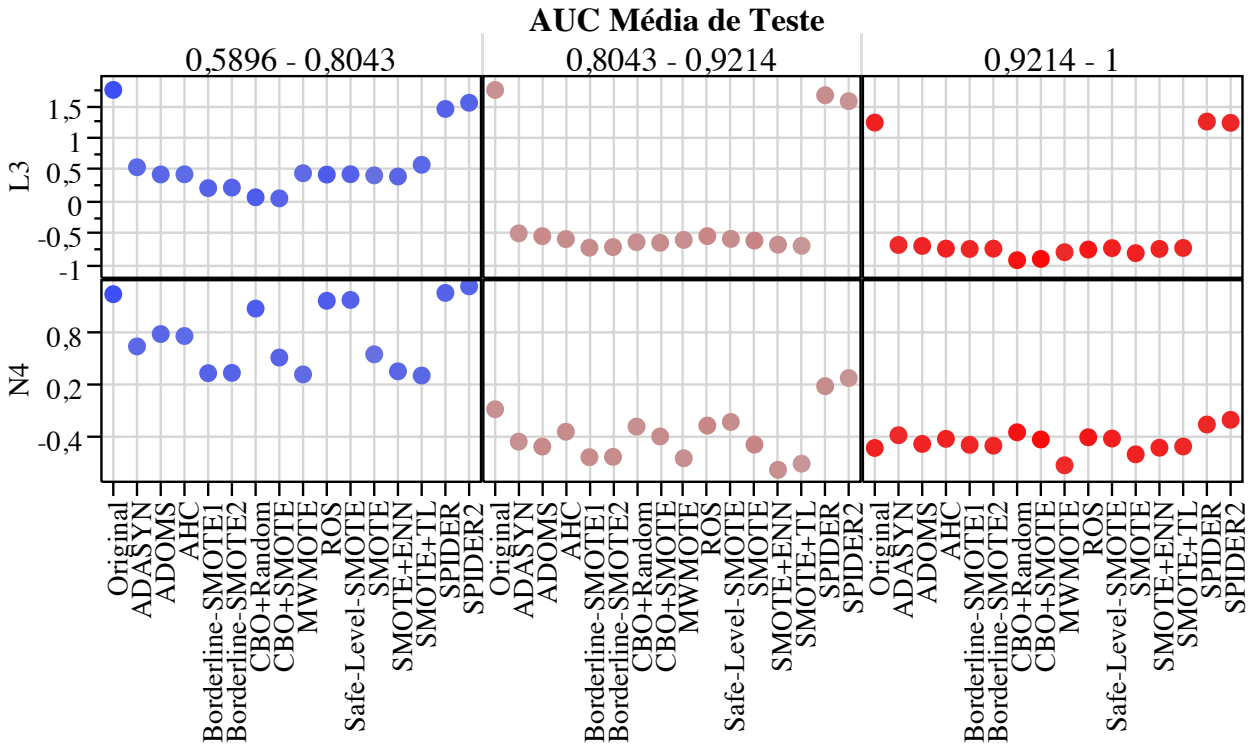


Figura 5.10: Evolução das métricas L3 e N4 (treino) em termos de AUC média (teste) para cada técnica de sobre-amostragem, considerando a Fase 2.

Analisando cada uma das métricas de complexidade, quando os valores de AUC aumentam, F1 e F3 aumentam, o que significa que os *datasets* com melhores AUC apresentam variáveis com maior poder discriminativo e com maior fração de pontos separáveis por cada variável (ou seja, variáveis mais "eficientes"). Por outro lado, o aumento dos valores de AUC está associado à diminuição dos valores de F2, o que corresponde a regiões de sobreposição entre classes menores. L3 e N4 também diminuem com o aumento da AUC, sugerindo fronteiras mais suaves e menos sobrepostas e uma geometria e topologia mais linear. N1, N2, N3, L1 e L2 também diminuem, o que sugere no global uma maior separabilidade entre classes.

As Figuras 5.12, 5.13 e 5.14 possibilitam avaliar o efeito dos métodos de sobre-amostragem na complexidade dos *datasets* de treino e associar esses resultados aos valores das AUCs médias obtidas nos conjuntos de teste. A análise das métricas de complexidade deve ser analisada como um todo, pois compreender o efeito dos melhores métodos de pré-processamento utilizados com base na informação de apenas uma métrica não seria nem correto nem conclusivo. Por exemplo os métodos *Borderline*-SMOTE1 e *Borderline*-SMOTE2 apresentam o melhor valor médio de F1 no treino, mas não obtêm os melhores resultados da AUC no teste. Contudo quando são observadas as restantes métricas, tais como F2 e F3, temos que estas duas técnicas apresentam valores superiores (F2) e inferiores (F3) que os melhores métodos,

tais como o SMOTE+TL. O SMOTE+TL e o SMOTE+ENN apresentam comportamentos semelhantes na complexidade, ambos se encontram entre os melhores resultados nas métricas F1, F2, L3, N4 e N3, e nas restantes métricas apresentam valores intermédios (equivalentes aos restantes métodos). O MWMOTE, que é dos melhores métodos de pré-processamento, também apresenta dos melhores resultados para as métricas F1, L3 e N4 (o melhor), nas métricas N1, N2 e N3 é o método que menos varia em relação aos valores dos *datasets* originais e nas restantes métricas apresenta valores intermédios.

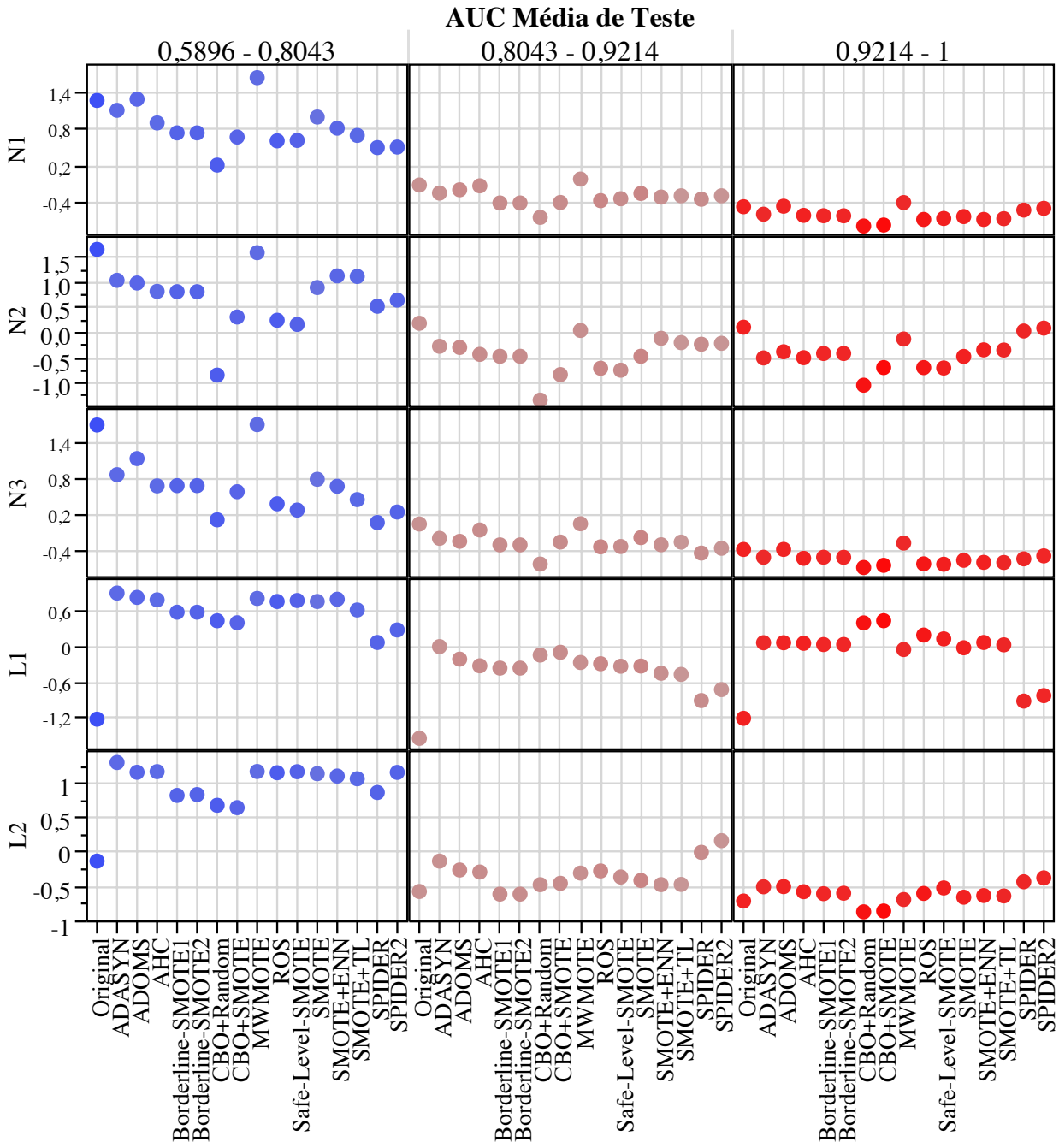


Figura 5.11: Evolução das métricas N1, N2, N3, L1 e L2 (treino) em termos de AUC média (teste) para cada técnica de sobre-amostragem, considerando a Fase 2.

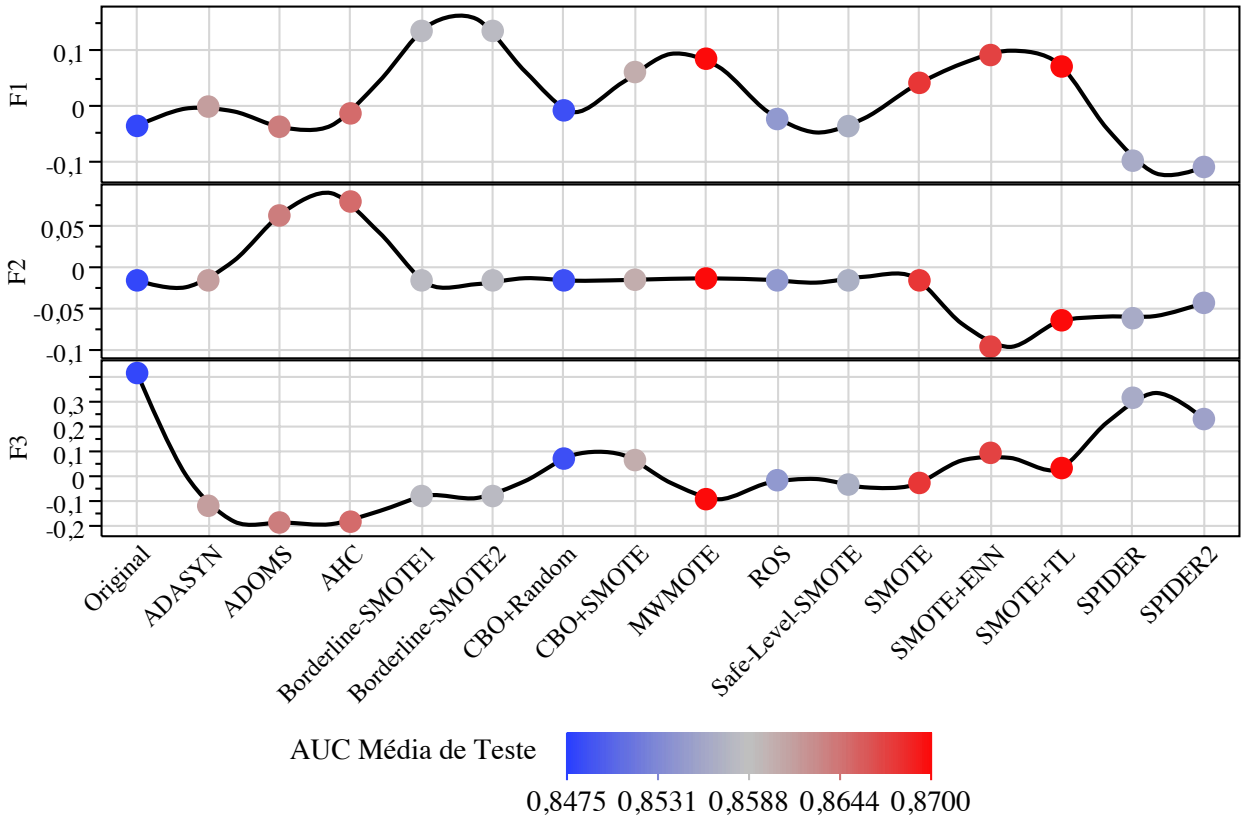


Figura 5.12: Valores médios das métricas F1, F2 e F3 nos conjuntos de treino por cada técnica de sobre-amostragem, onde a cor dos pontos é a média das AUCs de teste.

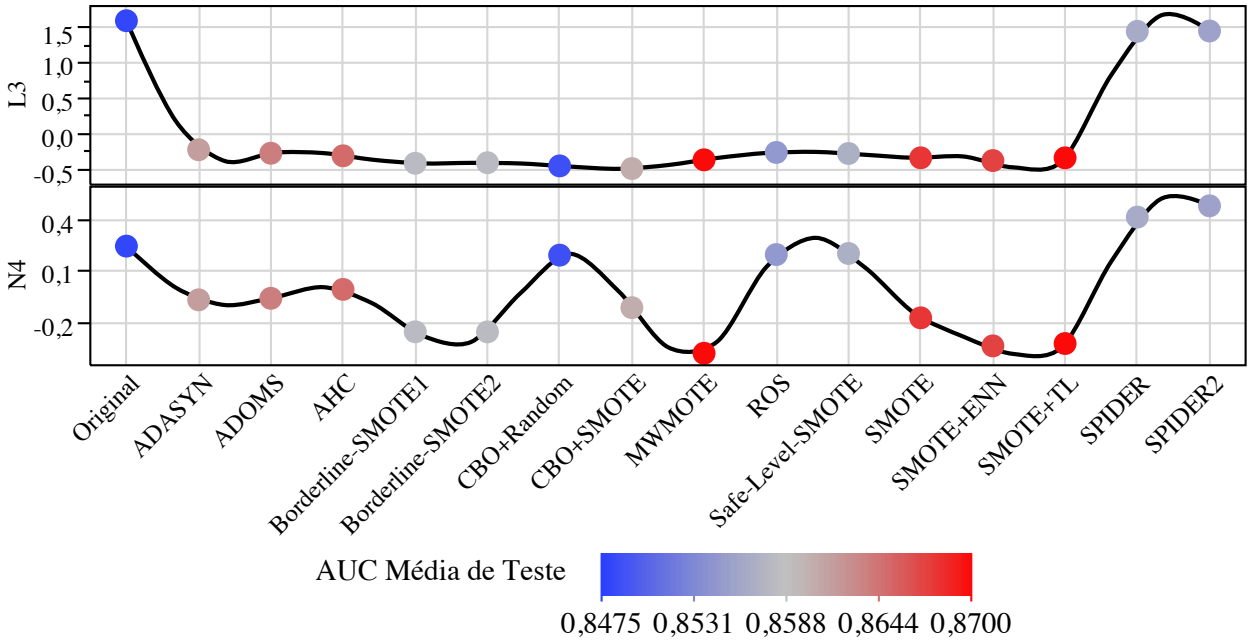


Figura 5.13: Valores médios das métricas L3 e N4 nos conjuntos de treino por cada técnica de sobre-amostragem, onde a cor dos pontos é a média das AUCs de teste.

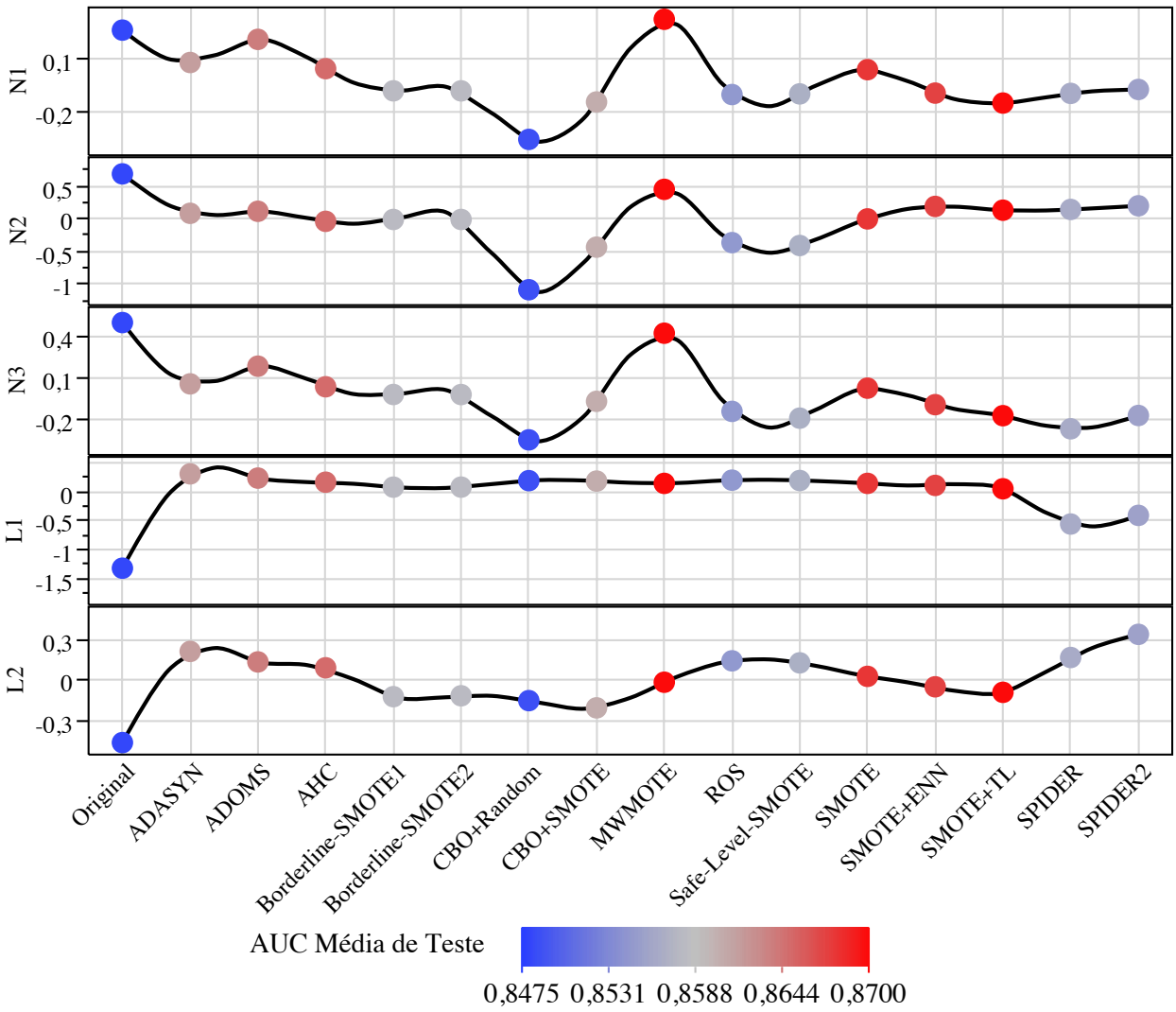
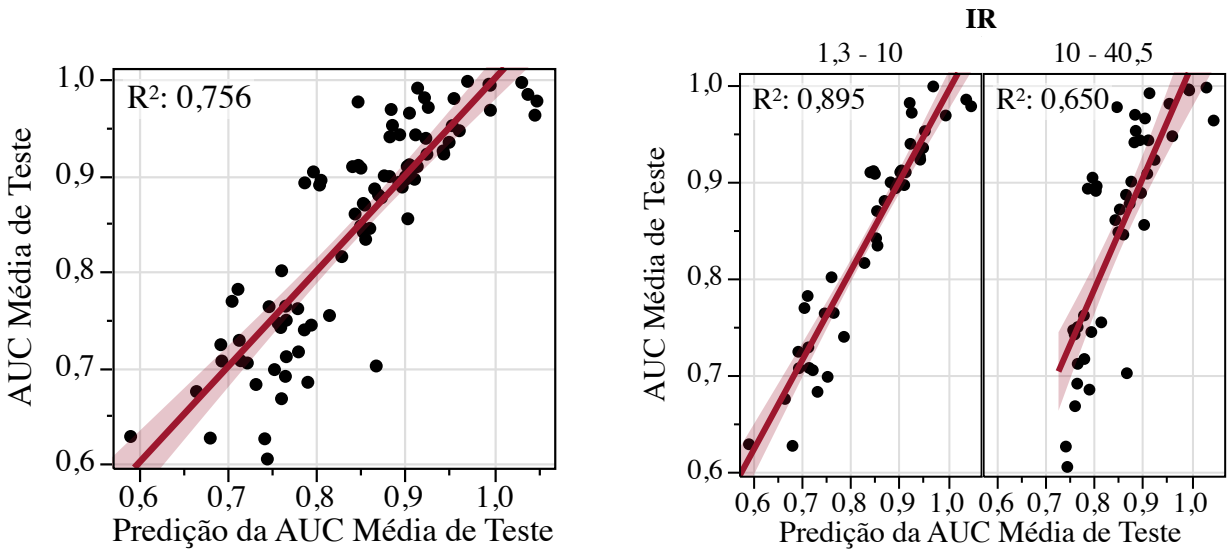


Figura 5.14: Valores médios das métricas N1, N2, N3, L1 e L2 nos conjuntos de treino por cada técnica de sobre-amostragem, onde a cor dos pontos é a média das AUCs de teste.

Esta aparente relação entre os resultados da AUC de teste e as métricas de complexidade dos datasets de treino correspondentes instigou a procura de um modelo de regressão que estimasse o valor da AUC (teste) a partir do valores das métricas de complexidade (treino). Primeiramente efetuou-se apenas esta análise para os 86 *datasets* originais, onde através dos verdadeiros valores da AUC e os estimados pela equação 5.1 foi possível obter uma regressão linear com um r^2 de 0.756 (Figura 5.15a). Agrupando os *datasets* por IR (2 grupos, *IR* inferior e superior a 10) notou-se que a estimativa dos resultados da AUC tem um r^2 muito superior (0.895) para os *datasets* com IR inferior a 10, em contraste com os *datasets* com IR

superior a 10 (0.650).

$$\begin{aligned}\text{Predição} = {} & 0.9439 - 0.002155 \times \text{F1} + 0.002485 \times \text{F2} + 0.02517 \times \text{F3} \\ & + 0.04802 \times \text{L1} + 0.04108 \times \text{L2} - 0.004407 \times \text{L3} \\ & - 0.1328 \times \text{N1} + 0.003983 \times \text{N2} + 0.04377 \times \text{N3} \\ & - 0.02455 \times \text{N4}\end{aligned} \tag{5.1}$$



(a) (b)

Figura 5.15: Regressões lineares entre os verdadeiros valores médios da AUC dos *datasets* de teste originais e os previstos, onde no canto superior esquerdo de cada ajuste está o valor do coeficiente de determinação (r^2). Na Figura (b) os resultados de predição estão agrupados por IR superiores e inferiores a 10.

Estes resultados mostram que estimar as AUC pelas complexidades é possível e o valor de correlação do modelo é sobretudo satisfatório para IRs inferiores a 10. Assim sendo procedeu-se a uma nova análise, na qual se engloba todos os *datasets* originais e os sujeitos às técnicas de pré-processamento. A equação para estimar a AUC de teste obtida está representada na equação 5.2 e as regressões lineares obtidas para os *datasets* processados de cada método de sobre-amostragem mostram-se na Figura 5.16.

$$\begin{aligned}\text{Predição} = {} & 0.8593 - 0.01077 \times \text{F1} - 0.006369 \times \text{F2} + 0.03737 \times \text{F3} \\ & + 0.02194 \times \text{L1} - 0.05654 \times \text{L2} + 0.0004768 \times \text{L3} \\ & - 0.01037 \times \text{N1} - 0.0008070 \times \text{N2} + 0.0004026 \times \text{N3} \\ & - 0.01931 \times \text{N4}\end{aligned} \tag{5.2}$$

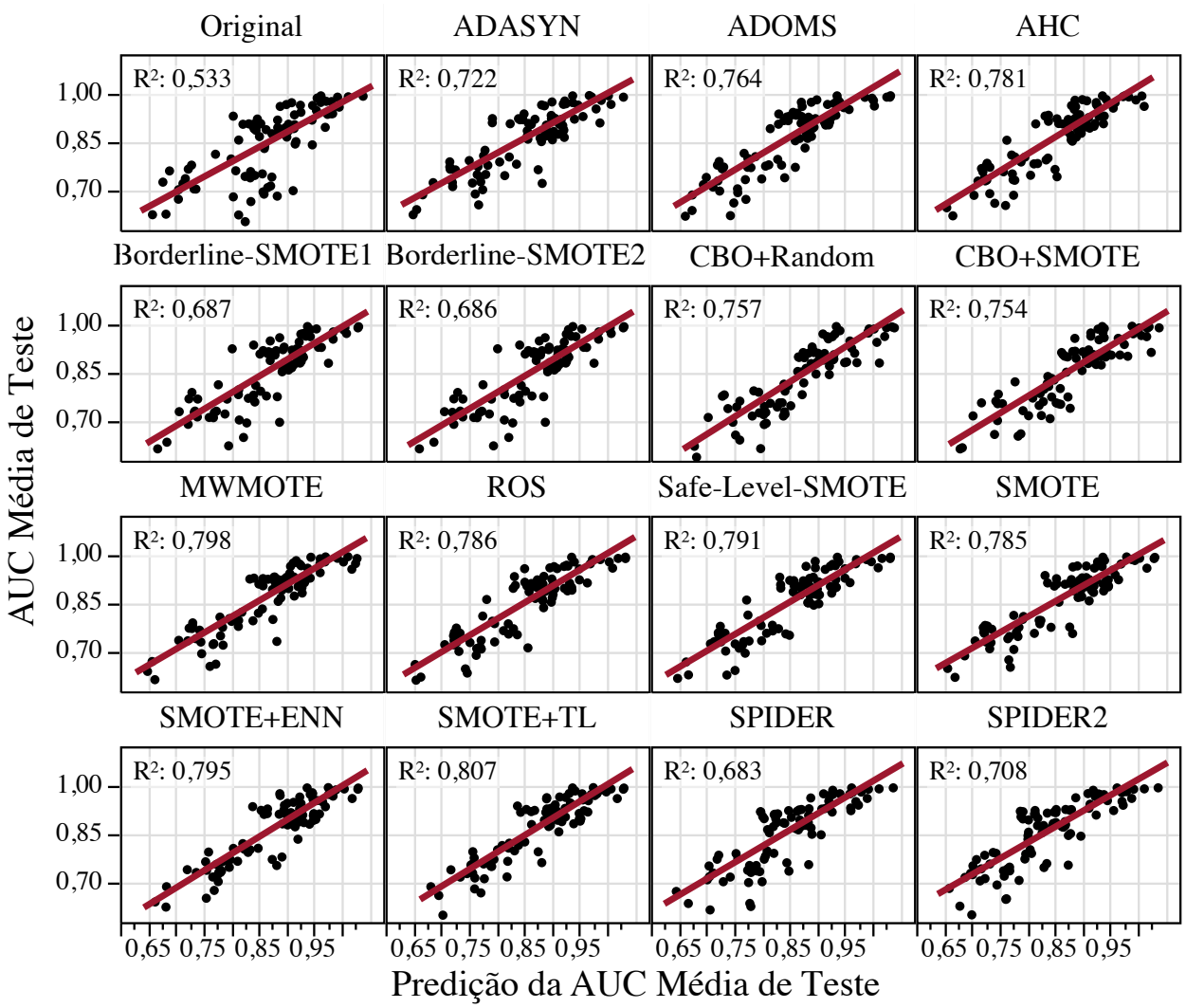


Figura 5.16: Regressões lineares entre todos verdadeiros valores médios da AUC de teste e os previstos, agrupados por método de pré-processamento. Os coeficientes de determinação (r^2) estão representados no canto superior esquerdo de cada ajuste.

Dos valores de r^2 temos que SMOTE+TL, MWMOTE e SMOTE+ENN são as técnicas com maior correlação entre os valores da AUC verdadeiros e os estimados, o que vai de encontro aos resultados das outras análises (onde estes três são considerados os melhores métodos de pré-processamento). A equação 5.2 prevê bem os 15 métodos de sobre-amostragem, ainda que para os resultados dos *datasets* originais o r^2 desça (0.533).

Numa linha de raciocínio semelhante, para melhor compreender a relação entre os valores das métricas de complexidade usadas e a sua influência nos resultados da classificação, foi aplicado o KMC a esses valores. Os resultados do KMC com $k = 4$ nos *datasets* originais mostram que os 6 melhores *datasets* na média da AUC (teste) estão contidos no mesmo *cluster* (na Figura 5.17 é o *cluster* vermelho): **ecoli-0vs1**, **iris0**, **shuttle-c0-vs-c4**, **dermatology-6**, **shuttle-6vs2-3** e **abalone-3vs11**.

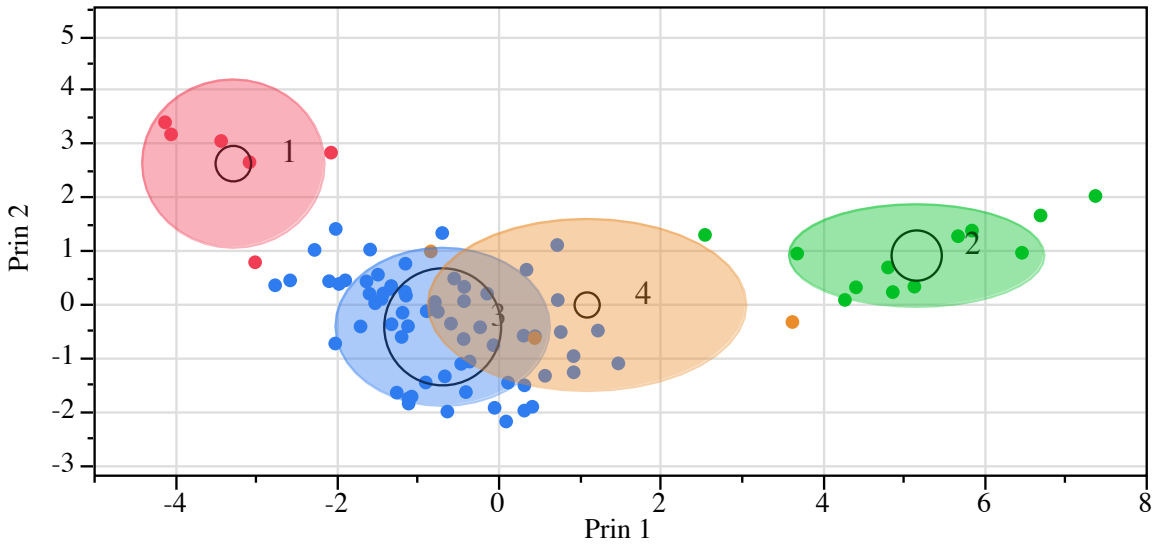


Figura 5.17: Representação das duas primeiras componentes principais dos *clusters* encontrados para as métricas de complexidade de todos os *datasets* de treino originais.

Por fim, quando aplicado o KMC com o mesmo k nos *datasets* originais e pré-processados (treino), os resultados mostram que há um *cluster* que contém os 70 *datasets* com os melhores resultados na AUC média de teste (na Figura 5.18 é o *cluster* laranja). Assim sendo, a avaliação através de *clustering* também permite identificar uma relação entre as melhores AUCs médias (teste) e a complexidade dos dados de treino.

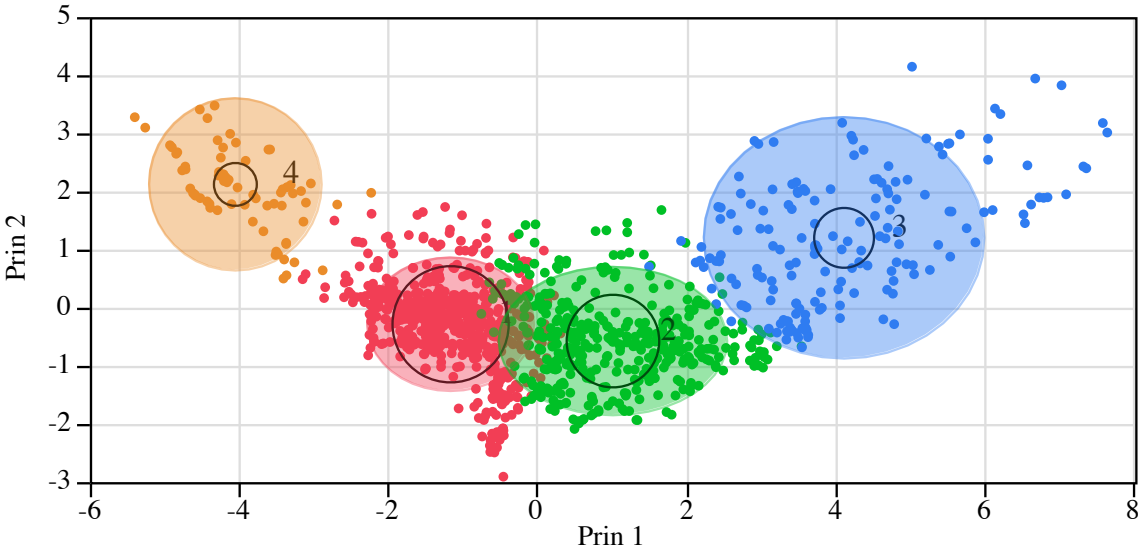


Figura 5.18: Representação das duas primeiras componentes principais dos *clusters* encontrados para as métricas de complexidade de todos os conjuntos de treino (pré-processados e originais).

