UNIVERSIDADE D
COIMBRA

Márcia Vanessa Pereira Dias

MODELOS DE PREVISÃO DE QUEBRAS DE FOLHA PARA A INDÚSTRIA DA PASTA E PAPEL

Dissertação no âmbito do Mestrado em Engenharia e Gestão Industrial orientada pelo Professor Doutor Samuel de Oliveira Moniz e apresentada ao Departamento de Engenharia Mecânica da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Julho de 2020

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

DEPARTAMENTO DE
ENGENHARIA MECÂNICA

Modelos de previsão de quebras de folha para a indústria da pasta e papel

Dissertação apresentada para a obtenção do grau de Mestre em Engenharia e Gestão Industrial

Web breaks prediction models for the pulp and paper industry

Autor
Márcia Vanessa Pereira Dias

Orientador
Professor Doutor Samuel Moniz

Júri

Presidente	**Professor Doutor Pedro Mariano Simões Neto** **Professor Auxiliar da Universidade de Coimbra**
Vogais	**Professor Doutor João Nuno G. C. Cavaleiro Correia** **Professor Auxiliar da Universidade de Coimbra** **Professor Doutor Samuel de Oliveira Moniz** **Professor Auxiliar da Universidade de Coimbra**
Orientador	**Professor Doutor Samuel de Oliveira Moniz** **Professor Auxiliar da Universidade de Coimbra**

Colaboração Institucional

The Navigator Company

Coimbra, julho, 2020

Agradecimentos

Esta dissertação não seria possível sem o apoio e colaboração de algumas pessoas, às quais deixo aqui o meu agradecimento.

Em primeiro lugar gostaria de agradecer ao meu orientador, Professor Doutor Samuel Moniz, pela orientação, apoio e conselhos dados ao longo do trabalho.

Ao Professor Doutor Nuno Lourenço gostaria de agradecer pela disponibilidade, pelo conhecimento transmitido e, mais ainda, pelo interesse demonstrado pelo trabalho, sendo uma ajuda fundamental no avanço do mesmo.

Ao Engenheiro Paulo Serrano da *The Navigator Company* pelo voto de confiança e autonomia dada na abordagem ao problema.

Ao Professor Doutor Cristóvão Silva pela disponibilidade demonstrada ao longo do trabalho.

À Engenheira Rita Sousa da *The Navigator Company* gostaria de agradecer pela disponibilidade, pela ajuda na compreensão do processo e pela ajuda fundamental na redução da complexidade do problema.

Por último, mas não menos importante, gostaria de agradecer à *The Navigator Company* pela oportunidade dada.

Resumo

As quebras de folha apresentam-se como um dos maiores problemas na indústria do papel, influenciando não só a *performance* do processo mas também a qualidade e valor do produto final. Com a colaboração da *The Navigator Company*, o presente trabalho visa construir um modelo de previsão do risco de ocorrer uma quebra de folha na máquina de papel *tissue*. Numa primeira fase, procede-se a uma análise do sistema de forma a identificar variáveis importantes ao problema. Segue-se uma análise de dados que procura conhecer a relação entre as variáveis selecionadas e que, para além disso, procura preparar a amostra que será usada na fase de aprendizagem do modelo. Por último, recorre-se a árvores de decisão de forma a construir o modelo de previsão. Este apresenta uma exatidão superior a 80% numa fase de teste com dados novos e desconhecidos para o mesmo. São identificadas sete variáveis com o maior poder decisivo: pH das águas brancas, lâmina de limpeza, gramagem, velocidade, percentagem de *broke* no *hood layer*, vácuo de sucção e pH da pasta *slush*.

Palavras-chave: Quebras de folha, Máquina de papel, Indústria da pasta e papel, Modelos de previsão, Análise de dados, Árvores de decisão.

Abstract

Web breaks are one of the biggest problems in the paper industry, not only affecting the performance of the process, but also the quality and value of the final product. In collaboration with The Navigator Company, this dissertation aims at creating a prediction model for the risk of web breaks in a tissue machine. Firstly, system analysis is made in order to identify the essential characteristics of the problem. This step is followed by a data analysis that seeks to understand the relationships between selected variables. Besides that, it seeks to prepare the dataset for the learning phase of the model. Finally, decision trees are used in order to build the prediction model, which has accuracy greater than 80%. A total of seven variables with the highest importance are identified: pH of white water, cleaning blade, grammage, speed, percentage of broke in the hood layer, suction vacuum, and pH of slush pulp.

Keywords Web breaks, Tissue machine, Pulp and paper industry, Prediction models, Data analysis, Decision trees.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

SIMBOLOGIA E SIGLAS

Simbologia

v'- novo valor da observação

v - valor da observação antiga

μ_A - média de todas as observações da variável A

σ_A - desvio padrão de todas as observações da variável A

Siglas

PC – *Principal component(s)*

PCA – *Principal componentes analysis*

DR – *Dimensionality reduction*

DT – *Decision trees*

FC – Fibra curta

FL – Fibra longa

HL – *Hood layer*

RBF – *Radial basis function*

SVM – *Support vector machine*

TM – *Tissue machine*

WB – *Web break(s)*

YL – *Yankee layer*

1. INTRODUÇÃO

O papel do tipo *tissue* desempenha uma função importante no dia-a-dia da sociedade moderna. É um produto vital para o conforto individual e para a saúde pública, para além de se apresentar como uma escolha sustentável (AF&PA, 2019). O *tissue* é um produto altamente tecnológico e que consegue combinar a suavidade, à capacidade de absorção, à resistência e à leveza.

Durante a sua produção, as quebras de folha (WB na sua denominação anglo-saxónica) são um dos maiores problemas na *performance* e rendimento da máquina de papel (Sorsa, 1992). Isto leva a que os produtores de *tissue* apostem cada vez mais na melhoria das propriedades de resistência, sem que para isso aumentem custos de produção ou ponham em causa a suavidade e ademais características do produto final (Uesaka, 2004).

Para além disso, uma WB traduz-se num aumento do número de defeitos no produto final (bobinas *tissue*), o que leva a que o mesmo tenha um menor valor de mercado. Assim, compreende-se que conseguir prever uma WB poderia trazer um aumento significativo na produtividade e, principalmente, na qualidade e valor das bobinas *tissue*. Contudo, a natureza complexa do processo de fabrico do *tissue* vem dificultar esta tarefa.

Sabe-se que a produção de *tissue* apresenta alta complexidade, que vai desde as características da matéria-prima até à elevada interação não-linear entre as variáveis de funcionamento das máquinas de papel (Niskanen, 2012). Para além disso, a previsão de falha em tempo real também é dificultada por algumas características do processo não poderem ser avaliadas de forma rápida e automática (Blanco et al. 2006).

Este trabalho procura perceber as WB e construir um modelo de previsão para as mesmas. O modelo visa evitar alterações de planos de produção quando não se consegue decifrar qual a causa da má *performance* da máquina, aumentar o tempo de resposta a uma falha de forma a evitá-la e, consequentemente, aumentar a qualidade e valor do produto final.

1.1. Metodologia

A metodologia proposta está dividida em 5 estágios principais (ver figura 1).

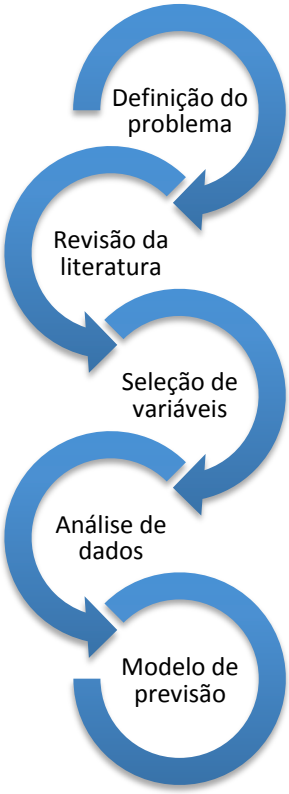


Figura 1. Metodologia proposta para a dissertação.

O **primeiro** estágio procura contextualizar e caracterizar o problema, permitindo uma correta identificação dos objetivos.

Num **segundo** estágio procura-se fazer uma revisão da literatura. Este ponto permite identificar e analisar trabalhos relevantes acerca das WB na indústria do papel. Comparam-se também ferramentas e métodos que são utilizados por outros autores, quais os seus resultados e como podem ser úteis para o caso em estudo.

O **terceiro** estágio foca-se numa seleção das variáveis que possam ter uma grande influência nas WB e que, por isso mesmo, devem ser consideradas no modelo de previsão. Esta seleção é baseada numa análise qualitativa, a qual é possível por:

- um estudo detalhado do processo produtivo;
- uma análise de trabalhos que procuram responder a questões semelhantes na indústria do papel;

- sessões de *brainstorming* com profissionais.

A correta identificação de variáveis é um dos passos fundamentais para a construção de um modelo de previsão viável e preciso (Ahola, 2005).

No **quarto** estágio é feita a análise de dados das variáveis selecionadas. A figura seguinte esquematiza, sucintamente, a metodologia adotada para esta fase.

Figura 2. Metodologia proposta para a análise de dados

O objetivo da análise de dados deste trabalho passa por identificar e compreender as relações que as variáveis têm entre si e com as WB. Para além disso, procura também reduzir a complexidade do problema sem pôr em causa a confiança e qualidade dos resultados finais.

No **quinto** estágio são construídos modelos de previsão de risco de WB para o caso em estudo. É feita uma análise de *performance* de modelos baseados em árvores de decisão e em máquinas de vetores de suporte. Comparar a *performance* de ambos os modelos permite perceber qual dos dois se adapta melhor à estrutura do problema.

2. ENQUADRAMENTO TEÓRICO

Neste capítulo são apresentados os conteúdos que serviram de apoio ao desenvolvimento do trabalho. Numa primeira parte, é feito um enquadramento acerca de aprendizagem computacional. Como se está perante um problema de alta complexidade, analisam-se técnicas de preparação de dados. Segue-se uma revisão dos algoritmos que irão auxiliar a construção do modelo de previsão. Na segunda parte, é feita uma análise de trabalhos desenvolvidos acerca das WB na indústria do papel.

2.1. Aprendizagem computacional

A aprendizagem computacional é um ramo da inteligência artificial e é baseado no conceito de aprender a partir de dados. Esta técnica permite recorrer a algoritmos que aprendem sem a necessidade de serem programados explicitamente. Um dos campos de aplicação da aprendizagem computacional passa pela previsão e análise de acontecimentos futuros. Este campo da ciência da computação procura analisar padrões e gerar hipóteses a partir de uma série de dados históricos, conseguindo alcançar modelos de previsão com desempenho considerável (Praveena e Jaiganesh, 2017).

2.1.1. Preparação de dados

Atualmente, os processos industriais multivariados complexos aliados à crescente aposta na monitorização e controlo dos sistemas resulta em conjuntos de dados cada vez maiores. Assim, há uma crescente preocupação em simplificar a análise desses conjuntos de dados, sem que para isso se modifique ou perca as suas características iniciais (Jollife, 2002).

2.1.1.1. Redução de dimensionalidade

A redução da dimensionalidade, *dimensionality reduction* (DR) na denominação anglo-saxónica, apresenta-se como uma ferramenta poderosa, facilitando, entre outros, a classificação, visualização e compreensão de conjuntos de dados de grandes

dimensões. A DR baseia-se na transformação de um determinado conjunto de dados de grande dimensão em um que apresente um número menor de variáveis e que, mesmo assim, continue a representá-lo fidedignamente (van der Maaten et al., 2009).

Resumidamente, citando a revisão proposta por van der Maaten et al. (2009), "os métodos de DR transformam um conjunto de dados X com dimensionalidade D em um novo conjunto de dados Y com dimensionalidade d, enquanto mantêm o máximo possível da geometria dos dados iniciais." Como nem a geometria dos dados iniciais, nem d intrínseco de X são conhecidas à partida, a DR só é conseguida pressupondo algumas dessas propriedades.

Tradicionalmente, a DR era feita usando métodos lineares tais como: (i) PCA, (ii) Análise de Fatores e (iii) Escalonamento Multidimensional. Porém, estes métodos não se adequam a conjuntos de dados não-lineares complexos, o que levou à necessidade de desenvolvimento de outros métodos que possam lidar com essas situações, tais como: (i) Kernal PCA, (ii) *Isomap*, (iii) Máxima Variância de Desdobramento, (iv) Mapeamento de Difusão, entre outros (van der Maaten et al., 2009). O mesmo autor defende ainda que os métodos não-lineares, apesar da sua grande variação, muitas vezes não são capazes de superar os métodos lineares tradicionais, tais como a PCA. Para além destes, existem outros métodos que procuram analisar dados com características distintas.

Assim, é importante realçar que mesmo que os métodos tenham como objetivo a DR, continua a ser necessário perceber o contexto em que serão aplicados, de forma a ser escolhido aquele que melhor se adequa ao conjunto de dados em análise.

2.1.1.2. Análise de componentes principais

A análise de componentes principais ou, em inglês, *principal componentes analysis* (PCA) apresenta-se como um dos métodos mais utilizados e, provavelmente, o mais antigo para a análise estatística multivariada, sendo utilizado nas mais diversas áreas científicas (Abdi, 2010). Karl Pearson desenvolveu a base da PCA no início do século XX (Pearson, K., 1901), porém só mais tarde é que Hotteling, H. (1933) criou o procedimento para a sua implementação geral.

Este método procura transformar um conjunto de dados de alta dimensão em um com uma dimensão inferior sem que haja uma perda significativa de informação. Para tal, como indicado por Wang e Du (2000), as variáveis correlacionadas são projetadas num

plano de variáveis não-correlacionadas, criando os chamados "componentes principais". Dheri et al. (2019) enumera os seguintes passos para elaborar uma PCA:

(1) Normalizar o conjunto de dados;

(2) Calcular a matriz de correlação/covariância;

(3) Calcular os valores e vetores próprios da matriz de correlação;

(4) Calcular os componentes principais (computacionalmente);

(5) Interpretação dos resultados e aplicação.

O vetor próprio associado ao maior valor próprio define o primeiro PC e assim consequentemente (Jollife, 2002), sendo que os primeiros PC encontrados possuem a maior parte da variabilidade presente no conjunto de dados original. Cao et al. (2003) refere as seguintes características inerentes aos PC: são não-correlacionados, têm sequencialmente variâncias máximas e, na representação das entradas originais pelos primeiros PC, o erro de aproximação ao quadrado médio é mínimo. Note-se ainda que a PCA é sensível à escala de medição, o que pode ser contornado utilizando variáveis normalizadas (Dheri et al., 2019).

A PCA apresenta-se como um método com tempo computacional rápido e fácil de implementar, com garantias de encontrar uma representação com uma dimensionalidade mais baixa num subespaço linear, caso exista (Gorban et al., 2008). Por outro lado, a PCA só consegue identificar a variabilidade bruta e não distinguir entre e dentro da variabilidade dos grupos (Gorban et al., 2008). Para além disso, também é sensível a erros grosseiros que possam estar presentes (Liang et al., 2020).

2.1.2. Algoritmos de aprendizagem

Existem vários algoritmos de aprendizagem que procuram dar resposta às características de um dado problema. Por exemplo, as redes neurais são um dos algoritmos que têm tido maior relevância e apresentado bons resultados, sendo um dos mais utilizados recentemente. Para além desse, pode-se recorrer também a algoritmos de florestas aleatórias, *K-nearest neighbors*, *Naive Bayes*, entre outros. Neste trabalho são construídos modelos de classificação baseados em árvores de decisão e em máquinas de vetores de suporte, os quais serão analisados em detalhe mais à frente.

Praveena e Jaiganesh (2017) referem três categorias de aprendizagem:

- supervisionada: onde a variável de resposta é conhecida;
- não-supervisionada: quando não é conhecida a variável de resposta;
- por *reinforcement*: onde se procura ensinar qual a ação que tem mais prioridade numa determinada situação.

A aprendizagem supervisionada é a mais comum quando se tratam de problemas de engenharia (Xu e Yang, 2013) e é a adotada no presente trabalho, visto que se pretendem mapear as ligações entre variáveis independentes e dependentes.

2.1.2.1. Árvores de decisão

As árvores de decisão, *Decision-Trees* (DT) na denominação anglo-saxónica, são usadas frequentemente na literatura enquanto método de aprendizagem computacional (Barak et al., 2017). Este método recorre à subdivisão dos dados em subconjuntos, os quais são separados de acordo com os valores dos dados originais até se alcançar a unidade básica de classificação (Henrique et al., 2019).

Por outras palavras, as DT funcionam como tabelas de decisão, onde se tenta mapear as várias alternativas possíveis e probabilidade de ocorrerem. Na sua utilização prática, as DT são treinadas/classificadas com um conjunto de dados de treino e, posteriormente, podem ser utilizadas para classificar novas medições (Breiman et al., 1984).

Musharraf et al. (2020) indicam que um dos pontos críticos no desenvolvimento de uma DT passa pela alocação do melhor atributo para cada subconjunto. Diferentes algoritmos de DT usam diferentes medidas para seleção dos atributos, sendo que alguns conseguem lidar com dados altamente multivariados.

Neste trabalho recorre-se ao *Gini index* de forma a selecionar os melhores atributos durante a fase de treino. Sucintamente, o *Gini index* procura medir a quantidade de vezes que um elemento seria identificado incorretamente dada uma possível divisão. Assim, percebe-se que quanto menor o seu valor, melhor será a divisão.