5.3 Conclusões

Atendendo ao objetivo primordial deste capítulo, às simulações realizadas e aos resultados obtidos podem-se tirar as seguintes conclusões:

- A divisão do *dataset* depois do pré-processamento para realizar a validação cruzada (Fase 1) leva a desempenhos sobre-otimistas e torna esta a abordagem menos apropriada para problemas de dados não balanceados (Figuras 5.3 e 5.4, Tabelas 5.2 e B.3–B.5);

- Num modo geral, a diferença entre a complexidade dos conjuntos de teste e treino entre as Fases 1 e 2 (divisão depois e antes) é inferior na Fase 1, ou seja, nessa Fase os conjuntos de treino e teste têm características semelhantes (Figuras 5.6–5.8). Este efeito advém da divisão ser realizada depois e torna os desempenhos sobre-otimistas;

- Os dois melhores métodos de sobre-amostragem são SMOTE+TL e MWMOTE (Tabela 5.3 e 5.4). Estas técnicas têm tendência a alterar as regiões de sobreposição e aumentar o poder discriminativo das variáveis (Figuras 5.12–5.14);

- Através dos valores das métricas de complexidade de um *dataset* é possível construir um modelo (regressão e *clustering*) capaz de inferir o desempenho da classificação (AUC de teste) desse *dataset* (Figuras 5.15–5.18).

6 Conclusões e Trabalho Futuro

Neste capítulo são descritas as conclusões principais desta dissertação e discutidas algumas direções para futuros trabalhos.

6.1 Conclusões

Atualmente a comunidade científica que estuda as técnicas de *Machine Learning*, no geral, está muito focada no desenvolvimento de novos algoritmos e na melhoria dos existentes. No entanto, a importância da qualidade dos dados não pode ser esquecida, pois é um factor crucial para a obtenção de bons desempenhos dos classificadores. Um processo da exploração de conhecimento que possibilita a melhoria dessa qualidade é o pré-processamento. Ao longo desta dissertação foram discutidos dois problemas de pré-processamento: dados incompletos e dados não balanceados.

Relativamente ao problema de dados incompletos, foi explorado o potencial da utilização das características intrínsecas dos dados (e.g. distribuição dos dados) para a escolha dos algoritmos de imputação mais apropriados. As SVM provaram ser um método robusto para várias distribuições enquanto que os outros algoritmos relevaram um comportamento heterogéneo para os diferentes contextos estudados (e.g. distribuição, Taxa de Valores em Falta e estratégia de geração de dados em falta). De uma forma geral, o SOM é o método mais apropriado no que diz respeito à preservação dos valores originais e o k-NN é o mais indicado para manter a distribuição original dos dados. Em particular, o SOM mostrou ser a melhor abordagem para as distribuições Weibull e Birnbaum-Saunders.

Para o problema de dados não balanceados, relacionaram-se os desempenhos sobre-otimistas com diferentes abordagens de validação cruzada (antes ou depois do pré-processamento), quando são aplicadas técnicas de sobre-amostragem para equilibrar o número de exemplos entre classes. A validação cruzada após o pré-processamento mostrou desempenhos sobre-otimistas, que se relacionam com a complexidade dos *datasets*. Além disso, as métricas de

complexidade utilizadas aparentam estar correlacionadas com o desempenho da classificação (AUC), podendo ser desenvolvidos modelos (regressão e *clustering*) que permitem antever os resultados da classificação. Para a abordagem mais apropriada de validação cruzada (antes do pré-processamento), o SMOTE+TL e MWMOTE provaram ser os métodos mais eficientes.

Na minha opinião, este trabalho foi uma experiência gratificante no meu percurso académico. Os dois desafios mais marcantes foram a necessidade aprender quase todos os conceitos associados à extração de conhecimento num curto espaço de tempo e conseguir atingir a qualidade esperada para a submissão dos artigos.

6.2 Trabalhos Futuros

No que diz respeito ao problema de dados incompletos, as sugestões para trabalhos futuros são:

- Estudar os parâmetros de SVM que obtêm os melhores resultados (análise de sensibilidade);
- Estender as simulações para lidar com variáveis discretas;
- Verificar se as melhores técnicas em termos de Exatidão Preditiva e Exatidão Distribucional são eficientes em termos do erro de classificação.

Nos problemas de dados não balanceados, duas direções interessantes para trabalhos futuros são:

- Comparar técnicas de sob-amostragem com técnicas de sobre-amostragem atendendo à complexidade gerada apelos métodos;
- Focar num ou em alguns dos sub-problemas de dados não balanceados (e.g. *small-disjuncts* e sobreposição de classes) e verificar qual das técnicas é mais adequada para a sua resolução.

Dentro desta última direção, existem tópicos cuja resolução também seria uma contribuição para trabalhos futuros, tal como a identificação desses sub-problemas em conjuntos de dados multi-dimensionais e/ou *multiclass*.

Bibliografia

[1] Edgar F. Codd. “A relational model of data for large shared data banks”. Em: *Communications of the ACM* 13.6 (1970), pp. 377–387.

[2] Gregory Piatetsky-Shapiro. “Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop”. Em: *AI Magazine* 11.4 (1990), p. 68.

[3] Usama Fayyad, G. Piatetsky-Shapiro e Padhraic Smyth. “From data mining to knowledge discovery in databases”. Em: *AI magazine* 17.3 (1996), pp. 37–54.

[4] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. 2ª ed. John Wiley & Sons, Inc., 1987.

[5] José M. Jerez et al. “Missing data imputation using statistical and machine learning methods in a real breast cancer problem”. Em: *Artificial Intelligence in Medicine* 50.2 (2010), pp. 105–115.

[6] Pedro J. García-Laencina, José-Luis Sancho-Gómez e Aníbal R. Figueiras-Vidal. “Pattern classification with missing data: a review”. Em: *Neural Computing and Applications* 19 (2010), pp. 263–282.

[7] Thomas M. Mitchell. *Machine Learning*. 1ª ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[8] L. Breiman et al. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[9] Darryl N. Davis e Rahman M. Mostafizur. “Missing Value Imputation Using Stratified Supervised Learning for Cardiovascular Data”. Em: *Journal of Informatics and Data Mining* 1 (2016), pp. 2–13.

[10] D. Randall Wilson e Tony R. Martinez. “Improved heterogeneous distance functions”. Em: *Journal of Artificial Intelligence Research* 6 (1997), pp. 1–34.

[11] Pedro J. García-Laencina, José Luis Sancho-Gómez e Aníbal R. Figueiras-Vidal. "Classifying patterns with missing values using Multi-Task Learning perceptrons". Em: *Expert Systems with Applications* 40.4 (2013), pp. 1333–1341.

[12] Pedro J. García-Laencina et al. "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values". Em: *Computers in Biology and Medicine* 59 (2015), pp. 125–133.

[13] Teuvo Kohonen. *Self-Organizing Maps*. Vol. 30. Springer Series in Information Sciences. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997.

[14] Françoise Fessant e Sophie Midenet. "Self-organising map for data imputation and correction in surveys". Em: *Neural Computing and Applications* 10.4 (2002), pp. 300–310.

[15] Bernhard E. Boser, Isabelle M. Guyon e Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classiiers". Em: *Proceedings of the fifth annual workshop on Computational learning theory* (1992), pp. 144–152.

[16] Corinna Cortes e Vladimir Naoumovitch Vapnik. "Support-vector networks". Em: *Machine Learning* 20.3 (1995), pp. 273–297.

[17] Vladimir Naoumovitch Vapnik. *Statistical Learning Theory*. 1998.

[18] Rahman M. Mostafizur e Darryl N. Davis. "Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data". Em: *Proceedings of the World Congress on Engineering* I (2012), pp. 391–394.

[19] Nazziwa Aisha, Mohd Bakri Adam e Shamarina Shohaimi. "Effect of Missing Value Methods on Bayesian Network Classification of Hepatitis Data". Em: 4.6 (2013), pp. 4–8.

[20] Pedro Henriques Abreu et al. "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review". Em: *ACM Computing Surveys* 49.3 (2016), 52:1–52:40.

[21] Stef Van Buuren. *Flexible Imputation of Missing Data*. Chapman e Hall/CRC, 2012.

[22] Ray L. Chambers. *Evaluation criteria for statistical editing and imputation*. National Statistics Methodological Series 28. Office for National Statistics, 2001.

[23] Heikki Junninen et al. “Methods for imputation of missing values in air quality data sets”. Em: *Atmospheric Environment* 38.18 (2004), pp. 2895–2907.

[24] Raul H. C. Lopes. “Kolmogorov-Smirnov Test”. Em: *International Encyclopedia of Statistical Science*. Ed. por Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 718–720.

[25] Ray L. Chambers e UN/ECE Statistical Division. “Evaluation criteria for editing and imputation in EUREDIT”. Em: *Statistical Data Editing: Impact on Data Quality* 3 (2006), pp. 17–27.

[26] Taeho Jo e Nathalie Japkowicz. “Class Imbalances versus Small Disjuncts”. Em: *SIGKDD Explorations* 6.1 (2004), pp. 40–49.

[27] Haibo He e Edwardo a. Garcia. “Learning from Imbalanced Data”. Em: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[28] Victoria López et al. “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. Em: *Information Sciences* 250 (2013), pp. 113–141.

[29] Gustavo E. A. P. A. Batista, Ronaldo C. Prati e Maria Carolina Monard. “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data”. Em: *SIGKDD Explorations* 6.1 (2004), p. 20.

[30] Nitesh V. Chawla et al. “SMOTE : Synthetic Minority Over-sampling Technique”. Em: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[31] Show Jane Yen e Yue Shi Lee. “Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset”. Em: *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China*. Vol. 344. Springer Berlin Heidelberg, 2006, pp. 731–740.

[32] Show-Jane Yen e Yue-Shi Lee. “Cluster-based under-sampling approaches for imbalanced data distributions”. Em: *Expert Systems with Applications* 36.3, Part 1 (2009), pp. 5718–5727.

[33] Tomasz MacIejewski e Jerzy Stefanowski. “Local neighbourhood extension of SMOTE for mining imbalanced data”. Em: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. 2011, pp. 104–111.

[34] Miriam Seoane Santos et al. “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients”. Em: *Journal of biomedical informatics* 58 (2015), pp. 49–59.

[35] Dennis L. Wilson. “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data”. Em: *IEEE Transactions on Systems, Man and Cybernetics* 2.3 (1972), pp. 408–421.

[36] Ivan Tomek. “Two Modification of CNN”. Em: *IEEE Transactions on Systems and Man and Cybernetics* 6 (1976), pp. 769–772.

[37] Hui Han, Wen-Yuan Wang e Bing-Huan Mao. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”. Em: *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I*. Springer Berlin Heidelberg, 2005.

[38] Chumpol Bunkhumpornpat, Krung Sinapiromsaran e Chidchanok Lursinsap. “Safe-Level-SMOTE:Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem”. Em: *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. PAKDD ’09. Springer-Verlag, 2009, pp. 475–482.

[39] Sukarna Barua et al. “MWMOTE - Majority weighted minority Oversampling Technique for Imbalanced Data Set Learning”. Em: *IEEE Transactions on Knowledge and Data Engineering* 26.2 (2014), pp. 405–425.

[40] Haibo He et al. “Adaptive Synthetic Sampling Approach for Imbalanced Learning”. Em: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328.

[41] Sheng Tang e Si Ping Chen. “The generation mechanism of synthetic minority class examples”. Em: *5th Int. Conference on Information Technology and Applications in Biomedicine, ITAB 2008 in conjunction with 2nd Int. Symposium and Summer School on Biomedical and Health Engineering, IS3BHE 2008*. 2008, pp. 444–447.

[42] Jerzy Stefanowski e Szymon Wilk. “Selective Pre-processing of Imbalanced Data for”. Em: *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings*. Springer Berlin Heidelberg, 2008, pp. 283–292.

[43] Krystyna Napierała, Jerzy Stefanowski e Szymon Wilk. "Learning from Imbalanced Data in Presence of Noisy and Borderline Examples". Em: *Rough Sets and Current Trends in Computing: 7th International Conference, RSCTC 2010, Warsaw, Poland, June 28-30,2010. Proceedings*. Springer Berlin Heidelberg, 2010, pp. 158–167.

[44] Anil K. Jain. "Data clustering: 50 years beyond K-means". Em: *Pattern Recognition Letters* 31 (2010), pp. 651–666.

[45] Gilles Cohen et al. "Learning from imbalanced data in surveillance of nosocomial infection". Em: *Artificial Intelligence in Medicine* 37.1 (2006), pp. 7–18.

[46] Joaquim P. Marques de Sá. *Pattern Recognition: Concepts, Methods and Applications*. Springer Science & Business Media, 2001.

[47] Kai Tan e Xiushan Cai. "Prediction of earthquake in Yunnan region based on the AHC over sampling". Em: *2010 Chinese Control and Decision Conference*. 2010, pp. 2449–2452.

[48] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[49] J. R. Quinlan. "Induction of Decision Trees". Em: *Machine Learning* 1.1 (1986), pp. 81–106.

[50] Jiawei Han e Micheline Kamber. *Data Mining: Concepts and Techniques*. 2ª ed. Vol. 54. 2006.

[51] Leo Breiman. "Technical note: Some properties of splitting criteria". Em: *Machine Learning* 24.1 (1996), pp. 41–47. DOI: 10.1007/BF00117831.

[52] N. S. Altman. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". Em: *The American Statistician* 46.3 (1992), pp. 175–185.

[53] Bernhard Schölkopf e Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[54] John Shawe-Taylor e Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[55] Stephen P. Luttrell. "Partitioned mixture distribution: An adaptive Bayesian network for low-level image processing." Em: *IEEE Proceedings on Vision, Image, and Signal Processing*. Vol. 141. 4. 1994, pp. 251–260.

[56] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval Introduction*. Cambridge University Press, 2008.

[57] Nancy Chinchor. "MUC-4 evaluation metrics". Em: *Proceedings of the 4th Conference on Message Understanding*. MUC4 '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 22–29.

[58] R. Barandela et al. "Strategies for learning in class imbalance problems". Em: *Pattern Recognition* 36 (2003), pp. 849–851.

[59] Charles E. Metz. "Basic principles of ROC analysis". Em: *Seminars in Nuclear Medicine* 8.4 (1978), pp. 283–298.

[60] Charles X Ling, Jin Huang e Harry Zhang. "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms". Em: *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 329–341.

[61] Tin Kam Ho e Mitra Basu. "Complexity measures of supervised classification problems". Em: *IEEE transactions on pattern analysis and* 24.3 (2002), pp. 289–300.

[62] Loris Nanni, Alessandra Lumini e Sheryl Brahnam. "A classifier ensemble approach for the missing feature problem". Em: *Artificial Intelligence in Medicine* 55.1 (2012), pp. 37–50.

[63] Pilsung Kang. "Locally linear reconstruction based missing value imputation for supervised learning". Em: *Neurocomputing* 118 (2013), pp. 65–78.

[64] Md. Geaur Rahman e Md Zahidul Islam. "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques". Em: *Knowledge-Based Systems* 53 (2013), pp. 51–65.

[65] Pedro Henriques Abreu et al. "Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data". Em: *XIII Mediterranean Conference on Medical and Biological Engineering and Computing*. 2014, pp. 1366–1369.

[66] Mehran Amiri e Richard Jensen. "Missing data imputation using fuzzy-rough methods". Em: *Neurocomputing* 205 (2016), pp. 152–164.

[67] Rupam Deb e Alan Wee-Chung Liew. "Missing value imputation for the analysis of incomplete traffic accident data". Em: *Information Sciences* 339 (2016), pp. 274–289.

[68] Nishith Kumar et al. “Metabolomic Biomarker Identification in Presence of Outliers and Missing Values”. Em: *BioMed Research International* 2017 (2017), p. 11.

[69] Jason Van Hulse, Taghi M. Khoshgoftaar e Amri Napolitano. “Experimental Perspectives on Learning from Imbalanced Data”. Em: *Proceedings of the 24th International Conference on Machine Learning*. ICML ’07. Corvalis, Oregon, USA: ACM, 2007, pp. 935–942.

[70] Jason Van Hulse e Taghi Khoshgoftaar. “Knowledge discovery from imbalanced and noisy data”. Em: *Data and Knowledge Engineering* 68.12 (2009), pp. 1513–1542.

[71] Nele Verbiest et al. “Improving SMOTE with Fuzzy Rough Prototype Selection to Detect Noise in Imbalanced Classification Data”. Em: *Advances in Artificial Intelligence – IBERAMIA 2012: 13th Ibero-American Conference on AI*. Vol. 7637. Lecture Notes in Artificial Intelligence. Cartagena de Indias, Colombia: Springer Berlin Heidelberg, 2012, pp. 169–178.

[72] V. García et al. “Surrounding neighborhood-based SMOTE for learning from imbalanced data sets”. Em: *Progress in Artificial Intelligence* 1.4 (2012), pp. 347–362.

[73] Othman Soufan et al. “Mining Chemical Activity Status from High-Throughput Screening Assays”. Em: *PLoS ONE* 10.12 (2015), pp. 1–16.

[74] Octavio Loyola-González et al. “Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases”. Em: *Neurocomputing* 175 (2016), pp. 935–947.

[75] Julien Ah-Pine e Edmundo-Pavel Soriano-Morales. “A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis”. Em: *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing, {DMNLP} 2016, co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, {ECML-PKDD} 2016*. Riva del Garda, Italy, 2016, pp. 17–24.

[76] Roberto Alejo et al. “A Selective Dynamic Sampling Back-Propagation Approach for Handling the Two-Class Imbalance Problem”. Em: *Applied Sciences* 6.7 (2016), p. 200.

[77] Bing Zhu et al. “Benchmarking sampling techniques for imbalance learning in churn prediction”. Em: *Journal of the Operational Research Society* (2017), pp. 1–17.

[78] Juan López-de Uralde et al. “Automatic Morphological Categorisation of Carbon Black Nano-aggregates”. Em: *Database and Expert Systems Applications: 21th International Conference, DEXA 2010, Bilbao, Spain, August 30 - September 3, 2010, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 185–193.

[79] Reda Al-Bahrani, Ankit Agrawal e Alok Choudhary. “Colon cancer survival prediction using ensemble data mining on SEER data”. Em: *2013 IEEE International Conference on Big Data*. Silicon Valley, CA, USA: IEEE, 2013, pp. 9–16.

[80] Mehdi Naseriparsa e Mohammad Mansour Riahi Kashani. “Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset”. Em: *International Journal of Computer Applications* 77.3 (2013), pp. 33–38.

[81] Paul Fergus et al. “Prediction of Preterm Deliveries from EHG Signals Using Machine Learning”. Em: *PLoS ONE* 8.10 (2013), pp. 1–16.

[82] Mustakim Al Helal et al. “Algorithms Efficiency Measurement on Imbalanced Data using Geometric Mean and Cross Validation”. Em: *2016 International Workshop on Computational Intelligence (IWCI)*. Dhaka, Bangladesh: IEEE, 2016, pp. 110–114.

[83] K. Usha Rani, G. Naga Ramadevi e D. Lavanya. “Performance of synthetic minority oversampling technique on imbalanced breast cancer data”. Em: *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. New Delhi, India: IEEE, 2016, pp. 1623–1627.

[84] Ketil Oppedal et al. “Classifying Alzheimer’s disease, Lewy body dementia, and normal controls using 3D texture analysis in magnetic resonance images”. Em: *Biomedical Signal Processing and Control* 33 (2017), pp. 19–29.

[85] Julían Luengo et al. “Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling”. Em: *Soft Computing* 15.10 (2011), pp. 1909–1936.

[86] Rok Blagus e Lara Lusa. “Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models”. Em: *BMC Bioinformatics* 16.1 (2015), p. 362.

[87] *Kaggle Inc.* 2017. URL: `https://www.kaggle.com` (acedido em 20/01/2017).

[88] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: `http://archive.ics.uci.edu/ml` (acedido em 26/10/2016).

[89] Catherine Forbes et al. *Statistical Distributions*. 4ª ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011.

[90] Harald Cramér. *Mathematical Methods of Statistics (PMS-9)*. Vol. 9. Princeton university press, 2016.

[91] A. Colin Cameron. *EXCEL Univariate: Histogram*. 2009. URL: `http://cameron.econ.ucdavis.edu/excel/ex11histogram.html` (acedido em 24/02/2017).

[92] S. K. Mangal e S. Mangal. *Research Methodology in Behavioural Sciences*. PHI Learning, 2013.

[93] L. Ogundipe, O. E. Hodgson e R. E. Hodgson. *Lecture Notes on Paper Critique: Research Methodology And Statistic for Critical Paper Reading in Psychiatry*. Trafford Publishing, 2005.

[94] Darren Blend e Tshilidzi Marwala. "Comparison of Data Imputation Techniques and their Impact". Em: (2008), p. 7. URL: `https://arxiv.org/pdf/0812.1539.pdf`.

[95] J. Alcalá-Fdez et al. "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework". Em: *Journal of Multiple-Valued Logic and Soft Computing* 17.2-3 (2011), pp. 255–287.

Apêndice A

Informações auxiliares ao Capítulo 4

Tabela A.1: Comparação do *dataset* **letter** (original) e o **letterhalf**, em termos da distribuição de melhor ajuste para cada variável e o seu respetivo valor de GoF.

letter		**letterhalf**	
Distribuição	**GoF**	**Distribuição**	**GoF**
Rayleigh	0.895765015	Rayleigh	0.88936875
Normal	0.903986189	Normal	0.904206665
Gama	0.901861539	Gama	0.896506836
Normal	0.89703187	Normal	0.896768712
Pareto Generalizada	0.884020113	Pareto Generalizada	0.883528437
Gama	0.89651455	Gama	0.896117117
Gama	0.887716508	Gama	0.883813268
Gama	0.89127203	Gama	0.89095681
Rayleigh	0.921408171	Rayleigh	0.921553576
Gama	0.884595645	Gama	0.88072052
Gama	0.893336551	Gama	0.8911331
Gama	0.885697149	Gama	0.883161321
Exponencial	0.861718263	Exponencial	0.859147402
Gama	0.877644438	Gama	0.871758466
Pareto Generalizada	0.883462542	Pareto Generalizada	0.87665534
Gama	0.882482231	Gama	0.875133459

Tabela A.2: Sumário da distribuição de melhor ajuste para cada variável e respetivo valor de GoF, agrupados por *datasets*.

spectf			backpain			ctg			parkinson			wdbc			hillvalley								
V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF	V	Distribuicao	GoF
1	Weibull	0,956535795	1	Gamma	0,912877198	1	Gaussiana Inversa	0,959378642	1	Valor Extremo Generalizado	0,841457337	1	Valor Extremo Generalizado	0,933440047	1	Birnbaum-Saunders	0,875049217	45	Birnbaum-Saunders	0,874772328	89	Birnbaum-Saunders	0,872726919
2	Weibull	0,965415117	2	t Location-scale	0,91922175	2	Gamma	0,702189803	2	Log-Logistica	0,860973751	2	Log-Normal	0,980080157	2	Birnbaum-Saunders	0,87509001	46	Valor Extremo Generalizado	0,87574206	90	Valor Extremo Generalizado	0,873667269
3	Weibull	0,944969636	3	Gamma	0,95362245	3	Valor Extremo Generalizado	0,399184671	3	Valor Extremo Generalizado	0,92323648	3	Valor Extremo Generalizado	0,930062562	3	Birnbaum-Saunders	0,875137884	47	Valor Extremo Generalizado	0,879530702	91	Birnbaum-Saunders	0,873508869
4	Valor Extremo	0,944434573	4	Nakagami	0,947140299	4	Normal	0,910535008	4	Valor Extremo Generalizado	0,957335385	4	Valor Extremo Generalizado	0,93504573	4	Birnbaum-Saunders	0,875193965	48	Birnbaum-Saunders	0,875061563	92	Birnbaum-Saunders	0,873775193
5	Weibull	0,941429495	5	t Location-scale	0,951875312	5	Gamma	0,624850906	5	Gaussiana Inversa	0,899776682	5	Gamma	0,970512196	5	Birnbaum-Saunders	0,875258741	49	Birnbaum-Saunders	0,875188126	93	Birnbaum-Saunders	0,873985351
6	Valor Extremo	0,929755032	6	Pareto Generalizada	0,821803982	6	t Location-scale	0,251218705	6	Valor Extremo Generalizado	0,943339341	6	Birnbaum-Saunders	0,962051189	6	Birnbaum-Saunders	0,87533387	50	Valor Extremo Generalizado	0,876374182	94	Birnbaum-Saunders	0,874151447
7	Valor Extremo	0,942015157	7	Beta	0,94672111	7	Normal	0,253599257	7	Valor Extremo Generalizado	0,956510396	7	Pareto Generalizada	0,937473294	7	Birnbaum-Saunders	0,875420312	51	Birnbaum-Saunders	0,87585237	95	Birnbaum-Saunders	0,874283725
8	Valor Extremo	0,931466254	8	Pareto Generalizada	0,942710192	8	Valor Extremo Generalizado	0,898665363	8	Valor Extremo Generalizado	0,942991423	8	Gamma	0,889321883	8	Birnbaum-Saunders	0,875517928	52	Birnbaum-Saunders	0,876037246	96	Birnbaum-Saunders	0,874388071
9	Valor Extremo	0,935815586	9	Pareto Generalizada	0,944472258	9	Birnbaum-Saunders	0,946914751	9	Valor Extremo Generalizado	0,925379728	9	Log-Logistica	0,967450329	9	Birnbaum-Saunders	0,875630039	53	Birnbaum-Saunders	0,875990091	97	Birnbaum-Saunders	0,874471613
10	Valor Extremo	0,950764142	10	Pareto Generalizada	0,931315229	10	Gamma	0,48284636	10	Valor Extremo Generalizado	0,932211661	10	Valor Extremo Generalizado	0,974714848	10	Birnbaum-Saunders	0,875760678	54	Birnbaum-Saunders	0,875819297	98	Birnbaum-Saunders	0,874538493
11	Weibull	0,952070942	11	Normal	0,858720377	11	t Location-scale	0,952234582	11	Pareto Generalizada	0,936015384	11	Valor Extremo Generalizado	0,964500124	11	Birnbaum-Saunders	0,875915489	55	Birnbaum-Saunders	0,87554423	99	Birnbaum-Saunders	0,874592202
12	Valor Extremo	0,945035479	12	Pareto Generalizada	0,941498243	12	Nakagami	0,910009004	12	Valor Extremo Generalizado	0,940154447	12	Log-Normal	0,975227606	12	Birnbaum-Saunders	0,875746665	56	Birnbaum-Saunders	0,875185505	100	Birnbaum-Saunders	0,874635516
13	Valor Extremo	0,943594501	**breast**			13	Weibull	0,830563714	13	Valor Extremo Generalizado	0,928959682	13	Valor Extremo Generalizado	0,970721139	13	Birnbaum-Saunders	0,875693054	57	Birnbaum-Saunders	0,874921305	**leaf**		
14	Valor Extremo	0,947584451	V	Distribuicao	GoF	14	Valor Extremo Generalizado	0,973598198	14	Pareto Generalizada	0,935962979	14	Valor Extremo Generalizado	0,923935685	14	Birnbaum-Saunders	0,875660805	58	Birnbaum-Saunders	0,874781527	V	Distribuicao	GoF
15	Valor Extremo	0,940595886	1	Valor Extremo Generalizado	0,841970858	15	Pareto Generalizada	0,914541617	15	Valor Extremo Generalizado	0,947700787	15	Valor Extremo Generalizado	0,96835205	15	Birnbaum-Saunders	0,875674052	59	Birnbaum-Saunders	0,874419461	1	Beta	0,896709503
16	Valor Extremo	0,933529733	2	Log-Normal	0,920645931	16	Logistica	0,424804583	16	Weibull	0,920131802	16	Gaussiana Inversa	0,971255997	16	Birnbaum-Saunders	0,875759865	60	Birnbaum-Saunders	0,87428202	2	Pareto Generalizada	0,92107964
17	Valor Extremo	0,935058006	3	Valor Extremo Generalizado	0,904381598	17	Logistica	0,955963284	17	Beta	0,891815247	17	Valor Extremo Generalizado	0,969203845	17	Birnbaum-Saunders	0,875896768	61	Birnbaum-Saunders	0,874084569	3	Nakagami	0,927874905
18	Valor Extremo	0,931526215	4	Birnbaum-Saunders	0,903566815	18	Weibull	0,970838245	18	Valor Extremo Generalizado	0,917774107	18	Valor Extremo Generalizado	0,944045134	18	Birnbaum-Saunders	0,876044702	62	Birnbaum-Saunders	0,873776719	4	Valor Extremo Generalizado	0,783071242
19	Weibull	0,952198723	5	Birnbaum-Saunders	0,914619413	19	Normal	0,969129989	19	Valor Extremo Generalizado	0,956479156	19	Valor Extremo Generalizado	0,974913065	19	Birnbaum-Saunders	0,876060674	63	Birnbaum-Saunders	0,873650335	5	Beta	0,318298324
20	Weibull	0,95675474	6	Pareto Generalizada	0,933644466	20	Gamma	0,943816634	20	Valor Extremo Generalizado	0,939953867	20	Valor Extremo Generalizado	0,965054653	20	Birnbaum-Saunders	0,876168257	64	Birnbaum-Saunders	0,873538047	6	Valor Extremo Generalizado	0,901884359
21	Weibull	0,942595183	7	Valor Extremo Generalizado	0,92205603	21	Pareto Generalizada	0,624778177	21	Gaussiana Inversa	0,955411837	21	Valor Extremo Generalizado	0,928617984	21	Birnbaum-Saunders	0,876194506	65	Birnbaum-Saunders	0,873598315	7	Birnbaum-Saunders	0,929999773
22	Valor Extremo	0,953043695	8	Pareto Generalizada	0,906964569	**letterhalf**			22	Gamma	0,948529525	22	Gamma	0,964771423	22	Birnbaum-Saunders	0,876142515	66	Birnbaum-Saunders	0,873794312	8	Log-Normal	0,915589064
23	Weibull	0,948364376	9	Valor Extremo Generalizado	0,881134489	V	Distribuicao	GoF	**pen**			23	Valor Extremo Generalizado	0,928692943	23	Birnbaum-Saunders	0,876087215	67	Birnbaum-Saunders	0,873901693	9	Pareto Generalizada	0,941125854
24	Valor Extremo	0,953451039	**bupa**			1	Rayleigh	0,88936875	V	Distribuicao	GoF	24	Valor Extremo Generalizado	0,929125962	24	Birnbaum-Saunders	0,876251061	68	Birnbaum-Saunders	0,873892085	10	Beta	0,926893263
25	Valor Extremo	0,938863992	V	Distribuicao	GoF	2	Normal	0,904206665	1	Valor Extremo Generalizado	0,691433529	25	t Location-scale	0,970328036	25	Birnbaum-Saunders	0,876014945	69	Birnbaum-Saunders	0,873844006	11	Pareto Generalizada	0,946905668
26	Valor Extremo	0,940144583	1	Birnbaum-Saunders	0,934015503	3	Gamma	0,896506836	2	Logistica	0,81733282	26	Valor Extremo Generalizado	0,95768883	26	Birnbaum-Saunders	0,875800265	70	Birnbaum-Saunders	0,873720578	12	Pareto Generalizada	0,949467248
27	Logistica	0,927876589	2	Gaussiana Inversa	0,964102143	4	Normal	0,896768712	3	Valor Extremo Generalizado	0,936121544	27	Gamma	0,929863824	27	Birnbaum-Saunders	0,87550145	71	Birnbaum-Saunders	0,873791667	13	Pareto Generalizada	0,949186926
28	Weibull	0,937520201	3	Log-Logistica	0,948903794	5	Pareto Generalizada	0,883528437	4	Logistica	0,686374721	28	Gamma	0,905013262	28	Birnbaum-Saunders	0,875228204	72	Birnbaum-Saunders	0,873823111	14	Rayleigh	0,963100545
29	Valor Extremo	0,886650738	4	Log-Logistica	0,941817481	6	Gamma	0,896117117	5	Logistica	0,840970981	29	Log-Logistica	0,969027097	29	Birnbaum-Saunders	0,875046415	73	Birnbaum-Saunders	0,874280494	**whitewine**		
30	Valor Extremo	0,836308977	5	Valor Extremo Generalizado	0,952847498	7	Gamma	0,883813268	6	Valor Extremo	0,922997681	30	Valor Extremo Generalizado	0,971495937	30	Birnbaum-Saunders	0,874844647	74	Birnbaum-Saunders	0,874609886	V	Distribuicao	GoF
31	Valor Extremo	0,936706941	6	Exponential	0,805039073	8	Gamma	0,89095681	7	Valor Extremo Generalizado	0,894526673	**relax**			31	Valor Extremo Generalizado	0,878269077	75	Birnbaum-Saunders	0,874979743	1	Log-Logistica	0,970992321
32	Valor Extremo	0,912886572	**redwine**			9	Rayleigh	0,921553576	8	Logistica	0,910878523	V	Distribuicao	GoF	32	Birnbaum-Saunders	0,874087922	76	Birnbaum-Saunders	0,875370859	2	Log-Logistica	0,97527416
33	Logistica	0,939614649	V	Distribuicao	GoF	10	Gamma	0,88072052	9	Logistica	0,738722865	1	t Location-scale	0,949873406	33	Birnbaum-Saunders	0,874234927	77	Birnbaum-Saunders	0,875496295	3	t Location-scale	0,94950006
34	Logistica	0,929196034	1	Valor Extremo Generalizado	0,956576055	11	Gamma	0,8911331	10	Logistica	0,833355767	2	Logistica	0,950547685	34	Birnbaum-Saunders	0,874166877	78	Birnbaum-Saunders	0,875470435	4	Pareto Generalizada	0,848591747
35	Valor Extremo	0,920786287	2	Nakagami	0,960349601	12	Gamma	0,883161321	11	Logistica	0,568574488	3	t Location-scale	0,957941199	35	Birnbaum-Saunders	0,874159836	79	Birnbaum-Saunders	0,875578438	5	Log-Logistica	0,956741485
36	Valor Extremo	0,892421677	3	Logistica	0,883263344	13	Exponential	0,859147402	12	Pareto Generalizada	0,88346684	4	Logistica	0,958276295	36	Birnbaum-Saunders	0,87420843	80	Birnbaum-Saunders	0,875156455	6	Nakagami	0,972712488
37	Weibull	0,957623243	4	Valor Extremo Generalizado	0,894611809	14	Gamma	0,871758466	13	Gamma	0,935103	5	Normal	0,944040904	37	Birnbaum-Saunders	0,874174392	81	Valor Extremo Generalizado	0,878527048	7	Nakagami	0,980884981
38	Valor Extremo	0,950616695	5	t Location-scale	0,936192398	15	Pareto Generalizada	0,87665534	14	Valor Extremo Generalizado	0,74136399	6	Logistica	0,9506503	38	Birnbaum-Saunders	0,874240103	82	Birnbaum-Saunders	0,874825781	8	Valor Extremo Generalizado	0,937250064
39	Valor Extremo	0,919360221	6	Birnbaum-Saunders	0,945321558	16	Gamma	0,875133459	15	Logistica	0,215924676	7	t Location-scale	0,958914191	39	Birnbaum-Saunders	0,874358426	83	Birnbaum-Saunders	0,874279644	9	Valor Extremo Generalizado	0,977266774
40	Valor Extremo	0,929276591	7	Birnbaum-Saunders	0,96065315	**iris**			16	Gamma	0,423959266	8	t Location-scale	0,95154181	40	Birnbaum-Saunders	0,874638497	84	Birnbaum-Saunders	0,873837063	10	Valor Extremo Generalizado	0,973186773
41	Valor Extremo	0,872716207	8	Log-Logistica	0,979619518	V	Distribuicao	GoF				9	Valor Extremo Generalizado	0,946186775	41	Birnbaum-Saunders	0,874926637	85	Birnbaum-Saunders	0,873486255	11	Valor Extremo Generalizado	0,888778951
42	Valor Extremo	0,876438129	9	t Location-scale	0,976751888	1	Valor Extremo Generalizado	0,921486151				10	t Location-scale	0,937129519	42	Birnbaum-Saunders	0,874872165	86	Birnbaum-Saunders	0,87330215			
43	Valor Extremo	0,919290488	10	Valor Extremo Generalizado	0,966149765	2	Gaussiana Inversa	0,921900482				11	t Location-scale	0,940972827	43	Birnbaum-Saunders	0,87471343	87	Birnbaum-Saunders	0,873050337			
44	Valor Extremo	0,893844091	11	Valor Extremo Generalizado	0,910869266	3	Valor Extremo	0,796740547				12	t Location-scale	0,941735905	44	Birnbaum-Saunders	0,874573388	88	Birnbaum-Saunders	0,872751881			
						4	Valor Extremo Generalizado	0,759833273															

Tabela A.3: Contagem de variáveis de cada *dataset* por distribuição de melhor ajuste.