Concluindo, dado o cariz de aplicação prática do presente trabalho e dado que este método de aprendizagem computacional é facilmente interpretável, ou seja, o resultado é de fácil compreensão e análise (Hu et al., 2019), apresenta-se como uma boa opção e será o usado para construção do modelo no capítulo 5.

2.1.2.2. Máquina de vetores de suporte

As máquinas de vetores de suporte (SVM na denominação anglo-saxónica) constituem uma técnica estatística não-paramétrica de aprendizagem supervisionada, o que leva a que não seja necessário fazer suposições acerca da distribuição subjacente dos dados a tratar (Mountrakis et al., 2011).

Como referido por Barakat e Bradley (2010), as SVM têm demonstrado uma *performance* de generalização superior comparativamente a muitas outras técnicas de classificação. Porém, os mesmos autores referem também a inabilidade que as SVM têm em dar uma justificação compreensível acerca das soluções que alcançam.

Sucintamente, uma SVM procura classificar um conjunto de variáveis de alta dimensão usando um conjunto de hiperplanos que representam a maior distância mínima que separa todos os pontos dentro de uma classe (Goh et al., 2015). De uma forma mais simples, Lauer e Bloch (2007) explicam que para um conjunto de treino, a margem é definida como a distância mínima entre pontos de duas classes (medidas perpendicularmente ao hiperplano de separação). Maximizar esta margem é uma forma do algoritmo controlar a capacidade da SVM e de selecionar os hiperplanos de divisão ótimos entre duas classes. Mountrakis et al. (2011) indica que a classificação por SVM é conhecida por encontrar o equilíbrio perfeito entre a precisão alcançada nos dados de treino e a capacidade de generalização perante novos dados.

2.2. Quebras de folha na indústria do papel

Esta secção apresenta uma análise de trabalhos relevantes para o caso em estudo e que, na sua maioria, procuram prever ou diagnosticar as WB na indústria da pasta e papel.

Nesta parte da análise da literatura considera-se importante compreender como é que é feita a seleção das variáveis e encontrar potenciais variáveis que podem influenciar as WB. Depois desta primeira seleção, é necessário compreender como é que as variáveis foram tratadas: quais os métodos utilizados, quais os relevantes e quais evitar. Por último, procura-se entender quais os métodos utilizados até agora para a construção de modelos de previsão, quais os mais promissores e quais se adequam ao caso em estudo.

2.2.1. Seleção de variáveis

Uma revisão da literatura dá uma boa indicação das variáveis normalmente utilizadas na construção de modelos de previsão para a indústria do papel. Para além disso, analisar trabalhos relativos a várias partes do processo produtivo dão uma ótima indicação das variáveis que têm maior influência na formação do *tissue* e nas WB.

Alonso et al. (2006) divide as variáveis associadas à folha de papel em três vertentes: propriedades estruturais (gramagem, espessura, porosidade, suavidade, etc.), propriedades mecânicas (tensão, resistência à tensão, alongamento, etc.) e propriedades de aparência (opacidade, transparência, cor, etc).

Uesaka (2005) indica que defeitos macroscópicos (como, por exemplo, os indicados por Niskanen (2012): buracos, rachaduras, cortes laterais, vincos, pontos finos, etc.) não provocam, por norma, WB a não ser que os mesmos ultrapassem um determinado tamanho ou que coincidam com picos de tensão. O mesmo autor sublinha ainda que a maior parte das quebras são um resultado de variações de tensão combinadas com variações da resistência, sendo que a resistência à tensão é um dos fatores mais consistentes na previsão de WB. Parola e Beletski (1999) também defendem a importância que a tensão tem nas quebras baseando-se na distribuição não uniforme da tensão ao longo da folha.

Esta não uniformidade ao longo da folha de papel *tissue* é transversal a outras características, tal como a humidade (Lo Cascio, 2001). Lo Cascio analisou a influência que o filtro da cabeça de máquina pode ter na distribuição de conteúdo de água ao longo da folha e, consequentemente, nível de humidade não-uniforme na zona de secagem.

Note-se ainda que resistência do papel *tissue* pode assumir várias formas, tais como a resistência à tensão, resistência ao rasgo, absorção de energia elástica e tensão de rotura (Niskanen, 2012).

É também relevante perceber a influência indireta que uma WB tem no processo. Ekvall (2004) estudou o efeito que uma WB tem na temperatura do *Yankee*. O *Yankee* é um cilindro aquecido a vapor e que, através da sua superfície quente, auxilia o processo de secagem e de crepagem do papel. O autor explica que a cada WB, a temperatura do yankee é afetada, o que irá influenciar negativamente o processo de secagem e, consequentemente, a humidade presente na folha nos momentos pós-quebra.

Este trabalho é um dos exemplos que sublinham a importância de entender e analisar não só as variáveis diretamente relacionadas com WB, mas também as que podem ser afetadas indiretamente.

No trabalho efetuado por Sorsa et al. (1992) há também a indicação de que a gramagem tem uma grande influência em todas as variáveis, dominando os resultados da análise estatística. Isto levou a que a mesma não fosse considerada pelos mesmos autores na interpretação de resultados.

Mesmo que no caso em estudo as WB ocorram no final da zona de secagem, continua a ser necessário fazer uma análise das zonas mais a montante na máquina. Como exemplo, Dahlquist (2008) sublinha que, mesmo as quebras ocorrendo em zonas de suspensão e baixa humidade, conhecer o nível de humidade ao longo de toda a zona de secagem mostra-se muito vantajoso, permitindo prever antecipadamente quebras que ocorrem mesmo quando o nível de humidade na zona final é o normal. Para além do nível de humidade da folha, a humidade do ambiente envolvente também pode afetar a *performance* da máquina de papel, sendo que Hiltunen et al. (2011) indica que existe uma maior ocorrência de quebras em alturas do ano menos húmidas.

Outros trabalhos funcionam ainda como uma boa base para conhecer o tipo de variáveis que têm sido usadas em modelos de previsão de quebras e qual a importância que lhes tem sido dada, tal como o trabalho realizado por Miyanishi e Shimada (1998) que demonstra a especial importância da consistência das águas brancas, velocidade da máquina e temperatura de entrada na cabeça de máquina, entre muitas outras transversais a todo o processo. Fernandes (2018) indica que é vantajoso não tratar variáveis que sejam calculadas a partir de outras variáveis já selecionadas de forma a evitar informações redundantes na construção do modelo.

Sabe-se que a combinação de todas as variáveis mecânicas e de processo têm uma influência tremenda na *performance* da máquina de papel. Porém, Niskanen (2012) refere aquele que é o fator mais difícil de caracterizar e que maior influência tem nas WB: a variável humana. O autor defende que o manuseamento do processo e o comportamento humano é o fator mais difícil de prever e controlar e que, por isso, mais consequências negativas pode trazer.

2.2.2. Análise de dados

Na análise de dados que precede um modelo de previsão e, principalmente, quando se lida com processos muito complexos, torna-se vantajoso fazer uma seleção e diminuição do número de variáveis a tratar, procurando simplificar o modelo a construir (Dahlquist, 2008).

Antes de efetuar qualquer análise estatística e interpretação de resultados, Chen et al. (2002) sublinham a importância de efetuar uma limpeza dos dados adquiridos do processo real, eliminando *outliers*, etc.

Posteriormente, a relação entre variáveis pode ser analisada recorrendo a um simples teste de correlação. Como indica Fernandes (2018), neste caso é necessário entender o tipo de dados com que se está a lidar, recorrendo-se ao coeficiente de correlação de *Pearson* quando os dados seguem uma distribuição normal. Nos casos em que não se conhece a distribuição, a mesma autora indica o coeficiente de *Spearman* como uma boa alternativa.

Podem ainda surgir outros problemas, tais como os indicados por Dahlquist (2008). O mesmo autor indica que quando dependências não lineares estão presentes ou quando as amostras não são independentes, tal como numa série temporal, o cálculo de correlação pode levar a um valor muito baixo, mesmo quando duas variáveis são altamente correlacionadas, ou então, pelo contrário, pode sobrestimar o mesmo valor. Para contornar este facto, sugere a utilização da teoria de informação ou da informação mútua.

Nagappan (2006), Chen et al. (2002), Blanco (2000), Myianishi (1998), Sorsa (1992), entre outros, efetuaram uma DR recorrendo à PCA. Destes trabalhos pode-se concluir que modelos com alta complexidade costumam ter boa *performance* na fase de treino, mas apresentam dificuldades em adaptarem-se a novas situações, ou seja, não apresentam robustez.

Para além de problemas com robustez e sobreajuste, o uso de muitas variáveis pode também tornar o modelo mais sensível às falhas dos sensores de medição (Chen et al., 2002). O mesmo autor defende ainda a vantagem de utilizar a PCA quando se procura entender qual o tipo de relação entre variáveis e quais as que têm maior influência nas WB, revelando muitas vezes relações que não eram suspeitas inicialmente.

2.2.3. Construção de um modelo de previsão de quebras

Da literatura é importante perceber quais as técnicas utilizadas na construção de modelos de previsão de WB e comparar a sua *performance*.

As redes neurais artificiais são as mais utilizadas para a construção de modelos de previsão confiáveis na indústria de papel, apresentando a melhor *performance* entre os métodos usados (Blanco, 2006; Dahlquist, 2008).

Sorsa et al. (1992) usou uma rede neural não supervisionada de forma a classificar os dados e encontrar situações sensíveis a quebras. Já Miyanishi e Shimada (1998) recorreram às redes neurais de forma a diagnosticar as causas de WB. Outros autores, tal como Bonissone et al. (1999) também recorreram às redes neurais. Estes autores indicam que uma das grandes vantagens das redes neurais é a possibilidade de descrever relações não-lineares entre variáveis, porém apresenta a desvantagem de ser uma estratégia complexa de desenvolvimento.

Porém, mesmo que a literatura aponte para uma melhor *performance* das redes neurais, outros autores recorreram a técnicas diferentes que continuam a apresentar um bom resultado. Chen et al. (2002) recorreram a árvores de decisão de forma a classificar e prever WB. Já outros autores recorreram a florestas aleatórias de forma a dar resposta a uma maior complexidade do problema: o trabalho apresentado por Amruthnath e Gupta (2019) procura fazer uma análise de variáveis recorrendo a um diagnóstico com florestas aleatórias; e Guo et al. (2004) recorreu ao método referido para construir uma previsão robusta de propensão a falhas.

As florestas aleatórias apresentam-se como uma boa escolha quando existem limitações de tempo, ao contrário das redes neurais. Também apresentam versatilidade, podendo ser usadas para classificação ou regressão, capturando a não linearidade entre variáveis dependentes e independentes. Para além disso, não ocorre *overfitting* do modelo se forem criadas árvores suficientes (Chen et al., 2002). Sublinha-se ainda que as árvores de decisão apresentam melhor *performance* no diagnóstico e as florestas aleatórias na previsão.

Por outro lado, quando as causas das falhas são conhecidas, podemos usar soluções antigas para combater os novos problemas que possam surgir. Esta é a ideia por detrás do raciocínio baseado em casos ou, em inglês, *case-based reasoning*. Nakamura

(2007) recorreu a este método porém concluiu que o mesmo apresenta baixa precisão na previsão de falhas. Esta foi também uma conclusão alcançada por Ahola (2006). Tal acontece por não existir informação disponível ou correta acerca das causas das falhas anteriores (Nakamura, 2007).

Devido à natureza binária da variável de resposta, ou seja, ocorrer ou não ocorrer WB, um modelo de regressão logística também pode ser indicado para o caso em estudo, tal como indica Nagappan (2006). O mesmo autor sublinha que este método não deve ser usado com variáveis independentes correlacionadas e, portanto, sugere que a regressão logística seja utilizada em conjunto com uma análise de componentes principais, o que também é defendido por Musa (2014).

Cada método tem as suas vantagens e desvantagens, o que faz com que muitas vezes sejam construídos modelos híbridos, ou seja, modelos que conjugam diversos métodos. Desta forma procura-se colmatar falhas e contornar limitações que possam existir (Dahlquist, 2006).

3. DESCRIÇÃO DO CASO

Neste capítulo é feita a apresentação da empresa onde o trabalho é desenvolvido, o enquadramento do processo produtivo presente no complexo industrial e a descrição do problema.

Compreender o processo de fabrico do *tissue* e discuti-lo com profissionais da área são passos fulcrais deste trabalho, pois leva à correta identificação de variáveis que podem influenciar as WB e, posteriormente, à correta interpretação de resultados.

3.1. A empresa

A *The Navigator Company* é uma presença forte na indústria da pasta e papel a nível mundial. O seu modelo de negócio integra I&D, a floresta, a pasta de celulose, o papel de impressão e escrita, o papel *tissue* e a energia renovável.

Este trabalho é desenvolvido no complexo industrial da *The Navigator Company* em Cacia. Porém, a empresa tem ainda mais três unidades fabris, as quais estão localizadas em Setúbal, Vila Velha de Rodão e Figueira da Foz.

Em Cacia existe produção de pasta e *tissue*, em Vila Velha de Rodão produz-se apenas *tissue* e nas restantes duas, em Setúbal e Figueira da Foz, são produzidos pasta e papel de escrita.

Num universo de mais de 3000 colaboradores diretos e para além das quatro unidades fabris, possui ainda um centro de investigação e três viveiros em solo português. Para além disso, a empresa gere uma área superior a 100 000 hectares de floresta certificada pelo *Forest Stewardship Council*.

3.2. O processo de fabrico

Como referido anteriormente, o complexo industrial da *The Navigator Company* em Cacia integra duas zonas produtivas — uma com a produção de pasta de papel e a outra com a produção e transformação do papel *tissue*.

Na zona de foco deste trabalho, ou seja, na máquina de papel (TM) são produzidas bobinas de *tissue* que se enquadram em 4 gamas:

- Doméstico
- Guardanapos
- Cozinha
- Industrial

Cada produto procura responder às exigências do cliente, sendo que diferem em características tais como o tamanho, a suavidade, número de quebras aceitáveis por bobina, entre outros.

3.2.1. Produção de pasta

Esta parte do complexo industrial produz pasta de fibra curta para venda a clientes externos, para transformação noutras fábricas do grupo e ainda para envio direto via *pipeline* para a parte de produção de *tissue*.

A figura 3 esquematiza, sucintamente, o método de produção da pasta de papel adotado em Cacia.

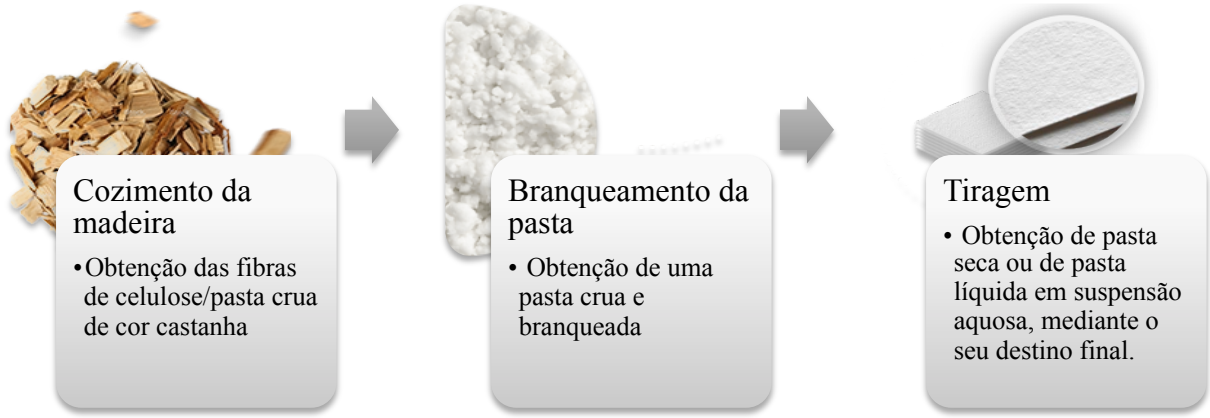


Figura 3. Produção de pasta de papel (adaptado de *thenavigatorcompany.com*)

3.2.2. Produção de papel *tissue*

3.2.2.1. O que é o papel *tissue*?

O *tissue* é um tipo de papel absorvente que se caracteriza por apresentar suavidade ao toque, resistência à tensão, elasticidade, leveza, entre outros.

Atualmente, a necessidade de produzir *tissue* com alta suavidade contrabalança com a necessidade do mesmo ser muito resistente, por exemplo, à tensão ou à humidade. Para combater este problema e como o *tissue* é naturalmente constituído por duas camadas (inferior e superior), procura-se dar suavidade a um dos lados do papel e dar resistência ao outro (Boudreau, 2013).

Como o restante papel, o *tissue* é produzido através de fibras vegetais virgens e fibras recicladas. Estas fibras (curtas ou longas) têm um papel fundamental nas características finais e nas condições de funcionamento da máquina.

A **fibra curta** (FC) está associada à suavidade e portanto é a fibra predileta para a camada superior do *tissue*, ou seja, a que estará em contacto com a pele — nesta camada, a percentagem de FC presente na pasta é superior a 70%, chegando a ser mais de 95% consoante o nível de suavidade que se procura alcançar (Ramaratnam et al., 2013). Esta é uma fibra que provém de árvores como o eucalipto, etc.

A **fibra longa** (FL) está associada à resistência e é o tipo de fibra mais utilizado nas camadas interiores do *tissue* (mais de 70%), as quais não estarão em contacto direto com a pele. A incorporação deste tipo de fibra aumenta a resistência a seco, a resistência ao alongamento e a resistência a húmido do produto final (Hiltunen, 2011). A FL é obtida através de árvores como o pinheiro, etc.