Dataset	**Distribuição**															
	Beta	Birnbaum-Saunders	Exponencial	Valor Extremo	Gama	Valor Extremo Generalizado	Pareto Generalizada	Gaussiana Inversa	Logistica	Log-Logistica	Normal	Nakagami	Log-Normal	Rayleigh	t Location-scale	Weibull
backpain	1				2		5				1	1			2	
breast		2				4	2						1			
bupa		1	1			1		1		2						
ctg		1			**4**	**3**	2	1	**2**		**3**	1			**2**	2
hillvalley		94				6										
iris				1		2		1								
leaf	**3**	1				2	5					1	1	1		
letterhalf			1		9		2				2			2		
parkinson	1				1	14	2	2		1						1
pen				1	**2**	4	1		**8**							
redwine		2				4			1	1		1			2	
relax						1			3		1				7	
spectf				30					3							11
wdbc		1			5	17	1	1		2			2		1	
whitewine						4	1			3		2			1	
Soma	**5**	**102**	**2**	**32**	**23**	**62**	**21**	**6**	**17**	**9**	**7**	**6**	**4**	**3**	**15**	**14**

Os valores sublinhados e vermelhos (e.g. **8**) têm como significado a existência de uma variável com GoF com valor entre $[0.2, 0.3]$.

Os valores azuis (e.g. **4**) têm como significado a existência de uma variável com GoF com valor entre $[0.3, 0.5]$.

Tabela A.4: Técnicas de imputação vencedoras com a votação para cada distribuição, métrica, MR e estratégia (excluindo SVM).

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Beta	MSE	5-10%	DT-KNN-(2/3)	DT-KNN-MM-(1/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-(2/3)	SOM-(2/3)	SOM-(2/3)
		15-20%	SOM-(2/3)	KNN-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)
		25%	SOM-(2/3)	KNN-(2/3)	MM-(2/3)	KNN-MM-SOM-(1/3)	SOM-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)
		Todas	DT-KNN-MM-SOM-(1/3)	KNN-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)
	r^2	5-10%	KNN-SOM-(2/3)	SOM-(2/3)	KNN-(3/3)	SOM-(3/3)	DT-(2/3)	KNN-(3/3)	KNN-(3/3)
		15-20%	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-SOM-(2/3)	KNN-(2/3)	SOM-(2/3)
		25%	SOM-(2/3)	KNN-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)
		Todas	KNN-SOM-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)	DT-SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)
	D_{KS}	5-10%	KNN-(3/3)	KNN-(2/3)	KNN-(3/3)	KNN-(3/3)	KNN-(2/3)	KNN-(2/3)	KNN-(3/3)
		15-20%	KNN-(3/3)	KNN-SOM-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	SOM-(2/3)	KNN-(2/3)	KNN-(2/3)
		25%	KNN-(2/3)	KNN-SOM-(2/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	KNN-(3/3)	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)
		Todas	KNN-(3/3)	KNN-(2/3)	KNN-(2/3)	KNN-(3/3)	KNN-SOM-(2/3)	KNN-(2/3)	KNN-(2/3)
Birnbaum-Saunders	MSE	5-10%	SOM-(6/7)	KNN-(5/7)	KNN-(6/7)	KNN-(4/7)	SOM-(4/7)	KNN-SOM-(4/7)	SOM-(4/7)
		15-20%	SOM-(5/7)	KNN-SOM-(4/7)	SOM-(7/7)	SOM-(6/7)	SOM-(5/7)	SOM-(5/7)	SOM-(6/7)
		25%	SOM-(6/7)	SOM-(4/7)	SOM-(5/7)	SOM-(5/7)	SOM-(4/7)	SOM-(7/7)	SOM-(6/7)
		Todas	SOM-(5/7)	KNN-SOM-(3/7)	SOM-(6/7)	SOM-(5/7)	SOM-(5/7)	SOM-(5/7)	SOM-(5/7)
	r^2	5-10%	KNN-(5/7)	KNN-(4/7)	KNN-(5/7)	SOM-(5/7)	SOM-(6/7)	KNN-(5/7)	KNN-(5/7)
		15-20%	SOM-(5/7)	MM-SOM-(3/7)	SOM-(6/7)	SOM-(6/7)	SOM-(5/7)	SOM-(6/7)	SOM-(6/7)
		25%	SOM-(6/7)	SOM-(5/7)	SOM-(5/7)	KNN-SOM-(3/7)	SOM-(4/7)	SOM-(6/7)	SOM-(6/7)
		Todas	KNN-(4/7)	SOM-(3/7)	SOM-(5/7)	SOM-(5/7)	SOM-(5/7)	SOM-(4/7)	SOM-(4/7)
	D_{KS}	5-10%	SOM-(4/7)	KNN-(4/7)	DT-KNN-(4/7)	KNN-(5/7)	DT-KNN-SOM-(3/7)	KNN-(4/7)	KNN-(4/7)
		15-20%	SOM-(5/7)	KNN-(5/7)	KNN-(5/7)	SOM-(4/7)	KNN-(4/7)	SOM-(5/7)	KNN-(4/7)
		25%	SOM-(4/7)	SOM-(4/7)	KNN-SOM-(3/7)	SOM-(5/7)	SOM-(4/7)	SOM-(4/7)	KNN-(4/7)
		Todas	SOM-(4/7)	KNN-(4/7)	KNN-(3/7)	SOM-(4/7)	SOM-(3/7)	SOM-(4/7)	KNN-(4/7)
Exponencial	MSE	5-10%	KNN-(2/2)	DT-KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	MM-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)
		15-20%	KNN-(2/2)	KNN-(2/2)	MM-SOM-(1/2)	DT-MM-SOM-(1/2)	KNN-(2/2)	KNN-MM-SOM-(1/2)	KNN-MM-SOM-(1/2)
		25%	KNN-MM-(1/2)	KNN-MM-(1/2)	MM-SOM-(1/2)	DT-MM-(1/2)	KNN-SOM-(1/2)	MM-SOM-(1/2)	MM-SOM-(1/2)
		Todas	KNN-(2/2)	KNN-(2/2)	MM-SOM-(1/2)	MM-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)
	r^2	5-10%	KNN-(2/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-MM-SOM-(1/2)	MM-(2/2)	KNN-(2/2)	KNN-(2/2)
		15-20%	DT-KNN-MM-SOM-(1/2)	DT-KNN-MM-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		25%	MM-SOM-(1/2)	MM-SOM-(1/2)	MM-SOM-(1/2)	KNN-MM-(1/2)	MM-SOM-(1/2)	MM-SOM-(1/2)	MM-SOM-(1/2)
		Todas	KNN-(2/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-MM-(1/2)	MM-(2/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)
	D_{KS}	5-10%	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		15-20%	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		25%	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		Todas	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
Valor Extremo	MSE	5-10%	KNN-(2/3)	SOM-(3/3)	SOM-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	KNN-(2/3)	SOM-(2/3)
		15-20%	KNN-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)
		25%	KNN-MM-SOM-(1/3)	KNN-MM-SOM-(1/3)	SOM-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	SOM-(3/3)	SOM-(2/3)
		Todas	KNN-(2/3)	SOM-(2/3)	KNN-MM-SOM-(1/3)	KNN-SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)
	r^2	5-10%	KNN-(2/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	KNN-(2/3)	KNN-(3/3)	SOM-(2/3)	KNN-(3/3)
		15-20%	SOM-(2/3)	KNN-(3/3)	SOM-(2/3)	KNN-SOM-(2/3)	KNN-(3/3)	SOM-(2/3)	SOM-(2/3)
		25%	SOM-(2/3)	SOM-(3/3)	KNN-MM-SOM-(1/3)	SOM-(2/3)	KNN-(2/3)	SOM-(2/3)	SOM-(2/3)
		Todas	KNN-MM-SOM-(1/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	KNN-(2/3)	KNN-(3/3)	SOM-(2/3)	SOM-(2/3)
	D_{KS}	5-10%	KNN-(2/3)	SOM-(2/3)	DT-KNN-(2/3)	KNN-(2/3)	SOM-(2/3)	DT-KNN-SOM-(1/3)	KNN-(2/3)
		15-20%	KNN-(2/3)	SOM-(2/3)	KNN-SOM-(2/3)	SOM-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	KNN-(2/3)
		25%	KNN-(2/3)	SOM-(2/3)	KNN-(2/3)	SOM-(2/3)	KNN-(2/3)	KNN-(2/3)	KNN-(2/3)
		Todas	KNN-(2/3)	SOM-(2/3)	KNN-(2/3)	KNN-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	KNN-(2/3)

Continua na página seguinte...

Tabela A.4: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Gama	MSE	5-10%	KNN-SOM-(3/6)	KNN-(5/6)	SOM-(3/6)	SOM-(3/6)	KNN-(3/6)	KNN-(4/6)	DT-(3/6)
		15-20%	SOM-(3/6)	KNN-SOM-(3/6)	SOM-(3/6)	SOM-(3/6)	KNN-SOM-(2/6)	SOM-(3/6)	SOM-(3/6)
		25%	SOM-(4/6)	SOM-(3/6)	SOM-(4/6)	SOM-(4/6)	SOM-(3/6)	SOM-(4/6)	SOM-(4/6)
		Todas	SOM-(3/6)	KNN-(3/6)	SOM-(3/6)	SOM-(3/6)	KNN-SOM-(2/6)	SOM-(3/6)	SOM-(3/6)
	r^2	5-10%	KNN-(4/6)	KNN-(5/6)	KNN-(4/6)	KNN-(3/6)	KNN-(3/6)	KNN-(4/6)	KNN-(4/6)
		15-20%	SOM-(3/6)	KNN-(4/6)	SOM-(3/6)	SOM-(3/6)	SOM-(3/6)	KNN-(3/6)	SOM-(3/6)
		25%	SOM-(4/6)	DT-KNN-SOM-(2/6)	SOM-(4/6)	SOM-(4/6)	SOM-(4/6)	SOM-(4/6)	SOM-(3/6)
		Todas	SOM-(3/6)	KNN-(3/6)	SOM-(3/6)	SOM-(3/6)	SOM-(3/6)	KNN-(3/6)	SOM-(3/6)
	D_{KS}	5-10%	KNN-(5/6)	KNN-SOM-(3/6)	KNN-(5/6)	KNN-(5/6)	KNN-(4/6)	KNN-(5/6)	KNN-(4/6)
		15-20%	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-(4/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)
		25%	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-(4/6)	KNN-SOM-(3/6)	SOM-(4/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)
		Todas	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-(4/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)
Valor Extremo Generalizado	MSE	5-10%	KNN-SOM-(5/12)	KNN-(7/12)	KNN-(7/12)	SOM-(7/12)	SOM-(7/12)	KNN-SOM-(7/12)	SOM-(5/12)
		15-20%	SOM-(8/12)	KNN-(7/12)	SOM-(6/12)	SOM-(7/12)	SOM-(7/12)	SOM-(7/12)	SOM-(8/12)
		25%	SOM-(10/12)	KNN-(6/12)	SOM-(7/12)	SOM-(10/12)	KNN-SOM-(5/12)	SOM-(6/12)	SOM-(8/12)
		Todas	SOM-(7/12)	KNN-(7/12)	SOM-(6/12)	SOM-(8/12)	SOM-(6/12)	SOM-(6/12)	SOM-(7/12)
	r^2	5-10%	KNN-(10/12)	KNN-(9/12)	KNN-(11/12)	KNN-(9/12)	KNN-(8/12)	KNN-(8/12)	KNN-(10/12)
		15-20%	KNN-(7/12)	KNN-(9/12)	SOM-(6/12)	KNN-(6/12)	SOM-(7/12)	SOM-(7/12)	SOM-(6/12)
		25%	SOM-(7/12)	SOM-(7/12)	SOM-(7/12)	SOM-(8/12)	KNN-SOM-(6/12)	SOM-(7/12)	SOM-(7/12)
		Todas	KNN-(6/12)	KNN-(9/12)	SOM-(6/12)	KNN-(7/12)	SOM-(7/12)	SOM-(6/12)	SOM-(6/12)
	D_{KS}	5-10%	KNN-(7/12)	KNN-(8/12)	KNN-(10/12)	SOM-(6/12)	KNN-(8/12)	KNN-(8/12)	KNN-(6/12)
		15-20%	SOM-(7/12)	KNN-(7/12)	KNN-(9/12)	KNN-(7/12)	KNN-SOM-(5/12)	KNN-SOM-(6/12)	KNN-(7/12)
		25%	SOM-(8/12)	SOM-(9/12)	KNN-(7/12)	SOM-(7/12)	KNN-(6/12)	SOM-(7/12)	KNN-(7/12)
		Todas	KNN-(6/12)	KNN-(7/12)	KNN-(8/12)	KNN-(6/12)	KNN-(6/12)	KNN-(6/12)	KNN-(7/12)
Pareto Generalizada	MSE	5-10%	KNN-(6/9)	KNN-(7/9)	KNN-(6/9)	KNN-(5/9)	SOM-(5/9)	KNN-(8/9)	DT-(5/9)
		15-20%	SOM-(6/9)	SOM-(5/9)	SOM-(6/9)	SOM-(6/9)	SOM-(5/9)	SOM-(5/9)	SOM-(7/9)
		25%	SOM-(6/9)	SOM-(7/9)	SOM-(7/9)	SOM-(6/9)	SOM-(7/9)	SOM-(7/9)	SOM-(7/9)
		Todas	KNN-(5/9)	SOM-(4/9)	SOM-(5/9)	KNN-SOM-(4/9)	SOM-(5/9)	SOM-(5/9)	SOM-(5/9)
	r^2	5-10%	KNN-(9/9)	KNN-(7/9)	KNN-(7/9)	KNN-(8/9)	DT-KNN-(4/9)	KNN-(9/9)	KNN-(5/9)
		15-20%	KNN-(7/9)	KNN-(4/9)	KNN-(5/9)	SOM-(6/9)	SOM-(5/9)	KNN-(5/9)	SOM-(5/9)
		25%	KNN-(7/9)	SOM-(6/9)	SOM-(7/9)	SOM-(5/9)	KNN-(4/9)	SOM-(8/9)	SOM-(6/9)
		Todas	KNN-(8/9)	KNN-(4/9)	KNN-(6/9)	KNN-(7/9)	SOM-(4/9)	KNN-(5/9)	SOM-(5/9)
	D_{KS}	5-10%	KNN-(6/9)	KNN-(8/9)	KNN-(8/9)	KNN-(7/9)	KNN-(6/9)	KNN-(7/9)	KNN-(5/9)
		15-20%	SOM-(6/9)	KNN-(8/9)	KNN-(6/9)	KNN-(5/9)	KNN-(7/9)	KNN-(5/9)	KNN-(5/9)
		25%	SOM-(8/9)	KNN-(6/9)	KNN-(5/9)	SOM-(7/9)	KNN-(6/9)	KNN-(6/9)	KNN-(5/9)
		Todas	KNN-(5/9)	KNN-(7/9)	KNN-(6/9)	KNN-(5/9)	KNN-(5/9)	KNN-(6/9)	KNN-(5/9)
Gaussiana Inversa	MSE	5-10%	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	SOM-(3/5)	KNN-SOM-(3/5)	SOM-(4/5)	SOM-(3/5)
		15-20%	SOM-(4/5)	KNN-(4/5)	SOM-(5/5)	SOM-(5/5)	KNN-(3/5)	SOM-(4/5)	SOM-(4/5)
		25%	SOM-(3/5)	KNN-(4/5)	SOM-(5/5)	SOM-(4/5)	KNN-SOM-(3/5)	SOM-(5/5)	SOM-(5/5)
		Todas	SOM-(3/5)	KNN-(4/5)	SOM-(5/5)	SOM-(5/5)	KNN-(3/5)	SOM-(5/5)	SOM-(4/5)
	r^2	5-10%	KNN-(5/5)	KNN-(3/5)	KNN-(4/5)	KNN-SOM-(4/5)	KNN-(5/5)	KNN-(3/5)	SOM-(4/5)
		15-20%	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	SOM-(5/5)	KNN-(3/5)	SOM-(5/5)	SOM-(3/5)
		25%	KNN-SOM-(3/5)	KNN-SOM-(2/5)	SOM-(4/5)	SOM-(3/5)	SOM-(3/5)	SOM-(4/5)	SOM-(5/5)
		Todas	SOM-(3/5)	KNN-SOM-(3/5)	KNN-(3/5)	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	SOM-(4/5)
	D_{KS}	5-10%	KNN-(3/5)	KNN-SOM-(3/5)	KNN-(4/5)	SOM-(4/5)	KNN-SOM-(3/5)	KNN-SOM-(3/5)	KNN-SOM-(2/5)
		15-20%	SOM-(4/5)	SOM-(4/5)	DT-KNN-SOM-(2/5)	KNN-SOM-(3/5)	SOM-(4/5)	KNN-SOM-(4/5)	KNN-(3/5)
		25%	SOM-(4/5)	SOM-(4/5)	SOM-(4/5)	SOM-(3/5)	SOM-(3/5)	KNN-SOM-(3/5)	KNN-(3/5)
		Todas	KNN-SOM-(3/5)	SOM-(4/5)	KNN-SOM-(2/5)	KNN-(3/5)	SOM-(4/5)	KNN-(4/5)	KNN-(3/5)

Continua na página seguinte...

Tabela A.4: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Logistica	MSE	5-10%	KNN-(3/5)	KNN-(5/5)	KNN-(4/5)	KNN-SOM-(2/5)	KNN-(3/5)	KNN-(3/5)	DT-SOM-(2/5)
		15-20%	SOM-(4/5)	KNN-(5/5)	SOM-(4/5)	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	SOM-(3/5)
		25%	SOM-(4/5)	KNN-(5/5)	SOM-(4/5)	SOM-(4/5)	KNN-(4/5)	SOM-(3/5)	SOM-(5/5)
		Todas	SOM-(4/5)	KNN-(5/5)	SOM-(4/5)	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	DT-SOM-(2/5)
	r^2	5-10%	KNN-(5/5)	KNN-(3/5)	KNN-(5/5)	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(4/5)	KNN-(4/5)
		15-20%	SOM-(4/5)	KNN-(3/5)	KNN-SOM-(3/5)	SOM-(4/5)	DT-KNN-(3/5)	SOM-(4/5)	DT-KNN-SOM-(2/5)
		25%	SOM-(4/5)	KNN-(3/5)	SOM-(4/5)	SOM-(4/5)	KNN-SOM-(2/5)	SOM-(4/5)	SOM-(4/5)
		Todas	SOM-(3/5)	KNN-(3/5)	SOM-(3/5)	SOM-(4/5)	KNN-(2/5)	SOM-(4/5)	KNN-(3/5)
	D_{KS}	5-10%	KNN-(4/5)	KNN-(5/5)	KNN-(4/5)	KNN-(5/5)	KNN-(5/5)	KNN-(4/5)	KNN-(4/5)
		15-20%	KNN-(5/5)	KNN-(4/5)	KNN-(4/5)	KNN-(3/5)	KNN-(5/5)	KNN-(5/5)	KNN-(3/5)
		25%	KNN-(4/5)	KNN-(4/5)	KNN-(3/5)	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(4/5)	KNN-(4/5)
		Todas	KNN-(4/5)	KNN-(4/5)	KNN-(3/5)	KNN-(3/5)	KNN-(5/5)	KNN-(5/5)	KNN-(3/5)
Log-Logistica	MSE	5-10%	KNN-SOM-(3/5)	KNN-(3/5)	SOM-(3/5)	KNN-(4/5)	KNN-(4/5)	KNN-(4/5)	SOM-(3/5)
		15-20%	SOM-(3/5)	KNN-(4/5)	SOM-(3/5)	SOM-(3/5)	KNN-(4/5)	DT-KNN-SOM-(3/5)	SOM-(4/5)
		25%	SOM-(5/5)	KNN-(5/5)	SOM-(3/5)	SOM-(4/5)	KNN-(4/5)	SOM-(4/5)	SOM-(4/5)
		Todas	SOM-(3/5)	KNN-(4/5)	SOM-(3/5)	SOM-(3/5)	KNN-(3/5)	KNN-SOM-(2/5)	SOM-(3/5)
	r^2	5-10%	KNN-(5/5)	KNN-(4/5)	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(4/5)	KNN-(5/5)	KNN-(5/5)
		15-20%	KNN-SOM-(3/5)	SOM-(3/5)	KNN-SOM-(3/5)	KNN-(3/5)	KNN-(4/5)	SOM-(3/5)	KNN-(3/5)
		25%	KNN-SOM-(3/5)	KNN-(4/5)	SOM-(3/5)	SOM-(3/5)	SOM-(3/5)	KNN-SOM-(3/5)	SOM-(3/5)
		Todas	KNN-SOM-(3/5)	KNN-(3/5)	SOM-(3/5)	SOM-(3/5)	SOM-(3/5)	SOM-(3/5)	KNN-(3/5)
	D_{KS}	5-10%	KNN-(4/5)	SOM-(3/5)	KNN-(3/5)	KNN-(4/5)	KNN-(3/5)	KNN-(3/5)	KNN-(4/5)
		15-20%	SOM-(3/5)	KNN-(4/5)	KNN-SOM-(3/5)	KNN-SOM-(3/5)	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(3/5)
		25%	SOM-(4/5)	SOM-(3/5)	SOM-(4/5)	KNN-(3/5)	KNN-SOM-(3/5)	SOM-(3/5)	KNN-(3/5)
		Todas	KNN-SOM-(3/5)	SOM-(3/5)	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(3/5)	KNN-(3/5)	KNN-(3/5)
Normal	MSE	5-10%	KNN-(4/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(4/4)	KNN-(3/4)	DT-(3/4)
		15-20%	KNN-SOM-(2/4)	KNN-(4/4)	SOM-(3/4)	KNN-SOM-(2/4)	KNN-(4/4)	SOM-(3/4)	DT-(2/4)
		25%	SOM-(2/4)	KNN-(4/4)	DT-KNN-MM-SOM-(1/4)	SOM-(2/4)	KNN-(3/4)	SOM-(2/4)	SOM-(2/4)
		Todas	KNN-(3/4)	KNN-(4/4)	SOM-(3/4)	KNN-SOM-(2/4)	KNN-(4/4)	KNN-(3/4)	DT-(2/4)
	r^2	5-10%	SOM-(3/4)	KNN-(2/4)	KNN-(4/4)	KNN-(3/4)	MM-(3/4)	KNN-(3/4)	KNN-(4/4)
		15-20%	SOM-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-SOM-(2/4)	KNN-(3/4)	KNN-(3/4)	KNN-(2/4)
		25%	KNN-(2/4)	KNN-(3/4)	KNN-(2/4)	KNN-(2/4)	DT-KNN-MM-SOM-(1/4)	SOM-(2/4)	DT-KNN-MM-SOM-(1/4)
		Todas	SOM-(3/4)	KNN-(2/4)	KNN-(3/4)	KNN-(2/4)	KNN-(2/4)	KNN-(3/4)	KNN-(3/4)
	D_{KS}	5-10%	KNN-(4/4)	KNN-SOM-(2/4)	KNN-(4/4)	KNN-(4/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)
		15-20%	KNN-(4/4)	KNN-SOM-(2/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(4/4)
		25%	KNN-(4/4)	SOM-(2/4)	DT-KNN-SOM-(2/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)
		Todas	KNN-(4/4)	KNN-SOM-(2/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)	KNN-(3/4)
Nakagami	MSE	5-10%	KNN-(4/5)	KNN-(5/5)	KNN-(4/5)	KNN-(3/5)	DT-KNN-(2/5)	KNN-(4/5)	DT-KNN-(3/5)
		15-20%	SOM-(3/5)	KNN-(5/5)	SOM-(3/5)	DT-KNN-SOM-(2/5)	KNN-(3/5)	SOM-(4/5)	DT-KNN-(2/5)
		25%	SOM-(4/5)	KNN-(3/5)	SOM-(4/5)	SOM-(3/5)	KNN-(3/5)	SOM-(4/5)	SOM-(4/5)
		Todas	SOM-(3/5)	KNN-(5/5)	SOM-(3/5)	KNN-SOM-(2/5)	DT-KNN-(2/5)	SOM-(3/5)	DT-KNN-(2/5)
	r^2	5-10%	KNN-(4/5)	KNN-SOM-(3/5)	KNN-(5/5)	KNN-(3/5)	KNN-(3/5)	KNN-(4/5)	KNN-(4/5)
		15-20%	KNN-SOM-(3/5)	KNN-(4/5)	KNN-(4/5)	KNN-(3/5)	KNN-(3/5)	KNN-SOM-(2/5)	KNN-(3/5)
		25%	KNN-(3/5)	KNN-(3/5)	KNN-SOM-(2/5)	DT-KNN-(2/5)	DT-(2/5)	KNN-SOM-(2/5)	KNN-SOM-(2/5)
		Todas	KNN-(3/5)	KNN-(4/5)	KNN-(3/5)	KNN-(3/5)	KNN-(2/5)	KNN-SOM-(2/5)	KNN-(3/5)
	D_{KS}	5-10%	DT-KNN-(3/5)	KNN-(4/5)	DT-KNN-(3/5)	KNN-(3/5)	SOM-(3/5)	KNN-(3/5)	KNN-(4/5)
		15-20%	KNN-SOM-(2/5)	KNN-(4/5)	DT-KNN-(3/5)	DT-KNN-(3/5)	SOM-(4/5)	KNN-(3/5)	KNN-(3/5)
		25%	SOM-(3/5)	SOM-(4/5)	KNN-(3/5)	KNN-SOM-(2/5)	SOM-(4/5)	KNN-(3/5)	KNN-(3/5)
		Todas	KNN-(2/5)	SOM-(3/5)	DT-KNN-(3/5)	KNN-(3/5)	SOM-(4/5)	DT-KNN-(2/5)	KNN-(3/5)

Continua na página seguinte...

Tabela A.4: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Log-Normal	MSE	5-10%	DT-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	KNN-(3/3)	KNN-SOM-(2/3)	DT-(2/3)	DT-(2/3)
		15-20%	DT-MM-SOM-(1/3)	MM-(2/3)	SOM-(3/3)	SOM-(2/3)	SOM-(2/3)	SOM-(3/3)	DT-SOM-(2/3)
		25%	SOM-(3/3)	KNN-MM-SOM-(1/3)	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	DT-MM-SOM-(1/3)	SOM-(2/3)
		Todas	DT-MM-SOM-(1/3)	MM-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)	DT-(2/3)
	r^2	5-10%	KNN-(3/3)	MM-SOM-(2/3)	KNN-(2/3)	KNN-(3/3)	MM-SOM-(2/3)	KNN-(2/3)	DT-KNN-(2/3)
		15-20%	SOM-(2/3)	DT-MM-(2/3)	DT-KNN-MM-SOM-(1/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	DT-SOM-(2/3)
		25%	SOM-(3/3)	DT-KNN-MM-SOM-(1/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)
		Todas	SOM-(3/3)	MM-(2/3)	KNN-(2/3)	KNN-(2/3)	SOM-(3/3)	SOM-(2/3)	DT-(2/3)
	D_{KS}	5-10%	SOM-(2/3)	DT-KNN-SOM-(2/3)	KNN-(2/3)	KNN-SOM-(2/3)	SOM-(2/3)	DT-SOM-(2/3)	DT-(2/3)
		15-20%	DT-MM-SOM-(1/3)	SOM-(2/3)	DT-KNN-(2/3)	DT-MM-SOM-(1/3)	SOM-(2/3)	MM-(2/3)	DT-(2/3)
		25%	SOM-(2/3)	SOM-(2/3)	DT-SOM-(2/3)	KNN-(2/3)	KNN-MM-SOM-(1/3)	DT-KNN-MM-SOM-(1/3)	SOM-(2/3)
		Todas	DT-MM-SOM-(1/3)	SOM-(2/3)	DT-KNN-(2/3)	DT-KNN-MM-SOM-(1/3)	SOM-(2/3)	DT-MM-SOM-(1/3)	DT-(2/3)
Rayleigh	MSE	5-10%	KNN-SOM-(1/2)	DT-KNN-SOM-(1/2)	DT-KNN-SOM-(1/2)	DT-KNN-SOM-(1/2)	DT-KNN-MM-SOM-(1/2)	DT-KNN-SOM-(1/2)	DT-KNN-SOM-(1/2)
		15-20%	SOM-(2/2)	DT-KNN-SOM-(1/2)	SOM-(2/2)	SOM-(2/2)	KNN-SOM-(1/2)	KNN-MM-SOM-(1/2)	KNN-MM-SOM-(1/2)
		25%	SOM-(2/2)	KNN-SOM-(1/2)	SOM-(2/2)	SOM-(2/2)	KNN-SOM-(1/2)	SOM-(2/2)	MM-SOM-(1/2)
		Todas	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)
	r^2	5-10%	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-MM-(1/2)	KNN-(2/2)	KNN-(2/2)
		15-20%	KNN-SOM-(1/2)	DT-KNN-SOM-(1/2)	KNN-SOM-(1/2)	MM-SOM-(1/2)	KNN-MM-SOM-(1/2)	KNN-MM-SOM-(1/2)	KNN-SOM-(1/2)
		25%	KNN-MM-SOM-(1/2)	KNN-MM-SOM-(1/2)	MM-(2/2)	KNN-MM-SOM-(1/2)	KNN-MM-(1/2)	KNN-MM-SOM-(1/2)	KNN-MM-SOM-(1/2)
		Todas	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)
	D_{KS}	5-10%	KNN-(2/2)	KNN-SOM-(1/2)	DT-KNN-(1/2)	DT-KNN-(1/2)	KNN-SOM-(1/2)	DT-KNN-(1/2)	DT-KNN-(1/2)
		15-20%	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-SOM-(1/2)
		25%	KNN-MM-SOM-(1/2)	SOM-(2/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-MM-SOM-(1/2)	KNN-SOM-(1/2)
		Todas	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-SOM-(1/2)	KNN-(2/2)	KNN-SOM-(1/2)
t Location-scale	MSE	5-10%	KNN-(4/6)	KNN-(6/6)	KNN-(6/6)	KNN-(4/6)	KNN-(6/6)	KNN-(4/6)	DT-KNN-(3/6)
		15-20%	KNN-SOM-(3/6)	KNN-(6/6)	KNN-SOM-(3/6)	KNN-SOM-(3/6)	KNN-(6/6)	DT-(3/6)	SOM-(4/6)
		25%	SOM-(3/6)	KNN-(6/6)	SOM-(6/6)	DT-SOM-(3/6)	KNN-(6/6)	SOM-(3/6)	SOM-(4/6)
		Todas	KNN-SOM-(2/6)	KNN-(6/6)	SOM-(4/6)	KNN-(3/6)	KNN-(6/6)	DT-KNN-(3/6)	KNN-(3/6)
	r^2	5-10%	KNN-(5/6)	KNN-(5/6)	KNN-(5/6)	KNN-(5/6)	DT-(3/6)	KNN-(5/6)	KNN-(5/6)
		15-20%	KNN-(5/6)	DT-KNN-(4/6)	KNN-(5/6)	KNN-(4/6)	KNN-(4/6)	KNN-(5/6)	DT-KNN-(3/6)
		25%	SOM-(4/6)	DT-(3/6)	SOM-(5/6)	SOM-(4/6)	DT-(4/6)	SOM-(3/6)	SOM-(3/6)
		Todas	KNN-(5/6)	DT-(3/6)	KNN-(3/6)	KNN-(4/6)	DT-(3/6)	KNN-(5/6)	DT-KNN-(3/6)
	D_{KS}	5-10%	KNN-(6/6)	KNN-(5/6)	KNN-(5/6)	KNN-(4/6)	KNN-(5/6)	KNN-(3/6)	KNN-(4/6)
		15-20%	DT-KNN-(3/6)	KNN-(5/6)	KNN-(6/6)	KNN-(4/6)	KNN-SOM-(4/6)	KNN-(6/6)	KNN-(4/6)
		25%	KNN-(3/6)	KNN-SOM-(3/6)	KNN-(4/6)	KNN-(5/6)	SOM-(4/6)	KNN-(5/6)	KNN-(4/6)
		Todas	DT-KNN-(3/6)	KNN-(5/6)	KNN-(5/6)	KNN-(4/6)	KNN-SOM-(3/6)	KNN-(5/6)	KNN-(4/6)
Weibull	MSE	5-10%	SOM-(2/3)	KNN-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	SOM-(2/3)	SOM-(2/3)
		15-20%	SOM-(3/3)	KNN-(2/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)
		25%	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)
		Todas	SOM-(3/3)	KNN-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)
	r^2	5-10%	KNN-SOM-(3/3)	DT-(2/3)	KNN-(3/3)	KNN-SOM-(2/3)	KNN-(2/3)	SOM-(2/3)	KNN-(3/3)
		15-20%	SOM-(3/3)	KNN-(2/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)
		25%	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)
		Todas	SOM-(3/3)	DT-KNN-SOM-(1/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)
	D_{KS}	5-10%	SOM-(2/3)	SOM-(3/3)	DT-KNN-(2/3)	SOM-(3/3)	SOM-(3/3)	DT-(3/3)	DT-KNN-SOM-(1/3)
		15-20%	SOM-(2/3)	SOM-(3/3)	SOM-(3/3)	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	DT-KNN-SOM-(1/3)
		25%	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)	SOM-(3/3)
		Todas	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	SOM-(2/3)	SOM-(3/3)	SOM-(2/3)	DT-KNN-SOM-(1/3)

Tabela A.5: Técnicas de imputação vencedoras com a votação para cada distribuição, métrica, MR e estratégia (incluindo SVM).

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Beta	MSE	5-10%	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
		25%	DT-KNN-MM-SOM-SVM-(1/3)	SVM-(3/3)	SVM-(2/3)	KNN-MM-SOM-(1/3)	SVM-(3/3)	MM-SOM-SVM-(1/3)	KNN-MM-SOM-(1/3)
		Todas	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)
	r^2	5-10%	KNN-SOM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	KNN-SVM-(2/3)	SVM-(3/3)
		15-20%	SVM-(3/3)	SVM-(2/3)	SOM-SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	KNN-(2/3)	SVM-(2/3)
		25%	DT-KNN-MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(2/3)	SOM-(2/3)	SVM-(2/3)	SOM-(2/3)	KNN-MM-SVM-(1/3)
		Todas	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	KNN-MM-SOM-SVM-(1/3)	SVM-(3/3)	KNN-(2/3)	SVM-(3/3)
	D_{KS}	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)
		25%	SVM-(3/3)	KNN-SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	MM-SOM-SVM-(1/3)	SVM-(3/3)
		Todas	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)
Birnbaum-Saunders	MSE	5-10%	SVM-(6/7)	SVM-(6/7)	SVM-(6/7)	SVM-(6/7)	SVM-(6/7)	SVM-(7/7)	SVM-(7/7)
		15-20%	SVM-(6/7)	SVM-(4/7)	SVM-(6/7)	SVM-(7/7)	SVM-(5/7)	SVM-(7/7)	SVM-(7/7)
		25%	SVM-(5/7)	SVM-(4/7)	SVM-(4/7)	SVM-(6/7)	SVM-(5/7)	SVM-(5/7)	SVM-(7/7)
		Todas	SVM-(5/7)	SVM-(4/7)	SVM-(6/7)	SVM-(6/7)	SVM-(5/7)	SVM-(7/7)	SVM-(7/7)
	r^2	5-10%	SVM-(7/7)	SVM-(5/7)	SVM-(7/7)	SVM-(7/7)	SVM-(5/7)	SVM-(7/7)	SVM-(7/7)
		15-20%	SVM-(6/7)	SVM-(5/7)	SVM-(6/7)	SVM-(7/7)	SVM-(5/7)	SVM-(7/7)	SVM-(7/7)
		25%	SVM-(7/7)	SVM-(4/7)	SVM-(5/7)	SVM-(6/7)	SVM-(4/7)	SVM-(6/7)	SVM-(7/7)
		Todas	SVM-(7/7)	SVM-(4/7)	SVM-(6/7)	SVM-(7/7)	SVM-(4/7)	SVM-(7/7)	SVM-(7/7)
	D_{KS}	5-10%	SVM-(7/7)	SVM-(7/7)	SVM-(6/7)	SVM-(7/7)	SVM-(7/7)	SVM-(5/7)	SVM-(6/7)
		15-20%	SVM-(6/7)	SVM-(6/7)	SVM-(5/7)	SVM-(6/7)	SVM-(4/7)	SVM-(6/7)	SVM-(5/7)
		25%	SVM-(4/7)	SVM-(5/7)	SVM-(6/7)	SVM-(6/7)	SVM-(4/7)	SVM-(5/7)	SVM-(5/7)
		Todas	SVM-(6/7)	SVM-(6/7)	SVM-(6/7)	SVM-(7/7)	SVM-(5/7)	SVM-(5/7)	SVM-(5/7)
Exponencial	MSE	5-10%	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	MM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)
		15-20%	KNN-SOM-SVM-(1/2)	SVM-(2/2)	MM-SOM-SVM-(1/2)	DT-MM-SOM-(1/2)	SVM-(2/2)	SVM-(2/2)	SOM-SVM-(1/2)
		25%	KNN-SVM-(1/2)	KNN-SVM-(1/2)	MM-SOM-(1/2)	DT-MM-(1/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)
		Todas	SVM-(2/2)	SVM-(2/2)	MM-SOM-(1/2)	DT-MM-SOM-(1/2)	SVM-(2/2)	SOM-SVM-(1/2)	SOM-SVM-(1/2)
	r^2	5-10%	SVM-(2/2)	KNN-SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	KNN-(2/2)	SVM-(2/2)
		15-20%	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	KNN-(2/2)	KNN-MM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)
		25%	MM-SOM-(1/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	SOM-SVM-(1/2)	SOM-SVM-(1/2)	SVM-(2/2)
		Todas	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)
	D_{KS}	5-10%	KNN-SOM-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)	SVM-(2/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		15-20%	KNN-SOM-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-(2/2)	KNN-(2/2)	KNN-(2/2)
		25%	KNN-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-(2/2)	KNN-(2/2)
		Todas	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-(2/2)	KNN-(2/2)
Valor Extremo	MSE	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)
		15-20%	KNN-SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)
		25%	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
	r^2	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	KNN-SOM-SVM-(1/3)	SVM-(3/3)
		15-20%	KNN-MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	KNN-SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
		25%	MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
		Todas	KNN-MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)	KNN-SOM-SVM-(1/3)	SVM-(2/3)
	D_{KS}	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)
		15-20%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	KNN-SVM-(2/3)	SVM-(2/3)	SVM-(2/3)
		25%	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	KNN-(2/3)	KNN-(2/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(2/3)

Continua na página seguinte...

Tabela A.5: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Gama	MSE	5-10%	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
		15-20%	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
		25%	SVM-(5/6)	SVM-(6/6)	SVM-(4/6)	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)
		Todas	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
	r^2	5-10%	KNN-SVM-(3/6)	SVM-(6/6)	SVM-(4/6)	SVM-(4/6)	SVM-(5/6)	KNN-SVM-(4/6)	SVM-(5/6)
		15-20%	SVM-(4/6)	SVM-(6/6)	SOM-SVM-(2/6)	SVM-(4/6)	SVM-(5/6)	SVM-(5/6)	SVM-(5/6)
		25%	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
		Todas	SVM-(3/6)	SVM-(6/6)	SVM-(3/6)	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)
	D_{KS}	5-10%	KNN-SVM-(3/6)	SVM-(4/6)	SVM-(4/6)	SVM-(5/6)	KNN-(4/6)	SVM-(4/6)	KNN-SVM-(3/6)
		15-20%	SVM-(4/6)	SVM-(4/6)	SVM-(4/6)	SVM-(5/6)	KNN-SVM-(3/6)	SVM-(4/6)	KNN-SVM-(3/6)
		25%	KNN-SVM-(3/6)	SVM-(4/6)	KNN-SVM-(3/6)	SVM-(5/6)	KNN-SVM-(3/6)	SVM-(4/6)	KNN-SVM-(3/6)
		Todas	KNN-SVM-(3/6)	SVM-(4/6)	SVM-(4/6)	SVM-(5/6)	KNN-SVM-(3/6)	KNN-SVM-(3/6)	KNN-SVM-(3/6)
Valor Extremo Generalizado	MSE	5-10%	SVM-(11/12)	SVM-(9/12)	SVM-(10/12)	SVM-(10/12)	SVM-(10/12)	SVM-(12/12)	SVM-(12/12)
		15-20%	SVM-(7/12)	SVM-(10/12)	SVM-(7/12)	SVM-(9/12)	SVM-(10/12)	SVM-(9/12)	SVM-(9/12)
		25%	SVM-(6/12)	SVM-(9/12)	SOM-(7/12)	SVM-(7/12)	SVM-(9/12)	SVM-(9/12)	SVM-(8/12)
		Todas	SVM-(7/12)	SVM-(9/12)	SVM-(7/12)	SVM-(10/12)	SVM-(9/12)	SVM-(10/12)	SVM-(10/12)
	r^2	5-10%	SVM-(11/12)	SVM-(12/12)	SVM-(9/12)	SVM-(10/12)	SVM-(11/12)	SVM-(10/12)	SVM-(11/12)
		15-20%	SVM-(7/12)	SVM-(9/12)	SVM-(8/12)	SVM-(9/12)	SVM-(10/12)	SVM-(7/12)	SVM-(9/12)
		25%	SVM-(8/12)	SVM-(7/12)	SVM-(9/12)	SVM-(8/12)	SVM-(7/12)	SVM-(9/12)	SVM-(9/12)
		Todas	SVM-(7/12)	SVM-(11/12)	SVM-(7/12)	SVM-(9/12)	SVM-(10/12)	SVM-(9/12)	SVM-(8/12)
	D_{KS}	5-10%	SVM-(11/12)	SVM-(12/12)	SVM-(10/12)	SVM-(11/12)	SVM-(11/12)	SVM-(9/12)	SVM-(8/12)
		15-20%	SVM-(8/12)	SVM-(11/12)	SVM-(11/12)	SVM-(9/12)	SVM-(9/12)	SVM-(9/12)	SVM-(7/12)
		25%	SVM-(9/12)	SVM-(12/12)	SVM-(10/12)	SVM-(10/12)	SVM-(11/12)	SVM-(10/12)	SVM-(7/12)
		Todas	SVM-(8/12)	SVM-(11/12)	SVM-(11/12)	SVM-(11/12)	SVM-(9/12)	SVM-(10/12)	SVM-(7/12)
Pareto Generalizada	MSE	5-10%	SVM-(9/9)	SVM-(8/9)	SVM-(7/9)	SVM-(9/9)	SVM-(9/9)	SVM-(8/9)	SVM-(9/9)
		15-20%	SVM-(7/9)	SVM-(7/9)	SVM-(7/9)	SVM-(9/9)	SVM-(9/9)	SVM-(7/9)	SVM-(9/9)
		25%	SVM-(9/9)	SVM-(9/9)	SVM-(7/9)	SVM-(9/9)	SVM-(9/9)	SVM-(8/9)	SVM-(9/9)
		Todas	SVM-(8/9)	SVM-(9/9)	SVM-(8/9)	SVM-(9/9)	SVM-(9/9)	SVM-(8/9)	SVM-(9/9)
	r^2	5-10%	SVM-(9/9)	SVM-(9/9)	SVM-(6/9)	SVM-(9/9)	SVM-(8/9)	SVM-(7/9)	SVM-(9/9)
		15-20%	SVM-(8/9)	SVM-(9/9)	SVM-(7/9)	SVM-(7/9)	SVM-(9/9)	SVM-(5/9)	SVM-(9/9)
		25%	SVM-(8/9)	SVM-(9/9)	SOM-(5/9)	SVM-(9/9)	SVM-(8/9)	SVM-(6/9)	SVM-(8/9)
		Todas	SVM-(8/9)	SVM-(9/9)	SVM-(6/9)	SVM-(7/9)	SVM-(9/9)	SVM-(6/9)	SVM-(9/9)
	D_{KS}	5-10%	SVM-(9/9)	SVM-(8/9)	SVM-(7/9)	SVM-(8/9)	SVM-(7/9)	SVM-(6/9)	SVM-(5/9)
		15-20%	SVM-(8/9)	SVM-(6/9)	SVM-(8/9)	SVM-(7/9)	SVM-(7/9)	SVM-(7/9)	SVM-(5/9)
		25%	SVM-(9/9)	SVM-(8/9)	SVM-(8/9)	SVM-(7/9)	SVM-(6/9)	SVM-(7/9)	SVM-(5/9)
		Todas	SVM-(8/9)	SVM-(7/9)	SVM-(7/9)	SVM-(6/9)	SVM-(7/9)	SVM-(6/9)	SVM-(5/9)
Gaussiana Inversa	MSE	5-10%	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		15-20%	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)
		25%	SVM-(4/5)	SVM-(4/5)	SOM-SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SOM-SVM-(3/5)	SVM-(4/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SOM-SVM-(3/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)
	r^2	5-10%	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
		15-20%	SVM-(3/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)
		25%	KNN-SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(4/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)
	D_{KS}	5-10%	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		15-20%	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)
		25%	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)	SVM-(3/5)
		Todas	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)

Continua na página seguinte...

Tabela A.5: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Logistica	MSE	5-10%	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		15-20%	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		25%	SOM-SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
	r^2	5-10%	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(5/5)
		15-20%	KNN-SOM-SVM-(2/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(5/5)
		25%	SOM-SVM-(2/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
		Todas	SOM-SVM-(2/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(5/5)
	D_{KS}	5-10%	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	KNN-(3/5)
		15-20%	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	KNN-(3/5)
		25%	SVM-(4/5)	SVM-(3/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	KNN-SVM-(3/5)
		Todas	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	KNN-(3/5)
Log-Logistica	MSE	5-10%	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		15-20%	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
		25%	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
	r^2	5-10%	SVM-(3/5)	SVM-(4/5)	KNN-(3/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
		15-20%	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		25%	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SVM-(3/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
	D_{KS}	5-10%	SVM-(5/5)	SVM-(5/5)	SVM-(3/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
		15-20%	SVM-(5/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)
		25%	SVM-(4/5)	SVM-(4/5)	SOM-SVM-(2/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)
		Todas	SVM-(5/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)
Normal	MSE	5-10%	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)
		15-20%	SVM-(3/4)	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)
		25%	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)
		Todas	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)
	r^2	5-10%	KNN-(3/4)	SVM-(4/4)	KNN-SVM-(2/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(4/4)
		15-20%	SOM-(3/4)	SVM-(3/4)	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)
		25%	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)	SVM-(3/4)	SVM-(4/4)	SVM-(3/4)
		Todas	SVM-(3/4)	SVM-(4/4)	KNN-SVM-(2/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(4/4)
	D_{KS}	5-10%	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)	SVM-(3/4)	KNN-SVM-(3/4)	SVM-(3/4)	KNN-SVM-(2/4)
		15-20%	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)
		25%	SVM-(4/4)	SVM-(2/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)
		Todas	SVM-(4/4)	SVM-(3/4)	SVM-(3/4)	SVM-(4/4)	SVM-(4/4)	SVM-(3/4)	KNN-SVM-(2/4)
Nakagami	MSE	5-10%	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)
		15-20%	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)
		25%	SVM-(4/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)
		Todas	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)	SVM-(4/5)
	r^2	5-10%	KNN-SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	KNN-SVM-(3/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)
		15-20%	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(4/5)	SVM-(3/5)	SVM-(3/5)
		25%	SVM-(4/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)
		Todas	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(4/5)	SVM-(3/5)	SVM-(4/5)
	D_{KS}	5-10%	SVM-(5/5)	SVM-(5/5)	SVM-(4/5)	SVM-(5/5)	SVM-(5/5)	SVM-(5/5)	SVM-(3/5)
		15-20%	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)
		25%	SVM-(4/5)	SVM-(5/5)	SVM-(4/5)	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)
		Todas	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)	SVM-(5/5)	SVM-(3/5)	SVM-(3/5)

Continua na página seguinte...

Tabela A.5: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Log-Normal	MSE	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(2/3)	SVM-(2/3)	SOM-SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SOM-SVM-(2/3)	SVM-(3/3)
		25%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
	r^2	5-10%	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(2/3)	SVM-(2/3)	KNN-MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(3/3)	SOM-SVM-(2/3)	SVM-(3/3)
		25%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
	D_{KS}	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(3/3)	KNN-MM-SOM-SVM-(1/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)
		25%	SVM-(2/3)	MM-SOM-SVM-(1/3)	SVM-(3/3)	KNN-MM-SOM-SVM-(1/3)	SVM-(3/3)	MM-SOM-SVM-(1/3)	SVM-(2/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)
Rayleigh	MSE	5-10%	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
		15-20%	SVM-(2/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
		25%	SVM-(2/2)	SVM-(2/2)	MM-SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
		Todas	SVM-(2/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
	r^2	5-10%	KNN-(2/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-(2/2)	MM-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)
		15-20%	KNN-SVM-(1/2)	SVM-(2/2)	KNN-MM-SOM-SVM-(1/2)	MM-SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
		25%	SVM-(2/2)	SVM-(2/2)	MM-(2/2)	KNN-MM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
		Todas	KNN-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	KNN-SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SVM-(2/2)
	D_{KS}	5-10%	KNN-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)
		15-20%	KNN-SOM-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)
		25%	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	SVM-(2/2)	SOM-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)
		Todas	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)	SVM-(2/2)	KNN-SVM-(1/2)	KNN-SVM-(1/2)
t Location-scale	MSE	5-10%	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
		15-20%	KNN-SVM-(3/6)	SVM-(6/6)	SVM-(4/6)	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
		25%	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
		Todas	SVM-(4/6)	SVM-(6/6)	SVM-(6/6)	SVM-(4/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)
	r^2	5-10%	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
		15-20%	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
		25%	SVM-(5/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
		Todas	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	SVM-(4/6)	SVM-(6/6)	SVM-(5/6)	SVM-(6/6)
	D_{KS}	5-10%	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(6/6)	SVM-(4/6)	KNN-SVM-(3/6)
		15-20%	SVM-(6/6)	SVM-(6/6)	SVM-(4/6)	SVM-(5/6)	SVM-(6/6)	SVM-(5/6)	KNN-SVM-(3/6)
		25%	SVM-(5/6)	SVM-(5/6)	SVM-(5/6)	SVM-(5/6)	SVM-(5/6)	SVM-(5/6)	KNN-SVM-(3/6)
		Todas	SVM-(5/6)	SVM-(5/6)	SVM-(4/6)	SVM-(5/6)	SVM-(5/6)	SVM-(4/6)	KNN-SVM-(3/6)
Weibull	MSE	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		25%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
	r^2	5-10%	KNN-SOM-(2/3)	SVM-(3/3)	KNN-SVM-(2/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SOM-(2/3)	KNN-SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		25%	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)	SOM-SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)
		Todas	SOM-(2/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)
	D_{KS}	5-10%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)
		15-20%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SOM-SVM-(2/3)	SVM-(3/3)	SVM-(3/3)
		25%	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)
		Todas	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(3/3)	SVM-(2/3)	SVM-(3/3)	SVM-(3/3)

Tabela A.6: Técnicas de imputação vencedoras com a sua média para cada distribuição, métrica, MR e estratégia (excluindo SVM).
O código de cor contêm informação dos empates: vermelho corresponde ao empate de todos os métodos, amarelo ao empate entre 3 métodos e verde ao empate entre 2 métodos.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Beta	MSE	5-10%	DT[0.035882±0.0060517]	DT[0.00022353±0.00018827]	SOM[0.053841±0.054059]	SOM[0.040822±0.026387]	KNN[0.076636±0.075282]	SOM[0.055708±0.021929]	SOM[0.027941±0.031969]
		15-20%	SOM[0.10993±0.059654]	KNN[0.13672±0.13294]	SOM[0.12703±0.1298]	MM[0.042514±0.0057678]	KNN[0.1079±0.090925]	MM[0.050106±0.019381]	SOM[0.07512±0.06778]
		25%	SOM[0.13408±0.08651]	KNN[0.15884±0.12366]	MM[0.17536±0.2294]	MM[0.062625±0.01789]	SOM[0.082621±0.070384]	SOM[0.17736±0.15882]	SOM[0.10788±0.10202]
		Todas	MM[0.026292±0.016263]	KNN[0.14489±0.1018]	SOM[0.088418±0.093859]	MM[0.049405±0.017943]	KNN[0.11092±0.092744]	MM[0.046586±0.017953]	MM[0.060943±0.0737]
	r^2	5-10%	KNN[0.79855±0.37056]	SOM[0.19428±0.11996]	KNN[0.73087±0.34426]	SOM[0.744±0.40377]	DT[0.35052±0.24689]	KNN[0.67639±0.38377]	KNN[0.72234±0.29822]
		15-20%	SOM[0.61621±0.40288]	SOM[0.56294±0.15488]	SOM[0.50698±0.42487]	SOM[0.8159±0.13431]	SOM[0.48519±0.26554]	KNN[0.19446±0.24908]	SOM[0.6015±0.35978]
		25%	SOM[0.81234±0.063703]	KNN[0.10033±0.061011]	SOM[0.53337±0.43429]	SOM[0.51036±0.35432]	SOM[0.36706±0.40121]	SOM[0.48613±0.43868]	MM[0.58653±0.46232]
		Todas	SOM[0.66367±0.33746]	SOM[0.4393±0.22218]	MM[0.92747±0.024277]	MM[0.76571±0.17243]	DT[0.43019±0.25681]	MM[0.83931±0.032833]	SOM[0.64156±0.33141]
	D_{KS}	5-10%	KNN[0.48934±0.33033]	KNN[0.52296±0.42159]	KNN[0.3719±0.12667]	KNN[0.37264±0.13004]	KNN[0.44087±0.20347]	KNN[0.33231±0.10743]	KNN[0.17244±0.055552]
		15-20%	KNN[0.47696±0.30718]	SOM[0.19279±0.062948]	KNN[0.26453±0.12089]	SOM[0.22577±0.090675]	SOM[0.168±0.068487]	KNN[0.23173±0.095855]	KNN[0.15344±0.037309]
		25%	KNN[0.51177±0.14653]	SOM[0.1856±0.11273]	KNN[0.25807±0.21523]	MM[0.18235±0.05823]	KNN[0.29789±0.14001]	MM[0.14881±0.0084146]	MM[0.12941±0.034688]
		Todas	KNN[0.48738±0.28659]	KNN[0.47997±0.33989]	KNN[0.31741±0.13645]	KNN[0.35379±0.11404]	SOM[0.21388±0.10171]	KNN[0.29583±0.1054]	KNN[0.16056±0.044999]
Birnbaum-Saunders	MSE	5-10%	SOM[0.020785±0.030717]	KNN[0.0001841±0.0020196]	KNN[0.080911±0.042088]	KNN[0.080392±0.070467]	SOM[0.00021987±0.0018644]	SOM[0.014267±0.013631]	SOM[0.0040318±0.010471]
		15-20%	SOM[0.026252±0.070216]	KNN[0.0011339±0.0095155]	SOM[0.021162±0.035475]	SOM[0.014337±0.045765]	SOM[0.0011062±0.0079416]	SOM[0.015805±0.025946]	SOM[0.010581±0.021858]
		25%	SOM[0.021465±0.049302]	SOM[0.0011127±0.0071926]	SOM[0.020244±0.033922]	SOM[0.017429±0.069685]	SOM[0.0010677±0.0064629]	SOM[0.018534±0.051243]	SOM[0.013734±0.036051]
		Todas	SOM[0.023103±0.053174]	KNN[0.0014±0.011468]	SOM[0.019678±0.027814]	SOM[0.016179±0.045302]	SOM[0.00074576±0.0059268]	SOM[0.015762±0.029679]	SOM[0.0085972±0.022543]
	r^2	5-10%	KNN[0.36447±0.23966]	KNN[0.45158±0.35409]	KNN[0.71642±0.18495]	SOM[0.90733±0.10375]	SOM[0.9718±0.1276]	KNN[0.84777±0.22039]	KNN[0.96816±0.098474]
		15-20%	SOM[0.95215±0.11041]	SOM[0.3308±0.18377]	SOM[0.95936±0.080163]	SOM[0.95165±0.13433]	SOM[0.96976±0.10881]	SOM[0.96349±0.076289]	SOM[0.95803±0.10782]
		25%	SOM[0.95039±0.13108]	SOM[0.80303±0.10018]	SOM[0.95945±0.090071]	SOM[0.97665±0.082168]	SOM[0.97119±0.10857]	SOM[0.9647±0.097472]	SOM[0.95975±0.11836]
		Todas	KNN[0.40405±0.24902]	SOM[0.75087±0.19335]	SOM[0.95882±0.065735]	SOM[0.93831±0.11583]	SOM[0.97086±0.1164]	SOM[0.96318±0.071067]	SOM[0.96501±0.10246]
	D_{KS}	5-10%	SOM[0.18736±0.095181]	KNN[0.77899±0.21092]	KNN[0.47717±0.032084]	KNN[0.45008±0.13404]	SOM[0.33937±0.094211]	KNN[0.28089±0.056917]	KNN[0.21365±0.032426]
		15-20%	SOM[0.074703±0.056041]	KNN[0.97603±0.11703]	KNN[0.48654±0.045399]	SOM[0.26779±0.099231]	KNN[0.30944±0.12324]	SOM[0.17732±0.025547]	KNN[0.21954±0.031646]
		25%	SOM[0.047235±0.028516]	SOM[0.78907±0.10993]	SOM[0.49008±0.050215]	SOM[0.23511±0.032482]	SOM[0.19154±0.079873]	SOM[0.19218±0.0536]	KNN[0.20577±0.028675]
		Todas	SOM[0.11404±0.093143]	KNN[0.87357±0.20076]	KNN[0.48379±0.0408]	SOM[0.18825±0.085376]	SOM[0.26732±0.108]	SOM[0.18783±0.03974]	KNN[0.21442±0.031742]
Exponencial	MSE	5-10%	KNN[0.099393±0.070874]	KNN[0.057877±0.082377]	KNN[0.058834±0.039903]	KNN[0.084086±0.06917]	SOM[0.012688±0.026483]	KNN[0.055082±0.035704]	DT[0.024717±0.027203]
		15-20%	SOM[0.11778±0.067827]	SOM[0.020735±0.021997]	SOM[0.17817±0.13331]	SOM[0.13229±0.10787]	SOM[0.017612±0.018688]	SOM[0.070583±0.046026]	SOM[0.080016±0.06346]
		25%	SOM[0.12429±0.058683]	SOM[0.039012±0.03132]	SOM[0.16785±0.14882]	SOM[0.12613±0.068573]	SOM[0.055824±0.07261]	SOM[0.14074±0.11049]	SOM[0.11509±0.083818]
		Todas	KNN[0.14832±0.083736]	SOM[0.026585±0.028754]	SOM[0.15961±0.13021]	SOM[0.11016±0.089041]	SOM[0.026101±0.045289]	SOM[0.090764±0.089906]	SOM[0.069044±0.067592]
	r^2	5-10%	KNN[0.34704±0.2168]	KNN[0.27934±0.18701]	KNN[0.72421±0.31827]	KNN[0.54643±0.28408]	DT[0.35109±0.34572]	KNN[0.68639±0.30483]	KNN[0.72023±0.28539]
		15-20%	KNN[0.36359±0.28192]	KNN[0.24829±0.17094]	KNN[0.55376±0.3573]	SOM[0.50797±0.29441]	SOM[0.35111±0.20932]	KNN[0.51955±0.34354]	SOM[0.57955±0.33978]
		25%	KNN[0.31494±0.21814]	SOM[0.30291±0.1795]	SOM[0.69369±0.29474]	SOM[0.56081±0.25662]	KNN[0.31921±0.24976]	SOM[0.65012±0.29932]	SOM[0.56729±0.32833]
		Todas	KNN[0.34721±0.23745]	KNN[0.26538±0.17567]	KNN[0.63201±0.34953]	KNN[0.47041±0.27484]	SOM[0.34132±0.21457]	KNN[0.59809±0.33175]	SOM[0.61191±0.3279]
	D_{KS}	5-10%	KNN[0.58817±0.25401]	KNN[0.58923±0.30027]	KNN[0.37071±0.12468]	KNN[0.34478±0.12544]	KNN[0.3221±0.096349]	KNN[0.28805±0.13284]	KNN[0.14271±0.071886]
		15-20%	SOM[0.33136±0.1788]	KNN[0.48305±0.24122]	KNN[0.34223±0.10977]	KNN[0.32556±0.099312]	KNN[0.2798±0.10288]	KNN[0.23562±0.10427]	KNN[0.12745±0.054671]
		25%	SOM[0.28667±0.21815]	KNN[0.37943±0.19232]	KNN[0.3183±0.091516]	SOM[0.20075±0.07564]	KNN[0.27166±0.10797]	KNN[0.21994±0.075433]	KNN[0.13338±0.053145]
		Todas	KNN[0.54074±0.24689]	KNN[0.51294±0.26973]	KNN[0.3508±0.11411]	KNN[0.31643±0.11943]	KNN[0.29707±0.10153]	KNN[0.25862±0.11746]	KNN[0.13481±0.061754]
Valor Extremo	MSE	5-10%	KNN[0.17558±0.10875]	KNN[0.001267±0.0011107]	KNN[0.11175±0.058643]	KNN[0.15054±0.086996]	KNN[0.0047826±0.003789]	KNN[0.11835±0.090429]	DT[0.031883±0.019104]
		15-20%	SOM[0.23042±0.092838]	KNN[0.0090903±0.0049036]	SOM[0.20911±0.047375]	KNN[0.26433±0.10801]	KNN[0.019182±0.014385]	DT[0.16593±0.078741]	SOM[0.10127±0.037534]
		25%	SOM[0.29993±0.048794]	KNN[0.025715±0.0107]	SOM[0.28576±0.10932]	DT[0.22448±0.082407]	KNN[0.046163±0.023216]	SOM[0.28202±0.058801]	SOM[0.16187±0.049711]
		Todas	SOM[0.23335±0.090397]	KNN[0.0095618±0.010442]	SOM[0.22315±0.093604]	KNN[0.21429±0.11501]	KNN[0.020046±0.021268]	DT[0.13887±0.078545]	KNN[0.085641±0.062906]
	r^2	5-10%	KNN[0.61273±0.23195]	KNN[0.20032±0.16577]	KNN[0.63237±0.23067]	KNN[0.73402±0.17866]	DT[0.38036±0.25658]	KNN[0.64464±0.23617]	KNN[0.63358±0.21393]
		15-20%	KNN[0.4112±0.18087]	DT[0.27675±0.17654]	KNN[0.3865±0.19552]	KNN[0.40473±0.22225]	KNN[0.19124±0.16707]	KNN[0.45022±0.22879]	DT[0.49038±0.1842]
		25%	SOM[0.42857±0.13188]	DT[0.20575±0.15913]	SOM[0.25327±0.17943]	SOM[0.30555±0.19805]	DT[0.28489±0.17519]	SOM[0.26923±0.17351]	SOM[0.38936±0.18568]
		Todas	KNN[0.4766±0.23327]	DT[0.25566±0.16944]	KNN[0.51073±0.24469]	KNN[0.57243±0.25499]	DT[0.36116±0.22265]	KNN[0.53526±0.25436]	DT[0.53702±0.20949]
	D_{KS}	5-10%	KNN[0.48573±0.14615]	KNN[0.25595±0.094595]	KNN[0.30682±0.077902]	KNN[0.42873±0.15319]	KNN[0.31187±0.091828]	KNN[0.25957±0.07297]	KNN[0.16442±0.063837]
		15-20%	KNN[0.39025±0.11042]	KNN[0.21414±0.064845]	KNN[0.25564±0.048942]	KNN[0.35527±0.11748]	SOM[0.21831±0.056053]	KNN[0.23192±0.053491]	KNN[0.1477±0.049264]
		25%	KNN[0.31457±0.077732]	SOM[0.15253±0.039634]	KNN[0.25225±0.05805]	KNN[0.31809±0.10014]	SOM[0.19467±0.069663]	KNN[0.21909±0.044202]	KNN[0.1505±0.04926]
		Todas	KNN[0.42647±0.13972]	KNN[0.22744±0.088427]	KNN[0.2774±0.06817]	KNN[0.37487±0.13416]	SOM[0.22976±0.074349]	KNN[0.24021±0.061625]	KNN[0.15495±0.055397]

Continua na página seguinte...