Posto isto, é fácil compreender que um produto de uso doméstico (como o papel higiénico) terá a maior percentagem possível de FC de forma a ser o mais suave possível. Para além disso, no caso em estudo, a matéria-prima da FL é importada maioritariamente de países nórdicos a preços elevados. Assim, existe uma grande preocupação em diminuir o seu consumo sem que para isso seja necessário descuidar a *performance* da TM.

3.2.2.2. O processo

Este ponto resume, de uma forma geral, a produção de papel *tissue* adotada no complexo industrial de Cacia. A pasta pode chegar à parte da produção do tissue através de fardos ou diretamente por *pipeline*. Esta última será denominada de pasta *slush* ao longo do trabalho.

Antes de ser reencaminhada para uma torre de armazenamento, a *slush* passa por um sistema de filtros que retiram o ClO_2 presente e o devolvem à parte da pasta para

reaproveitamento. A torre de armazenamento de *slush* funciona como um *buffer* e combate possíveis quebras de produção na parte da pasta.

Caso chegue na forma de fardos de pasta de FC e FL, os mesmos serão processados nos *pulpers* e seguir para as respetivas torres de armazenamento de FC e FL. Estas duas torres encontram-se conectadas à torre *slush* de forma a ajustar as características das pastas.

Importa referir que a torre de FC alimenta maioritariamente a camada exterior do papel, à qual iremos chamar *yankee layer* (YL), com mais suavidade. Por outro lado, a torre de FL alimenta principalmente a camada interior, a *hood layer* (HL), com mais resistência.

Ainda antes de alcançar a máquina de papel ou, em inglês, *tissue machine* (TM), a pasta passa por um processo de refinação, sendo que na fibra longa este processo pode ser duplo de forma a alcançar as propriedades pretendidas.

O processo de refinação da pasta funciona como um tratamento mecânico com a finalidade de criar mais pontos de ligação entre as fibras – o aumento da superfície de contacto irá traduzir-se num aumento de resistência e numa maior adesão na zona de secagem, entre outras características (Boudreau, 2013). Assim, a pasta de FL é sujeita a dois momentos de refinação de forma a aumentar consideravelmente a sua resistência, enquanto que a fibra curta é sujeita apenas a um. Compreende-se assim que, caso a pasta seja refinada em demasia, a suavidade do produto final estará comprometida.

A pasta é ainda sujeita a um processo de humidificação, do qual resulta uma "pasta" com um nível de humidade de aproximadamente 98%.

Na fase do *blending* — antes da chegada aos *machine chest* da TM — definem-se as percentagens de cada fibra de forma a alcançar as características pretendidas no final do processo.

Depois deste ponto, a pasta está pronta para seguir para as *machine chests* da TM que se dividem conforme a camada de papel que alimentam — a YL ou o HL.

Antes de avançar, importa ainda referir a existência de uma grande preocupação em recuperar toda a fibra possível, quer seja por razões ambientais, quer seja pelos custos do desperdício. Assim, a água utilizada ao longo do processo é sempre reaproveitada. Caso seja necessário desperdiçar alguma água (por limite da capacidade dos

depósitos, por exemplo), será sempre descartada a água mais purificada, pois é aquela que contém menos quantidade de fibra.

3.2.2.3. A máquina de papel *tissue*

A TM é o foco do trabalho desenvolvido, daí ser imprescindível apresentar detalhadamente o seu funcionamento. O modelo utilizado no complexo industrial da *The Navigator Company* em Cacia é o *XCelLine* da *Voith* (figura 4).

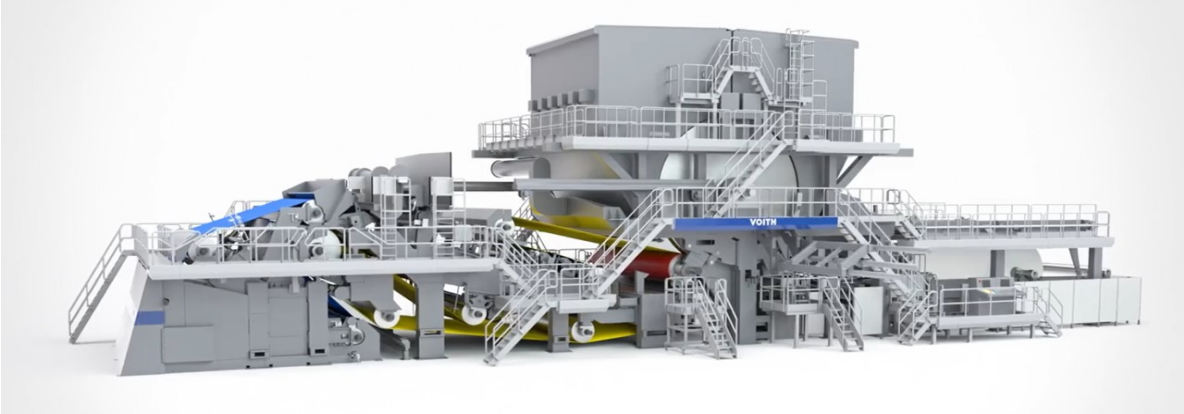


Figura 4. Máquina de papel *tissue XCelLine* da *Voith* (fonte: *voith.com*)

Nesta máquina a pasta é processada em três fases distintas (figura 5).

- Formação (a)
- Secagem (b)
- Bobinagem (c)

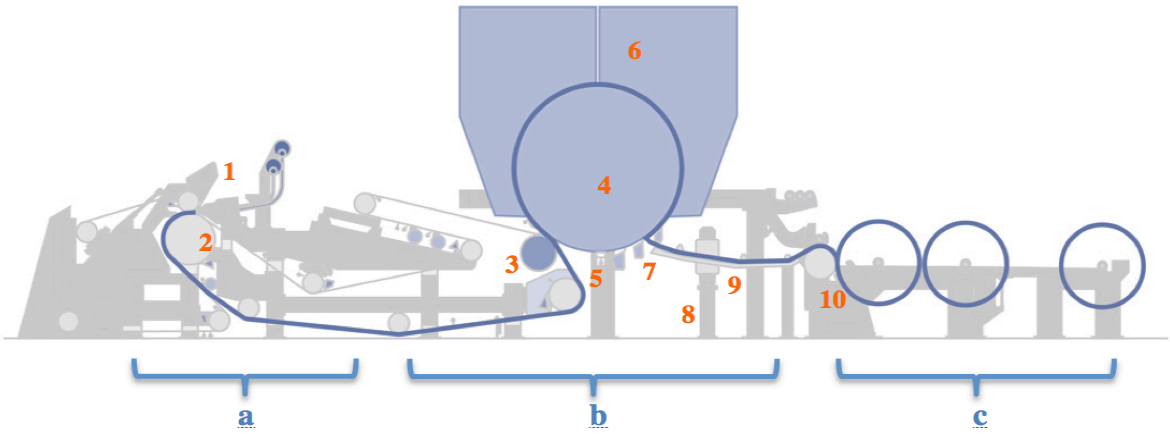


Figura 5. Esquema máquina de papel (adaptado de: *voith.com*)

Uma explicação detalhada da figura anterior é abordada na tabela 1.

Tabela 1. Processo na máquina de papel (adaptado de *voith.com*)

Fase	Componente da TM	Função
Formação	Caixa de entrada *MasterJet Pro T* (1)	Ponto de entrada de pasta na TM, sendo essencial para uma boa formação da folha.
	Crescent Former (2)	A pasta passa pelo rolo de formação, comprimindo a suspensão e iniciando o processo de drenagem.
Secagem	*NipcoFlex T shoe press* (3)	Este ponto é constituído por um processo mecânico de desidratação, apresentando alta eficiência de secagem.
	Cilindro EcoDry Y Yankee (4)	O cilindro *yankee* é aquecido a vapor e leva a que a água evapore da folha de papel quando entra em contacto com o mesmo.
	Químicos (5)	Os químicos funcionam como um revestimento. São aplicados direta e continuamente no *yankee* e têm um papel fundamental para a proteção do mesmo e para a aderência e lançamento do papel. Quatro tipos: lançamento, adesivo, MAP e *edge*.
	Ecohood T (6)	Combinado com o *yankee*, este componente minimiza consideravelmente o consumo de energia durante o processo de secagem.
	Lâminas *Doctor* (7)	Três tipos: Crepe, limpeza e corte. A lâmina de crepe remove a folha do *yankee*; A lâmina de limpeza, tal como o nome indica, limpa a superfície do *yankee*. A lâmina de corte efetua os cortes central e lateral da folha de papel em formação.
	Vácuo de poeiras (8)	Sistema de remoção de poeiras. A remoção das poeiras resultantes do processo resulta num aumento de segurança e melhor ambiente operacional.
Bobinagem	Estabilizadores (9)	Garantem uma marcha uniforme da folha.
	MasterReel T (10)	Enrola a folha de papel em bobinas com diâmetro até 3000mm.

3.3. Descrição do problema

No caso em estudo, as WB trazem muitos problemas ao nível do planeamento de produção e da qualidade do produto final.

Cada cliente especifica quantas quebras por bobina está disposto a aceitar – enquanto alguns clientes têm tolerância zero, outros aceitam até 3 quebras/bobina. Em todos os casos, a presença de quebras numa bobina leva à desvalorização da mesma. Caso a bobina ultrapasse o número de quebras aceitável é "descartada" para a zona dos *pulpers* de forma a ser novamente incorporada no processo, o que se traduz em perda de capacidade e num aumento dos custos de produção.

Para além destes problemas, enquanto a produção não é normalizada, o encaminhamento de papel para os *broke pulpers* leva a que haja necessidade, no futuro, de maior incorporação de *broke* na pasta — influenciando a *performance* da máquina e as características finais do papel.

Sucintamente, as WB podem traduzir-se em perda de tempo de produção, na diminuição do valor de mercado das bobinas produzidas (mais quebras leva a menor valor €) e também na diminuição de qualidade.

Torna-se fácil de compreender que prever e evitar uma WB pode aumentar significativamente a produtividade e eficiência do processo. Porém, dado o *tissue* ser um produto altamente tecnológico e com um processo de fabrico muito complexo, as WB podem ter inúmeras causas. Estas causas podem estar relacionadas com variáveis de:

- funcionamento de máquina (velocidade, etc.);
- receita do papel (gramagem, % FL e FC incorporadas, etc.);
- tratamento do papel (pH, energia utilizada na refinação, etc.);
- manuseamento e influência humana;
- entre outros.

3.4. Análise da situação atual

Os profissionais da máquina de papel do caso em estudo indicam que as WB ocorrem no final da zona de secagem, entre o *Yankee* e a Bobinagem. De acordo com os mesmos, cada WB corresponde, habitualmente, a 5 minutos de paragem. Importa realçar

que as quebras, mesmo ocorrendo só no final da fase de secagem, podem ser provocadas por variáveis mais a montante no processo.

A TM tem capacidade para produzir, aproximadamente, 9.5 ton/hora de papel. A taxa de produção é influenciada por:

- velocidade da máquina [m/s]: capacidade máxima de 2000m/min ≈ 33,3m/s;
- diâmetro da bobina [m]: capacidade máxima suportada de 3m;
- gramagem do *tissue* [g/m^2]: atualmente, existem gamas de 14,70 a 22,00g/m^2;
- largura da bobina [m]: máximo suportado pela máquina é de 5,670m. Atualmente são produzidas bobinas com pelo menos 5,120m;
- comprimento [m]: varia consoante o processo de crepagem, sendo que duas bobinas com o mesmo diâmetro podem ter comprimentos de folha distintos.

Desta forma, podemos quantificar (1) o pior cenário de perdas de produção por quebra (tendo como referência o valor teórico da taxa de produção e valor aproximado de tempo de paragem):

$$Produção\ perdida = 5\ [min] \times \frac{1}{60\ [min]} \times 9.5\ \left[\frac{ton}{h}\right] = 0{,}79\ [ton] \qquad (1)$$

4. ANÁLISE DE DADOS

Neste capítulo pretende-se conhecer quais as variáveis a usar e fazer uma preparação dos dados a usar no modelo de previsão. Para além disso, procura-se fazer uma análise de componentes principais de forma a reduzir a dimensão do problema.

4.1. Seleção das variáveis

Esta secção visa expor o resultado de uma seleção inicial de variáveis baseada numa análise qualitativa do processo. Nesta fase foi dada muita relevância à opinião dos profissionais que convivem diariamente com o processo. Os mesmos dão maior importância às variáveis que estão diretamente ligadas à TM. Assim, variáveis associadas à parte do complexo industrial da produção de pasta não são tidas em conta neste trabalho.

Numa primeira parte, por estudo do processo e por uma cuidada análise do estado da arte (secções apresentadas anteriormente), selecionam-se as variáveis presentes no anexo A. Esta lista representa um conjunto de possíveis fatores que estão diretamente relacionados com a máquina de papel e que, teoricamente, podem ter uma grande influência na *performance* da máquina.

Numa segunda abordagem, procura-se simplificar o problema com a ajuda dos profissionais (especialistas de domínio). Para tal, é feita uma análise qualitativa da lista referida anteriormente e selecionam-se apenas as variáveis que, teoricamente, apresentam um maior poder discriminatório relativamente às WB.

A lista final apresentada na tabela 2 é alcançada através de uma combinação do conhecimento de profissionais experientes acerca do processo, da máquina e do dia a dia no chão de fábrica. Chega-se a uma pré-seleção final de 24 variáveis: 1 variável de resposta (sinal de WB) e 23 variáveis independentes.

De notar que as lâminas de crepe e de limpeza são abordadas de acordo com a oscilação da corrente nos seus motores. Esta abordagem facilita a perceção de quando é que a lâmina de limpeza está ou não em uso ou quando é que a lâmina de crepe é trocada (valores iguais a zero).

Tabela 2. Pré-seleção de variáveis

Variável		Unidades	Zona da máquina
Condutividade da pasta *slush*		mS/cm	Tratamento da pasta
pH pasta *slush*		-	
pH águas brancas		-	
FC – refinação		kWh/ton	
FL – refinação 1		kWh/ton	
FL – refinação 2		kWh/ton	
Velocidade da máquina		m/min	Formação/cabeça de máquina
Yankee layer	FC	%	
	Broke	%	
	Consistência	%	
	Caudal	L/min	
Hood layer	FC	%	
	Broke	%	
	Consistência	%	
	Caudal	L/min	
Gramagem		g/m^2	*Yankee*/Zona final de secagem
Humidade		%	
Químicos	Adesivo	mg	
	Lançamento	mg	
	Fosfato	A	
Lâmina de crepe (corrente do motor)		A	
Lâmina de limpeza (corrente do motor)		A	
Vácuo do rolo de sucção		kPa	

Nesta fase do trabalho são encontrados alguns obstáculos. Estes estão na sua maioria relacionadas com a indisponibilidade de dados de algumas variáveis do sistema que se consideram de extrema importância para o problema. Algumas das variáveis que não foram tidas em conta e respetiva razão:

- tensão ao longo da folha: este é um dos principais fatores a considerar na WB (Uesaka, 2005). Porém, não existem medições viáveis durante o processo;
- resistências da folha: apenas são conhecidas na avaliação de qualidade do produto final;
- buracos na folha: existem sensores de buracos porém os dados não são armazenados de uma forma viável;
- qualidade do corte lateral: de acordo com as estatísticas, uma grande parte das quebras tem início na margem da folha (Ahola, 2005). Porém, medições não são viáveis.

Conclui-se assim que a seleção de variáveis efetuadas está dependente da existência de sensores no processo e da disponibilidade de dados recolhidos no chão de fábrica.

4.2. Seleção da amostra

O conjunto de dados a tratar foi selecionado com o apoio dos profissionais da TM, os quais propuseram um intervalo de tempo em que o funcionamento da máquina esteve, maioritariamente, dentro de valores normais.

A amostra selecionada corresponde a julho e agosto do ano de 2019 e o intervalo entre medições é de 1.5 minutos — o que corresponde a aproximadamente 60.000 observações.

Para o mesmo intervalo, é importante conhecer o sinal de quebra, o qual caracteriza a variável de resposta do modelo de previsão. Este sinal encontra-se num formato binário: 1 para WB e 0 para funcionamento normal.

Ao longo dos dois meses definidos anteriormente foram registadas 1342 WB, o que se traduz numa média de aproximadamente 22 quebras/dia. Porém, esta média não define apropriadamente a realidade, sendo que existem dias com o dobro desse valor de

quebras e outros dias com apenas metade desse valor. Este facto revela que as quebras não se encontram distribuídas uniformemente ao longo da amostra. Este facto será tratado na secção 4.3.2.

4.3. Pré-processamento

Neste trabalho os dados irão sofrer um processamento inicial de redução, o qual procura excluir observações que não traduzem um funcionamento normal do processo. Este ponto inclui a limpeza e a segmentação da amostra em classes de risco, apresentados de seguida.

4.3.1. Limpeza

A "*raw-data*" que serve de base a qualquer análise de dados e, principalmente, sendo obtida de processos reais, pode apresentar inconsistências. Estas podem estar relacionados a erros de medição (falha de sensor) ou, por outro lado, podem corresponder a situações anómalas e pontuais no processo. Assim, de forma a aumentar a confiança e a qualidade dos resultados obtidos, é de máxima importância fazer uma limpeza inicial dos dados (Han, 2012).

Numa primeira parte da limpeza de dados são eliminadas observações que constituem *outliers* ou valores tecnicamente impossíveis. Posteriormente, são eliminadas observações que correspondem a quebras com causas conhecidas.

4.3.1.1. *Outliers* e valores tecnicamente impossíveis

Nesta fase foram feitas as seguintes operações:

- Eliminação de valores que não tenham sido registados pelo sensor com o máximo de confiança;
- Eliminar medições que correspondam a máquina parada, ou seja, velocidade = 0 m/min. Importante pois alguns sensores continuam a assumir valores mesmo a máquina estando parada);
- Eliminar medições de velocidade que correspondam a arranque de máquina ou que ultrapassem a velocidade técnica máxima: velocidade < 1300 m/min e velocidade > 2000 m/min;

- Eliminar todos os valores de caudal (L/min) e consistência (%) que sejam iguais a zero, pois correspondem a período de produção parada (pois não está a ser alimentada pasta à máquina);
- Cruzar os valores da refinação com os valores de "alimentação de pasta à máquina", de forma a compreender se os valores nulos na refinação correspondem a paragem de máquina ou ao não uso/não refinação da pasta. Eliminar todos os valores que correspondam a paragem de máquina. Os restantes são úteis para compreender a influência que a não refinação terá nas WB.
- Eliminar longos períodos de tempo em que, após uma quebra, não está a ocorrer enrolamento. Isto porque nestes períodos de tempo não é possível avaliar possíveis quebras.

A redução efetuada sobre os dados após as operações referidas anteriormente pode ser observada na tabela seguinte:

Tabela 3. Resultados da 1ª parte da limpeza de dados

	Dados originais [unid]	Após 1ª limpeza [unid]	Redução [%]
Nº de observações na amostra	59521	46627	22%
Nº de quebras na amostra	1342	1151	14%

4.3.1.2. Quebras com causas conhecidas

Foram efetuadas as seguintes operações:

- Eliminar situações de quebra que correspondam a paragens programadas (limpeza de máquina e manutenção). Torna-se benéfico excluir esta quebras pois podem ser provocadas propositadamente por intervenção humana.
- Analisar quais as WB que estão diretamente relacionadas com falha na lâmina de crepe e excluir as mesmas.

Considera-se que as quebras devido à lâmina de crepe são bem conhecidas no chão de fábrica e não têm uma influência significativa no funcionamento da máquina e

variáveis a analisar — a maioria das variáveis em análise correspondem a zonas mais a montante no processo comparativamente à lâmina de crepe. Apenas os químicos são influenciados por esta falha (uma lâmina nova retira parte do revestimento do *yankee*) levando aproximadamente 10min para estabilizarem os seus valores após troca de lâmina.

Assim, para além de serem eliminadas as quebras devido a falha da lâmina de crepe, também são eliminadas as observações nos 10 minutos que se seguem às mesmas. Isto porque se considera que durante esse período a máquina não tem o funcionamento normalizado.

Neste caso de estudo não é possível eliminar todas as quebras com causas conhecidas devido à pouca viabilidade que os relatórios de produção apresentam. Assim, a título de exemplo, não se conseguem eliminar da amostra falhas por erro humano, por ocorrência de uma situação pontual, etc. A partir deste ponto tomam-se todas as quebras restantes como tendo uma causa desconhecida.

4.3.1.3. Resultados da limpeza de dados

Na tabela 4 é possível observar o impacto que a limpeza tem na amostra do caso em estudo.

Tabela 4. Resultados finais da limpeza de dados

	Dados originais [unid]	Após limpeza [unid]	Redução [%]
Nº de observações na amostra	59521	46085	23 %
Nº de quebras na amostra	1342	998	26 %

Note-se que a contagem de WB decresce 26% com esta fase do trabalho.

4.3.2. Segmentação por classes de risco

Como não é possível definir de forma viável quais as causas de todas as quebras, define-se que apenas é vantajoso abordar os dados que correspondam a um funcionamento normalizado da máquina. Esta decisão visa aumentar a qualidade dos dados em tratamento e evitar o *overfitting* do modelo que se procura construir.

Com uma análise exploratória dos dados é possível compreender qual a frequência de quebras. No caso em estudo, existem períodos de tempo em que as quebras

ocorrem separadas por poucos minutos — as chamadas "avalanches de quebras"; e outros períodos em que ocorrem apenas um par de vezes ao dia.

Neste ponto do trabalho, procura-se definir um mínimo espaço temporal entre quebras que consiga traduzir um funcionamento normal da máquina. O objetivo será obter um conjunto de dados onde a TM esteve a funcionar de forma correta e onde, posteriormente, algo mudou e provocou uma quebra. Com base na literatura, considerou-se um espaço temporal mínimo de 3h entre quebras. A amostra é analisada e dividida em partes consoante o tempo entre quebras. Finalmente, apenas as partes com um tempo superior a 3h são selecionadas. Na tabela 5 consegue-se perceber o resultado e a influência que esta operação tem na amostra, reduzindo as quebras a analisar em 89%.

Tabela 5. Impacto da segmentação temporal nos dados

	Dados após limpeza [unid]	Dados após segmentação [unid]	Redução [%]
Nº de observações na amostra	46085	19555	58 %
Nº de quebras na amostra	998	108	89 %

Pelo desenvolvimento que o trabalho toma até este ponto e pela análise de trabalhos presentes na literatura, tais como o de Ahola (2005), considera-se importante definir qual o risco associado a cada observação — a classe de risco. Assim, são definidos níveis de risco: nulo, baixo, médio e alto; e dependem do afastamento temporal que têm da quebra. Notando que apenas estão a ser considerados períodos de tempo entre quebras superiores a 3h, toma-se por:

- Risco alto: observações que ocorrem desde a 1h anterior à quebra e até à mesma, inclusive;
- Risco médio: observações correspondentes ao espaço entre as 2h e 1h anteriores à quebra;
- Risco baixo: observações correspondentes ao espaço entre as 3h e 2h anteriores à quebra;

- Risco nulo: todas as observações que ocorrem para lá das 3h anteriores à quebra.

Esta classe de risco funciona como uma variável de resposta e irá auxiliar a classificação durante a construção do modelo do capítulo 5. A figura 6 apresenta a distribuição dos dados por classe de risco depois da segmentação.

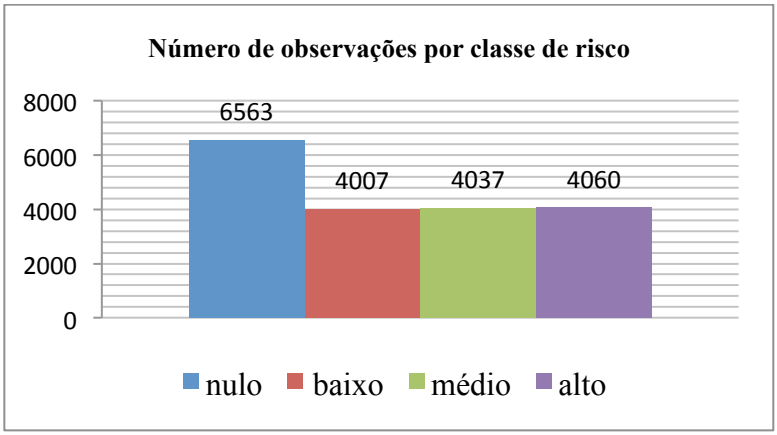


Figura 6. Distribuição dos dados por classes de risco.

4.4. Análise exploratória

Fazer uma análise exploratória ajuda a perceber algumas relações que possam existir entre as WB e as variáveis.

No anexo B encontram-se histogramas para cada uma das variáveis com a percentagem de quebras que ocorrem dentro dos seus valores. Para além disso, são também apresentados gráficos de densidade das observações. É importante analisar estas duas ferramentas em conjunto de forma a compreender se a maior presença de quebras num certo intervalo de valores se deve também à sua maior utilização durante o funcionamento da máquina.

Repare-se, por exemplo, no caso da velocidade. A maior parte das quebras ocorrem por volta dos 1700m/s. Porém, isto não pode ser um indicador definitivo de que esta é uma velocidade mais propícia à ocorrência de quebras. Isto porque coincide com a maior densidade de observações para a mesma variável. Observando-se todo o anexo B consegue-se perceber que, na sua maioria, a ocorrência de quebras por valor da variável está relacionada com o número de observações registadas para esse mesmo valor.

Porém, atente-se à variável do vácuo do rolo de sucção. Repare-se que para uma pressão de, aproximadamente, -42[kPa], as ocorrências de quebra não são proporcionais à sua densidade quando comparadas com o resto dos seus valores. Isto sugere que um valor mais baixo de pressão no rolo de sucção equivale a um menor risco de ocorrer uma quebra. Para as outras variáveis não se consegue tirar conclusões acerca da relação que um valor específico possa ter com as WB, pois a ocorrência de quebras é transversal a qualquer valor que possam apresentar.

Com esta análise define-se também que o químico de proteção do *yankee*, ou seja, o fosfato, não tem qualquer poder explicativo relativamente às quebras pois assume um valor constante (1mg) ao longo do tempo. Assim, não pode melhorar o poder do modelo que se pretende construir dado que apresenta variância igual a zero (Raj e Raman, 2018).

Neste ponto do trabalho torna-se também benéfico excluir da análise a lâmina de crepe, a qual foi muito útil na identificação das quebras que ocorreram por falha da mesma (secção 4.3.), mas que não consegue adicionar mais informação ao problema.

Assim, a partir deste ponto do trabalho são excluídos do conjunto de variáveis o fosfato e a lâmina de crepe. Passam a ser consideradas 21 variáveis independentes, ao invés das 23 que foram apresentadas e estavam a ser alvo de análise.

4.5. Normalização da amostra

Como as variáveis em análise têm unidades de medida e escalas muito distintas — velocidade com m/s, condutividade com mS/cm, gramagem com g/m^2, FL e FC com %, entre outros; torna-se necessário proceder a uma normalização dos dados (Bjorklund, 2009). É adotada a normalização Z-score de forma a preservar a forma da distribuição dos dados, sendo efetuada a seguinte operação sobre a amostra:

$$v' = \frac{v - \mu_A}{\sigma_A} \tag{2}$$

Onde:

v'representa o novo valor;

v é o valor da observação antiga;

μ_A é a média de todas as observações da variável A;

σ_A é o desvio padrão de todas as observações da variável A.

Esta normalização transforma a amostra numa equivalente que apresenta média igual a zero e desvio padrão igual a 1.

4.6. Análise de correlação estratificada

A análise de correlação estratificada procura perceber se existe alguma alteração na relação entre variáveis quando se altera a classe de risco.

No anexo C são apresentadas as matrizes de correlação estratificada para cada classe de risco recorrendo ao coeficiente de *Pearson*. De seguida faz-se uma interpretação dos seus resultados.

Como esperado teoricamente, existe uma forte correlação positiva entre o químico adesivo e o químico de lançamento, a qual é transversal a qualquer classe de risco. O mesmo acontece para os caudais do HL e do YL. Existe também uma forte correlação negativa entre o químico adesivo e a velocidade em todas as classes de risco.

Por outro lado, repara-se que a correlação entre o pH e a condutividade da pasta *slush* vai diminuindo à medida que o risco aumenta.

Note-se ainda na forte correlação negativa entre o vácuo do rolo de sucção e a velocidade, demonstrando que é necessária uma menor pressão no rolo de sucção quando a velocidade é mais elevada.

Teoricamente, são esperados altos coeficientes de correlação entre algumas variáveis, tais como a gramagem e a velocidade. Porém, tal como indica Dahlquist (2008), quando dependências não lineares estão presentes ou quando as amostras não são independentes, tal como numa série temporal, o cálculo de correlação pode levar a um valor muito baixo mesmo quando duas variáveis são altamente correlacionadas.

4.7. Redução da dimensionalidade

Esta fase do trabalho procura reduzir a dimensionalidade recorrendo a uma análise de componentes principais e compreender se a sua utilização é benéfica para o problema.

4.7.1. Análise de componentes principais

O código em linguagem *python* desenvolvido e que permite alcançar os PC da amostra em análise encontra-se no anexo D.

Os resultados alcançados nesta PCA encontram-se no anexo E, onde se apresentam os auto-valores e as variâncias explicadas por cada PC. Para além disso, o gráfico da figura 7 apresenta visualmente qual a percentagem de variância explicada por cada PC e respetiva variância acumulada.

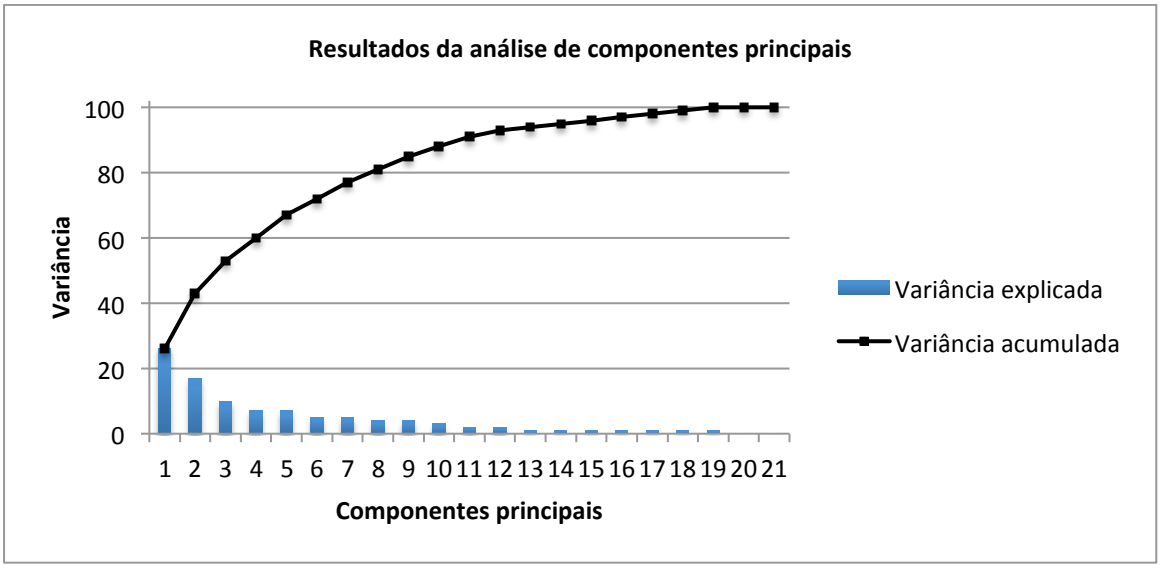


Figura 7. Percentagem de variância explicada por cada componente principal.

Brown (2009) refere várias técnicas para decidir quantos PC usar. Neste trabalho, sugere-se uma combinação de três técnicas (Hubert, 2009):

- Auto-valor > 1, pelo critério de Kaiser-Guttman;
- Variância acumulada > 90%;
- Observação do gráfico dos auto-valores;

Analisando os resultados alcançados, percebe-se que do critério apresentado no primeiro ponto (Guttman, 1954) poderão ser aceites os primeiros seis PC, pois apresentam um auto-valor superior a 1.

Relativamente à segunda técnica referida, note-se que a variância acumulada a aceitar depende das características do problema. Por exemplo, é aceitável uma variância acumulada de apenas 60% quando se lida com problemas das ciências sociais. Porém, quando se está perante problemas mais complexos, como por exemplo nas ciências da vida ou ciências exatas, torna-se benéfico aceitar apenas os PC que representam uma variância acumulada elevada. Para este problema, assume-se que será vantajoso selecionar PC que consigam explicar pelo menos 90% da variância do sistema. Assim, imposta esta condição, pelo anexo E ou pela figura 7 percebe-se que será necessário selecionar 11 PC.

Relativamente à observação do gráfico de auto-valores (figura 8), procura-se perceber visualmente em que momento existe uma descida brusca no auto-valor seguido de uma estabilização da reta. Os PC anteriores à estabilização devem ser os escolhidos (Brown, 2009).

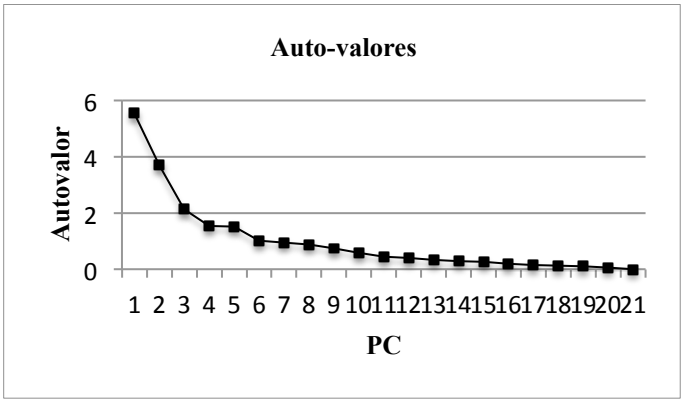


Figura 8. Gráfico dos auto-valores dos componentes principais.

Pela figura consegue-se compreender que existe uma descida brusca entre os PC 5 e 6, sendo que a reta estabiliza de seguida. Assim, compreende-se que a condição sugere a seleção de 6 PC.

Dadas as três condições propostas anteriormente, considera-se que não será benéfico utilizar os PC neste caso de estudo. Seria necessário selecionar 11 PC, o que significa uma redução de dimensionalidade pouco eficiente. Para além disso, a utilização dos PC na construção do modelo levaria a que o mesmo fosse mais difícil de interpretar.

Assim, dada a aplicação prática do problema, torna-se benéfico construir um modelo facilmente explicável, de forma a que a sua interpretação não seja muito complexa.