Tabela A.6: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Gama	MSE	5-10%	KNN[0.11248±0.084352]	KNN[0.0079157±0.0073656]	KNN[0.050127±0.019198]	KNN[0.066428±0.042359]	MM[2.5855e-06±1.3912e-06]	KNN[0.043264±0.010078]	KNN[0.037725±0.025976]
		15-20%	KNN[0.16942±0.10855]	KNN[0.073144±0.054706]	MM[0.14311±0.016504]	MM[0.14109±0.031424]	KNN[0.044859±0.054589]	KNN[0.12511±0]	MM[0.074189±0.017722]
		25%	MM[0.18015±0]	MM[0.069898±0]	MM[0.17848±0]	MM[0.18433±0]	KNN[0.06419±0]	MM[0.16401±0]	MM[0.10974±0]
		Todas	KNN[0.15215±0.087714]	KNN[0.061169±0.058521]	MM[0.1549±0.02352]	MM[0.1555±0.033421]	KNN[0.049692±0.045608]	KNN[0.070546±0.047788]	KNN[0.086837±0.067835]
	r^2	5-10%	KNN[0.30851±0.01563]	KNN[0.22463±0.19068]	KNN[0.83938±0.062947]	KNN[0.39056±0]	MM[0.69618±0.52623]	KNN[0.62695±0.089893]	KNN[0.56683±0.27193]
		15-20%	KNN[0.44839±0]	KNN[0.47452±0]	KNN[0.49459±0.31641]	KNN[0.34635±0.21331]	KNN[0.14781±0.098222]	KNN[0.29067±0.25468]	KNN[0.43661±0.22047]
		25%	MM[0.37543±0]	MM[0.43676±0]	MM[0.59969±0]	MM[0.599±0]	MM[0.44518±0]	MM[0.41161±0]	MM[0.56095±0]
		Todas	KNN[0.33649±0.064004]	KNN[0.27461±0.19939]	KNN[0.63251±0.29447]	MM[0.46468±0.16593]	MM[0.53895±0.44148]	KNN[0.49244±0.23277]	KNN[0.46641±0.24782]
	D_{KS}	5-10%	KNN[0.4581±0.2221]	KNN[0.45504±0.2967]	KNN[0.32731±0.11297]	KNN[0.38286±0.083648]	KNN[0.50447±0.29096]	KNN[0.29034±0.10423]	KNN[0.11872±0.084485]
		15-20%	KNN[0.22567±0.068356]	KNN[0.49538±0.29476]	KNN[0.31964±0.033964]	KNN[0.2875±0.02192]	KNN[0.41259±0.25117]	KNN[0.23916±0.037599]	KNN[0.099282±0.053055]
		25%	KNN[0.3088±0]	KNN[0.34739±0.19938]	KNN[0.31242±0.051506]	KNN[0.32629±0.11411]	KNN[0.28562±0.089407]	KNN[0.22394±0.077973]	KNN[0.093701±0.044531]
		Todas	KNN[0.35574±0.18482]	KNN[0.44965±0.25688]	KNN[0.32127±0.070483]	KNN[0.33945±0.080513]	KNN[0.42395±0.23943]	KNN[0.25659±0.075142]	KNN[0.10594±0.060527]
Valor Extremo Generalizado	MSE	5-10%	KNN[0.031918±0.057599]	SOM[0.0052123±0.0040884]	SOM[0.11813±0.066164]	KNN[0.0073832±0.0057008]	KNN[0.0072573±0.0062623]	KNN[0.02668±0.025922]	SOM[0.029432±0.017325]
		15-20%	KNN[0.027228±0.01787]	SOM[0.02186±0.0085461]	SOM[0.17114±0.065924]	SOM[0.17841±0.081099]	SOM[0.031119±0.016616]	SOM[0.1506±0.060151]	SOM[0.079685±0.030819]
		25%	MM[0.026669±0]	MM[0.014987±0]	SOM[0.19185±0.06539]	SOM[0.19409±0.067024]	MM[0.0058113±0]	SOM[0.16335±0.080371]	SOM[0.12099±0.040778]
		Todas	KNN[0.034001±0.04332]	SOM[0.019786±0.015771]	KNN[0.05177±0.056529]	KNN[0.0098638±0.0074256]	SOM[0.025636±0.02005]	SOM[0.13439±0.066281]	SOM[0.068259±0.04525]
	r^2	5-10%	KNN[0.63974±0.18196]	KNN[0.21409±0.12084]	MM[0.99401±0.0023476]	KNN[0.67538±0.20976]	KNN[0.26912±0.2213]	SOM[0.76634±0.13343]	KNN[0.66906±0.12857]
		15-20%	SOM[0.55678±0.13798]	KNN[0.1674±0.14242]	SOM[0.64302±0.14921]	KNN[0.6014±0.23562]	KNN[0.25929±0.32378]	SOM[0.63667±0.15615]	SOM[0.57731±0.13487]
		25%	SOM[0.5332±0.15342]	SOM[0.19942±0.1588]	MM[0.92802±0]	SOM[0.50906±0.1425]	KNN[0.18003±0.1947]	SOM[0.59974±0.12799]	SOM[0.54942±0.15527]
		Todas	MM[0.95597±0.047054]	KNN[0.18363±0.12691]	MM[0.96989±0.027293]	KNN[0.64762±0.21075]	KNN[0.24905±0.25173]	SOM[0.66908±0.15728]	SOM[0.6044±0.14058]
	D_{KS}	5-10%	KNN[0.58352±0.26188]	SOM[0.36856±0.1123]	KNN[0.40758±0.086552]	KNN[0.54129±0.2444]	SOM[0.35451±0.10472]	SOM[0.33002±0.097093]	KNN[0.20965±0.038392]
		15-20%	KNN[0.39318±0.22376]	SOM[0.30281±0.10064]	KNN[0.27504±0.091844]	SOM[0.3816±0.1048]	KNN[0.31257±0.1217]	SOM[0.25609±0.054107]	KNN[0.19595±0.036834]
		25%	KNN[0.26206±0.14809]	SOM[0.25031±0.080384]	KNN[0.26921±0.12392]	SOM[0.30317±0.092503]	KNN[0.23745±0.11245]	KNN[0.24378±0.083372]	KNN[0.20141±0.038849]
		Todas	KNN[0.47081±0.25836]	SOM[0.31277±0.10922]	KNN[0.37471±0.10627]	KNN[0.46275±0.24357]	KNN[0.32597±0.12509]	SOM[0.27618±0.082891]	KNN[0.20257±0.03813]
Pareto Generalizada	MSE	5-10%	KNN[0.054901±0.049353]	KNN[0.008397±0.0088846]	SOM[0.055464±0.026111]	SOM[0.059117±0.023446]	KNN[0.0039424±0.0044662]	KNN[0.039695±0.029418]	DT[0.019971±0.01157]
		15-20%	SOM[0.11324±0.047124]	SOM[0.019438±0.0097678]	SOM[0.098531±0.036625]	SOM[0.10987±0.045988]	SOM[0.019434±0.010679]	SOM[0.092485±0.044854]	SOM[0.054574±0.026108]
		25%	SOM[0.12736±0.063196]	SOM[0.049688±0.033262]	SOM[0.10914±0.054397]	SOM[0.11258±0.053337]	SOM[0.037903±0.012048]	SOM[0.10874±0.060419]	SOM[0.086792±0.040838]
		Todas	SOM[0.10497±0.051229]	KNN[0.018548±0.016969]	SOM[0.088953±0.044151]	SOM[0.093705±0.048024]	KNN[0.017815±0.020588]	SOM[0.084836±0.049002]	SOM[0.050437±0.037146]
	r^2	5-10%	KNN[0.82157±0.11268]	KNN[0.32094±0.28362]	KNN[0.78438±0.18448]	KNN[0.78471±0.1527]	KNN[0.25365±0.2984]	KNN[0.83912±0.10675]	KNN[0.77427±0.11548]
		15-20%	SOM[0.73792±0.096843]	KNN[0.21653±0.23991]	SOM[0.72984±0.17283]	SOM[0.68209±0.12952]	SOM[0.29316±0.14821]	KNN[0.7101±0.13057]	SOM[0.70807±0.14327]
		25%	SOM[0.73078±0.11682]	DT[0.33936±0.13607]	SOM[0.73804±0.16349]	SOM[0.60188±0.17556]	SOM[0.44291±0.24076]	SOM[0.77797±0.1181]	SOM[0.68672±0.15271]
		Todas	SOM[0.75758±0.12978]	KNN[0.26776±0.26503]	SOM[0.75885±0.16431]	SOM[0.65575±0.21999]	SOM[0.3041±0.20325]	KNN[0.78098±0.13742]	SOM[0.73343±0.14265]
	D_{KS}	5-10%	KNN[0.28549±0.18633]	KNN[0.18695±0.096743]	KNN[0.18356±0.11291]	KNN[0.33613±0.16167]	KNN[0.28261±0.096674]	KNN[0.20266±0.079138]	KNN[0.08946±0.043087]
		15-20%	KNN[0.21623±0.11115]	KNN[0.22024±0.1148]	KNN[0.1561±0.075884]	SOM[0.26244±0.09571]	SOM[0.19971±0.082783]	KNN[0.1897±0.041843]	KNN[0.085143±0.037337]
		25%	KNN[0.19477±0.10583]	KNN[0.18932±0.097087]	KNN[0.14612±0.053574]	SOM[0.18894±0.068229]	SOM[0.17785±0.07807]	KNN[0.19285±0.038374]	KNN[0.092541±0.037828]
		Todas	KNN[0.24462±0.15338]	KNN[0.20121±0.10461]	KNN[0.16622±0.091496]	SOM[0.28901±0.14342]	SOM[0.22351±0.10255]	KNN[0.19583±0.059788]	KNN[0.088381±0.03973]
Gaussiana Inversa	MSE	5-10%	SOM[0.09084±0.091661]	KNN[0.005652±0.0058568]	KNN[0.093233±0.10507]	SOM[0.081504±0.0801]	SOM[0.0052291±0.0084082]	SOM[0.054702±0.066271]	SOM[0.01706±0.018474]
		15-20%	SOM[0.14855±0.12071]	KNN[0.021877±0.019426]	SOM[0.12247±0.10232]	SOM[0.10765±0.092418]	SOM[0.015757±0.020301]	SOM[0.080487±0.07916]	SOM[0.053127±0.043628]
		25%	SOM[0.1869±0.13039]	KNN[0.053667±0.021234]	SOM[0.14656±0.14037]	SOM[0.12604±0.09554]	SOM[0.024677±0.023336]	SOM[0.097441±0.08591]	SOM[0.088373±0.063958]
		Todas	SOM[0.13937±0.1199]	KNN[0.020724±0.023076]	SOM[0.11659±0.11079]	SOM[0.10214±0.09006]	SOM[0.014341±0.019465]	SOM[0.075248±0.077887]	SOM[0.047048±0.050044]
	r^2	5-10%	KNN[0.60654±0.2732]	KNN[0.23743±0.23188]	KNN[0.78318±0.20879]	KNN[0.67669±0.24484]	KNN[0.37165±0.2961]	KNN[0.78802±0.20445]	KNN[0.80457±0.19592]
		15-20%	KNN[0.54328±0.26957]	KNN[0.17887±0.14634]	SOM[0.75861±0.21656]	KNN[0.62154±0.27804]	SOM[0.51208±0.30951]	SOM[0.79489±0.1852]	SOM[0.74984±0.22097]
		25%	SOM[0.66046±0.23784]	SOM[0.42441±0.21898]	SOM[0.73116±0.24966]	SOM[0.66999±0.25577]	SOM[0.52606±0.28185]	SOM[0.72944±0.25243]	SOM[0.6849±0.25464]
		Todas	KNN[0.58072±0.26903]	KNN[0.20855±0.18805]	SOM[0.77233±0.2271]	KNN[0.64632±0.26059]	SOM[0.5228±0.31977]	SOM[0.79178±0.20814]	SOM[0.75938±0.22199]
	D_{KS}	5-10%	KNN[0.49335±0.17422]	KNN[0.50083±0.21561]	KNN[0.34351±0.10102]	SOM[0.34379±0.16971]	KNN[0.33736±0.11902]	KNN[0.27649±0.090408]	KNN[0.14431±0.060216]
		15-20%	SOM[0.27917±0.1433]	KNN[0.51203±0.31331]	KNN[0.30874±0.11403]	KNN[0.23981±0.10763]	KNN[0.23961±0.10527]	SOM[0.19683±0.065872]	KNN[0.13937±0.058312]
		25%	SOM[0.23269±0.10727]	SOM[0.31133±0.19252]	KNN[0.29019±0.14247]	SOM[0.22138±0.077356]	KNN[0.2293±0.09854]	SOM[0.18171±0.058409]	KNN[0.13938±0.056299]
		Todas	KNN[0.4104±0.18178]	KNN[0.482±0.26034]	KNN[0.32377±0.11349]	KNN[0.28136±0.12867]	KNN[0.29148±0.12216]	KNN[0.24197±0.086797]	KNN[0.14148±0.058635]

Continua na página seguinte...

Tabela A.6: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Logistica	MSE	5-10%	SOM[0.13512±0.09813]	KNN[0.0041825±0.0039363]	SOM[0.1112±0.069354]	SOM[0.12305±0.068216]	KNN[0.01113±0.011147]	SOM[0.097197±0.05806]	SOM[0.033217±0.023143]
		15-20%	SOM[0.14933±0.039587]	KNN[0.018154±0.011665]	SOM[0.16183±0.047467]	SOM[0.19501±0.12242]	KNN[0.038269±0.012208]	SOM[0.10912±0.081383]	SOM[0.084439±0.04437]
		25%	SOM[0.23383±0.059971]	KNN[0.044446±0.013928]	SOM[0.19315±0.11143]	SOM[0.16507±0.038429]	KNN[0.045314±0.015153]	SOM[0.15629±0.10149]	SOM[0.13122±0.060476]
		Todas	SOM[0.16001±0.076979]	KNN[0.017824±0.017722]	SOM[0.14958±0.0783]	SOM[0.16531±0.095126]	KNN[0.031544±0.018449]	SOM[0.11819±0.080841]	SOM[0.074689±0.054946]
	r^2	5-10%	KNN[0.70029±0.25178]	KNN[0.26867±0.22069]	KNN[0.61401±0.29025]	SOM[0.6883±0.23475]	KNN[0.24203±0.17301]	KNN[0.6073±0.35883]	SOM[0.61962±0.24525]
		15-20%	SOM[0.53764±0.29502]	KNN[0.28201±0.22042]	SOM[0.58593±0.24679]	SOM[0.54801±0.29065]	KNN[0.21995±0.10652]	SOM[0.58878±0.26125]	SOM[0.55472±0.25537]
		25%	SOM[0.54262±0.10406]	SOM[0.28262±0.2885]	SOM[0.5119±0.27536]	SOM[0.56831±0.15544]	SOM[0.34373±0.21208]	SOM[0.68728±0.095103]	SOM[0.52208±0.25557]
		Todas	SOM[0.56553±0.2878]	KNN[0.25429±0.20317]	KNN[0.57241±0.25706]	SOM[0.59658±0.24706]	KNN[0.21691±0.14636]	SOM[0.66973±0.21902]	SOM[0.5709±0.24543]
	D_{KS}	5-10%	KNN[0.55316±0.18834]	SOM[0.34274±0.089299]	KNN[0.31444±0.10283]	SOM[0.49972±0.21297]	KNN[0.31103±0.15911]	KNN[0.28151±0.082488]	KNN[0.16629±0.05651]
		15-20%	SOM[0.33518±0.087746]	SOM[0.36528±0.11124]	SOM[0.27649±0.028831]	KNN[0.30649±0.13619]	SOM[0.3127±0.13042]	KNN[0.19311±0.062219]	KNN[0.14958±0.041089]
		25%	SOM[0.32265±0.095683]	SOM[0.35918±0.12032]	SOM[0.26614±0.052122]	SOM[0.23986±0.039366]	SOM[0.20724±0.10274]	SOM[0.17503±0.047405]	KNN[0.14974±0.036365]
		Todas	SOM[0.36807±0.12415]	SOM[0.35516±0.09923]	SOM[0.2846±0.053584]	KNN[0.35774±0.136]	SOM[0.31216±0.1451]	KNN[0.23411±0.079957]	KNN[0.15629±0.046352]
Log-Logistica	MSE	5-10%	KNN[0.058375±0.097127]	KNN[0.0027298±0.004944]	KNN[0.051594±0.081962]	KNN[0.056737±0.089122]	KNN[0.004423±0.0046965]	KNN[0.033279±0.046357]	DT[0.029667±0.024278]
		15-20%	SOM[0.22754±0.092706]	KNN[0.014946±0.012972]	SOM[0.19239±0.11136]	SOM[0.17582±0.099029]	KNN[0.023196±0.014313]	SOM[0.16764±0.10875]	SOM[0.083893±0.044989]
		25%	SOM[0.23067±0.080653]	KNN[0.037983±0.020412]	SOM[0.18437±0.10059]	SOM[0.17763±0.089864]	KNN[0.048814±0.017073]	SOM[0.19027±0.089282]	SOM[0.10895±0.054652]
		Todas	SOM[0.208±0.085634]	KNN[0.014681±0.017733]	SOM[0.17366±0.097326]	SOM[0.17084±0.095452]	KNN[0.020994±0.019949]	SOM[0.16026±0.093808]	DT[0.070185±0.054418]
	r^2	5-10%	KNN[0.61331±0.31426]	KNN[0.2397±0.13971]	KNN[0.76432±0.20732]	KNN[0.70898±0.19231]	KNN[0.32687±0.32667]	KNN[0.79687±0.21899]	KNN[0.75841±0.22748]
		15-20%	SOM[0.6019±0.2552]	KNN[0.38639±0.19044]	KNN[0.65113±0.23167]	SOM[0.52538±0.2847]	KNN[0.68409±0.35536]	SOM[0.52828±0.27258]	KNN[0.66621±0.23999]
		25%	SOM[0.53328±0.24893]	KNN[0.46098±0.1659]	SOM[0.56769±0.16375]	SOM[0.45302±0.24809]	KNN[0.67106±0.29456]	SOM[0.52671±0.25561]	SOM[0.59128±0.23163]
		Todas	SOM[0.55163±0.30134]	KNN[0.34714±0.18689]	SOM[0.53447±0.21613]	SOM[0.54819±0.2668]	KNN[0.53281±0.37045]	SOM[0.56627±0.25201]	KNN[0.6874±0.24167]
	D_{KS}	5-10%	KNN[0.37953±0.20738]	KNN[0.17997±0.069783]	KNN[0.27996±0.070832]	KNN[0.34468±0.20738]	KNN[0.25632±0.11982]	KNN[0.22307±0.099308]	KNN[0.11516±0.079565]
		15-20%	KNN[0.25566±0.14476]	KNN[0.16409±0.042624]	KNN[0.21884±0.064022]	KNN[0.21174±0.11209]	KNN[0.20193±0.097132]	KNN[0.20309±0.10643]	KNN[0.10649±0.06975]
		25%	KNN[0.20816±0.10337]	KNN[0.12741±0.027172]	KNN[0.20227±0.038381]	KNN[0.15171±0.044841]	KNN[0.19256±0.099547]	KNN[0.18094±0.073059]	KNN[0.11194±0.065154]
		Todas	KNN[0.29907±0.18056]	KNN[0.16355±0.057868]	KNN[0.24229±0.071057]	KNN[0.26963±0.17837]	KNN[0.22312±0.1099]	KNN[0.2077±0.098146]	KNN[0.11105±0.072265]
Normal	MSE	5-10%	KNN[0.19796±0.080098]	KNN[0.0051879±0.0043681]	SOM[0.14978±0.065193]	KNN[0.19085±0.066874]	KNN[0.00588±0.0055186]	KNN[0.15743±0.079126]	SOM[0.04353±0.018913]
		15-20%	SOM[0.30186±0.04751]	KNN[0.018405±0.0083098]	SOM[0.23758±0.075667]	SOM[0.25304±0.079388]	KNN[0.021831±0.014262]	DT[0.18696±0.030601]	SOM[0.10553±0.024674]
		25%	SOM[0.31509±0.082302]	KNN[0.034257±0.0079697]	SOM[0.21868±0.071822]	SOM[0.2857±0.077598]	KNN[0.043462±0.017891]	SOM[0.28844±0.075417]	SOM[0.16862±0.041144]
		Todas	SOM[0.28419±0.069167]	KNN[0.017522±0.012838]	SOM[0.20283±0.07929]	SOM[0.24266±0.086317]	KNN[0.019716±0.018621]	SOM[0.20843±0.076244]	SOM[0.096923±0.055113]
	r^2	5-10%	KNN[0.53761±0.24478]	KNN[0.087662±0.10638]	SOM[0.63251±0.14033]	KNN[0.52333±0.21704]	KNN[0.17647±0.22938]	KNN[0.56027±0.15746]	KNN[0.5113±0.14248]
		15-20%	SOM[0.43241±0.10766]	SOM[0.076246±0.056041]	SOM[0.50918±0.1592]	KNN[0.40066±0.15901]	KNN[0.10136±0.074234]	SOM[0.37395±0.13832]	KNN[0.3986±0.15025]
		25%	SOM[0.39855±0.15047]	KNN[0.073867±0.031859]	SOM[0.38364±0.14318]	SOM[0.39765±0.15794]	SOM[0.15785±0.10411]	SOM[0.34119±0.088436]	SOM[0.3259±0.11861]
		Todas	SOM[0.43801±0.12789]	KNN[0.073624±0.080666]	SOM[0.5203±0.17408]	SOM[0.45465±0.13302]	SOM[0.13906±0.10937]	SOM[0.42886±0.14955]	KNN[0.43021±0.16279]
	D_{KS}	5-10%	KNN[0.49615±0.14918]	SOM[0.39629±0.092725]	KNN[0.27485±0.082922]	KNN[0.41167±0.07281]	KNN[0.29568±0.14201]	KNN[0.2556±0.05031]	KNN[0.12476±0.065577]
		15-20%	SOM[0.43498±0.078611]	KNN[0.25521±0.10047]	KNN[0.22649±0.02644]	KNN[0.33008±0.058918]	KNN[0.23411±0.090347]	KNN[0.22972±0.04348]	KNN[0.11304±0.059516]
		25%	SOM[0.33902±0.045861]	SOM[0.20249±0.072087]	SOM[0.19179±0.027843]	KNN[0.27252±0.052247]	SOM[0.22925±0.10404]	SOM[0.17269±0.025815]	KNN[0.11157±0.061786]
		Todas	KNN[0.43194±0.12644]	SOM[0.3283±0.10897]	KNN[0.24658±0.065403]	KNN[0.35664±0.084222]	KNN[0.26352±0.11698]	KNN[0.23674±0.051441]	KNN[0.11743±0.061327]
Nakagami	MSE	5-10%	KNN[0.074144±0.069191]	KNN[0.0021901±0.0018821]	KNN[0.058609±0.053812]	KNN[0.046442±0.023212]	KNN[0.029159±0.036364]	KNN[0.066958±0.053881]	DT[0.031848±0.026626]
		15-20%	KNN[0.14394±0.035063]	KNN[0.014973±0.0093485]	SOM[0.13158±0.06466]	SOM[0.12226±0.046825]	KNN[0.055943±0.063791]	SOM[0.092767±0.03244]	DT[0.084173±0.039873]
		25%	SOM[0.17257±0.060297]	KNN[0.040122±0.015429]	SOM[0.11973±0.019687]	SOM[0.17183±0.047606]	KNN[0.10064±0.11499]	SOM[0.15173±0.071929]	SOM[0.12877±0.061526]
		Todas	KNN[0.11279±0.10398]	KNN[0.015768±0.016563]	SOM[0.12383±0.054811]	KNN[0.079173±0.057877]	KNN[0.051723±0.0648]	KNN[0.10288±0.078258]	DT[0.069049±0.050767]
	r^2	5-10%	SOM[0.67899±0.31335]	KNN[0.19786±0.17034]	KNN[0.65299±0.28476]	KNN[0.54376±0.31659]	MM[0.76335±0.40743]	KNN[0.72141±0.15903]	KNN[0.63948±0.26847]
		15-20%	SOM[0.67629±0.25402]	KNN[0.2476±0.15359]	KNN[0.55491±0.24486]	SOM[0.63153±0.34355]	KNN[0.1944±0.13951]	KNN[0.62407±0.29525]	KNN[0.54766±0.26007]
		25%	KNN[0.64405±0.08992]	KNN[0.20481±0.18339]	KNN[0.58434±0.065367]	KNN[0.58985±0.074843]	MM[0.6106±0.5507]	SOM[0.3763±0.38325]	SOM[0.46283±0.30281]
		Todas	SOM[0.64812±0.30121]	KNN[0.2186±0.15401]	KNN[0.60992±0.24942]	KNN[0.57473±0.23756]	KNN[0.25424±0.13444]	KNN[0.68593±0.2008]	KNN[0.57772±0.26485]
	D_{KS}	5-10%	KNN[0.36813±0.19201]	KNN[0.19547±0.089799]	KNN[0.24849±0.11656]	KNN[0.29651±0.1431]	KNN[0.33546±0.16993]	KNN[0.30391±0.15641]	KNN[0.11214±0.081326]
		15-20%	KNN[0.36505±0.15219]	SOM[0.16395±0.074144]	KNN[0.19269±0.070964]	KNN[0.23886±0.13045]	KNN[0.23872±0.088472]	KNN[0.23049±0.15469]	KNN[0.10549±0.061385]
		25%	KNN[0.33406±0.14964]	SOM[0.18423±0.065444]	KNN[0.15712±0.077283]	KNN[0.26325±0.09656]	KNN[0.17995±0.048675]	KNN[0.20226±0.063552]	KNN[0.11168±0.056446]
		Todas	KNN[0.36144±0.16615]	KNN[0.16962±0.086652]	KNN[0.21595±0.099223]	KNN[0.27016±0.12901]	KNN[0.27433±0.13976]	KNN[0.2579±0.14532]	KNN[0.10939±0.067397]

Continua na página seguinte...

Tabela A.6: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Log-Normal	MSE	5-10%	KNN[0.096807±0.063624]	KNN[0.0044026±0.0032892]	KNN[0.083862±0.04554]	KNN[0.11451±0.067799]	DT[0.0086468±0.0060818]	KNN[0.074937±0.066037]	DT[0.028171±0.023468]
		15-20%	SOM[0.20819±0.095043]	KNN[0.018813±0.00793]	SOM[0.1513±0.080521]	DT[0.095055±0.06164]	KNN[0.03526±0.012336]	SOM[0.16507±0.08418]	DT[0.095912±0.041305]
		25%	SOM[0.24573±0.092469]	KNN[0.042796±0.0047218]	SOM[0.22353±0.071786]	SOM[0.26123±0.015421]	KNN[0.055836±0.0041939]	SOM[0.17784±0.069329]	SOM[0.14782±0.052548]
		Todas	SOM[0.22321±0.090798]	KNN[0.016406±0.014475]	SOM[0.18285±0.078323]	KNN[0.18215±0.09641]	DT[0.013271±0.0090994]	SOM[0.17126±0.07073]	DT[0.07244±0.058264]
	r^2	5-10%	KNN[0.6154±0.16383]	SOM[0.23086±0.1774]	KNN[0.64435±0.20193]	KNN[0.67237±0.18563]	KNN[0.13593±0.1579]	KNN[0.6873±0.22252]	KNN[0.69027±0.20504]
		15-20%	SOM[0.60776±0.12259]	KNN[0.2096±0.20243]	KNN[0.51284±0.17464]	KNN[0.55473±0.15817]	KNN[0.25576±0.27514]	SOM[0.77576±0.16909]	KNN[0.55385±0.21228]
		25%	KNN[0.4804±0.179]	KNN[0.26447±0.19387]	SOM[0.52124±0.26789]	DT[0.55749±0.18803]	DT[0.17038±0.076754]	SOM[0.61095±0.21368]	SOM[0.49023±0.18889]
		Todas	KNN[0.5527±0.16749]	KNN[0.18375±0.18032]	KNN[0.55374±0.19726]	KNN[0.56153±0.19144]	KNN[0.20201±0.21746]	SOM[0.67078±0.16456]	KNN[0.58721±0.22069]
	D_{KS}	5-10%	KNN[0.41095±0.078661]	KNN[0.35274±0.063325]	KNN[0.26033±0.067774]	KNN[0.41419±0.12224]	SOM[0.32688±0.12925]	KNN[0.23248±0.042803]	KNN[0.089894±0.043876]
		15-20%	KNN[0.34857±0.053337]	KNN[0.25013±0.063825]	KNN[0.20999±0.028879]	DT[0.29587±0.099355]	SOM[0.24583±0.068421]	KNN[0.22909±0.043475]	KNN[0.083732±0.036008]
		25%	SOM[0.30013±0.064635]	SOM[0.18488±0.059985]	KNN[0.19663±0.041729]	KNN[0.24377±0.036715]	SOM[0.17012±0.019519]	KNN[0.17729±0.0047481]	KNN[0.083537±0.037072]
		Todas	KNN[0.37605±0.069198]	SOM[0.26661±0.11238]	KNN[0.22837±0.056247]	KNN[0.34313±0.10813]	SOM[0.25662±0.10402]	KNN[0.21943±0.043033]	KNN[0.086158±0.038322]
Rayleigh	MSE	5-10%	DT[0.12886±0.066762]	KNN[0.01149±0.0093595]	SOM[0.051878±0.042]	KNN[0.11174±0.079967]	SOM[0.00487±0.0051132]	DT[0.067942±0.054183]	DT[0.026594±0.028764]
		15-20%	SOM[0.16919±0.15812]	MM[0.016257±0.027675]	SOM[0.2015±0.14248]	SOM[0.093466±0.084382]	SOM[0.044343±0.032367]	SOM[0.11296±0.10359]	SOM[0.068828±0.051688]
		25%	SOM[0.19458±0.14522]	SOM[0.0043283±0]	SOM[0.18993±0.089619]	SOM[0.17849±0.20893]	SOM[0.037919±0.046549]	SOM[0.057392±0]	SOM[0.10904±0.074506]
		Todas	SOM[0.16588±0.1316]	MM[0.024807±0.034988]	SOM[0.14905±0.11997]	SOM[0.11724±0.12946]	SOM[0.029758±0.032024]	SOM[0.091233±0.093116]	DT[0.053814±0.046421]
	r^2	5-10%	KNN[0.51695±0.28308]	MM[0.39337±0.12939]	KNN[0.58517±0.28236]	KNN[0.56944±0.24553]	MM[0.26044±0.058457]	KNN[0.62415±0.20583]	KNN[0.73021±0.22686]
		15-20%	SOM[0.44362±0.1966]	DT[0.21943±0.025017]	MM[0.89519±0]	SOM[0.5607±0.26138]	SOM[0.13672±0.041769]	SOM[0.58618±0.32587]	SOM[0.63959±0.28861]
		25%	SOM[0.46588±0.25821]	DT[0.30277±0]	SOM[0.59458±0.21153]	SOM[0.48211±0.29491]	SOM[0.26528±0.04575]	SOM[0.71904±0.15083]	SOM[0.59213±0.30249]
		Todas	SOM[0.48134±0.25623]	MM[0.29097±0.12789]	KNN[0.58041±0.27674]	KNN[0.56565±0.23773]	SOM[0.17647±0.071541]	SOM[0.65787±0.25703]	DT[0.64378±0.25915]
	D_{KS}	5-10%	SOM[0.34532±0.046267]	SOM[0.38147±0.067512]	KNN[0.38148±0.10364]	SOM[0.31711±0.19777]	SOM[0.38187±0.14776]	DT[0.34286±0.14252]	DT[0.17401±0.045396]
		15-20%	MM[0.28186±0.099831]	SOM[0.34813±0.10507]	KNN[0.29388±0.072815]	MM[0.16422±0.045064]	SOM[0.25389±0.16477]	MM[0.17375±0.019445]	DT[0.16953±0.030972]
		25%	SOM[0.37963±0.17023]	SOM[0.33933±0.14392]	SOM[0.25461±0.088098]	KNN[0.29773±0.10679]	SOM[0.25±0.18424]	MM[0.10714±0]	SOM[0.19127±0.019963]
		Todas	MM[0.26206±0.099372]	SOM[0.35482±0.098882]	KNN[0.34644±0.098899]	MM[0.24654±0.16981]	SOM[0.30431±0.15671]	MM[0.22034±0.15899]	DT[0.17293±0.035819]
t Location-scale	MSE	5-10%	KNN[0.049792±0.020202]	DT[0.0030724±0]	DT[0.012625±0]	SOM[0.023999±0]	MM[0.00025394±0]	DT[0.0074508±0]	SOM[0.011432±0.0065786]
		15-20%	SOM[0.12471±0.015293]	DT[0.015602±0]	SOM[0.069435±0.029169]	SOM[0.086168±0.055734]	KNN[0.018484±0.004054]	SOM[0.058464±0.019502]	SOM[0.045821±0.017799]
		25%	SOM[0.11221±0.045803]	SOM[0.0413±0]	SOM[0.11405±0.041306]	SOM[0.12382±0.03153]	SOM[0.038277±0]	SOM[0.1122±0.047357]	SOM[0.070807±0.025519]
		Todas	KNN[0.068844±0.033734]	SOM[0.025071±0.016268]	KNN[0.051201±0.030578]	KNN[0.066259±0.039872]	KNN[0.021867±0.01309]	KNN[0.061255±0.035966]	KNN[0.033177±0.026051]
	r^2	5-10%	KNN[0.83743±0.063008]	KNN[0.21621±0.087309]	KNN[0.8552±0.071131]	KNN[0.74064±0.2155]	MM[1±0]	KNN[0.85383±0.06712]	KNN[0.82753±0.070125]
		15-20%	SOM[0.76866±0.017098]	DT[0.32765±0]	SOM[0.81971±0.049992]	SOM[0.83758±0.025702]	MM[0.36936±0]	SOM[0.80078±0.017218]	SOM[0.8151±0.03042]
		25%	SOM[0.81836±0]	KNN[0.44686±0]	MM[0.64475±0.2241]	KNN[0.74258±0]	MM[0.55671±0]	SOM[0.76522±0]	SOM[0.79478±0]
		Todas	SOM[0.77378±0.081995]	SOM[0.26235±0.08276]	KNN[0.80547±0.1045]	SOM[0.82768±0.08797]	KNN[0.23469±0.14057]	KNN[0.82437±0.077433]	SOM[0.83714±0.047303]
	D_{KS}	5-10%	KNN[0.43949±0.21254]	KNN[0.2155±0.053923]	KNN[0.2454±0.10164]	KNN[0.3285±0.12603]	SOM[0.19117±0.020796]	KNN[0.2256±0.050108]	KNN[0.077514±0.046132]
		15-20%	KNN[0.32156±0.0114]	SOM[0.20343±0.017331]	KNN[0.16717±0.013236]	SOM[0.19608±0.027733]	SOM[0.19363±0.0034648]	KNN[0.20373±0.019084]	KNN[0.071637±0.031057]
		25%	MM[0.21176±0]	SOM[0.20334±0.13783]	SOM[0.16667±0]	SOM[0.18824±0]	SOM[0.11765±0]	SOM[0.15476±0]	KNN[0.081958±0.032482]
		Todas	KNN[0.38944±0.17137]	SOM[0.22873±0.074077]	KNN[0.2047±0.078961]	SOM[0.19347±0.020126]	SOM[0.17745±0.035074]	KNN[0.21021±0.03858]	KNN[0.076052±0.035665]
Weibull	MSE	5-10%	SOM[0.1784±0.10292]	KNN[0.0050438±0.0024482]	SOM[0.13444±0.069267]	SOM[0.15881±0.081248]	SOM[0.0073923±0.0055479]	SOM[0.11589±0.056405]	SOM[0.041462±0.020679]
		15-20%	SOM[0.24337±0.12028]	KNN[0.022734±0.0099898]	SOM[0.19424±0.10873]	SOM[0.20802±0.087099]	SOM[0.022241±0.0090073]	SOM[0.16677±0.073613]	SOM[0.10327±0.042893]
		25%	SOM[0.28053±0.12525]	SOM[0.041129±0.012089]	SOM[0.2176±0.087602]	SOM[0.23355±0.11254]	SOM[0.044181±0.012351]	SOM[0.1905±0.085556]	SOM[0.1532±0.055469]
		Todas	SOM[0.23019±0.12059]	KNN[0.020284±0.019258]	SOM[0.18192±0.097296]	SOM[0.19547±0.09435]	SOM[0.017436±0.015518]	SOM[0.15535±0.076236]	SOM[0.090641±0.05769]
	r^2	5-10%	KNN[0.67768±0.23303]	DT[0.34728±0.13974]	KNN[0.73485±0.22847]	KNN[0.68098±0.19665]	KNN[0.28489±0.24186]	SOM[0.58747±0.20912]	KNN[0.56846±0.18952]
		15-20%	SOM[0.48995±0.25551]	KNN[0.10131±0.10922]	SOM[0.53391±0.18747]	SOM[0.48879±0.18234]	SOM[0.33256±0.25622]	SOM[0.53146±0.21941]	SOM[0.43696±0.19194]
		25%	SOM[0.46796±0.24024]	SOM[0.252±0.21566]	SOM[0.48059±0.2107]	SOM[0.4568±0.21188]	SOM[0.43957±0.30679]	SOM[0.43421±0.22608]	SOM[0.41666±0.20816]
		Todas	SOM[0.52894±0.24027]	DT[0.32925±0.19459]	SOM[0.55861±0.20053]	SOM[0.50487±0.20098]	SOM[0.34562±0.23612]	SOM[0.52144±0.22185]	SOM[0.46684±0.18576]
	D_{KS}	5-10%	SOM[0.56697±0.16915]	SOM[0.36916±0.11758]	KNN[0.36277±0.087945]	SOM[0.54745±0.15595]	SOM[0.33859±0.11208]	DT[0.36708±0.069745]	KNN[0.19983±0.061118]
		15-20%	SOM[0.46267±0.10747]	SOM[0.26239±0.06349]	SOM[0.28171±0.077781]	SOM[0.44374±0.11371]	SOM[0.23636±0.092481]	SOM[0.2782±0.061985]	KNN[0.18848±0.052554]
		25%	SOM[0.38619±0.075149]	SOM[0.25789±0.085089]	SOM[0.23304±0.043799]	SOM[0.38756±0.11014]	SOM[0.18594±0.053646]	SOM[0.22152±0.058612]	SOM[0.18806±0.054252]
		Todas	SOM[0.47874±0.14149]	SOM[0.29606±0.10185]	SOM[0.30882±0.09133]	SOM[0.4639±0.14054]	SOM[0.27163±0.1132]	SOM[0.29824±0.091923]	KNN[0.19414±0.055591]