Desta PCA surge também o anexo F, onde são apresentadas as contribuições que cada variável tem em cada PC. Podemos ver quais as variáveis que mais contribuem para os primeiros PC, ou seja, para os PC que representam mais variância do sistema, e também quais as variáveis que menos influência têm nos mesmos.

Do mesmo anexo observa-se que, relativamente ao PC1, existe uma grande contribuição da gramagem, da humidade, da refinação da FC e das percentagens de FC e de *broke* no YL. Por outro lado, variáveis tais como os químicos de lançamento e adesivo, o caudal do YL, o vácuo de sucção e o pH da pasta *slush* não apresentam uma grande contribuição.

Ao analisar o conjunto dos primeiros 3 PC, repara-se que existe uma prevalência elevada de contribuição por parte da humidade e das percentagens de FC e *broke* no YL.

Atente-se ainda ao pH da pasta *slush*, que apresenta uma contribuição baixa no primeiro PC mas que, por outro lado, tem uma forte presença desde o 2º até ao 7º PC.

4.8. Considerações finais

Durante o capítulo 4 compreende-se o comportamento das variáveis e as relações que têm entre si. Percebe-se que, por exemplo, a correlação entre o pH e a condutividade da pasta *slush* sofre uma diminuição à medida que o risco de quebra aumenta. Percebe-se também que existem variáveis das quais é esperado um coeficiente de correlação elevado e tal não acontece, o que pode ser explicado por dependências temporais, condição que se aplica ao conjunto de dados sob análise. Para além disso, por análise exploratória, descartam-se duas variáveis iniciais — o fosfato e a lâmina de crepe.

Deste capítulo conclui-se também que não é benéfico utilizar PC na construção do modelo de previsão. Como se parte de um número pequeno de variáveis (apenas 21) e os resultados demonstram que são necessários 11 PC para representar 90% da variância, decide-se avançar com a construção do modelo do capítulo seguinte recorrendo às variáveis originais e descartando a utilização de PC.

Para além desta decisão ser benéfica do ponto de vista de preservar informação importante contida nas variáveis originais, é igualmente importante para facilitar a interpretação de resultados das árvores de decisão que se pretendem construir. Parte-se assim para a construção do modelo de previsão (capítulo 5) recorrendo às 21 variáveis originais.

5. MODELOS DE PREVISÃO E DISCUSSÃO DE RESULTADOS

Neste capítulo é desenvolvido o modelo de previsão de risco de quebra e são analisados os seus resultados. Por fim, a *performance* da DT será comparada, enquanto método de classificação, com a *performance* que uma SVM teria se aplicada aos mesmos dados.

5.1. Árvores de decisão

O código desenvolvido em *python* para a construção da DT e para a avaliação da sua *performance* encontra-se no anexo G. Um esquema do algoritmo adotado é apresentado na figura 9.

Figura 9. Esquema da construção das árvores de decisão.

Como métricas de *performance* são tidas em conta a exatidão, precisão, *recall* e matriz de confusão das previsões. Torna-se importante analisar a exatidão que o modelo tem consoante a profundidade máxima da DT. Os resultados são apresentados na tabela 6.

Tabela 6. *Performance* da árvore de decisão dependendo da profundidade máxima.

Profundidade máxima [unid]	Exatidão [%]
3	46,55
5	50,82
10	63,39
15	78,13
18	82,75
21	86,88

Decide-se construir uma DT com 21 níveis de profundidade dado que esta é a que apresenta melhor exatidão na fase de teste do modelo.

5.1.1. Análise da *performance*

Torna-se agora importante analisar detalhadamente a *performance* do modelo adotado. A matriz de confusão é apresentada na tabela seguinte. Consegue-se perceber que existe mais confusão entres os riscos baixo, médio e alto. Porém, na globalidade, o modelo não apresenta muito confusão nas suas previsões. A exatidão associada a estes resultados é de 86,55%

Tabela 7. Matriz de confusão do modelo (4 classes e 21 variáveis).

		Risco real [unid]			
		nulo	baixo	médio	alto
Risco previsto [unid]	nulo	1866	51	6	7
	baixo	48	1034	148	95
	médio	14	96	952	126
	alto	12	37	114	988

Para auxiliar esta análise, na tabela 8 são também apresentados os valores da precisão e o do *recall* associados à previsão de cada classe

Tabela 8. Precisão e *recall* do modelo (4 classes e 21 variáveis).

		precisão [%]	*recall* [%]
Classe de risco	nulo	97	96
	baixo	78	85
	médio	80	78
	alto	86	81
	média	87	87

Pelos dados apresentados nas figuras consegue-se perceber que a classe do risco nulo tem a melhor *performance*, apresentando uma precisão de 97% e *recall* de 96%.

5.1.2. Integração em escada

Do ponto 5.1.1. segue o valor quase perfeito na previsão de risco nulo e a existência de alguma confusão na previsão das restantes classes. Desta forma, decide-se construir uma DT que classifique apenas os riscos baixo, médio e alto.

Esta abordagem procura entender se a combinação com uma segunda árvore pode ser um bom auxiliar para a previsão. Um esquema do modelo possível é apresentado na figura 10.

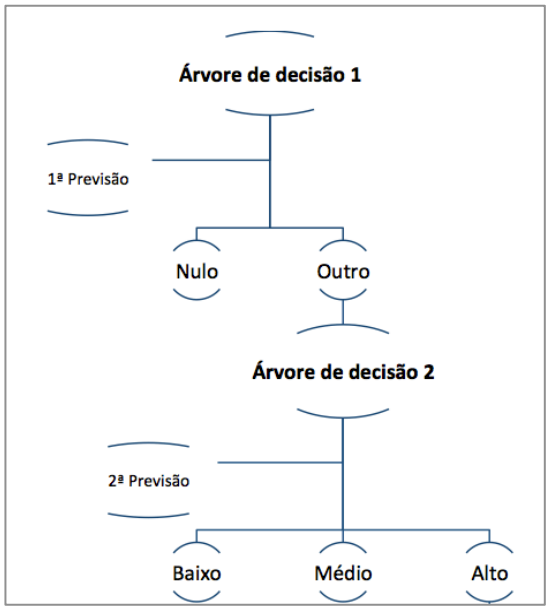


Figura 10. Integração em "escada" de duas árvores de decisão.

Caso a DT nº1 preveja um risco nulo, aceita-se a previsão. Caso contrário, a DT nº2 funciona como um auxiliar na previsão dos riscos baixo, médio e alto. A nova DT é construída com uma profundidade máxima de 21 níveis e a sua matriz de confusão e *performance* são apresentadas nas tabelas abaixo, sendo que a exatidão associada é de 71,86%.

Tabela 9. Matriz de confusão do modelo sem a classe de risco nulo.

		Risco real [unid]		
		baixo	médio	alto
Risco previsto [unid]	baixo	960	198	140
	médio	299	984	259
	alto	76	163	954

Tabela 10. Precisão e *recall* do modelo sem a classe de risco nulo.

		precisão [%]	*recall* [%]
Classe de risco	baixo	74	72
	médio	64	73
	alto	80	71
	média	73	72

Atente-se aos valores de *recall*. Pelas tabelas consegue-se entender que construir uma segunda DT não beneficia a previsão dos riscos baixo, médio e alto. Assim, a proposta é descartada e segue-se com a análise do modelo com uma DT.

5.1.3. Importância das variáveis

Numa árvore de decisão, as variáveis mais próximas da raiz são as mais importantes, exercendo maior poder de decisão no modelo. A tabela 11 apresenta as variáveis em análise por ordem de importância.

Tabela 11. Ordem de importância das variáveis na árvore de decisão.

Variável	Importância
pH águas brancas	0,18
Lâmina de limpeza	0,10
Gramagem	0,09
Velocidade	0,07
Broke HL	0,06
Vácuo de sucção	0,06
pH slush	0,06
Refinação FC	0,05
Humidade	0,05
Caudal HL	0,05
Condutividade slush	0,05
Caudal YL	0,04
Adesivo	0,03
Broke YL	0,03
FC HL	0,03
Consistência HL	0,02
FC YL	0,02
Consistência YL	0,02
Refinação FL 1	0,01
Refinação FL 2	0,00
Lançamento	0,00

Observa-se que o pH das águas brancas é o que apresenta mais poder de decisão no nosso modelo, seguido pela lâmina de limpeza, gramagem, velocidade, percentagem de *broke* no HL, o vácuo de sucção e o pH da pasta *slush*. Estas são consideradas as variáveis mais importantes do nosso sistema e são analisadas em detalhe no ponto 5.3. Com menos poder decisivo encontram-se o químico de lançamento e as duas refinações da FL.

Conhecendo a ordem de importância e de forma a ter uma visão alargada do comportamento do modelo de previsão, são retiradas, consecutivamente, as variáveis desde a menos até à mais importante. O nível de profundidade máxima é considerado como sendo igual ao número de variáveis em análise. Os resultados alcançados são apresentados na tabela 12.

Tabela 12. *Performance* do modelo com eliminação de variáveis (da menos à mais importante).

Nº de variáveis eliminadas	Nº de variáveis em análise	Exatidão	Precisão
0	21 (as iniciais)	86,16	86
1	20	81,67	83
2	19	87,17	87
3	18	82,08	83
4	17	79,72	81
5	16	78,96	80
6	15	76,87	79
7	14	77,22	77
8	13	74,38	75
9	12	68,23	69
10	11	63,86	67
11	10	62,89	68
12	9	60,02	59
13	8	56,48	55
14	7	54,36	52
15	6	49,72	51
16	5	49,1	49
17	4	46,94	41
18	3	46,79	45
19	2	45,67	34
20	1	44,49	25

Repare-se na *performance* do modelo quando se usam apenas as sete variáveis mais importantes, a qual apresenta uma exatidão de apenas 54,36%. Por outro lado, existe uma boa *performance* quando são eliminadas apenas as duas variáveis menos importantes. Esta melhoria pode ser explicada pela possível existência de elevado ruído associado a esses dados (Abraham et al., 2006) ou por continuarem a existir *outliers* nas observações dos mesmos, o que pode ser melhorado revisitando a limpeza de dados das duas variáveis. Por outro lado, esta melhoria pode também estar associada ao facto das duas variáveis em questão não apresentarem qualquer correlação com a variável de resposta.

Para além da melhoria na *performance*, a eliminação de duas variáveis apresenta-se também muito positiva em termos de eficiência e de facilidade de utilização prática do modelo (Tulu et al., 2019).

Define-se assim que será vantajoso usar um modelo com as 19 variáveis mais importantes, eliminando da análise as variáveis correspondentes ao químico de lançamento e à refinação 2 da FL.

5.1.4. Proposta final do modelo de previsão

Dada a conclusão alcançada na secção anterior, apresentam-se nas tabelas seguintes a matriz de confusão e a *performance* da DT com apenas 19 variáveis. A sua exatidão é de 87,17%.

Tabela 13. Matriz de confusão do modelo (4 classes e 19 variáveis).

		Risco real [unid]			
		nulo	baixo	médio	alto
Risco previsto [unid]	nulo	1263	18	6	6
	baixo	31	694	113	58
	médio	11	69	629	85
	alto	2	20	59	663

Tabela 14. Precisão e recall do modelo (4 classes e 19 variáveis).

		precisão [%]	*recall* [%]
Classe de risco	nulo	98	97
	baixo	77	87
	médio	79	78
	alto	89	82
	média	87	87

Esta constitui a proposta final de modelo de previsão. Indica-se ainda que esta é uma árvore constituída por 1827 nós.

5.1.5. Dinâmica das variáveis

Neste ponto do trabalho torna-se benéfico fazer uma análise exploratória às sete variáveis com o maior poder de decisão no modelo. Esta análise procura perceber se é

possível identificar padrões e relações que possam levar a uma mudança de risco. As figuras 11 e 12 exemplificam esta análise.

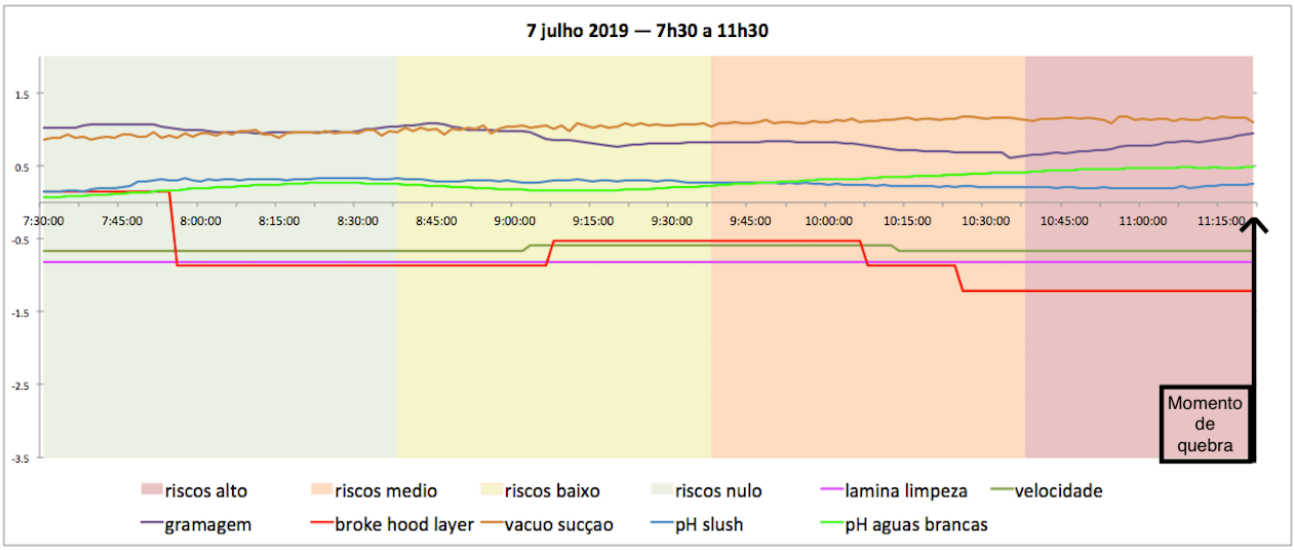


Figura 11. Observações das sete variáveis com maior importância – 7 julho.

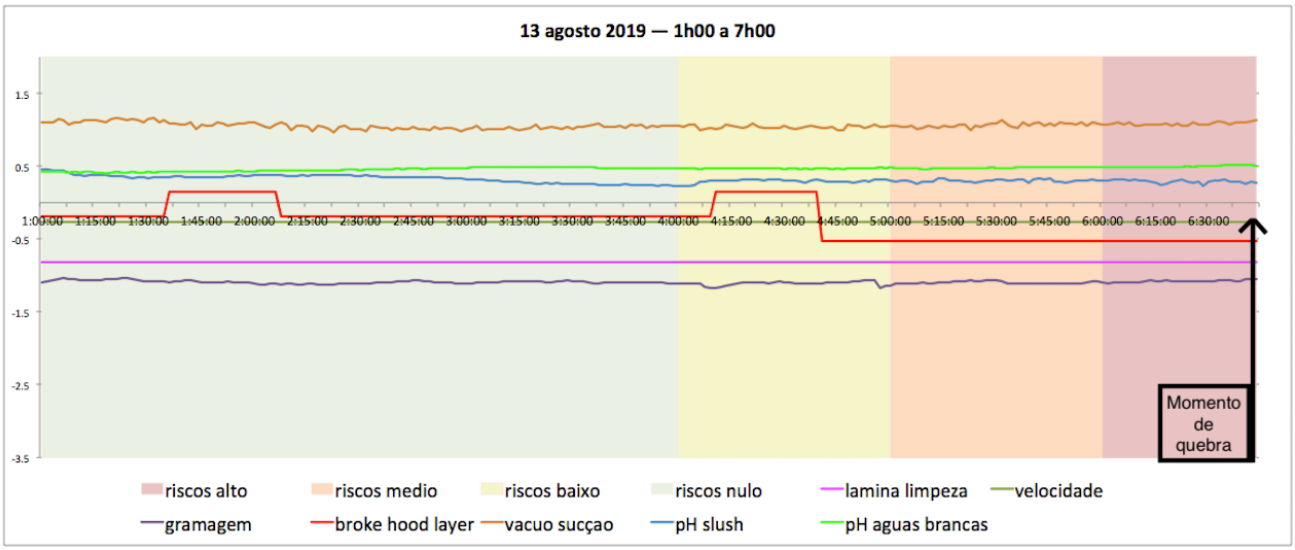


Figura 12. Observações das sete variáveis com maior importância – 13 agosto.

Nas figuras observam-se dois momentos de quebras. Repare-se que seis das variáveis têm comportamento semelhante em ambas as figuras, estando estabilizadas ao longo das mudanças de risco. Porém, atente-se também ao único fator diferenciador: o teor de *broke* no HL (linha a vermelho). Em ambas as quebras observa-se uma ligeira redução no seu valor e que, após isso, a máquina entra num trajeto de risco crescente.

Teoricamente, esta conclusão faz sentido, visto que a utilização de *broke* confere mais resistência à folha. Percebe-se assim que reduzir o seu valor provoca alterações nas características da folha, pelo que se torna necessário conjugar e alterar valores de outras variáveis de forma a manter a máquina e a folha estabilizadas. Estas duas figuras conseguem traduzir o que acontece regularmente quando se altera a percentagem de utilização de *broke* no HL e se mantêm as outras seis variáveis constantes. Porém, para além deste padrão, não se consegue identificar mais nenhuma dinâmica entre as restantes que consiga explicar a mudança de risco.