Tabela A.7: Técnicas de imputação vencedoras com a sua média para cada distribuição, métrica, MR e estratégia (incluindo SVM).
O código de cor contêm informação dos empates: vermelho corresponde ao empate de todos os métodos, laranja entre 4 métodos, amarelo entre 3 métodos e verde entre 2 métodos.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Beta	MSE	5-10%	SVM[0.053711±0.055208]	SVM[0.039642±0.059119]	SVM[0.040275±0.053098]	SVM[0.041078±0.040139]	SVM[0.036624±0.055852]	SVM[0.04054±0.030474]	SVM[0.014156±0.017889]
		15-20%	SVM[0.07428±0.06627]	SVM[0.062585±0.092194]	SVM[0.16412±0.14254]	SVM[0.10897±0.080428]	SVM[0.069255±0.086599]	SVM[0.14941±0.14969]	SVM[0.064507±0.066383]
		25%	MM[0.038902±0]	SVM[0.074861±0.080488]	SVM[0.27453±0.2312]	MM[0.055141±0.017437]	SVM[0.10199±0.10952]	MM[0.053197±0.01902]	SOM[0.10225±0.10391]
		Todas	SVM[0.059693±0.054636]	SVM[0.057665±0.076594]	SVM[0.11664±0.14135]	SVM[0.076195±0.06865]	SVM[0.064931±0.081053]	SVM[0.083096±0.092313]	SVM[0.049032±0.064667]
	r^2	5-10%	KNN[0.79855±0.37056]	SVM[0.68419±0.2878]	SVM[0.79207±0.28343]	SVM[0.60541±0.41302]	SVM[0.51448±0.26098]	SVM[0.87503±0.12195]	SVM[0.77901±0.27828]
		15-20%	SVM[0.55874±0.47849]	SVM[0.61771±0.40033]	SOM[0.62649±0.38141]	SVM[0.38565±0.45346]	SVM[0.39621±0.30732]	KNN[0.27198±0.29669]	SVM[0.6554±0.34812]
		25%	SVM[0.91288±0]	SVM[0.54392±0.068377]	SVM[0.39624±0.41441]	SOM[0.36497±0.35249]	SVM[0.59394±0.26174]	SOM[0.48613±0.43868]	SVM[0.65073±0.35868]
		Todas	SVM[0.57677±0.42639]	SVM[0.63865±0.29104]	SVM[0.59761±0.3568]	MM[0.76571±0.17243]	SVM[0.46988±0.2747]	KNN[0.45974±0.36979]	SVM[0.70593±0.31434]
	D_{KS}	5-10%	SVM[0.46559±0.26641]	SVM[0.34751±0.24823]	SVM[0.25665±0.075384]	SVM[0.25193±0.13304]	SVM[0.32982±0.1886]	SVM[0.21366±0.047963]	SVM[0.14388±0.044212]
		15-20%	SVM[0.3661±0.19133]	SVM[0.32387±0.24268]	SVM[0.22159±0.059914]	SVM[0.17288±0.062311]	SVM[0.24163±0.046259]	SVM[0.16497±0.02881]	SVM[0.12351±0.028582]
		25%	SVM[0.34221±0.24824]	SVM[0.36891±0.34855]	SVM[0.2326±0.057215]	SVM[0.19186±0.078577]	SVM[0.23686±0.13237]	MM[0.14881±0.0084146]	SVM[0.11843±0.031255]
		Todas	SVM[0.40634±0.229]	SVM[0.33984±0.23783]	SVM[0.23809±0.065182]	SVM[0.21037±0.10379]	SVM[0.28123±0.14308]	SVM[0.19536±0.04447]	SVM[0.13064±0.036433]
Birnbaum-Saunders	MSE	5-10%	SVM[0.022695±0.042878]	SVM[0.0024046±0.0032563]	SVM[0.024324±0.023564]	SVM[0.0097367±0.022724]	SVM[0.0025744±0.0029604]	SVM[0.0092506±0.016826]	SVM[0.0010324±0.0048761]
		15-20%	SVM[0.034154±0.058964]	SVM[0.022509±0.020239]	SVM[0.034969±0.056163]	SVM[0.011604±0.049546]	SVM[0.018808±0.015864]	SVM[0.015677±0.03957]	SVM[0.005498±0.020125]
		25%	SVM[0.096635±0.12241]	SVM[0.037228±0.022173]	SVM[0.053888±0.075605]	SVM[0.013425±0.046404]	SVM[0.024384±0.018781]	SVM[0.0090514±0.032082]	SVM[0.0091299±0.03223]
		Todas	SVM[0.032876±0.061245]	SVM[0.015085±0.019795]	SVM[0.033332±0.050168]	SVM[0.011219±0.040648]	SVM[0.012966±0.015563]	SVM[0.011807±0.03091]	SVM[0.0044381±0.019663]
	r^2	5-10%	SVM[0.92797±0.13821]	SVM[0.20204±0.092409]	SVM[0.97246±0.061304]	SVM[0.93832±0.1379]	SVM[0.50387±0.32079]	SVM[0.97321±0.072469]	SVM[0.98259±0.082547]
		15-20%	SVM[0.94772±0.14956]	SVM[0.11589±0.10099]	SVM[0.96128±0.1078]	SVM[0.96407±0.13453]	SVM[0.51109±0.33229]	SVM[0.96518±0.10951]	SVM[0.97131±0.1112]
		25%	SVM[0.95109±0.1377]	SVM[0.35534±0.25842]	SVM[0.95736±0.12754]	SVM[0.96664±0.11659]	SVM[0.58721±0.29839]	SVM[0.96496±0.11952]	SVM[0.96632±0.12103]
		Todas	SVM[0.93992±0.14261]	SVM[0.15803±0.11396]	SVM[0.96513±0.096266]	SVM[0.95361±0.13357]	SVM[0.52457±0.31084]	SVM[0.96846±0.098066]	SVM[0.97483±0.10295]
	D_{KS}	5-10%	SVM[0.15903±0.10337]	SVM[0.68875±0.14843]	SVM[0.32515±0.040041]	SVM[0.15201±0.07069]	SVM[0.25243±0.037617]	SVM[0.14321±0.029736]	SVM[0.16556±0.021113]
		15-20%	SVM[0.24406±0.13491]	SVM[0.69619±0.11519]	SVM[0.33928±0.057448]	SVM[0.19603±0.062083]	SVM[0.27617±0.033912]	SVM[0.1319±0.021661]	SVM[0.17649±0.025532]
		25%	SVM[0.20393±0.067284]	SVM[0.60416±0.10252]	SVM[0.3415±0.048237]	SVM[0.19393±0.026639]	SVM[0.28624±0.034678]	SVM[0.14617±0.049275]	SVM[0.17988±0.026953]
		Todas	SVM[0.1732±0.11051]	SVM[0.67171±0.12784]	SVM[0.33276±0.047188]	SVM[0.17276±0.062565]	SVM[0.26862±0.038111]	SVM[0.13922±0.03255]	SVM[0.1728±0.024868]
Exponencial	MSE	5-10%	SVM[0.11465±0.04893]	SVM[0.017191±0.011709]	SVM[0.085708±0.0552]	SVM[0.11041±0.059706]	MM[2.5855e-06±1.3912e-06]	SVM[0.098668±0.076738]	SVM[0.023347±0.015524]
		15-20%	SVM[0.10209±0.037535]	SVM[0.061456±0.053025]	MM[0.14311±0.016504]	MM[0.14109±0.031424]	SVM[0.045012±0.045573]	SVM[0.15652±0.0083934]	SVM[0.092584±0.049515]
		25%	SVM[0.16915±0]	SVM[0.032892±0]	MM[0.17848±0]	MM[0.18433±0]	SVM[0.074595±0.055219]	SVM[0.13975±0]	SVM[0.14407±0.08234]
		Todas	SVM[0.11955±0.043218]	SVM[0.041286±0.04166]	MM[0.1549±0.02352]	MM[0.1555±0.033421]	SVM[0.046312±0.041861]	SVM[0.13002±0.048538]	SVM[0.075185±0.063515]
	r^2	5-10%	SVM[0.42622±0.088755]	SVM[0.44356±0.25108]	SVM[0.59684±0.34052]	SVM[0.4111±0.21072]	SVM[0.70669±0.50803]	KNN[0.63482±0.12566]	SVM[0.61284±0.23085]
		15-20%	SVM[0.23416±0.19725]	SVM[0.33661±0.33812]	SVM[0.52238±0.32004]	KNN[0.34635±0.21331]	SVM[0.37667±0.042752]	SVM[0.30376±0.10809]	SVM[0.49293±0.25493]
		25%	MM[0.37543±0]	SVM[0.32243±0.4039]	SVM[0.60553±0]	SVM[0.40235±0.2988]	SVM[0.63399±0]	SVM[0.49047±0]	SVM[0.4315±0.3223]
		Todas	SVM[0.33019±0.17257]	SVM[0.3598±0.28986]	SVM[0.56106±0.25684]	SVM[0.4076±0.21106]	SVM[0.58457±0.36101]	SVM[0.41527±0.18274]	SVM[0.52861±0.23829]
	D_{KS}	5-10%	KNN[0.33±0.014142]	SVM[0.50494±0.23033]	KNN[0.29917±0.11995]	SVM[0.30741±0.030593]	KNN[0.40596±0.26222]	KNN[0.29034±0.10423]	KNN[0.11426±0.079954]
		15-20%	KNN[0.22567±0.068356]	KNN[0.24634±0.067408]	SVM[0.26793±0.019361]	SVM[0.27718±0.038431]	KNN[0.29371±0.09918]	KNN[0.22085±0.010354]	KNN[0.097232±0.050728]
		25%	KNN[0.3088±0]	SVM[0.2032±0]	SVM[0.26612±0.0018668]	SVM[0.31466±0.097666]	KNN[0.2224±0]	KNN[0.22394±0.077973]	KNN[0.092541±0.042891]
		Todas	KNN[0.28403±0.064276]	KNN[0.28426±0.1478]	SVM[0.29538±0.064454]	KNN[0.28115±0.027971]	KNN[0.33163±0.17796]	KNN[0.25242±0.078463]	KNN[0.1031±0.057345]
Valor Extremo	MSE	5-10%	SVM[0.092281±0.047232]	SVM[0.0018952±0.0013598]	SVM[0.05277±0.033827]	SVM[0.067498±0.042523]	SVM[0.0027528±0.0023948]	SVM[0.048663±0.028763]	SVM[0.010891±0.0055081]
		15-20%	KNN[0.017697±0.0096732]	SVM[0.013399±0.0075878]	SVM[0.12166±0.056988]	SVM[0.12721±0.060841]	SVM[0.017408±0.009567]	SVM[0.096587±0.051001]	SVM[0.052606±0.024059]
		25%	SVM[0.17734±0.069667]	SVM[0.026934±0.013144]	SVM[0.15283±0.055212]	SVM[0.14178±0.055302]	SVM[0.033196±0.016623]	SVM[0.1282±0.054549]	SVM[0.088996±0.027102]
		Todas	SVM[0.13408±0.070016]	SVM[0.011849±0.012136]	SVM[0.099988±0.063043]	SVM[0.10566±0.061659]	SVM[0.015006±0.015141]	SVM[0.083643±0.053804]	SVM[0.04291±0.03539]
	r^2	5-10%	SVM[0.7186±0.14381]	SVM[0.33027±0.13534]	SVM[0.87265±0.082545]	SVM[0.76899±0.13631]	SVM[0.36159±0.17063]	KNN[0.91413±0.045934]	SVM[0.81291±0.085603]
		15-20%	MM[0.98425±0.0021637]	SVM[0.20785±0.12936]	SVM[0.73633±0.12199]	SVM[0.67026±0.13]	KNN[0.34271±0.3691]	SVM[0.73232±0.11796]	SVM[0.7146±0.10922]
		25%	MM[0.9418±0]	SVM[0.22736±0.1425]	SVM[0.66743±0.11897]	SVM[0.63746±0.12999]	SVM[0.25954±0.17315]	SVM[0.68759±0.092232]	SVM[0.66319±0.10739]
		Todas	MM[0.9474±0.049626]	SVM[0.25901±0.14518]	SVM[0.76796±0.13547]	SVM[0.69871±0.14229]	SVM[0.30808±0.17273]	KNN[0.91601±0.043336]	SVM[0.74415±0.11586]
	D_{KS}	5-10%	SVM[0.35917±0.14126]	SVM[0.302±0.11076]	SVM[0.2529±0.079118]	SVM[0.3047±0.1448]	SVM[0.25031±0.08876]	SVM[0.21633±0.069627]	SVM[0.14671±0.039138]
		15-20%	SVM[0.32167±0.082013]	SVM[0.24786±0.06765]	SVM[0.22643±0.045349]	SVM[0.27647±0.088596]	SVM[0.24064±0.084055]	SVM[0.19865±0.04355]	SVM[0.12434±0.020035]
		25%	SVM[0.26497±0.059909]	SVM[0.18971±0.056148]	SVM[0.18081±0.043548]	SVM[0.20537±0.060021]	SVM[0.18666±0.051668]	KNN[0.17764±0.049062]	KNN[0.11973±0.021465]
		Todas	SVM[0.32551±0.1129]	SVM[0.26035±0.095953]	SVM[0.22945±0.066823]	SVM[0.27531±0.11738]	SVM[0.23487±0.08411]	SVM[0.20048±0.059219]	SVM[0.13177±0.031191]

Continua na página seguinte...

Tabela A.7: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Gama	MSE	5-10%	SVM[0.048714±0.036515]	SVM[0.0035996±0.0031917]	SVM[0.031509±0.024183]	SVM[0.052487±0.038512]	SVM[0.0016591±0.0026924]	SVM[0.033654±0.01837]	SVM[0.0090895±0.0061856]
		15-20%	SVM[0.11468±0.064701]	SVM[0.017291±0.013365]	SVM[0.086787±0.049072]	SVM[0.096943±0.068236]	SVM[0.011891±0.011998]	SVM[0.087641±0.045669]	SVM[0.04777±0.026705]
		25%	SVM[0.13645±0.057553]	SVM[0.030537±0.019807]	SVM[0.13329±0.073872]	SVM[0.12325±0.073601]	SVM[0.022554±0.012321]	SVM[0.11649±0.066682]	SVM[0.078216±0.037596]
		Todas	SVM[0.0899±0.063863]	SVM[0.014608±0.016021]	SVM[0.070511±0.059096]	SVM[0.083407±0.064574]	SVM[0.0097434±0.012152]	SVM[0.069739±0.052203]	SVM[0.037682±0.035213]
	r^2	5-10%	KNN[0.83986±0.11003]	SVM[0.49113±0.2723]	SVM[0.7967±0.1367]	SVM[0.62075±0.23074]	SVM[0.6946±0.30583]	KNN[0.85468±0.11453]	SVM[0.83963±0.11124]
		15-20%	SVM[0.72614±0.088536]	SVM[0.44436±0.25953]	SOM[0.76711±0.16826]	SVM[0.66069±0.17321]	SVM[0.50583±0.31636]	SVM[0.73694±0.12737]	SVM[0.73326±0.14014]
		25%	SVM[0.65453±0.1637]	SVM[0.49901±0.23045]	SVM[0.65863±0.17369]	SVM[0.65867±0.13601]	SVM[0.49192±0.24837]	SVM[0.67723±0.1572]	SVM[0.69023±0.14984]
		Todas	SVM[0.72873±0.1332]	SVM[0.47338±0.25773]	SVM[0.73313±0.15104]	SVM[0.64867±0.18145]	SVM[0.5668±0.30816]	SVM[0.75838±0.13618]	SVM[0.76793±0.14365]
	D_{KS}	5-10%	KNN[0.19557±0.089916]	SVM[0.21102±0.11381]	SVM[0.21186±0.062194]	SVM[0.2579±0.12685]	KNN[0.25228±0.079112]	SVM[0.18239±0.055577]	KNN[0.07261±0.028105]
		15-20%	SVM[0.2242±0.072545]	SVM[0.20183±0.087309]	SVM[0.1856±0.028643]	SVM[0.18972±0.075357]	SVM[0.17438±0.11336]	SVM[0.15848±0.047807]	KNN[0.069243±0.022787]
		25%	SVM[0.17719±0.068627]	SVM[0.1428±0.072046]	KNN[0.1306±0.044314]	SVM[0.17233±0.08028]	SVM[0.13383±0.038959]	SVM[0.14614±0.032773]	KNN[0.074805±0.023711]
		Todas	KNN[0.19641±0.095478]	SVM[0.19298±0.097232]	SVM[0.19237±0.048418]	SVM[0.21156±0.10433]	SVM[0.17003±0.092394]	SVM[0.16533±0.04976]	KNN[0.071719±0.025221]
Valor Extremo Generalizado	MSE	5-10%	SVM[0.074792±0.080035]	SVM[0.0014328±0.001586]	SVM[0.068665±0.079137]	SVM[0.043051±0.050692]	SVM[0.0021832±0.003685]	SVM[0.035755±0.047923]	SVM[0.006541±0.007884]
		15-20%	SVM[0.12884±0.10725]	SVM[0.010888±0.0088377]	SVM[0.12978±0.09452]	SVM[0.088242±0.095782]	SVM[0.014145±0.017163]	SVM[0.068037±0.072715]	SVM[0.037847±0.040525]
		25%	SVM[0.1665±0.11116]	SVM[0.020678±0.017731]	SOM[0.093508±0.106]	SVM[0.092949±0.084382]	SVM[0.027726±0.02966]	SVM[0.093724±0.087465]	SVM[0.064631±0.060135]
		Todas	SVM[0.1119±0.10285]	SVM[0.0091382±0.012209]	SVM[0.10486±0.097836]	SVM[0.070122±0.080735]	SVM[0.012045±0.019604]	SVM[0.058937±0.069965]	SVM[0.030437±0.043124]
	r^2	5-10%	SVM[0.75343±0.23976]	SVM[0.37951±0.22009]	SVM[0.85403±0.17377]	SVM[0.80309±0.19727]	SVM[0.55165±0.26631]	SVM[0.8727±0.16317]	SVM[0.88519±0.14789]
		15-20%	SVM[0.71624±0.23477]	SVM[0.29716±0.18747]	SVM[0.72678±0.23388]	SVM[0.74704±0.22813]	SVM[0.49465±0.21657]	SVM[0.78492±0.22768]	SVM[0.79527±0.2175]
		25%	SVM[0.6765±0.24739]	SVM[0.45581±0.19139]	SVM[0.7017±0.24578]	SVM[0.71872±0.24996]	SVM[0.55282±0.22302]	SVM[0.72911±0.2539]	SVM[0.75332±0.23778]
		Todas	SVM[0.72267±0.2403]	SVM[0.3617±0.20976]	SVM[0.77048±0.2248]	SVM[0.76335±0.22278]	SVM[0.52872±0.23915]	SVM[0.80671±0.2181]	SVM[0.82319±0.20355]
	D_{KS}	5-10%	SVM[0.29762±0.15091]	SVM[0.30355±0.16148]	SVM[0.25676±0.072801]	SVM[0.22926±0.12982]	SVM[0.20806±0.090915]	SVM[0.18323±0.067554]	SVM[0.11293±0.049011]
		15-20%	SVM[0.22997±0.09578]	SVM[0.30856±0.20451]	SVM[0.22299±0.05618]	SVM[0.16925±0.073367]	SVM[0.19529±0.076833]	SVM[0.16043±0.060564]	SVM[0.10243±0.042503]
		25%	SVM[0.18922±0.072495]	SVM[0.24268±0.16915]	SVM[0.21311±0.071259]	SVM[0.15394±0.056488]	SVM[0.16461±0.080875]	SVM[0.14905±0.050157]	SVM[0.098542±0.041174]
		Todas	SVM[0.25101±0.12669]	SVM[0.29341±0.18117]	SVM[0.23628±0.068997]	SVM[0.19045±0.103]	SVM[0.19395±0.084765]	SVM[0.1676±0.062978]	SVM[0.10591±0.045244]
Pareto Generalizada	MSE	5-10%	SVM[0.068697±0.038803]	SVM[0.028552±0.050014]	SVM[0.060902±0.046361]	SVM[0.04939±0.035198]	SVM[0.0067438±0.010923]	SVM[0.040509±0.069598]	SVM[0.013635±0.017292]
		15-20%	SVM[0.12014±0.070697]	SVM[0.046099±0.069978]	SVM[0.14763±0.12363]	SVM[0.11469±0.095071]	SVM[0.031715±0.052653]	SVM[0.06392±0.048018]	SVM[0.06205±0.061194]
		25%	SVM[0.11615±0.060966]	SVM[0.050636±0.062301]	SVM[0.151±0.11287]	SVM[0.13791±0.097444]	SVM[0.046953±0.064378]	SVM[0.14244±0.099582]	SVM[0.09794±0.083208]
		Todas	SVM[0.09678±0.060877]	SVM[0.040265±0.060942]	SVM[0.11069±0.1029]	SVM[0.09385±0.085096]	SVM[0.024944±0.046781]	SVM[0.070569±0.080949]	SVM[0.0494±0.062699]
	r^2	5-10%	SVM[0.43715±0.20978]	SVM[0.49171±0.26422]	SVM[0.67883±0.29363]	SVM[0.55777±0.27946]	SVM[0.62551±0.28463]	SVM[0.80792±0.25129]	SVM[0.77259±0.27753]
		15-20%	SVM[0.42519±0.26351]	SVM[0.46545±0.23642]	SVM[0.68688±0.26554]	SVM[0.59317±0.27599]	SVM[0.46742±0.28144]	SVM[0.78314±0.23579]	SVM[0.67133±0.31642]
		25%	SVM[0.47379±0.23571]	SVM[0.48893±0.22889]	SOM[0.70802±0.32361]	SVM[0.55547±0.27603]	SVM[0.59914±0.1769]	SVM[0.70242±0.20876]	SVM[0.62037±0.31744]
		Todas	SVM[0.44224±0.23162]	SVM[0.48076±0.24339]	SVM[0.67196±0.26465]	SVM[0.57031±0.27361]	SVM[0.55232±0.27292]	SVM[0.77665±0.23577]	SVM[0.70174±0.30482]
	D_{KS}	5-10%	SVM[0.38712±0.23511]	SVM[0.47232±0.21616]	SVM[0.29631±0.077905]	SVM[0.26137±0.15077]	SVM[0.22239±0.071249]	SVM[0.23481±0.095393]	SVM[0.13172±0.054532]
		15-20%	SVM[0.31057±0.18669]	SVM[0.40448±0.20958]	SVM[0.26311±0.06703]	SVM[0.19086±0.092896]	SVM[0.2288±0.088298]	SVM[0.19896±0.064225]	SVM[0.11439±0.035407]
		25%	SVM[0.23402±0.18807]	SVM[0.322±0.14369]	SVM[0.2145±0.048062]	SVM[0.14442±0.054789]	SVM[0.20102±0.035447]	SVM[0.18688±0.055071]	SVM[0.10901±0.03467]
		Todas	SVM[0.32924±0.21523]	SVM[0.41802±0.20683]	SVM[0.26729±0.074204]	SVM[0.21249±0.12465]	SVM[0.2212±0.073795]	SVM[0.21203±0.079421]	SVM[0.12018±0.044531]
Gaussiana Inversa	MSE	5-10%	SVM[0.076559±0.053807]	SVM[0.0010847±0.0008585]	SVM[0.052081±0.047451]	SVM[0.056159±0.056098]	SVM[0.001292±0.0014179]	SVM[0.048198±0.040274]	SVM[0.015335±0.013481]
		15-20%	SVM[0.12755±0.11017]	SVM[0.010271±0.0051163]	SVM[0.084628±0.026842]	SVM[0.091247±0.061205]	SVM[0.017804±0.011475]	SVM[0.08783±0.076251]	SVM[0.065251±0.054174]
		25%	SVM[0.21127±0.16522]	SVM[0.025946±0.013157]	SVM[0.14311±0.072588]	SVM[0.16417±0.17144]	SVM[0.031135±0.016303]	SVM[0.062935±0.024073]	SVM[0.10628±0.079076]
		Todas	SVM[0.11959±0.10814]	SVM[0.0091722±0.010681]	SVM[0.079631±0.054892]	SVM[0.088556±0.091583]	SVM[0.013382±0.01503]	SVM[0.066067±0.055992]	SVM[0.053491±0.059131]
	r^2	5-10%	SVM[0.67218±0.32905]	SVM[0.37575±0.1892]	SVM[0.77849±0.25701]	SVM[0.79036±0.26976]	SVM[0.47873±0.25443]	SVM[0.77496±0.25338]	SVM[0.74337±0.25248]
		15-20%	SVM[0.60621±0.30141]	SVM[0.27478±0.2089]	SVM[0.63398±0.29575]	SVM[0.67682±0.27786]	SVM[0.49818±0.24032]	SVM[0.69742±0.28183]	SVM[0.64455±0.29671]
		25%	SVM[0.64936±0.064467]	SVM[0.48194±0.21908]	SVM[0.53388±0.31045]	SVM[0.62145±0.31863]	SVM[0.49746±0.20918]	SVM[0.81417±0.056591]	SVM[0.60637±0.30645]
		Todas	SVM[0.64487±0.27788]	SVM[0.36232±0.20502]	SVM[0.66454±0.28588]	SVM[0.71068±0.27901]	SVM[0.48875±0.23209]	SVM[0.75134±0.23875]	SVM[0.67644±0.27759]
	D_{KS}	5-10%	SVM[0.38429±0.20136]	SVM[0.24585±0.082261]	SVM[0.2341±0.062183]	SVM[0.28571±0.1747]	SVM[0.2644±0.17626]	SVM[0.21442±0.074084]	SVM[0.13806±0.058826]
		15-20%	SVM[0.22593±0.060597]	SVM[0.26207±0.085448]	SVM[0.20925±0.040961]	SVM[0.20697±0.091654]	SVM[0.20544±0.13098]	SVM[0.17593±0.044622]	SVM[0.11589±0.03827]
		25%	SVM[0.20658±0.0493]	SVM[0.1654±0.051037]	SVM[0.22058±0.050642]	SVM[0.20294±0.10158]	SVM[0.14978±0.056337]	SVM[0.11493±0.021517]	SVM[0.10987±0.03673]
		Todas	SVM[0.2892±0.15771]	SVM[0.23748±0.083778]	SVM[0.22055±0.050719]	SVM[0.23704±0.13386]	SVM[0.21841±0.14383]	SVM[0.18594±0.065535]	SVM[0.12355±0.047442]

Continua na página seguinte...