5.2. Máquina de vetores de suporte

Esta análise de comparação surge devido às SVM terem cada vez mais relevância e visibilidade no mundo da ML quando aplicadas a problemas de classificação (Barakat e Bradley, 2010).

É analisada a *performance* de um classificador SVM quando aplicado aos mesmos dados (19 variáveis e 4 classes de risco). Como métricas de *performance* tomam-se, novamente, a matriz de confusão, exatidão, precisão e *recall* das previsões.

Para a classificação, elabora-se um código *python* (anexo H). Recorre-se ao *kernel* de função base radial (RBF). Sucintamente, as SVM usam o *kernel* para tornarem um conjunto de dados que não são separáveis linearmente, em um que o seja.

Adota-se um parâmetro C=1 (que pode ser interpretado como a punição por classificar um valor errado; é tomado um valor baixo de forma a prevenir *overffiting*) e compara-se a sua *performance* com diferentes valores do parâmetro $\boldsymbol{\gamma}$, ou seja, com diferentes propagações da região de decisão (Hsu et al., 2016).

Tabela 15. *Performance* SVM com diferentes parâmetros de $\boldsymbol{\gamma}$.

	Parâmetro $\boldsymbol{\gamma}$			
	0.01	1	10	100
Exatidão [%]	49	88	85	38
Precisão [%]	45	88	86	73
Recall [%]	49	88	85	38

Da tabela 15 consegue-se entender que, dos valores de γ analisados, aquele que leva a uma melhor *performance* do modelo é o de γ =1. Percebe-se assim que este valor é o que se molda melhor aos dados em análise e parte-se para uma análise mais detalhada do modelo.

Com uma exatidão de 88,24%, a matriz de confusão e restante *performance* deste classificador são apresentadas nas tabelas seguintes.

Tabela 16. Matriz de confusão com classificador SVM.

		Risco real [unid]			
		nulo	baixo	médio	alto
Risco previsto [unid]	nulo	1274	37	7	17
	baixo	26	684	79	16
	médio	1	62	628	76
	alto	6	18	76	703

Tabela 17. Precisão e recall com classificador SVM.

		precisão [%]	*recall* [%]
Classe de risco	nulo	95	97
	baixo	85	85
	médio	82	78
	alto	86	82
	média	88	87

Na tabela 18 encontra-se um resumo dos resultados alcançados para as DT e para as SVM, de forma a comparar as suas *performances*.

Tabela 18. Resumo dos resultados com classificador DT e SVM.

Risco	**DT**			**SVM**		
	Exatidão [%]	Precisão [%]	*Recall* [%]	Exatidão [%]	Precisão [%]	*Recall* [%]
nulo	-	98	97	-	95	97
baixo	-	77	87	-	85	85
médio	-	79	78	-	82	78
alto	-	89	82	-	86	82
média	87,17	87	87	88,24	88	87

Comparando os dois modelos, nota-se que em ambos existe maior dificuldade em classificar o risco médio. Note-se ainda que a exatidão de ambos os modelos é semelhante. Desta forma, conclui-se que as SVM também se apresentam como um bom método para este caso prático. Porém, tal como indicado por Barakat e Bradley (2010), as SVM apresentam a inabilidade de dar uma justificação compreensível acerca das soluções que alcançam quando comparadas com as DT.

5.3. Considerações finais

Depois do modelo de previsão construído, torna-se importante clarificar como é que o mesmo poder ser útil. Fundamentalmente, o modelo funciona como um sistema de aviso. Caso o valor devolvido seja nulo ou baixo, os operadores têm a perceção que o sistema está a funcionar bem; caso seja médio, é necessário estar alerta ao que mudou e ajustar valores; ao devolver risco alto, pode existir perigo de quebra eminente, pelo que é necessário perceber o que mudou e agir rapidamente.

De forma a auxiliar esta ação identificam-se as variáveis com maior poder de decisão. O pH das águas brancas, a utilização ou não da lâmina de limpeza, a velocidade da máquina, a gramagem do papel, a percentagem de *broke* usado no HL, o pH da pasta *slush* e o nível do vácuo de sucção podem ser uns bons indicadores do risco de ocorrer uma quebra e deverão sofrer uma análise mais cuidada quando se está perante um risco alto.

Por outro lado, é difícil compreender qual a dinâmica entre variáveis dado o seu comportamento muito complexo. Torna-se assim difícil indicar medidas preventivas. Isto vem sublinhar mais uma vez a importância da adoção da aprendizagem computacional neste problema, de forma a identificar padrões de funcionamento.

Não obstante, o modelo apresenta alguns constrangimentos. Indica-se a ausência de dependências temporais; claramente as observações não são independentes entre si, sendo que um dado valor depende do imediatamente anterior e influencia o seguinte. Isto leva a que não seja possível prever qual o tempo de mudança entre classes de risco. Por último, torna-se importante realçar a boa *performance* que o modelo apresenta mesmo quando confrontado com um conjunto de dados de teste que não conhece.

6. CONCLUSÕES E TRABALHOS FUTUROS

Na indústria da pasta e do papel, conseguir prever e evitar uma WB pode significar ganhos significativos de produtividade e eficiência. Porém, fazer uma previsão sólida é dificultada pela complexidade do fabrico de *tissue* e consequente dificuldade em identificar padrões. Mesmo que teoricamente algumas variáveis tenham uma alta dependência entre si, confirma-se que na prática essa relação é refutada pela complexidade do processo e que cada condição não consegue ser explicada por um pequeno conjunto de variáveis.

Desta forma, visto que o objetivo do trabalho passou pela criação de um modelo de previsão, tornou-se essencial desenvolver uma metodologia focada na aprendizagem computacional. As árvores de decisão mostraram-se como a opção indicada, dada a maior facilidade de interpretação dos resultados e das regras criadas.

O principal objetivo do trabalho é alcançado com a construção de um modelo que consegue prever qual o risco de ocorrer uma WB com uma exatidão superior a 80%. Tal como indicado, este modelo é baseado em árvores de decisão e a sua versão final apresenta apenas 19 variáveis independentes.

O modelo permite também ordenar as principais variáveis pelo seu poder de decisão. Estas funcionam como um indicativo de qual deve ser o foco de atenção perante o crescimento de risco de quebra. Assim, sugere-se uma especial atenção ao pH das águas brancas, à lâmina de limpeza, à gramagem, à velocidade, ao teor de *broke* no *hood layer*, ao vácuo de sucção e ao pH da pasta *slush*.

Sublinha-se particularmente a relação que a lâmina de limpeza tem com as WB, observando-se que a sua utilização assídua se mostra como uma boa opção de melhoria do sistema. Por análise exploratória, percebe-se ainda a relação que o teor de *broke* no HL tem com as outras seis variáveis mais importantes e com as quebras. De forma a manter a estabilização da máquina, é sugerido que se conjugue a redução de *broke* com uma alteração, a ser estudada, noutras variáveis. Para além disso, observa-se uma menor tendência em ocorrer quebras quando se usa um valor mais baixo de pressão no rolo de sucção a vácuo.

A elaboração do modelo apresentou ainda alguns constrangimentos, tais como a indisponibilidade de dados que se consideram essenciais numa análise de WB. Sugere-se assim uma melhoria na monitorização de alguns fatores, tais como as tensões e os buracos ao longo da folha ou a qualidade do corte lateral.

Finalmente, este trabalho apresenta-se como um alicerce de um modelo mais robusto e completo. Assim, como trabalhos futuros sugere-se que se incluam as variáveis que não estavam disponíveis durante a realização deste trabalho. Para além disso, sugere-se que sejam consideradas as dependências temporais entre observações. Esta última sugestão procura alcançar conclusões acerca do tempo que a máquina tem até ocorrer uma mudança de risco. Por último, sugere-se a construção de uma interface *online* de forma a criar previsões em tempo real e avaliar a sua *performance* em chão de fábrica.

REFERÊNCIAS BIBLIOGRÁFICAS

Abraham, A. et al. (2006), "Swarm intelligence in data mining", Studies in computational intelligence, Vol.34, Springer Berlin.

AF&PA (2019), *"Tissue: Making Everyday Life Convenient, Clean and Healthy"* , Acedido a 20 de Abril de 2020 em: https://www.afandpa.org/media.

Ahola, T. (2006). "Intelligent estimation of web break sensitivity in paper machine", University of Oulu, Sweden.

Alonso A., Blanco A. et al. (2006), "Application of advanced data treatment to predict paper properties", Proceeding of the 5th MathMod Conference, Vienna, Austria, 8-10 Feb.

Amruthnath, N. e Gupta, T. (2019), "Factor Analysis in Fault Diagnostics Using Random Forest", Western Michigan University, MI, USA.

Barak, S. et al. (2017), "Fusion of multiple diverse predictor in stock market", Information Fusion, Vol. 36 pp. 90-102.

Barakat, N. (2010), "Rule extraction from support vector machines: a review", Neurocomputing, Vol.74 pp. 178-190.

Berrar, D. (2019), "Cross-validation", Encyclopedia of bioinformatics and computational biology, Vol.1 pp. 542-545.

Björklund, K. e Svedjebrant, J. (2009), "Productivity improvements of a newsprint paper machine by reduction of web breaks", Luleå University of Technology, Sweden.

Blanco, A. et al. (2006), "Use of modelling and simulation in the pulp and paper industry", Math Comput Model Dynam Syst, 15, 409-423.

Bonissone et al. (1999), "System and method for predicting a web break in paper machine", USA.

Boudreau, J. (2013), "New methods for evaluation of tissue creping and the importance of coating, paper and adhesion", Karlstad University, Sweden.

Breiman, L. et al. (1984), "Classification and regression trees", 1st Ed., Chapman and Hall.

Brown, J. (2009), "Choosing the right number of components or factors in PCA and EFA", *JALT Testing & Evaluation SIG Newsletter,* Vol.13 pp.19-23.

Burges, C (1998), "A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery", Kluwer Academic Publishers.

Chen, Y. e Bonissone, P. (2002), "Paper web breakage prediction using principal components analysis and classification and regression trees", USA.

Dahlquist, E. (2008), "Use of modeling and simulation in pulp and paper industry". Malardalen University.

Ekvall, J. (2004), "Dryer Section Control in Paper Machines During Web Breaks." Department of Automatic Control, Lund Institute of Technology, Sweden.

Fernandes, R. (2018), "Deteção de falhas em processos industriais operando em múltiplas regiões via análise externa com múltiplos modelos lineares". Universidade Federal de Espírito Santo, Brasil.

Goh, K. e Singh, A. (2015) "Comprehensive literature review on machine learning

structures for web spam classification", Procedia Computer Science, Vol.70 pp.434-441.
Guo et al. (2004). "Robust prediction of fault-proneness by random forests", Proc. 15th Int'l Symp. Software Reliability Eng., France.
Guttman, L. (1954), "Some necessary conditions for common factor analysis", Psychometrika, Vol.19 pp. 149–161.
Han, J. et al. (2012) "Data mining: concepts and techniques", 3rd Ed., Elsevier Inc., USA.
Henrique, B. et al. (2019), "Literature review: machine learning techniques applied to financial market prediction", Expert Systems with Applications. Vol.124 pp.226-251.
Hiltunen, E. e Paulapuro, H. (2011), "Effect of long-fibred reinforcement pulp on mechanical properties of short fibred-based paper", O Papel, Vol.72 pp.42-48.
Hsu, C. et al. (2003), "A Practical Guide to Support Vector Classification", National Taiwan University, Taiwan.
Hu et al. (2019), "Optimal sparse decision trees", Cornell University, USA.
Hubert, M. (2020), "1.06 - Robust methods for high-dimensional data", Comprehensive chemometrics, 2nd Ed. Pp, 149-171.
Lauer, F. e Blochm G. (2008), "Incorporating prior knowledge in support vector machines for classification: a review", Neurocomputing, Vol.71 pp.1578-1594.
Lo Cascio, D. (2001), "Modelling the influence of the press felt on the moisture distribution in the paper web". Lanaken, Belgium.
Mandal, I. e Sairam, N. (2014) "New machine-learning algorithms for prediction of Parkinson's disease", *International Journal of Systems Science*, Vol.45 pp. 647-666.
Miyanishi, T. e Shimada, H. (1998), "Using neural networks to diagnose web breaks on a newsprint paper machine", Tappi Journal, 81, 163-170.
Mountrakis, G. et al. (2011), "Support vector machines in remote sensing: a review", Journal of photogrammetry and remote sensing, Vol.66 pp. 247-259.
Musa, A. (2014), "A comparison of l1-regularizion, PCA, KPCA and ICA for dimensionality deduction in logistic regression", International Journal of Machine Learning and Cybernetics, 5, 861-873.
Musharraf, M. et al. (2020), "Identifying route selection strategies in offshore emergency situations using decision trees", Reliability Engineering and System Safety, Vol.194 pp.106-179
Nagappan, N. et al. (2006), "Mining metrics to predict component failures." Proceedings of the 28th international conference on Software engineering, pp. 452–461.
Nakamura, J. (2007), "Predicting Time-to-Failure of Industrial Machines with Temporal Data Mining", University of Washington, Seattle, WA, USA.
Niskanen, K. (2012), "Mechanics of Paper Products", 1ª Ed., Gruyter, Boston.
Niuniu, X. et al. (2010)
Parola, M. e Beletski, N. (1999), "Tension across the paper web – a new important property", Proceedings of the 27th EUCEPA Conference, October, Grenoble, France.
Praveena, M. e Jaiganesh, V. (2017), "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process". *International Journal of Computer Applications.* Vol.169. pp. 32-35.
Raj, P. e Raman, A. (2018) "Handbook of research on cloud and fog computing

infrastructures for data science", IGI Global, USA.

Ramaratnam, K. et al (2016), "Soft through air dried tissue." USA

Sorsa, T., Koivo, H. e Korhonen, R. (1992), "Application of neural networks in the detection of breaks in a paper machine." IFAC Symp. on On-Line Fault Detection and Supervision in the Chemical Process Industries, pp.162-167. Abril.

Tulu, B. et al. (2019), "Extending the boundaries of design science theory and practice", 14th International conference on design science research in information systems and technology, MA, USA, Junho.

Uesaka, T. (2005), "Principal factors controlling web breaks in pressrooms - Quantitative evaluation". *Appita Journal*. Vol.58. pp. 425-432.

Xu, X. e Yang, G. (2013), "Robust manifold classification based on semi supervised learning". *International Journal of Advancements in Computing Technolog*y, Vol.5. pp. 174-183.

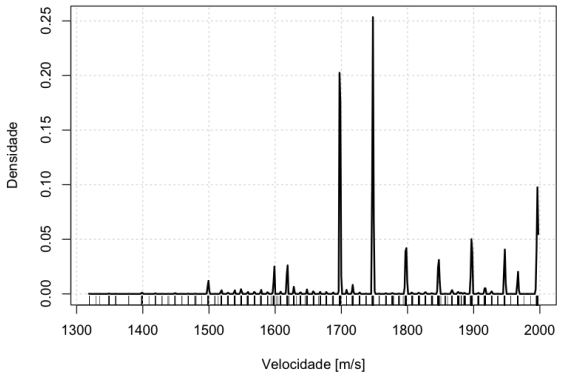
ANEXO A – PRÉ-SELEÇÃO DE VARIÁVEIS

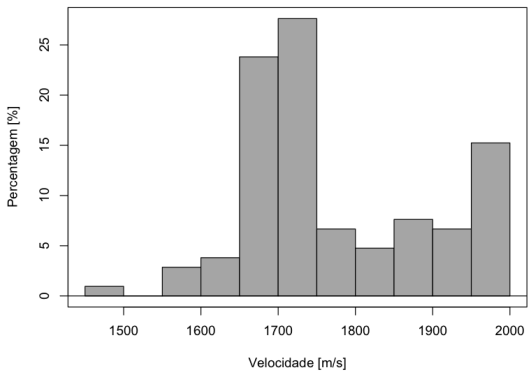
Variáveis gerais	Observações
Matéria-prima (características físicas e químicas; tratamentos químicos)	Pasta *slush*
	Pasta fibra longa
	Pasta fibra curta
Receita Receita=[(FL+FC)+*broke]*	YL
	HD
Águas	Brancas
	Purificada
Aditivos	—
HD *cleaner*	—
HCC *cleaner*	—
Recuperação de fibras	—
Broke	% de incorporação
	Deflaker (condições químicas, físicas e temporais)
Refinação (uma de FC e duas de FL)	pH, energia, tempo e esforço
Cabeça de entrada	—
Teia	—
Filtro	—
Caudal à entrada	YL
	HL
Consistência à entrada	YL
	HL
Chuveiros	—
Hood	—
Yankee	Temperatura
	Pressão
	Condensado
	Lubrificação
Químicos/*coating*	Lançamento
	Adesivo
	Proteção (fosfato)
	Limpeza
Lâminas (Aço e cerâmica)	Corte
	Crepagem
	Limpeza
Corte do meio	—

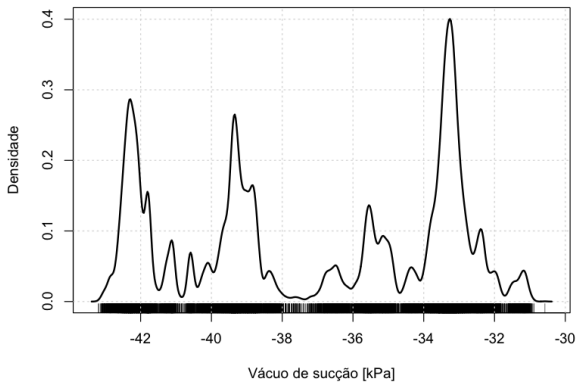
Corte lateral	—
Buracos	Tamanho, frequência
Gramagem	—
Humidade	—
Velocidade	—
Poeiras	—
Vácuo de sucção	—
Estabilizador	—
Enrolador	—

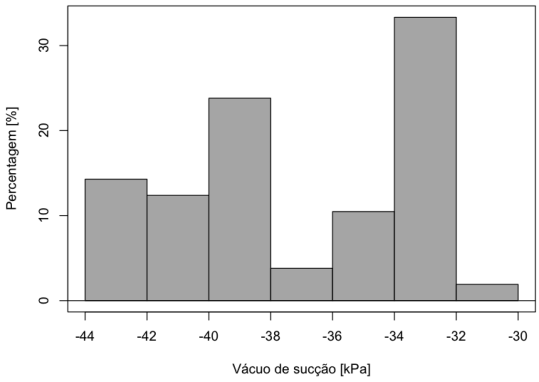
ANEXO B – ANÁLISE EXPLORATÓRIA DOS DADOS

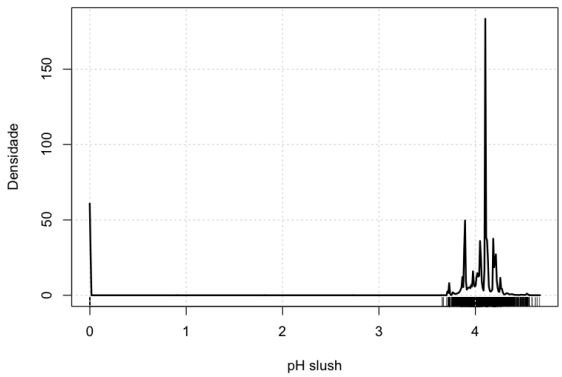
Do **lado esquerdo** apresentam-se as densidades da distribuição de observações associada a cada variável. No **lado direito** encontram-se as ocorrências de quebras [em %] por variável.

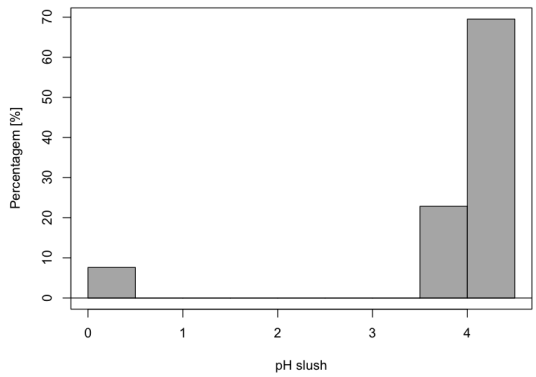


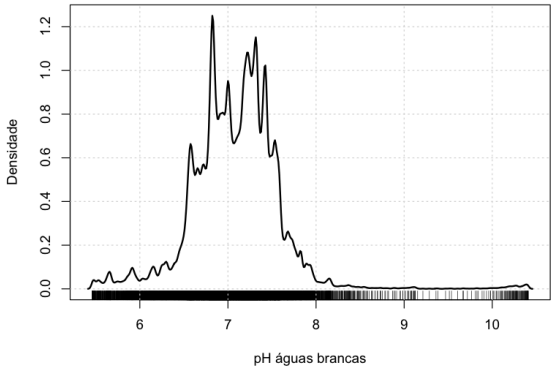


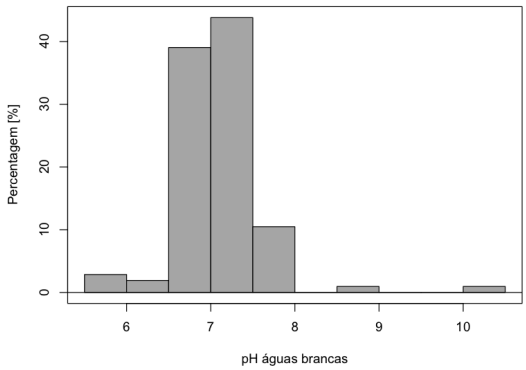


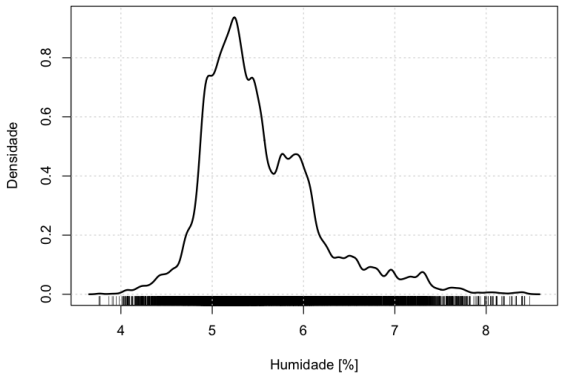


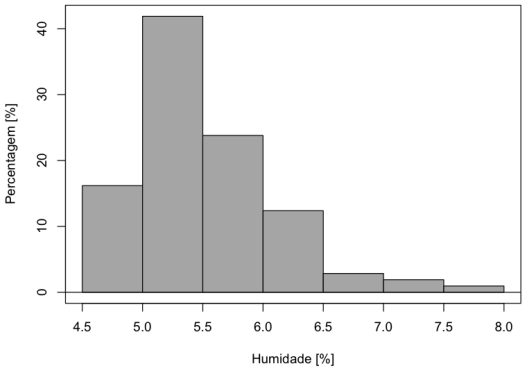


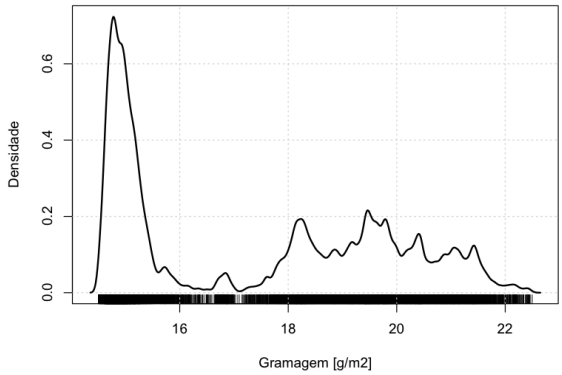


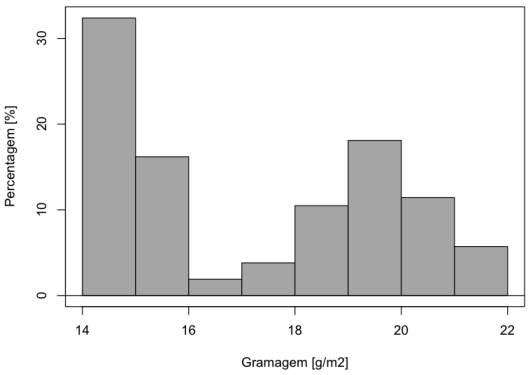

Densidade
pH águas brancas

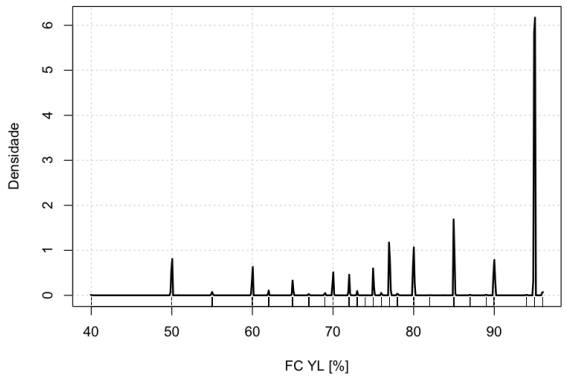

Percentagem [%]
pH águas brancas

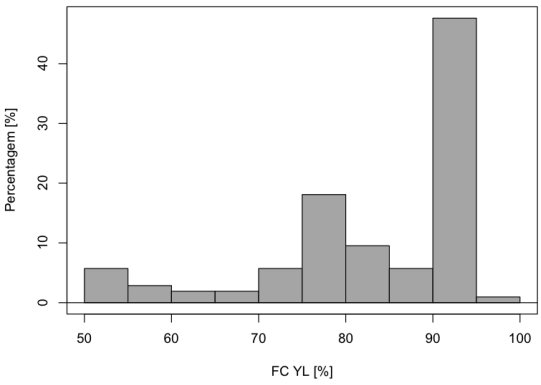

Densidade
Humidade [%]

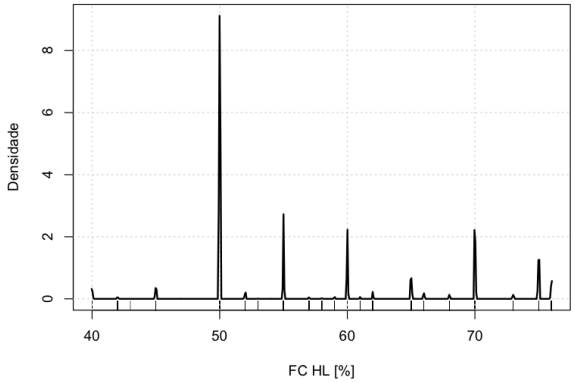

Percentagem [%]
Humidade [%]

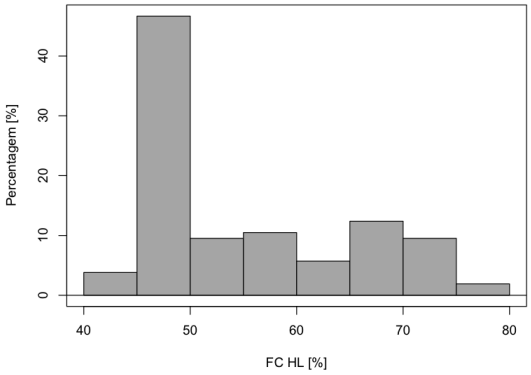

Densidade
Gramagem [g/m2]

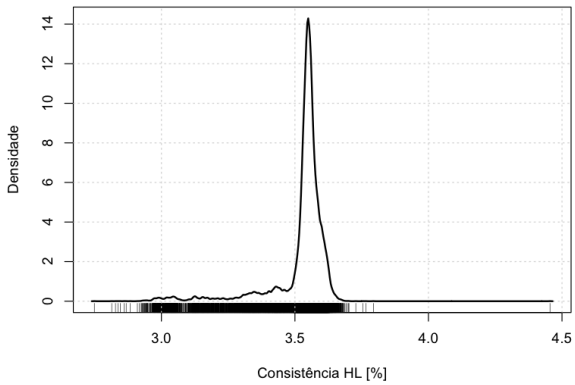

Percentagem [%]
Gramagem [g/m2]

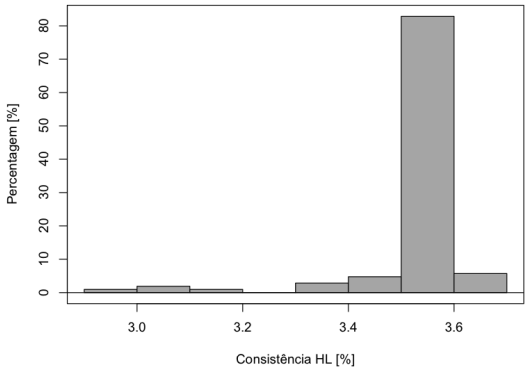

Densidade
FC YL [%]

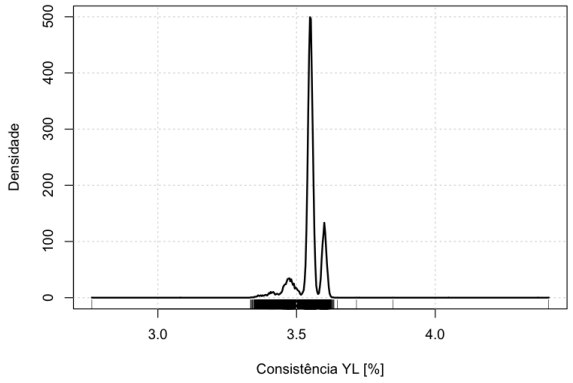

Percentagem [%]
FC YL [%]

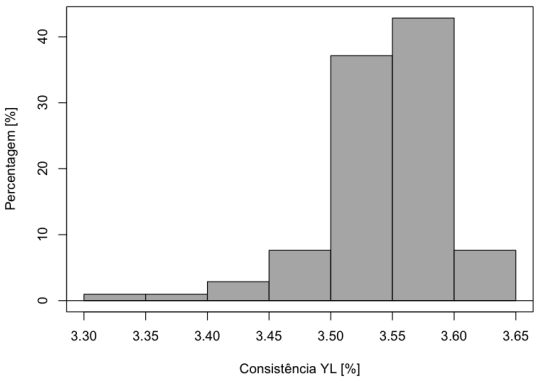

Densidade
0
2
4
6
8
40
50
60
70
FC HL [%]

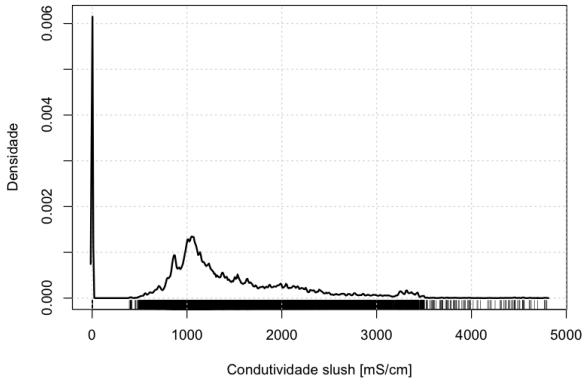

Percentagem [%]
0
10
20
30
40
40
50
60
70
80
FC HL [%]

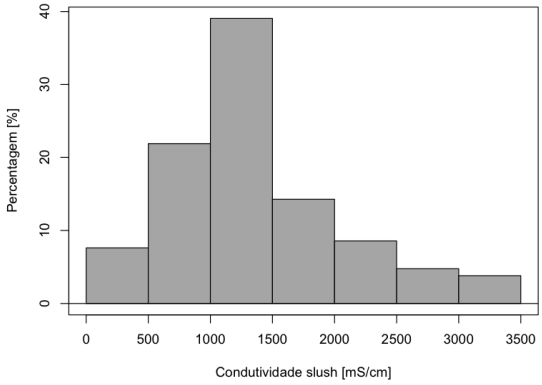

Densidade
0
2
4
6
8
10
12
14
3.0
3.5
4.0
4.5
Consistência HL [%]

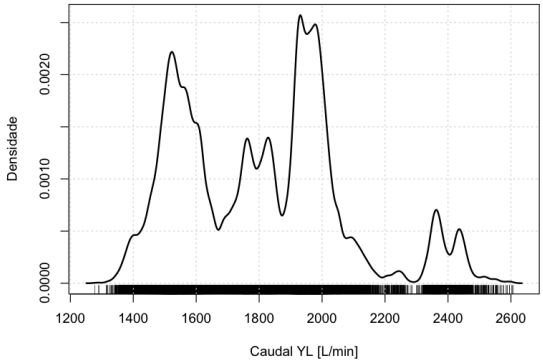

Percentagem [%]
0
10
20
30
40
50
60
70
80
3.0
3.2
3.4
3.6
Consistência HL [%]

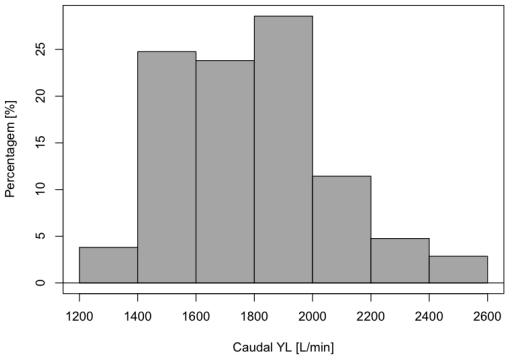

Densidade
0
100
200
300
400
500
3.0
3.5
4.0
Consistência YL [%]

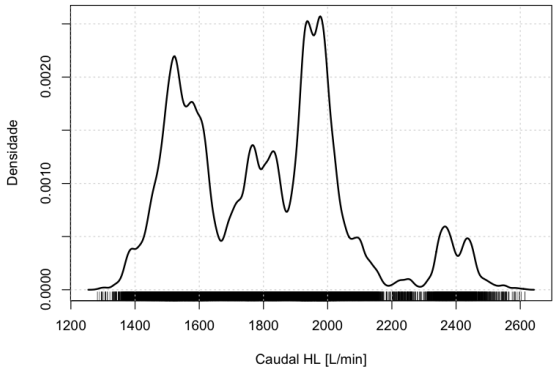

Percentagem [%]
0
10
20
30
40
3.30
3.35
3.40
3.45
3.50
3.55
3.60
3.65
Consistência YL [%]

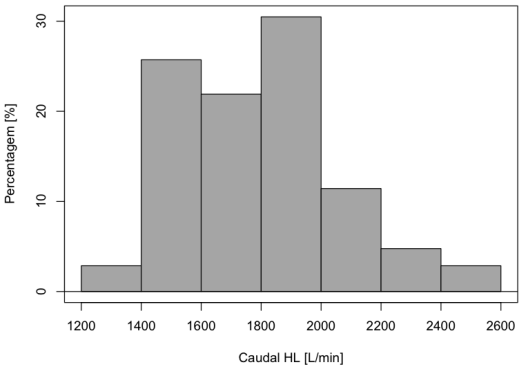

Densidade
0.000
0.002
0.004
0.006
0
1000
2000
3000
4000
5000
Condutividade slush [mS/cm]