Tabela A.7: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Logistica	MSE	5-10%	SVM[0.07932±0.068137]	SVM[0.0012606±0.0025025]	SVM[0.050129±0.070918]	SVM[0.059103±0.05639]	SVM[0.00386±0.003615]	SVM[0.029547±0.039216]	SVM[0.0097565±0.0094086]
		15-20%	SVM[0.12899±0.087345]	SVM[0.01194±0.010836]	SVM[0.10091±0.098499]	SVM[0.082925±0.068195]	SVM[0.019008±0.01526]	SVM[0.12259±0.08482]	SVM[0.049102±0.038483]
		25%	SVM[0.14348±0.074306]	SVM[0.024362±0.018086]	SVM[0.11825±0.082478]	SVM[0.098941±0.064337]	SVM[0.031938±0.022474]	SVM[0.11595±0.073325]	SVM[0.083775±0.052945]
		Todas	SVM[0.10858±0.08016]	SVM[0.010672±0.01375]	SVM[0.084059±0.088888]	SVM[0.076533±0.063986]	SVM[0.015936±0.017751]	SVM[0.082609±0.079296]	SVM[0.040298±0.044068]
	r^2	5-10%	SVM[0.545±0.29524]	SVM[0.43536±0.21803]	SVM[0.81118±0.11565]	SVM[0.74355±0.19989]	SVM[0.57977±0.32005]	SVM[0.84844±0.1997]	SVM[0.8267±0.17392]
		15-20%	KNN[0.75107±0.30116]	SVM[0.49536±0.2461]	SVM[0.69403±0.17618]	SVM[0.67769±0.19949]	SVM[0.57439±0.33902]	SVM[0.64136±0.21113]	SVM[0.72865±0.20392]
		25%	SVM[0.68194±0.11742]	SVM[0.57875±0.23631]	SVM[0.71307±0.12234]	SVM[0.646±0.20101]	SVM[0.67585±0.24339]	SVM[0.65804±0.2231]	SVM[0.67092±0.21051]
		Todas	SOM[0.65593±0.31728]	SVM[0.4903±0.23671]	SVM[0.73785±0.15421]	SVM[0.68846±0.20026]	SVM[0.59951±0.31014]	SVM[0.7226±0.22769]	SVM[0.75549±0.20128]
	D_{KS}	5-10%	SVM[0.40248±0.11082]	SVM[0.15334±0.068397]	SVM[0.21171±0.057057]	SVM[0.24949±0.15789]	SVM[0.23123±0.088329]	SVM[0.24018±0.07555]	KNN[0.091685±0.054206]
		15-20%	SVM[0.3035±0.088204]	SVM[0.11862±0.038568]	SVM[0.18527±0.059728]	SVM[0.18698±0.10425]	SVM[0.19359±0.077537]	SVM[0.20412±0.077612]	KNN[0.08121±0.041339]
		25%	SVM[0.26556±0.086014]	SVM[0.094805±0.035549]	SVM[0.15158±0.032047]	SVM[0.13205±0.062332]	SVM[0.14984±0.034454]	SVM[0.15642±0.064477]	KNN[0.08648±0.035985]
		Todas	SVM[0.33779±0.11227]	SVM[0.13066±0.056863]	SVM[0.18785±0.057654]	SVM[0.20435±0.1314]	SVM[0.20097±0.080665]	SVM[0.20991±0.079053]	KNN[0.086586±0.046018]
Log-Logistica	MSE	5-10%	SVM[0.15493±0.065598]	SVM[0.00078687±0.00079428]	SVM[0.11866±0.076128]	SVM[0.12689±0.065747]	SVM[0.0020745±0.0029263]	SVM[0.1053±0.056355]	SVM[0.018397±0.0088733]
		15-20%	SVM[0.25925±0.066347]	SVM[0.008737±0.0048349]	SVM[0.23651±0.090791]	SVM[0.20274±0.078699]	SVM[0.014332±0.010379]	SVM[0.18531±0.07404]	SVM[0.091397±0.036857]
		25%	SVM[0.28456±0.088251]	SVM[0.021446±0.0083057]	SVM[0.23553±0.072495]	SVM[0.25488±0.099332]	SVM[0.03242±0.013176]	SVM[0.19564±0.099697]	SVM[0.14533±0.056564]
		Todas	SVM[0.21915±0.090173]	SVM[0.0082424±0.0091472]	SVM[0.1869±0.098448]	SVM[0.18496±0.093205]	SVM[0.013272±0.014392]	SVM[0.15224±0.081629]	SVM[0.072984±0.059711]
	r^2	5-10%	SVM[0.69781±0.21752]	SVM[0.16721±0.14739]	KNN[0.431±0.20205]	SVM[0.63389±0.16214]	SVM[0.2416±0.20224]	SVM[0.65241±0.16881]	SVM[0.67182±0.1663]
		15-20%	SVM[0.53231±0.164]	SVM[0.15116±0.12328]	SVM[0.5241±0.24182]	SVM[0.53971±0.13215]	SVM[0.20122±0.13184]	SVM[0.47262±0.19606]	SVM[0.50403±0.18272]
		25%	SVM[0.46585±0.19781]	SVM[0.18675±0.10245]	SVM[0.40438±0.20524]	SVM[0.49017±0.19992]	SVM[0.25536±0.15402]	SVM[0.47797±0.21523]	SVM[0.43022±0.19404]
		Todas	SVM[0.57426±0.20692]	SVM[0.16397±0.12691]	SVM[0.56024±0.23482]	SVM[0.56802±0.16709]	SVM[0.22856±0.16453]	SVM[0.53591±0.20394]	SVM[0.55638±0.20064]
	D_{KS}	5-10%	SVM[0.37681±0.10809]	SVM[0.24993±0.13006]	SVM[0.25108±0.060103]	SVM[0.31205±0.097439]	SVM[0.21639±0.098938]	SVM[0.20324±0.058077]	SVM[0.11621±0.049294]
		15-20%	SVM[0.33152±0.070321]	SVM[0.23061±0.079274]	SVM[0.20736±0.036675]	SVM[0.27906±0.057726]	SVM[0.1925±0.054631]	SVM[0.19559±0.065593]	SVM[0.098922±0.035164]
		25%	SVM[0.25845±0.071676]	SVM[0.14275±0.056112]	SVM[0.16964±0.015532]	SVM[0.22865±0.056123]	SVM[0.17941±0.097579]	SVM[0.14571±0.038943]	SVM[0.099117±0.030034]
		Todas	SVM[0.33502±0.095896]	SVM[0.21959±0.1049]	SVM[0.22005±0.053974]	SVM[0.28198±0.078606]	SVM[0.19719±0.081631]	SVM[0.18939±0.06028]	SVM[0.10598±0.04048]
Normal	MSE	5-10%	SVM[0.059813±0.042905]	SVM[0.00081306±0.00080066]	SVM[0.050892±0.039803]	SVM[0.05181±0.034664]	SVM[0.012088±0.020388]	SVM[0.040611±0.037281]	SVM[0.016698±0.016835]
		15-20%	SVM[0.13554±0.080171]	SVM[0.0095901±0.011225]	SVM[0.09604±0.054777]	SVM[0.11709±0.057818]	SVM[0.040071±0.057399]	SVM[0.093531±0.071632]	SVM[0.070626±0.056764]
		25%	SVM[0.12284±0.032759]	SVM[0.026216±0.026313]	SVM[0.16291±0.08839]	SVM[0.17133±0.0654]	SVM[0.067891±0.087103]	SVM[0.14223±0.088855]	SVM[0.11176±0.07619]
		Todas	SVM[0.098714±0.067317]	SVM[0.0090535±0.01582]	SVM[0.094245±0.070304]	SVM[0.10183±0.067767]	SVM[0.034949±0.057221]	SVM[0.08295±0.072688]	SVM[0.057282±0.061052]
	r^2	5-10%	KNN[0.86668±0.069733]	SVM[0.47696±0.3508]	KNN[0.86655±0.042005]	SVM[0.668±0.2385]	SVM[0.63711±0.29039]	SVM[0.69862±0.2686]	SVM[0.71641±0.27003]
		15-20%	SOM[0.71649±0.25748]	SVM[0.50713±0.35915]	SVM[0.53001±0.29506]	SVM[0.51197±0.27854]	SVM[0.57584±0.28438]	SVM[0.47833±0.28351]	SVM[0.61±0.29507]
		25%	SVM[0.56435±0.28669]	SVM[0.46638±0.36862]	SVM[0.60423±0.20461]	SVM[0.56119±0.25267]	SVM[0.69141±0.27767]	SVM[0.51456±0.26671]	SVM[0.5595±0.29925]
		Todas	SVM[0.5922±0.22386]	SVM[0.48555±0.34578]	KNN[0.79638±0.11248]	SVM[0.57694±0.25775]	SVM[0.62088±0.27863]	SVM[0.5564±0.27891]	SVM[0.64246±0.28474]
	D_{KS}	5-10%	SVM[0.33151±0.11029]	SVM[0.17427±0.05078]	SVM[0.24797±0.089451]	SVM[0.29599±0.098344]	SVM[0.18338±0.060741]	SVM[0.27261±0.13106]	KNN[0.087441±0.056437]
		15-20%	SVM[0.29498±0.06677]	SVM[0.14916±0.068581]	SVM[0.17972±0.052409]	SVM[0.20464±0.044227]	SVM[0.20066±0.077903]	SVM[0.20438±0.10568]	SVM[0.10066±0.048528]
		25%	SVM[0.28654±0.064039]	SVM[0.15014±0.056435]	SVM[0.18093±0.038729]	SVM[0.18805±0.048651]	SVM[0.13839±0.052184]	SVM[0.16012±0.044641]	SVM[0.099581±0.045155]
		Todas	SVM[0.30611±0.083487]	SVM[0.15814±0.058695]	SVM[0.20377±0.070533]	SVM[0.22989±0.07947]	SVM[0.18005±0.069658]	SVM[0.22308±0.1133]	KNN[0.089736±0.045428]
Nakagami	MSE	5-10%	SVM[0.064389±0.03714]	SVM[0.00092917±0.00074833]	SVM[0.045535±0.035986]	SVM[0.088981±0.057847]	SVM[0.005752±0.0037738]	SVM[0.041327±0.03045]	SVM[0.012367±0.010714]
		15-20%	SVM[0.17932±0.084264]	SVM[0.014117±0.0090002]	SVM[0.1468±0.073404]	SVM[0.12763±0.089402]	SVM[0.022764±0.013146]	SVM[0.1311±0.080922]	SVM[0.06363±0.045082]
		25%	SVM[0.23279±0.11438]	SVM[0.027363±0.011288]	SVM[0.18072±0.085652]	SVM[0.19392±0.11208]	SVM[0.039912±0.015576]	SVM[0.16774±0.08129]	SVM[0.10014±0.063284]
		Todas	SVM[0.14827±0.10281]	SVM[0.010709±0.01224]	SVM[0.11307±0.08301]	SVM[0.12874±0.090473]	SVM[0.018864±0.016645]	SVM[0.10678±0.08266]	SVM[0.050427±0.051844]
	r^2	5-10%	SVM[0.73957±0.041521]	SVM[0.38928±0.28712]	SVM[0.75487±0.16801]	SVM[0.7524±0.19116]	SVM[0.44994±0.35451]	SVM[0.69325±0.16789]	SVM[0.76776±0.20289]
		15-20%	SVM[0.67143±0.16383]	SVM[0.40188±0.29201]	SVM[0.64473±0.19761]	SVM[0.68831±0.18831]	SVM[0.4131±0.30414]	SVM[0.6598±0.18842]	SVM[0.64974±0.23409]
		25%	SVM[0.5433±0.19631]	SVM[0.39401±0.28035]	SVM[0.52369±0.21472]	SVM[0.54972±0.2125]	SVM[0.4192±0.28788]	SVM[0.56802±0.22863]	SVM[0.60586±0.24996]
		Todas	SVM[0.65378±0.16242]	SVM[0.39552±0.2764]	SVM[0.6616±0.20025]	SVM[0.66102±0.20059]	SVM[0.42986±0.31168]	SVM[0.64871±0.18484]	SVM[0.68817±0.22761]
	D_{KS}	5-10%	SVM[0.31458±0.083547]	SVM[0.18367±0.092426]	SVM[0.19335±0.069777]	SVM[0.25258±0.11118]	SVM[0.15848±0.04485]	SVM[0.18516±0.052004]	SVM[0.071686±0.027088]
		15-20%	SVM[0.22044±0.054293]	SVM[0.14947±0.061673]	SVM[0.16841±0.032665]	SVM[0.17904±0.075668]	SVM[0.14466±0.035708]	SVM[0.15677±0.032922]	SVM[0.063272±0.02009]
		25%	SVM[0.21525±0.071559]	SVM[0.11832±0.048736]	SVM[0.1495±0.01445]	SVM[0.12035±0.057837]	SVM[0.11804±0.031171]	SVM[0.11134±0.025745]	SVM[0.068804±0.022231]
		Todas	SVM[0.26647±0.085618]	SVM[0.156±0.074783]	SVM[0.17303±0.048392]	SVM[0.20329±0.10169]	SVM[0.14325±0.039972]	SVM[0.16211±0.048626]	SVM[0.067707±0.023084]

Continua na página seguinte...

Tabela A.7: Continuação da página anterior.

Distribuição	Métrica	MR	T1	T2	T3	T4	T5	T6	T7
Log-Normal	MSE	5-10%	SVM[0.084648±0.049758]	SVM[0.0012902±0.0010828]	SVM[0.060877±0.046185]	SVM[0.071755±0.038959]	SVM[0.0015065±0.0012569]	SVM[0.029964±0.030195]	SVM[0.0074334±0.0062818]
		15-20%	SVM[0.17924±0.11062]	SVM[0.022832±0.016041]	SOM[0.078406±0.016097]	SVM[0.13238±0.10525]	SVM[0.017597±0.020235]	SOM[0.030467±0.01183]	SVM[0.048441±0.027711]
		25%	SVM[0.16938±0.090409]	SVM[0.029228±0.02048]	SVM[0.15796±0.025706]	SVM[0.18825±0.14948]	SVM[0.026227±0.021987]	SVM[0.10132±0.044074]	SVM[0.086527±0.040314]
		Todas	SVM[0.13501±0.090681]	SVM[0.015467±0.017751]	SVM[0.11815±0.07257]	SVM[0.11526±0.095918]	SVM[0.011884±0.017391]	SVM[0.059736±0.047049]	SVM[0.039655±0.038552]
	r^2	5-10%	SVM[0.7061±0.25159]	SVM[0.30239±0.16098]	SVM[0.74006±0.20158]	SVM[0.7291±0.22285]	SVM[0.52053±0.15222]	SVM[0.86705±0.13933]	SVM[0.88949±0.093873]
		15-20%	SVM[0.59395±0.18742]	SVM[0.29149±0.1094]	MM[0.89519±0]	SVM[0.63412±0.14466]	SVM[0.38523±0.13415]	SOM[0.86661±0.01825]	SVM[0.74021±0.14777]
		25%	SVM[0.62102±0.15041]	SVM[0.39265±0.10193]	SVM[0.67306±0.11764]	SVM[0.61305±0.25664]	SVM[0.41295±0.039651]	SVM[0.63405±0.17958]	SVM[0.6722±0.16322]
		Todas	SVM[0.63855±0.19506]	SVM[0.31978±0.13102]	SVM[0.65621±0.18099]	SVM[0.67172±0.19599]	SVM[0.43788±0.13919]	SVM[0.74484±0.18528]	SVM[0.78632±0.15368]
	D_{KS}	5-10%	SVM[0.32706±0.13932]	SVM[0.35261±0.11668]	SVM[0.26982±0.058042]	SVM[0.27896±0.15755]	SVM[0.21455±0.095366]	SVM[0.1959±0.07347]	SVM[0.1292±0.045962]
		15-20%	SVM[0.25571±0.10007]	SVM[0.25152±0.086376]	SVM[0.23134±0.046772]	SVM[0.23554±0.10237]	SVM[0.25218±0.15417]	SVM[0.17093±0.05072]	SVM[0.11373±0.035676]
		25%	SVM[0.23405±0.10136]	SVM[0.20071±0.064736]	SVM[0.19387±0.031059]	MM[0.15294±0]	SVM[0.20022±0.094418]	MM[0.10714±0]	SVM[0.11374±0.038711]
		Todas	SVM[0.27985±0.11724]	SVM[0.2948±0.11306]	SVM[0.23881±0.055945]	SVM[0.24981±0.12792]	SVM[0.226±0.11689]	SVM[0.17857±0.055734]	SVM[0.11992±0.039297]
Rayleigh	MSE	5-10%	SVM[0.055097±0.0046583]	SVM[0.0023889±0.001099]	SVM[0.02278±0.018645]	SVM[0.036804±0.020037]	SVM[0.0017865±0.0013682]	SVM[0.019027±0.016081]	SVM[0.0081982±0.0049771]
		15-20%	SVM[0.1053±0.024432]	SVM[0.017131±0.011713]	SOM[0.053278±0.011632]	SVM[0.071614±0.04031]	SVM[0.014116±0.0042292]	SVM[0.068351±0.018027]	SVM[0.044029±0.024887]
		25%	SVM[0.10666±0.057784]	SVM[0.032687±0.010397]	SOM[0.084845±0]	SVM[0.10432±0.021302]	SVM[0.026615±0.0080185]	SVM[0.097761±0.024971]	SVM[0.075228±0.040811]
		Todas	SVM[0.093088±0.037517]	SVM[0.014345±0.014197]	SVM[0.048424±0.037088]	SVM[0.066736±0.037967]	SVM[0.011684±0.010418]	SVM[0.057037±0.036027]	SVM[0.035936±0.034009]
	r^2	5-10%	KNN[0.86665±0.028877]	SVM[0.3793±0.11885]	SVM[0.90234±0.093404]	KNN[0.74064±0.2155]	SVM[0.6628±0.25792]	SVM[0.89003±0.12175]	SVM[0.87±0.093813]
		15-20%	SVM[0.7345±0.087422]	SVM[0.3054±0.085691]	SOM[0.81971±0.049992]	SVM[0.87925±0]	SVM[0.36203±0.08936]	SVM[0.72945±0.081045]	SVM[0.75386±0.12559]
		25%	SVM[0.83236±0.09331]	SVM[0.40893±0.15166]	MM[0.64475±0.2241]	SVM[0.76469±0]	SVM[0.54075±0.16916]	SVM[0.71614±0.083238]	SVM[0.70158±0.16384]
		Todas	KNN[0.81271±0.078082]	SVM[0.35566±0.1135]	SVM[0.83377±0.11491]	SOM[0.85974±0.063718]	SVM[0.49582±0.20851]	SVM[0.77363±0.11622]	SVM[0.78413±0.13318]
	D_{KS}	5-10%	KNN[0.27067±0.075507]	SVM[0.19853±0.051994]	SVM[0.22546±0.061789]	SVM[0.28841±0.13326]	SVM[0.17795±0.027626]	SVM[0.18623±0.045613]	KNN[0.063997±0.035646]
		15-20%	SOM[0.29412±0]	SVM[0.18087±0.079827]	KNN[0.16556±0.015723]	SVM[0.20912±0.054964]	SVM[0.19878±0.068445]	SVM[0.16282±0.057234]	KNN[0.06008±0.019186]
		25%	SVM[0.19876±0.048147]	SOM[0.10588±0]	SVM[0.14548±0.020541]	SVM[0.19332±0.0071842]	SOM[0.11765±0]	SVM[0.14868±0.025067]	KNN[0.072155±0.017803]
		Todas	SVM[0.29221±0.10102]	SVM[0.19904±0.057146]	KNN[0.15165±0.0191]	SVM[0.23204±0.086071]	SVM[0.18893±0.045381]	SVM[0.16807±0.043368]	KNN[0.064373±0.024939]
t Location-scale	MSE	5-10%	SVM[0.096597±0.061118]	SVM[0.00019395±0.00011087]	SVM[0.055302±0.049141]	SVM[0.075373±0.06605]	SVM[0.0015385±0.0019435]	SVM[0.059147±0.064612]	SVM[0.013957±0.012177]
		15-20%	SVM[0.21798±0.066218]	SVM[0.0065036±0.0038399]	SVM[0.17346±0.075159]	SVM[0.1638±0.077919]	SVM[0.011054±0.010451]	SVM[0.16864±0.082282]	SVM[0.076994±0.042441]
		25%	SVM[0.25207±0.064022]	SVM[0.018177±0.010384]	SVM[0.22424±0.069472]	SVM[0.24064±0.098338]	SVM[0.026917±0.014863]	SVM[0.1843±0.10891]	SVM[0.12812±0.059772]
		Todas	SVM[0.15164±0.091241]	SVM[0.0065092±0.0087019]	SVM[0.12913±0.093544]	SVM[0.13774±0.10158]	SVM[0.01042±0.013147]	SVM[0.12615±0.09809]	SVM[0.062004±0.057905]
	r^2	5-10%	SVM[0.67286±0.18184]	SVM[0.46672±0.2887]	SVM[0.7694±0.22317]	SVM[0.74266±0.21896]	SVM[0.52195±0.32741]	SVM[0.82721±0.21008]	SVM[0.77571±0.21202]
		15-20%	SVM[0.56028±0.15062]	SVM[0.37208±0.26762]	SVM[0.51416±0.24666]	SVM[0.60556±0.25109]	SVM[0.51059±0.30262]	SVM[0.54621±0.25576]	SVM[0.59051±0.22286]
		25%	SVM[0.46762±0.16273]	SVM[0.3763±0.28034]	SVM[0.37557±0.20645]	SVM[0.46651±0.19985]	SVM[0.48382±0.2958]	SVM[0.51606±0.24884]	SVM[0.50837±0.22958]
		Todas	SVM[0.58867±0.18233]	SVM[0.41108±0.27757]	SVM[0.58164±0.2761]	SVM[0.61904±0.24778]	SVM[0.50912±0.30612]	SVM[0.63875±0.27308]	SVM[0.64816±0.24281]
	D_{KS}	5-10%	SVM[0.3397±0.083316]	SVM[0.19136±0.067166]	SVM[0.23486±0.059962]	SVM[0.28996±0.10594]	SVM[0.21439±0.088079]	SVM[0.21589±0.068909]	SVM[0.1248±0.046392]
		15-20%	SVM[0.29115±0.066427]	SVM[0.14337±0.035804]	SVM[0.19595±0.03891]	SVM[0.24984±0.069748]	SVM[0.15956±0.048178]	SVM[0.18573±0.053802]	SVM[0.10919±0.028482]
		25%	SVM[0.24425±0.050263]	SVM[0.13734±0.038715]	SVM[0.1707±0.035315]	SVM[0.20949±0.064121]	SVM[0.14225±0.062457]	SVM[0.15443±0.027792]	SVM[0.10339±0.027098]
		Todas	SVM[0.30059±0.078585]	SVM[0.1643±0.057645]	SVM[0.20618±0.053451]	SVM[0.25871±0.089711]	SVM[0.18088±0.075642]	SVM[0.19232±0.061031]	SVM[0.11427±0.037078]
Weibull	MSE	5-10%	SVM[0.089596±0.060086]	SVM[0.0022866±0.0013118]	SVM[0.063725±0.042515]	SVM[0.081401±0.047511]	SVM[0.0026766±0.0024594]	SVM[0.05163±0.033744]	SVM[0.015136±0.0076285]
		15-20%	SVM[0.19473±0.090827]	SVM[0.015048±0.0065241]	SVM[0.15153±0.083063]	SVM[0.17189±0.080006]	SVM[0.016364±0.0084273]	SVM[0.11937±0.051552]	SVM[0.07385±0.034926]
		25%	SVM[0.24684±0.11338]	SVM[0.036931±0.014383]	SVM[0.15853±0.065907]	SVM[0.15795±0.066511]	SVM[0.039561±0.017493]	SVM[0.13888±0.066087]	SVM[0.11578±0.046258]
		Todas	SVM[0.1631±0.10521]	SVM[0.014297±0.014988]	SVM[0.11674±0.078372]	SVM[0.13186±0.07791]	SVM[0.016096±0.016822]	SVM[0.092903±0.060153]	SVM[0.058751±0.049271]
	r^2	5-10%	KNN[0.95156±0.032486]	SVM[0.45711±0.19105]	KNN[0.89165±0.079927]	SVM[0.72811±0.11904]	SVM[0.39491±0.23981]	SVM[0.75969±0.14785]	SVM[0.7516±0.11779]
		15-20%	SOM[0.89253±0.033417]	SVM[0.32732±0.27837]	SVM[0.64166±0.14929]	SVM[0.604±0.16426]	SVM[0.47482±0.28536]	SVM[0.63452±0.16202]	SVM[0.60182±0.16791]
		25%	SVM[0.51118±0.17813]	SVM[0.28608±0.24823]	SVM[0.59477±0.16951]	SOM[0.56757±0.31106]	SVM[0.29628±0.27522]	SVM[0.56069±0.15807]	SVM[0.55828±0.17375]
		Todas	SOM[0.80792±0.15584]	SVM[0.35713±0.2483]	SVM[0.69371±0.16997]	SVM[0.63961±0.15651]	SVM[0.39447±0.26926]	SVM[0.66564±0.17214]	SVM[0.65302±0.17023]
	D_{KS}	5-10%	SVM[0.37422±0.11876]	SVM[0.28985±0.12923]	SVM[0.24213±0.074405]	SVM[0.30438±0.11997]	SVM[0.24013±0.096835]	SVM[0.23635±0.085908]	SVM[0.14024±0.050062]
		15-20%	SVM[0.33072±0.078594]	SVM[0.20522±0.08279]	SVM[0.2196±0.060574]	SVM[0.29587±0.11134]	SOM[0.19217±0.09445]	SVM[0.20837±0.069337]	SVM[0.12046±0.030843]
		25%	SVM[0.2885±0.059905]	SVM[0.17073±0.060711]	SVM[0.17717±0.038868]	SVM[0.25857±0.088825]	SVM[0.14771±0.066154]	SVM[0.17426±0.063374]	SVM[0.11239±0.025745]
		Todas	SVM[0.33994±0.098799]	SVM[0.23248±0.11161]	SVM[0.22076±0.066973]	SVM[0.29096±0.1102]	SVM[0.21058±0.089335]	SVM[0.21332±0.078346]	SVM[0.12676±0.040119]

Apêndice B

Informações auxiliares ao Capítulo 5

Tabela B.1: Sumário dos parâmetros utilizados nas técnicas de sobre-amostragem implementadas.

Algoritmo	***Software***	**Parâmetro**	**Valores**
ADASYN	KEEL	Número de Vizinhos	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]
		Tipo de Interpolação	Padrão
		Alpha	0.5
		Mu	0.5
ADOMS	KEEL	Número de Vizinhos	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Tipo de Interpolação	Padrão
		Alpha	0.5
		Mu	0.5
AHC	KEEL	N.A.	N.A
Borderline-SMOTE 1 e 2	KEEL	Número de Vizinhos SMOTE	5
		Número de Vizinhos para considerar uma instancia BORDER	3
		Tipo de Borderline SMOTE	1 e 2
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]
		Tipo de Interpolação	Padrão
		Alpha	0.5
		Mu	0.5

Continua na página seguinte...

Tabela B.1: Continuação da página anterior.

Algoritmo	***Software***	**Parâmetro**	**Valores**
CBO+SMOTE	MatLab	Número de Vizinhos (SMOTE)	5
		Número de repetições da avaliação do K ótimo	50
		Número de repetições do K-Means	1
		Limite (dimensão minima de cada cluster)	6
		K ótimo máximo	20
		Critérios de avaliação do K ótimo	CalinskiHarabasz DaviesBouldin Silhouette
CBO+Random	MatLab	Número de repetições da avaliação do K ótimo	5
		Número de repetições do K-Means	1
		Limite (dimensão minima de cada cluster)	6
		K ótimo máximo	20
		Critérios de avaliação do K ótimo	CalinskiHarabasz DaviesBouldin Silhouette
MWMOTE	MatLab	Número de vizinhos para preditar ruído (k1)	5
		Número de vizinhos majoritários (k2)	5
		Número de vizinhos Minoritários (k3)	5
ROS	KEEL	N.A.	N.A.
Safe-Level-SMOTE	KEEL	Número de Vizinhos	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]
		Tipo de Interpolação	Padrão
		Alpha	0.5
		Mu	0.5
SMOTE	KEEL	Número de Vizinhos	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]
		Tipo de Interpolação	Padrão
		Alpha	0.5
		Mu	0.5
SMOTE+ENN	KEEL	Número de Vizinhos ENN	3
		Número de Vizinhos SMOTE	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]
SMOTE+TL	KEEL	Número de Vizinhos	5
		Tipo de SMOTE	Minoritário
		Balanceamento	Sim
		Função de Distância	HVDM[1]

Continua na página seguinte...

Tabela B.1: Continuação da página anterior.

Algoritmo	***Software***	**Parâmetro**	**Valores**
SPIDER	KEEL	Número de Vizinhos	5
		Função de Distância	HVDM[1]
		Amplificação	Fraca
SPIDER2	KEEL	Número de Vizinhos	5
		Função de Distância	HVDM[1]
		Reetiquetação	Sim
		Amplificação	Forte

[1] *Heterogeneous Value Difference Metric* – Esta técnica é similar à HEOM. A sua grande diferença encontra se nas funções correspondestes às das equações 2.3 e 2.4 de HEOM. Nessas métricas HDVM efetua um cálculo mais complexo mas que é menos susceptível a *outliers* [10].

Tabela B.2: *Rank* médio, para cada classificador, dos valores da AUC de teste em cada *dataset* por técnica de pré-processamento. Os valores a negrito correspondem aos 3 melhores resultados.

Metodo	**CART**	**C4.5**	**k-NN**	**NB**	**SVM Linear**	**SVM RBF**
Original	10,064±4,684	12,391±4,487	11,475±5,099	9,923±5,470	12,440±4,788	12,531±4,579
ADASYN	8,462±4,077	8,320±4,033	7,738±4,389	9,004±4,507	8,275±4,086	7,488±4,077
ADOMS	7,580±4,317	7,493±4,156	7,599±4,203	8,528±4,469	7,619±3,872	6,818±3,980
AHC	8,815±4,223	8,547±3,958	7,523±3,675	7,625±3,462	8,147±3,588	7,874±3,801
Borderline-SMOTE1	9,426±3,994	8,263±3,85	8,356±3,916	8,968±4,800	8,285±3,992	9,161±3,959
Borderline-SMOTE2	9,426±3,994	8,263±3,85	8,356±3,916	8,968±4,800	8,285±3,992	9,161±3,959
CBO+*Random*	9,933±4,159	10,472±4,638	10,782±3,871	9,522±4,882	9,403±4,591	8,817±4,012
CBO+SMOTE	7,535±4,372	8,025±4,574	7,772±4,148	11,298±4,240	9,416±4,198	8,423±4,194
MWMOTE	**6,993±4,074**	**6,020±4,392**	**6,276±4,804**	8,510±3,937	7,251±3,844	8,144±4,326
ROS	10,892±4,208	9,840±3,885	10,433±3,113	7,988±4,028	7,669±3,747	**7,291±3,606**
Safe-Level-SMOTE	10,049±4,342	9,262±3,965	9,969±3,463	**7,020±3,883**	7,276±3,718	6,888±3,602
SMOTE	7,480±3,733	7,247±3,458	6,568±3,512	7,557±3,369	**7,174±3,215**	**7,308±3,389**
SMOTE+ENN	**6,897±4,149**	**7,131±4,030**	**6,237±3,730**	**7,404±3,755**	**7,007±3,811**	8,532±4,565
SMOTE+TL	**5,532±3,859**	**6,356±3,915**	**5,892±4,043**	**7,313±3,519**	**7,115±4,141**	**7,000±4,700**
SPIDER	8,371±4,207	8,906±4,337	10,400±4,098	8,201±4,681	10,304±5,157	10,291±4,867
SPIDER2	8,545±4,548	9,465±4,393	10,622±3,979	8,172±4,815	10,334±5,174	10,273±5,052

Tabela B.3: *G-Mean* de teste e treino média para todos os algoritmos de sobre-amostragem e métodos de classificação, considerando ambas as fases. Os melhores valores para cada coluna encontram-se a negrito.

		CART		**C4.5**		**k-NN**		**SVM Linear**		**SVM RBF**		**NB**	
	Método	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2
Teste	Original	0,6192±0,2626	0,6192±0,2626	0,6156±0,2939	0,6085±0,2996	0,6135±0,3201	0,6135±0,3201	0,4972±0,3531	0,4278±0,3702	0,4298±0,3755	0,3498±0,3778	0,5945±0,2483	**0,5806±0,2612**
	ADASYN	0,9315±0,0807	0,6288±0,2394	0,9343±0,0776	0,6335±0,2422	0,9346±0,0812	0,6486±0,2394	0,8863±0,1055	0,6081±0,2225	0,8692±0,178	**0,6251±0,255**	0,8213±0,1347	0,5288±0,2435
	ADOMS	0,9296±0,0826	0,6271±0,2586	0,9327±0,0796	0,6418±0,2421	0,9304±0,0839	0,6466±0,2443	0,8854±0,0978	0,6148±0,2151	0,8792±0,1757	**0,6241±0,2364**	0,8353±0,1251	0,5362±0,2533
	AHC	0,9332±0,0785	0,6168±0,253	0,9353±0,0776	0,639±0,2416	0,9368±0,0864	0,659±0,2441	0,8947±0,0989	0,6077±0,2305	0,897±0,175	0,6161±0,2385	0,8404±0,1169	0,5617±0,237
	Borderline-SMOTE1	0,9441±0,0754	0,6313±0,2479	0,9451±0,0747	**0,6539±0,2423**	0,9468±0,0752	**0,6683±0,2484**	**0,9095±0,1068**	**0,627±0,2161**	**0,9115±0,1452**	0,6038±0,256	**0,8699±0,115**	0,5579±0,2439
	Borderline-SMOTE2	0,9432±0,0747	0,6313±0,2479	0,9459±0,0707	**0,6539±0,2423**	0,9477±0,0727	**0,6683±0,2484**	0,907±0,1074	**0,627±0,2161**	0,9087±0,1532	0,6038±0,256	**0,8707±0,1145**	0,5579±0,2439
	CBO+*Random*	**0,9701±0,0507**	0,6364±0,2483	**0,9697±0,0492**	0,6297±0,2498	**0,9651±0,0732**	0,6468±0,2506	0,9033±0,0989	0,5923±0,2585	0,6339±0,2726	0,6022±0,256	0,8508±0,1103	0,5202±0,2418
	CBO+SMOTE	0,9483±0,0648	0,6361±0,2438	0,9496±0,0659	0,6439±0,2349	0,9504±0,0769	0,658±0,2422	0,909±0,0857	0,592±0,2404	0,9021±0,1986	0,5821±0,2522	0,8602±0,1031	0,5235±0,2351
	MWMOTE	0,9166±0,086	0,622±0,2485	0,9258±0,0767	0,6306±0,2397	0,9255±0,078	0,6403±0,2446	0,904±0,0862	0,6232±0,2274	0,9052±0,1335	0,6179±0,2534	0,8629±0,1073	0,5638±0,2457
	ROS	**0,9554±0,0677**	0,6176±0,2661	**0,9537±0,0698**	0,6396±0,2435	0,9518±0,0855	0,6394±0,2557	0,8909±0,0955	0,6197±0,221	0,8291±0,2157	0,6193±0,2522	0,8341±0,1219	0,5428±0,2418
	Safe-Level-SMOTE	0,9551±0,0716	0,6315±0,2523	0,9513±0,0756	0,6413±0,2409	0,9552±0,0826	0,6424±0,2536	0,8861±0,1271	0,6137±0,2253	0,7444±0,22	0,6203±0,2516	0,833±0,1209	0,5521±0,2437
	SMOTE	0,9348±0,0751	**0,639±0,2411**	0,9352±0,0771	0,6484±0,2374	0,9368±0,08	0,6623±0,2455	0,9003±0,0965	0,6254±0,2201	0,8896±0,1606	0,6104±0,2427	0,8523±0,1199	0,5696±0,2451
	SMOTE+ENN	0,9411±0,0643	0,6277±0,2502	0,9435±0,0603	0,6478±0,2397	0,9515±0,051	0,6595±0,242	0,907±0,0913	**0,6362±0,2194**	0,8901±0,1947	0,5354±0,3158	0,8569±0,1183	0,5733±0,2458
	SMOTE+TL	0,9447±0,054	**0,6443±0,2372**	0,9468±0,0539	0,6457±0,24	**0,9579±0,0455**	0,656±0,2411	**0,914±0,073**	0,6139±0,2288	**0,9256±0,121**	0,5614±0,2862	0,8678±0,1083	0,5707±0,2439
	SPIDER	0,8892±0,0572	0,6291±0,2426	0,8998±0,0583	0,6461±0,2407	0,9155±0,1179	0,6454±0,2437	0,6535±0,3071	0,4991±0,3418	0,7631±0,3431	0,4833±0,3397	0,6865±0,1752	**0,579±0,2336**
	SPIDER2	0,8996±0,0612	0,6232±0,2434	0,9001±0,058	0,6356±0,2363	0,9076±0,129	0,6436±0,2458	0,676±0,2693	0,5443±0,3067	0,7878±0,3218	0,4862±0,3364	0,688±0,1656	0,5657±0,2363
Treino	Original	0,8654±0,1147	0,8654±0,1147	0,7885±0,2488	0,7885±0,2488	0,7156±0,2682	0,7156±0,2682	0,5311±0,3751	0,4632±0,3929	0,5442±0,3924	0,4429±0,4128	0,6251±0,2486	0,6178±0,2562
	ADASYN	0,9807±0,0281	0,9823±0,0259	0,9728±0,041	0,9749±0,0384	0,961±0,053	0,9614±0,0469	0,8922±0,1031	0,8783±0,1174	0,939±0,1469	0,9103±0,1238	0,8261±0,1329	0,8328±0,1269
	ADOMS	0,9795±0,0271	0,9805±0,0257	0,9707±0,0459	0,973±0,0424	0,9591±0,0498	0,9589±0,0493	0,8932±0,0933	0,8826±0,11	0,936±0,1521	0,9141±0,1148	0,841±0,1248	0,8416±0,117
	AHC	0,9791±0,0293	0,9809±0,0267	0,9732±0,0428	0,9741±0,0397	0,9627±0,0617	0,9587±0,0597	0,9012±0,0931	0,8781±0,1318	0,9568±0,1285	0,9181±0,1048	0,8454±0,1134	0,8505±0,1107
	Borderline-SMOTE1	0,9829±0,0255	0,9836±0,0248	0,9753±0,038	0,9765±0,0356	0,9646±0,0495	0,9666±0,0433	**0,9155±0,1017**	**0,9147±0,0896**	**0,9671±0,1029**	**0,9323±0,0894**	**0,8734±0,112**	**0,878±0,1089**
	Borderline-SMOTE2	0,9828±0,0259	0,9836±0,0248	0,9751±0,0394	0,9765±0,0356	0,9646±0,0513	0,9666±0,0433	0,913±0,1033	**0,9147±0,0896**	**0,9627±0,1202**	**0,9323±0,0894**	**0,873±0,1126**	**0,878±0,1089**
	CBO+*Random*	**0,9892±0,0225**	**0,9918±0,0186**	**0,9874±0,0248**	**0,9898±0,0198**	**0,9822±0,0611**	0,9707±0,0697	0,9069±0,0951	0,8914±0,1081	0,654±0,2832	0,9249±0,0988	0,8538±0,1078	0,8637±0,0953
	CBO+SMOTE	0,9841±0,0247	0,9857±0,023	0,9799±0,0343	0,9819±0,0306	0,9668±0,0619	0,9676±0,0601	0,9131±0,0821	0,894±0,0941	0,9417±0,1788	0,8922±0,1384	0,8621±0,1022	0,8683±0,0948
	MWMOTE	0,9764±0,0277	0,9774±0,0274	0,969±0,0432	0,9694±0,0433	0,9489±0,0526	0,9534±0,0513	0,9107±0,0818	0,8973±0,1274	0,936±0,114	0,9057±0,1345	0,8673±0,1054	0,8698±0,107
	ROS	**0,9862±0,028**	0,9886±0,0251	0,9811±0,0382	0,9835±0,0334	0,9773±0,0652	0,9594±0,0687	0,8969±0,0936	0,8824±0,1129	0,8698±0,1877	0,9302±0,1025	0,839±0,1201	0,8434±0,1196
	Safe-Level-SMOTE	0,986±0,029	**0,9888±0,0259**	0,9803±0,0421	0,9836±0,0341	**0,9824±0,0627**	0,9587±0,0716	0,8934±0,1224	0,8815±0,1103	0,7803±0,2224	0,9277±0,1039	0,8388±0,1199	0,8434±0,1179
	SMOTE	0,9808±0,027	0,9825±0,0255	0,9741±0,0415	0,9762±0,0384	0,9612±0,0524	0,9634±0,0463	0,9066±0,0922	0,8965±0,1087	0,9464±0,1331	0,9092±0,1057	0,8559±0,118	0,8573±0,1145
	SMOTE+ENN	0,9854±0,018	0,9868±0,0169	**0,9823±0,0231**	**0,985±0,0197**	0,9719±0,0287	**0,9754±0,0272**	0,9134±0,0884	0,9029±0,1043	0,9371±0,1788	0,7881±0,3391	0,8611±0,1165	0,8647±0,1141
	SMOTE+TL	0,9844±0,0187	0,9863±0,016	0,9818±0,0247	0,9836±0,0225	0,9748±0,0291	**0,9762±0,0256**	**0,9212±0,0697**	0,9125±0,0833	0,9514±0,1201	0,8901±0,1328	0,8716±0,1072	0,8747±0,0998
	SPIDER	0,9642±0,034	0,974±0,027	0,9672±0,0434	0,9743±0,0318	0,9616±0,1038	0,9384±0,0733	0,6758±0,3089	0,5734±0,3567	0,8306±0,3669	0,6429±0,4128	0,7013±0,1742	0,7128±0,1695
	SPIDER2	0,9698±0,0285	0,977±0,024	0,9679±0,0349	0,9726±0,0287	0,9506±0,12	0,9435±0,0819	0,7009±0,2713	0,63±0,3162	0,8467±0,3413	0,6562±0,4003	0,7055±0,1659	0,7198±0,1626

Tabela B.4: SEN de teste e treino média para todos os algoritmos de sobre-amostragem e métodos de classificação, considerando ambas as fases. Os melhores valores para cada coluna encontram-se a negrito.

		CART		**C4.5**		**k-NN**		**SVM Linear**		**SVM RBF**		**NB**	
	Método	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2
Teste	Original	0,6154±0,2649	0,6154±0,2649	0,6013±0,3044	0,5943±0,3095	0,5679±0,3246	0,5679±0,3246	0,4624±0,3514	0,3979±0,3634	0,3971±0,3712	0,3232±0,3688	0,6701±0,243	0,6545±0,2607
	ADASYN	0,9359±0,0848	0,682±0,2176	0,9515±0,0724	0,7196±0,2212	0,9648±0,0898	0,781±0,1859	0,9094±0,1146	0,8564±0,1266	0,8526±0,2046	0,7676±0,2341	0,8338±0,1708	0,7663±0,1777
	ADOMS	0,9329±0,0839	0,6853±0,2503	0,94±0,081	0,7247±0,2398	0,9516±0,0934	0,77±0,1877	0,8921±0,1047	0,8541±0,1407	0,8762±0,1786	**0,7713±0,225**	0,8424±0,1551	0,7965±0,1759
	AHC	0,9363±0,0811	0,6646±0,2486	0,9424±0,0785	0,71±0,2379	0,9559±0,1056	0,7633±0,211	0,9112±0,1074	0,8405±0,1616	0,883±0,1991	0,7349±0,2438	0,8525±0,1425	0,7953±0,1438
	Borderline-SMOTE1	0,9464±0,0757	0,6468±0,244	0,9519±0,0734	0,6971±0,2433	0,9613±0,0848	0,7243±0,2319	**0,9265±0,1197**	0,8331±0,1478	0,8962±0,1639	0,688±0,2832	**0,8864±0,1335**	0,7266±0,1746
	Borderline-SMOTE2	0,9438±0,0779	0,6468±0,244	0,952±0,0683	0,6971±0,2433	0,962±0,0802	0,7243±0,2319	0,9211±0,1227	0,8331±0,1478	0,8946±0,1685	0,688±0,2832	**0,8871±0,1335**	0,7266±0,1746
	CBO+*Random*	**0,98±0,0449**	0,6412±0,247	**0,9822±0,0442**	0,6775±0,2432	**0,9709±0,0898**	0,6984±0,239	0,9131±0,1127	0,8085±0,2187	0,6489±0,2748	0,7105±0,2568	0,8666±0,1309	0,776±0,1867
	CBO+SMOTE	0,9515±0,067	0,6949±0,2182	0,9579±0,066	0,7374±0,1915	0,9621±0,0943	0,7685±0,1931	0,9199±0,0974	0,8167±0,1948	0,9007±0,2146	0,7059±0,2588	0,8741±0,1304	0,7491±0,1816
	MWMOTE	0,9186±0,09	**0,7108±0,2165**	0,9349±0,0783	**0,781±0,1762**	0,9591±0,0698	**0,8209±0,1613**	0,9257±0,0906	**0,86±0,1482**	**0,9089±0,135**	0,7427±0,2489	0,8789±0,1223	0,766±0,1506
	ROS	0,9714±0,0648	0,6249±0,265	**0,9738±0,0651**	0,682±0,2423	0,9635±0,1015	0,7071±0,2418	0,9067±0,108	0,8483±0,1568	0,8287±0,2117	0,7357±0,2579	0,8465±0,1493	0,7925±0,1622
	Safe-Level-SMOTE	**0,9728±0,0673**	0,6368±0,2536	0,9731±0,0726	0,691±0,2376	0,9674±0,0995	0,7153±0,2368	0,9002±0,1299	0,8504±0,1533	0,7801±0,1823	0,7563±0,2454	0,8456±0,1474	**0,8068±0,1551**
	SMOTE	0,9397±0,0759	0,6875±0,2249	0,9449±0,0771	0,7296±0,2134	0,9598±0,0903	0,7745±0,1909	0,9158±0,1065	0,855±0,1325	0,874±0,1791	0,7323±0,2498	0,8665±0,1391	0,7851±0,1476
	SMOTE+ENN	0,9412±0,0667	0,6896±0,2335	0,9497±0,0603	0,7281±0,2181	0,9695±0,045	0,777±0,1934	0,9149±0,1037	0,8484±0,1456	0,8817±0,2081	0,6359±0,331	0,8689±0,1393	0,779±0,1524
	SMOTE+TL	0,9477±0,0543	**0,7374±0,2012**	0,9542±0,0506	**0,7573±0,2036**	**0,9745±0,0461**	**0,8174±0,1536**	**0,9273±0,0806**	**0,8778±0,1206**	**0,9573±0,0871**	**0,834±0,1921**	0,876±0,1335	**0,7987±0,1423**
	SPIDER	0,9291±0,0523	0,6686±0,2403	0,9435±0,0607	0,7022±0,241	0,9361±0,1377	0,7199±0,2157	0,6354±0,3274	0,5581±0,3728	0,7469±0,3451	0,5089±0,3661	0,7349±0,2104	0,7514±0,1877
	SPIDER2	0,9299±0,0595	0,667±0,2445	0,9352±0,0616	0,6977±0,2349	0,9288±0,1474	0,7157±0,2218	0,6619±0,307	0,6323±0,3379	0,7844±0,3262	0,5515±0,3773	0,7344±0,2061	0,768±0,186
Treino	Original	0,8376±0,1469	0,8376±0,1469	0,7511±0,2711	0,7511±0,2711	0,6547±0,2958	0,6547±0,2958	0,4876±0,3705	0,4253±0,3825	0,4991±0,3944	0,4063±0,4055	0,7202±0,2452	0,7119±0,2559
	ADASYN	0,9811±0,0295	0,9826±0,0263	0,9845±0,0308	0,9852±0,029	0,9818±0,0629	**0,9865±0,0451**	0,9143±0,1128	0,8961±0,1408	0,9451±0,1392	0,9148±0,1531	0,8391±0,1705	0,8464±0,1621
	ADOMS	0,9792±0,0296	0,9797±0,0283	0,9735±0,0494	0,9756±0,0456	0,9731±0,0621	0,9772±0,047	0,899±0,1018	0,8925±0,1274	0,9431±0,133	0,9154±0,1429	0,8496±0,1564	0,8676±0,1439
	AHC	0,9773±0,0312	0,9783±0,0291	0,9763±0,0416	0,9772±0,0371	0,972±0,0855	0,9773±0,0718	0,9163±0,1027	0,8915±0,1516	0,9575±0,1336	0,9156±0,1306	0,8578±0,1405	0,8622±0,1337
	Borderline-SMOTE1	0,9803±0,0294	0,9813±0,0269	0,979±0,0367	0,9793±0,0347	0,9715±0,0652	0,9778±0,036	0,9302±0,1157	**0,9368±0,1008**	**0,9727±0,0883**	0,9316±0,1208	**0,89±0,1315**	**0,8951±0,1251**
	Borderline-SMOTE2	0,9808±0,0278	0,9813±0,0269	0,9782±0,0402	0,9793±0,0347	0,9724±0,063	0,9778±0,036	0,9252±0,1193	**0,9368±0,1008**	0,9676±0,1046	0,9316±0,1208	**0,8898±0,1319**	**0,8951±0,1251**
	CBO+*Random*	**0,9916±0,0202**	**0,9941±0,015**	**0,9927±0,0207**	**0,9944±0,0174**	0,9813±0,0825	0,9778±0,09	0,9159±0,1102	0,8951±0,1308	0,6744±0,2857	0,9279±0,1128	0,8692±0,129	0,8787±0,1092
	CBO+SMOTE	0,9853±0,024	0,9869±0,0218	0,985±0,0294	0,9875±0,0259	0,9739±0,0825	0,9757±0,0767	0,9231±0,0943	0,8989±0,1197	0,9415±0,1885	0,8853±0,1647	0,8755±0,1301	0,8801±0,1087
	MWMOTE	0,9762±0,0307	0,9771±0,03	0,9763±0,0443	0,9769±0,0457	0,9757±0,0447	0,9767±0,0443	**0,9306±0,086**	0,9222±0,1366	0,943±0,1037	0,8998±0,1637	0,8838±0,1213	0,8873±0,1242
	ROS	0,9892±0,0265	0,9913±0,022	**0,9886±0,0336**	0,9909±0,0296	0,9769±0,0868	0,9731±0,0829	0,9112±0,1062	0,8923±0,1364	0,8725±0,169	0,9343±0,1199	0,8523±0,1488	0,8586±0,1468
	Safe-Level-SMOTE	**0,9898±0,026**	**0,9918±0,0225**	0,9884±0,038	**0,9923±0,0313**	0,9795±0,0871	0,9742±0,0848	0,9064±0,1263	0,8921±0,1325	0,8248±0,1715	**0,9367±0,12**	0,852±0,1483	0,8589±0,1439
	SMOTE	0,9813±0,0274	0,9825±0,0267	0,9803±0,0412	0,984±0,0337	0,9767±0,0652	0,9836±0,0457	0,9214±0,1037	0,9154±0,1246	0,9449±0,1286	0,9036±0,142	0,8704±0,1386	0,8707±0,1347
	SMOTE+ENN	0,9844±0,0199	0,9875±0,0163	0,9859±0,0248	0,9886±0,0201	**0,9818±0,0288**	0,985±0,0233	0,9203±0,1022	0,9125±0,1248	0,9366±0,1797	0,7848±0,3428	0,8729±0,1386	0,8744±0,1367
	SMOTE+TL	0,9851±0,018	0,9864±0,0168	0,9859±0,0249	0,9875±0,0234	**0,9847±0,0343**	**0,9871±0,0231**	**0,9335±0,0776**	0,9317±0,0956	**0,9776±0,083**	**0,9417±0,113**	0,8794±0,1329	0,8806±0,1258
	SPIDER	0,9775±0,0272	0,9862±0,0199	0,9805±0,044	0,9879±0,0293	0,9597±0,1254	0,9738±0,0966	0,6514±0,3302	0,546±0,3699	0,8322±0,3673	0,6335±0,4219	0,7502±0,2076	0,7603±0,1973
	SPIDER2	0,9778±0,0251	0,9815±0,0191	0,9793±0,0304	0,9827±0,0231	0,9547±0,1382	0,969±0,1109	0,6799±0,3084	0,6062±0,3393	0,8561±0,343	0,655±0,4103	0,7524±0,2047	0,7664±0,1963

Tabela B.5: *F-1* de teste e treino média para todos os algoritmos de sobre-amostragem e métodos de classificação, considerando ambas as fases. Os melhores valores para cada coluna encontram-se a negrito.

		CART		**C4.5**		**k-NN**		**SVM Linear**		**SVM RBF**		**NB**	
	Método	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2
Teste	Original	0,7207±0,2283	0,7207±0,2283	0,6924±0,2864	0,6843±0,2943	0,6651±0,3075	0,6651±0,3075	0,5487±0,368	0,4721±0,391	0,4767±0,3852	0,388±0,394	0,7417±0,1962	0,7245±0,2241
	ADASYN	0,9312±0,08	0,7709±0,1715	0,9313±0,0814	0,7885±0,1737	0,9302±0,0819	0,8273±0,1393	0,8811±0,1072	0,8504±0,1132	0,8593±0,2137	**0,7995±0,2024**	0,812±0,1395	0,7679±0,1592
	ADOMS	0,9291±0,0828	0,7636±0,2087	0,9316±0,0809	0,7888±0,1964	0,9273±0,0846	0,8241±0,144	0,8829±0,1002	0,8526±0,1208	0,8643±0,2211	**0,8027±0,1912**	0,8261±0,1347	0,774±0,1722
	AHC	0,9329±0,0781	0,753±0,2066	0,9345±0,0773	0,7853±0,1885	0,935±0,0817	0,819±0,1584	0,8904±0,101	0,8382±0,1457	0,8962±0,1809	0,7806±0,2008	0,8291±0,129	**0,8011±0,1345**
	Borderline-SMOTE1	0,9436±0,0765	0,745±0,1994	0,9439±0,0764	0,772±0,2046	0,9451±0,0737	0,7942±0,1814	0,9051±0,1164	0,8448±0,1225	**0,9058±0,1717**	0,7407±0,2447	**0,8602±0,127**	0,7628±0,1514
	Borderline-SMOTE2	0,943±0,0745	0,745±0,1994	0,9447±0,0731	0,772±0,2046	0,9456±0,0728	0,7942±0,1814	0,9028±0,1172	0,8448±0,1225	**0,9022±0,1842**	0,7407±0,2447	**0,8616±0,1249**	0,7628±0,1514
	CBO+*Random*	**0,9691±0,0528**	0,7413±0,2024	**0,9685±0,0513**	0,758±0,2074	**0,9653±0,0685**	0,7765±0,1877	0,9012±0,0971	0,8031±0,2101	0,551±0,323	0,7452±0,2428	0,843±0,1187	0,7599±0,1693
	CBO+SMOTE	0,9482±0,0644	0,7781±0,1726	0,9488±0,0665	0,8011±0,1562	0,9497±0,0722	0,8245±0,1462	**0,9058±0,0897**	0,8131±0,1865	0,9001±0,199	0,7372±0,2366	0,8513±0,1144	0,7478±0,1586
	MWMOTE	0,9166±0,0852	**0,7847±0,174**	0,9246±0,077	**0,8201±0,1503**	0,9187±0,0848	**0,8492±0,1256**	0,899±0,0902	**0,8532±0,1367**	0,8954±0,165	0,7849±0,2133	0,8519±0,1217	0,7838±0,1408
	ROS	**0,9536±0,0696**	0,7235±0,2242	**0,9511±0,0735**	0,7664±0,1948	0,9512±0,0815	0,7803±0,1883	0,8869±0,0977	0,8481±0,1364	0,792±0,2825	0,7804±0,2204	0,824±0,1322	0,7901±0,1463
	Safe-Level-SMOTE	0,953±0,0744	0,7358±0,2065	0,9488±0,078	0,771±0,1894	**0,9546±0,0786**	0,7846±0,1841	0,8824±0,1256	0,8479±0,1334	0,6465±0,3228	0,7918±0,21	0,8224±0,1327	**0,8023±0,1435**
	SMOTE	0,9341±0,0755	0,7769±0,1732	0,934±0,0782	0,7998±0,1676	0,9338±0,0782	0,8284±0,1443	0,8966±0,0987	**0,8542±0,1199**	0,8821±0,1925	0,7761±0,2068	0,841±0,1333	0,7912±0,1407
	SMOTE+ENN	0,9424±0,0622	0,7684±0,1881	0,9443±0,0592	0,7991±0,172	0,95±0,0541	0,8294±0,1448	**0,9082±0,0866**	0,8502±0,1471	0,8884±0,2081	0,6701±0,337	0,8501±0,1282	0,7909±0,1398
	SMOTE+TL	0,9382±0,0671	**0,7994±0,1619**	0,9388±0,0717	**0,8098±0,1604**	0,9498±0,059	**0,8405±0,1313**	0,8956±0,1258	0,8403±0,1692	0,8896±0,2314	0,6831±0,3447	0,849±0,1298	0,7916±0,141
	SPIDER	0,9306±0,0587	0,7538±0,1934	0,9366±0,0658	0,7733±0,1893	0,9398±0,1065	0,7838±0,1671	0,6887±0,3094	0,5777±0,3782	0,7763±0,3479	0,5448±0,3707	0,7754±0,1471	0,7734±0,1633
	SPIDER2	0,9292±0,0657	0,7491±0,1934	0,9288±0,0713	0,7661±0,1829	0,9262±0,1277	0,7819±0,1713	0,7088±0,2766	0,63±0,3553	0,8007±0,3295	0,5557±0,3786	0,7671±0,1458	0,7749±0,1637
Treino	Original	0,9045±0,0877	0,9045±0,0877	0,8212±0,2389	0,8212±0,2389	0,7651±0,2356	0,7651±0,2356	0,5752±0,3811	0,5017±0,4047	0,5805±0,3933	0,4725±0,421	0,782±0,1826	0,7729±0,2001
	ADASYN	0,9807±0,0282	0,9823±0,0259	0,9714±0,0443	0,9737±0,0412	0,959±0,0523	0,9583±0,0506	0,888±0,1035	0,8742±0,1169	0,923±0,1999	0,9111±0,1159	0,8176±0,1352	0,8246±0,1305
	ADOMS	0,9796±0,0269	0,9805±0,0255	0,9706±0,0455	0,9729±0,0423	0,9579±0,0496	0,9569±0,0518	0,8916±0,0937	0,8789±0,1108	0,9181±0,2123	0,9143±0,1073	0,8325±0,1315	0,821±0,1422
	AHC	0,9792±0,0292	0,981±0,0264	0,9729±0,0434	0,9736±0,0411	0,9625±0,0564	0,9568±0,0589	0,8979±0,0938	0,8749±0,1258	0,9514±0,1482	0,9187±0,0974	0,8346±0,1239	0,841±0,12
	Borderline-SMOTE1	0,983±0,0253	0,9837±0,0247	0,9746±0,0401	0,9761±0,0365	0,9642±0,0477	0,9651±0,0463	**0,9119±0,1104**	**0,91±0,0948**	**0,957±0,1464**	**0,9322±0,087**	**0,8648±0,121**	**0,8697±0,1178**
	Borderline-SMOTE2	0,9829±0,0259	0,9837±0,0247	0,9748±0,0403	0,9761±0,0365	0,9639±0,0507	0,9651±0,0463	0,9097±0,1116	**0,91±0,0948**	**0,9526±0,1653**	**0,9322±0,087**	**0,8644±0,1216**	**0,8697±0,1178**
	CBO+*Random*	**0,9891±0,0229**	**0,9917±0,019**	**0,987±0,0258**	**0,9896±0,0203**	**0,9831±0,0541**	0,9711±0,0631	0,9052±0,0926	0,8901±0,1045	0,5693±0,3354	0,9232±0,0993	0,8468±0,1141	0,857±0,1016
	CBO+SMOTE	0,984±0,025	0,9857±0,0233	0,9794±0,0363	0,9814±0,0319	0,967±0,0557	0,9675±0,0557	0,9102±0,0856	0,892±0,0946	0,9402±0,1823	0,8923±0,1343	0,8539±0,1109	0,8615±0,1022
	MWMOTE	0,9764±0,0275	0,9774±0,0272	0,9685±0,0437	0,9689±0,0433	0,945±0,0572	0,9501±0,0556	0,9066±0,085	0,8918±0,1244	0,9251±0,1551	0,908±0,1222	0,857±0,1171	0,8601±0,1169
	ROS	**0,986±0,0283**	0,9884±0,0255	0,9804±0,0399	0,983±0,0347	0,9782±0,0593	0,9585±0,0654	0,8936±0,0951	0,8791±0,1126	0,8296±0,2679	0,9302±0,0972	0,8297±0,1275	0,8345±0,1259
	Safe-Level-SMOTE	0,9858±0,0295	**0,9886±0,0265**	0,9797±0,0436	0,9831±0,0352	**0,9835±0,0558**	0,9576±0,0687	0,8909±0,1174	0,8782±0,1094	0,679±0,3344	0,9265±0,0993	0,8296±0,1273	0,834±0,1261
	SMOTE	0,9808±0,0271	0,9825±0,0255	0,9736±0,042	0,9754±0,0401	0,96±0,0508	0,9612±0,0487	0,9036±0,0929	0,8907±0,1128	0,9353±0,1795	0,9101±0,0986	0,8457±0,1278	0,8474±0,1243
	SMOTE+ENN	0,9858±0,0173	0,9872±0,0164	**0,9829±0,0221**	**0,9854±0,0189**	0,9717±0,0289	**0,9756±0,0266**	**0,915±0,0831**	0,8971±0,1347	0,9319±0,2	0,7873±0,3442	0,8552±0,123	0,8597±0,1185
	SMOTE+TL	0,9825±0,0226	0,9848±0,0192	0,979±0,0305	0,9812±0,0274	0,971±0,0348	**0,9712±0,0364**	0,9037±0,1247	0,8871±0,158	0,9262±0,2256	0,7774±0,3511	0,8546±0,1233	0,8584±0,1163
	SPIDER	0,9777±0,0269	0,9844±0,0212	0,978±0,0415	0,9826±0,0328	0,9713±0,0851	0,9595±0,0704	0,7072±0,3057	0,592±0,3632	0,8335±0,3673	0,6462±0,4148	0,7897±0,1426	0,7956±0,1401
	SPIDER2	0,9788±0,0253	0,983±0,0216	0,9759±0,0377	0,979±0,0332	0,9603±0,1023	0,9588±0,0753	0,7282±0,2764	0,6327±0,3403	0,8519±0,3458	0,6448±0,4141	0,7828±0,1432	0,7912±0,1416

Tabela B.6: Média das métricas de complexidade para os conjuntos de treino e teste, fase e técnica de sobre-amostragem, antes da normalização. Os melhores valores para cada coluna encontram-se a negrito.

		F1		F2		F3		L1		L2		L3		N1		N2		N3		N4	
	Método	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2	Fase 1	Fase 2
Treino	Original	2.1388	2.1388	**0.0432**	0.0432	**0.4835**	**0.4835**	**0.1788**	**0.1788**	**0.0501**	**0.0501**	0.4784	0.4784	0.0638	0.0638	0.3283	0.3283	0.0371	0.0371	0.2111	0.2111
	ADASYN	**2.8724**	2.6703	0.0692	0.0432	0.2986	0.3154	0.6324	0.6317	0.1631	0.1550	0.1351	0.1283	0.0670	0.0623	0.2434	0.2395	0.0312	0.0286	0.1512	0.1494
	ADOMS	2.0806	2.1621	0.0918	0.0630	0.2961	0.3062	0.6399	0.6412	0.1529	0.1515	0.1267	0.1257	0.0753	0.0715	0.2371	0.2345	0.0345	0.0323	0.1577	0.1580
	AHC	2.1836	2.2035	0.0997	0.0873	0.2718	0.2754	0.6280	0.6225	0.1495	0.1454	0.1227	0.1191	0.0653	0.0617	0.2277	0.2266	0.0318	0.0306	0.1739	0.1699
	Borderline-SMOTE1	**2.9052**	**3.0664**	0.0585	0.0432	0.3256	0.3301	0.6027	0.5933	0.1287	0.1214	0.1071	**0.1015**	**0.0491**	0.0454	0.2246	0.2224	0.0242	0.0220	**0.1317**	**0.1283**
	Borderline-SMOTE2	**2.9065**	**3.0664**	0.0586	0.0432	0.3229	0.3301	0.6020	0.5934	0.1290	0.1214	0.1068	0.1017	**0.0489**	**0.0453**	0.2245	0.2224	0.0242	0.0220	**0.1317**	**0.1274**
	CBO+*Random*	2.2135	2.3043	0.0662	0.0432	0.3331	0.3643	0.6722	0.6591	0.1324	**0.1204**	**0.1052**	**0.0963**	**0.0389**	**0.0335**	**0.1194**	**0.1029**	**0.0174**	**0.0167**	0.1972	0.1982
	CBO+SMOTE	2.4706	2.6546	0.0666	0.0432	0.3537	0.3592	0.6683	0.6543	**0.1177**	**0.1128**	**0.0938**	**0.0898**	0.0579	0.0565	**0.1929**	**0.1865**	0.0311	0.0297	0.1528	0.1462
	MWMOTE	2.7732	**2.8668**	0.0648	0.0438	0.3058	0.3189	0.6202	0.6140	0.1335	0.1328	0.1095	0.1085	0.0978	0.0898	0.2990	0.2841	0.0504	0.0457	**0.1167**	**0.1115**
	ROS	2.1228	2.2066	0.0678	0.0432	0.3203	0.3394	0.6388	0.6366	0.1589	0.1537	0.1336	0.1284	0.0517	**0.0453**	0.1930	0.1916	0.0239	0.0209	0.2007	0.2040
	Safe-Level-SMOTE	2.0719	2.1335	0.0690	0.0432	0.3151	0.3358	0.6342	0.6334	0.1543	0.1506	0.1281	0.1243	0.0525	0.0458	**0.1925**	**0.1906**	0.0238	**0.0207**	0.2017	0.2040
	SMOTE	2.4545	2.4999	0.0661	0.0432	0.3218	0.3371	0.6239	0.6236	0.1396	0.1368	0.1154	0.1146	0.0643	0.0597	0.2334	0.2303	0.0293	0.0266	0.1478	0.1430
	SMOTE+ENN	2.4663	2.5775	**0.0559**	**0.0334**	0.3443	0.3611	0.6334	0.6290	0.1336	0.1306	**0.1063**	0.1040	0.0686	0.0657	0.2796	0.2791	0.0284	0.0271	0.1377	0.1314
	SMOTE+TL	2.4804	2.5437	0.0590	**0.0372**	0.3264	0.3420	0.6186	0.6167	0.1320	0.1295	0.1138	0.1130	0.0632	0.0623	0.2705	0.2713	0.0270	0.0262	0.1355	0.1311
	SPIDER	1.9720	2.0141	**0.0558**	**0.0388**	**0.4323**	**0.4547**	**0.3309**	**0.3422**	**0.1212**	0.1239	0.4789	0.4725	0.0625	0.0555	0.2992	0.2935	**0.0205**	**0.0179**	0.2229	0.2266
	SPIDER2	1.9281	1.9923	0.0587	0.0441	**0.4097**	**0.4342**	**0.3814**	**0.3880**	0.1418	0.1461	0.4748	0.4698	0.0644	0.0569	0.2993	0.2931	**0.0233**	**0.0207**	0.2331	0.2345
Teste	Original	**3.4017**	3.4017	**0.0030**	0.0030	**0.6996**	0.6996	**0.1273**	0.1273	**0.0518**	0.0518	0.4930	0.4930	**0.0842**	0.0842	0.3788	0.3788	0.0464	0.0464	0.1800	0.1800
	ADASYN	3.1674	3.4017	0.0484	0.0030	0.3461	0.6996	0.6193	0.1273	0.1817	0.0518	0.1570	0.4930	0.1323	0.0842	0.3359	0.3788	0.0555	0.0464	0.1365	0.1800
	ADOMS	2.2503	3.4017	0.0704	0.0030	0.3483	0.6996	0.6037	0.1273	0.1699	0.0518	0.1416	0.4930	0.1454	0.0842	0.3383	0.3788	0.0590	0.0464	0.1447	0.1800
	AHC	2.3252	3.4017	0.0624	0.0030	0.3500	0.6996	0.5962	0.1273	0.1673	0.0518	0.1393	0.4930	0.1322	0.0842	0.2976	0.3788	0.0535	0.0464	0.1485	0.1800
	Borderline-SMOTE1	3.0859	3.4017	0.0248	0.0030	0.4097	0.6996	0.5710	0.1273	**0.1395**	0.0518	**0.1202**	0.4930	**0.0959**	0.0842	0.3014	0.3788	**0.0387**	0.0464	**0.1145**	0.1800
	Borderline-SMOTE2	3.1301	3.4017	**0.0222**	0.0030	0.4086	0.6996	0.5721	0.1273	0.1432	0.0518	0.1243	0.4930	0.0961	0.0842	0.2996	0.3788	**0.0391**	0.0464	**0.1147**	0.1800
	CBO+*Random*	2.2975	3.4017	0.0573	0.0030	0.3566	0.6996	0.5673	0.1273	0.1573	0.0518	**0.1235**	0.4930	**0.0949**	0.0842	**0.1879**	0.3788	**0.0326**	0.0464	0.1659	0.1800
	CBO+SMOTE	2.5545	3.4017	0.0479	0.0030	0.3906	0.6996	0.5558	0.1273	0.1522	0.0518	**0.1234**	0.4930	0.1027	0.0842	0.2619	0.3788	0.0466	0.0464	0.1425	0.1800
	MWMOTE	2.9473	3.4017	0.0336	0.0030	0.3629	0.6996	0.5945	0.1273	0.1536	0.0518	0.1314	0.4930	0.1539	0.0842	0.4004	0.3788	0.0756	0.0464	**0.1175**	0.1800
	ROS	2.3109	3.4017	0.0510	0.0030	0.3569	0.6996	0.6020	0.1273	0.1705	0.0518	0.1446	0.4930	0.1220	0.0842	**0.2570**	0.3788	0.0438	0.0464	0.1640	0.1800
	Safe-Level-SMOTE	2.2416	3.4017	0.0597	0.0030	0.3518	0.6996	0.5989	0.1273	0.1724	0.0518	0.1451	0.4930	0.1265	0.0842	**0.2561**	0.3788	0.0440	0.0464	0.1642	0.1800
	SMOTE	2.6435	3.4017	0.0368	0.0030	0.3792	0.6996	0.5896	0.1273	0.1587	0.0518	0.1354	0.4930	0.1250	0.0842	0.3237	0.3788	0.0513	0.0464	0.1363	0.1800
	SMOTE+ENN	2.7024	3.4017	0.0331	0.0030	0.3901	0.6996	0.5903	0.1273	0.1681	0.0518	0.1490	0.4930	0.1281	0.0842	0.3631	0.3788	0.0513	0.0464	0.1292	0.1800
	SMOTE+TL	2.6912	3.4017	0.0347	0.0030	0.3872	0.6996	0.5828	0.1273	0.1724	0.0518	0.1575	0.4930	0.1259	0.0842	0.3557	0.3788	0.0514	0.0464	0.1269	0.1800
	SPIDER	**6.1705**	3.4017	**0.0154**	0.0030	**0.5429**	0.6996	**0.2723**	0.1273	**0.1237**	0.0518	0.4982	0.4930	0.1427	0.0842	0.4133	0.3788	0.0671	0.0464	0.1871	0.1800
	SPIDER2	**3.7246**	3.4017	0.0309	0.0030	**0.5031**	0.6996	**0.3176**	0.1273	0.1461	0.0518	0.4978	0.4930	0.1529	0.0842	0.4088	0.3788	0.0649	0.0464	0.1938	0.1800