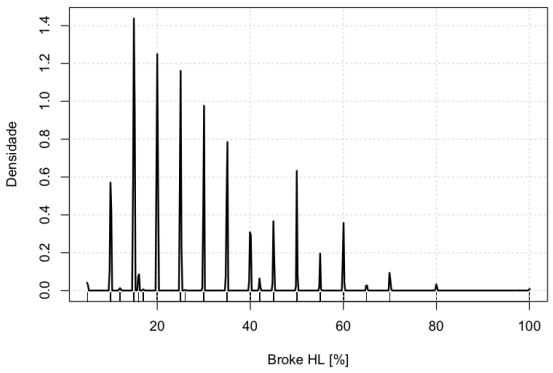

Percentagem [%]
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500
Condutividade slush [mS/cm]

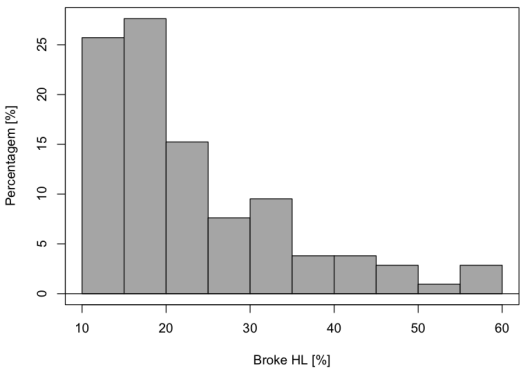

Densidade
0.0000
0.0010
0.0020
1200
1400
1600
1800
2000
2200
2400
2600
Caudal YL [L/min]


Percentagem [%]
0
5
10
15
20
25
1200
1400
1600
1800
2000
2200
2400
2600
Caudal YL [L/min]


Densidade
0.0000
0.0010
0.0020
1200
1400
1600
1800
2000
2200
2400
2600
Caudal HL [L/min]


Percentagem [%]
0
10
20
30
1200
1400
1600
1800
2000
2200
2400
2600
Caudal HL [L/min]


Densidade
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
20
40
60
80
100
Broke HL [%]


Percentagem [%]
0
5
10
15
20
25
10
20
30
40
50
60
Broke HL [%]

ANEXO C – ANÁLISE DE CORRELAÇÃO ESTRATIFICADA

CLASSE risco alto	Adesivo	Broke HL	Broke YL	Caudal HL	Caudal YL	Condutividade slush	Consistência HL	Consistência YL	FC HL	Ref. FC	FC YL	Gramagem	Humidade	Lamina de limpeza	Lançamento	Ref. 1 FL	Ref. 2 FL	ph águas brancas	ph slush	Vacuo sucção	Velocidade
Adesivo	1																				
Broke HL	-0,2	1																			
Broke YL	-0	0,5	1																		
Caudal HL	-0,3	-0	0,2	1																	
Caudal YL	-0,3	-0,1	0,2	1	1																
Condutividade slush	0,3	0,1	0	0,1	0,1	1															
Consistência HL	-0,3	0,1	0	-0,1	-0,1	-0,3	1														
Consistência YL	-0,3	-0,2	0,1	0,4	0,4	0	0,2	1													
FC HL	0,4	0	-0,2	-0,4	-0,4	0	-0,3	-0,6	1												
Ref. FC	0,4	-0,2	0,1	0,2	0,2	0,3	-0,2	0,1	-0,2	1											
FC YL	0,3	-0,1	-0,2	-0,7	-0,7	0	-0,1	-0,2	0,5	-0,1	1										
Gramagem	0,3	0.05	0,1	0,4	0,4	0,3	-0,3	-0,1	0,3	0,1	-0,1	1									
Humidade	-0,2	0,1	0,2	0,6	0,6	-0,1	0,2	0	-0,3	-0	-0,6	0,3	1								
Lamina de limpeza	-0,3	-0,1	0	0	0	-0,4	0,2	0	-0,3	-0,2	-0,2	-0,1	0,2	1							
Lançamento	0,9	0,2	-0	-0,4	-0,4	0,2	-0,3	-0,6	0,5	0,3	0,3	0,2	-0,3	-0,2	1						
Ref. 1 FL	-0,6	0,2	0	-0,1	-0,1	-0,2	0,3	0,2	-0,2	-0,5	-0,2	-0,3	0,1	0,4	-0,5	1					
Ref. 2 FL	0,6	-0,2	0,1	0,2	0,2	0,3	-0,3	0,1	-0,1	0,7	-0	0,2	-0,1	-0,4	0,4	-0,8	1				
ph águas brancas	-0	-0,1	-0,1	-0,2	-0,2	0	-0	-0,1	0,2	-0,3	0,3	-0	-0,1	0	-0	0	-0,2	1			
ph slush	-0,2	0,2	0,1	0,1	0,1	0,5	0,1	-0	-0	-0,2	-0,1	0,1	0,2	-0,2	-0,2	0,1	-0,2	0,1	1		
Vacuo sucção	0,7	-0,2	-0	-0,2	-0,2	0,4	-0,4	-0,5	0,5	0,3	0,2	0,3	-0,2	-0,4	-0,8	-0,6	0,5	-0,1	-0,1	1	
Velocidade	-0,7	-0	-0	0,3	0,3	-0,2	0,3	0,3	-0,3	-0,3	-0,1	-0,2	0,3	0,1	-0,8	0,3	-0,3	0,2	0,2	-0,6	1

CLASSE médio	Adesivo	Broke HL	Broke YL	Caudal HL	Caudal YL	Condutividade slush	Consistência HL	Consistência YL	FC HL	Ref. FC	FC YL	Gramagem	Humidade	Lamina de limpeza	Lançamento	Ref. 1 FL	Ref. 2 FL	ph águas brancas	ph slush	Vacuo sucção	Velocidade
Adesivo	1																				
Broke HL	-0,2	1																			
Broke YL	-0,1	0,5	1																		
Caudal HL	-0,2	0	0,2	1																	
Caudal YL	-0,2	0	0,2	1	1																
Condutividade slush	0,3	-0,1	-0,1	0,1	0,1	1															
Consistência HL	-0,3	0,1	0	-0,1	-0,1	-0,3	1														
Consistência YL	-0,3	0,1	0,1	0,4	0,4	0,1	0,1	1													
FC HL	0,4	-0,2	-0,2	-0,4	-0,4	0,1	-0,3	-0,5	1												
Ref. FC	0,4	0	0,1	0,2	0,2	0,2	-0	0,1	-0,3	1											
FC YL	0,3	-0,2	-0,3	-0,7	-0,7	0,1	-0,1	-0,2	0,6	-0,2	1										
Gramagem	0,3	-0,1	0	0,4	0,4	0,3	-0,3	-0,1	0,3	0	-0	1									
Humidade	-0,2	0,1	0,2	0,6	0,6	-0,1	0,2	0,1	-0,4	0,1	-0,6	0,2	1								
Lamina de limpeza	-0,3	0,1	0,1	0	0	-0,4	0,3	0	-0,3	0,1	-0,2	-0,1	0,1	1							
Lançamento	0,9	-0,1	-0,1	-0,3	-0,3	0,2	-0,4	-0,5	0,5	0,3	0,3	0,2	-0,2	-0,2	1						
Ref. 1 FL	-0,6	0,2	0,1	-0,1	-0,1	-0,3	0,3	0,2	-0,2	-0,4	-0,1	-0,3	0,1	0,4	-0,5	1					
Ref. 2 FL	0,6	-0,2	-0	0,2	0,2	0,4	-0,3	0	-0	0,6	-0	0,2	-0	-0,4	0,4	-0,8	1				
ph águas brancas	-0,1	-0	-0	-0,2	-0,2	-0	-0	-0,1	0,2	-0,3	0,2	-0,1	-0,1	0,1	-0,1	0,1	-0,1	1			
ph slush	-0,2	0,1	0	0,1	0,1	0,5	0	0	0	-0,2	0	0,1	0,1	-0,2	-0,2	0,1	-0,1	0	1		
Vacuo sucção	0,7	-0,2	-0,1	-0,1	-0,1	0,5	-0,4	-0,5	0,5	0,2	0,2	0,3	-0,2	-0,5	0,8	-0,6	0,5	0,1	-0,1	1	
Velocidade	-0,7	-0,1	-0	0,3	0,3	-0,1	0,3	0,2	-0,3	-0,3	-0,1	-0,2	0,2	0,1	-0,8	0,3	-0,3	0,2	0,2	-0,6	1

CLASSE baixo	Adesivo	Broke HL	Broke YL	Caudal HL	Caudal YL	Condutividade slush	Consistência HL	Consistência YL	FC HL	Ref. FC	FC YL	Gramagem	Humidade	Lamina de limpeza	Lançamento	Ref. 1 FL	Ref. 2 FL	ph águas brancas	ph slush	Vacuo sucção	Velocidade
Adesivo	1																				
Broke HL	-0,1	1																			
Broke YL	-0,1	0,5	1																		
Caudal HL	-0,2	-0,1	-0	1																	
Caudal YL	-0,2	-0,1	-0	1	1																
Condutividade slush	0,3	-0,1	-0,1	0,1	0,1	1															
Consistência HL	-0,3	0,1	0,1	-0,1	-0,1	-0,3	1														
Consistência YL	-0,3	0,2	0,1	0,3	0,3	0	0,1	1													
FC HL	0,4	-0,3	-0,1	-0,4	-0,4	0,2	-0,3	-0,5	1												
Ref. FC	0,4	0	0,1	0,1	0,1	0,3	-0,2	0,1	-0,2	1											
FC YL	0,2	-0,2	-0,1	-0,6	-0,6	0,1	-0,1	-0,2	0,5	-0,1	1										
Gramagem	0,3	-0,2	0	0,5	0,5	0,2	-0,3	-0	0,3	0,1	-0,1	1									
Humidade	-0,2	0,1	0,1	0,6	0,6	-0,1	0,2	0,1	-0,3	-0	-0,6	0,3	1								
Lamina de limpeza	-0,3	0,1	0,1	0	0	-0,4	0,2	-0	-0,2	-0,2	-0,2	-0,1	0,2	1							
Lançamento	0,9	-0	-0,1	-0,3	-0,3	0,2	-0,4	-0,5	0,4	0,3	0,3	0,2	-0,3	-0,2	1						
Ref. 1 FL	-0,6	0,2	0,1	-0,1	-0,1	-0,2	0,3	0,2	-0,3	-0,5	-0,2	-0,3	0,1	0,4	-0,5	1					
Ref. 2 FL	0,6	-0,1	-0,1	0,2	0,2	0,4	-0,3	0,1	0	0,7	-0	0,2	-0,1	-0,4	0,5	-0,8	1				
ph águas brancas	-0,1	0	0,1	-0,3	-0,3	-0,1	-0,1	-0,1	0,1	-0,2	0,2	-0,2	-0	0,2	-0,1	0,1	-0,2	1			
ph slush	-0,2	0	-0	0,1	0,1	0,5	0,1	0	0,1	-0,2	0	0,1	0,1	-0,1	-0,2	0,1	-0,1	0	1		
Vacuo sucção	0,7	-0,2	-0,1	-0,1	-0,1	0,5	-0,4	-0,5	0,5	0,3	0,2	0,3	-0,3	-0,5	0,8	-0,5	0,5	-0,2	-0,1	1	
Velocidade	0,6	-0,3	-0,1	0,3	0,3	-0,1	0,2	0,2	-0,2	-0,3	-0,1	-0,2	0,3	0,1	-0,7	0,3	-0,3	0,2	0,3	-0,5	1

CLASSE nulo	Adesivo	Broke HL	Broke YL	Caudal HL	Caudal YL	Condutividade slush	Consistência HL	Consistência YL	FC HL	Ref. FC	FC YL	Gramagem	Humidade	Lamina de limpeza	Lançamento	Ref. 1 FL	Ref. 2 FL	ph águas brancas	ph slush	Vacuo sucção	Velocidade
Adesivo	1																				
Broke HL	0,2	1																			
Broke YL	0,2	0,5	1																		
Caudal HL	-0,3	-0,2	-0	1																	
Caudal YL	-0,3	-0,2	-0	1	1																
Condutividade slush	0,3	0	0,1	0	0	1															
Consistência HL	-0,3	0,1	0	-0,1	-0,1	-0,2	1														
Consistência YL	-0,3	0,1	0	0,3	0,3	-0	0,1	1													
FC HL	0,4	-0,2	-0	-0,4	-0,4	0,2	-0,3	-0,5	1												
Ref. FC	0,4	0,2	0,1	0,1	0,1	0	-0,1	0,1	-0,2	1											
FC YL	0,3	-0,2	-0,1	-0,6	-0,6	0,1	-0,1	-0,1	0,5	-0,1	1										
Gramagem	0,3	-0,1	0	0,5	0,5	0,3	-0,2	-0,1	0,3	-0	-0,1	1									
Humidade	-0,3	0	0	0,6	0,6	-0,1	0,2	0	-0,4	0	-0,7	0,2	1								
Lamina de limpeza	0,2	0,1	0,1	0,1	0,1	0,1	-0,1	-0	-0,2	0,2	-0,2	0,1	0,1	1							
Lançamento	0,8	0,2	0,1	-0,3	-0,3	0,2	-0,4	-0,5	0,5	0,3	0,3	0,2	-0,3	0,1	1						
Ref. 1 FL	-0,6	0	-0,2	-0,1	-0,1	-0,2	0,3	0,3	-0,3	-0,4	-0,1	-0,3	0,1	-0,2	-0,6	1					
Ref. 2 FL	0,6	0,1	0,3	0,1	0,1	0,2	-0,3	-0	0,1	0,6	0	0,1	-0,1	0,3	0,5	-0,8	1				
ph águas brancas	-0,1	0	0,1	-0,3	-0,3	0,1	0,1	-0,2	0,2	-0,3	0,1	-0	-0,1	0	-0,1	0,1	-0,3	1			
ph slush	-0,3	-0,1	-0,2	0	0	0,6	0,1	0	0,1	-0,3	0,1	0,1	0	-0,2	-0,2	0,3	-0,4	0,2	1		
Vacuo sucção	0,7	0	0,2	-0,2	-0,2	0,3	-0,4	-0,5	0,5	0,3	0,1	0,2	-0,2	0,2	0,8	-0,6	0,6	-0,1	-0,3	1	
Velocidade	-0,7	-0,3	-0,1	0,3	0,3	0	0,2	0,2	-0,2	-0,3	-0,1	-0,2	0,1	-0,2	-0,8	0,4	-0,4	0,3	0,3	-0,5	1

ANEXO D – PCA – CÓDIGO PYTHON

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.decomposition import PCA
import pandas as pd
from sklearn.preprocessing import StandardScaler
df=pd.read_excel('/Users/xxx.xlsx', encoding='latin1')
X = df.iloc[:,0:21].values
y = df.iloc[:,21].values
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = None)
med = np.mean(X_train, axis=0)
mat_cov = (X_train - med).T.dot((X_train - med)) / (X_train.shape[0]-1)
print('\nMatriz de covariância \n%s' %mat_cov)
pca = PCA(n_components = 21)
X_train_pca = pca.fit_transform(X_train)
X_test_pca = pca.transform(X_test)
cov_mat = np.cov(X_train_pca.T)
auto_vals, vecs_proprios = np.linalg.eig(cov_mat)
print('\vetoresproprios: \n%s' %vecs_proprios)
auto_pairs = [(np.abs(auto_vals[i]), vecs_proprios[:,i]) for i in range(len(auto_vals))]
print('\autovalores em ordem decrescente:')
for i in auto_pairs:
    print(i[0])
variancia_explicada = pca.explained_variance_ratio_
print ('\nVariancia explicada por cada PC: \n%s' %variancia_explicada)
per_var = np.round(pca.explained_variance_ratio_* 100, decimals=1)
labels = [str(x) for x in range(1, len(per_var)+1)]
plt.bar(left=range(1, len(per_var)+1), height=per_var, tick_label=labels)
plt.ylabel('Percentagem de variância explicada')
plt.xlabel('Componente principal')
plt.title('Plot')
plt.show()
variancia_acumulada=np.cumsum(np.round(pca.explained_variance_ratio_, decimals=4)*100)
print ('\nVariancia acumulada: \n%s' %variancia_acumulada)
plt.plot(variancia_acumulada)
plt.show()
a=np.abs(pca.components_[0:22]).argsort()[::1][:22]
print ('\nMatriz contribuicao de cada variavel a cada PC: \n%s' %a)
```

ANEXO E – RESULTADOS DA ANÁLISE DE COMPONENTES PRINCIPAIS

PC	Autovalor	Contribuição [%]	Acumulada [%]
1	5.54	26	26
2	3.67	17	43
3	2.16	10	53
4	1.55	7	60
5	1.50	7	67
6	1.02	5	72
7	0.95	5	77
8	0.88	4	81
9	0.75	4	85
10	0.59	3	88
11	0.45	2	91
12	0.40	2	93
13	0.32	1	94
14	0.30	1	95
15	0.27	1	96
16	0.19	1	97
17	0.15	1	98
18	0.14	1	99
19	0.12	1	100
20	0.06	0	100
21	0.00	0	100

ANEXO F – CONTRIBUIÇÃO DE CADA VARIÁVEL AOS COMPONENTES PRINCIPAIS

PC / Variável	lamina de limpeza	velocidade	gramagem	humidade	adesivo	lançamento	refinação FC	refinação FL 1	refinação FL 2	FC YL	broke YL	FC HL	broke HL	vacuo sucção	consistência HL	caudal HL	consistência YL	caudal YL	condutividade slush	pH slush	pH águas brancas
1	10	12	20	19	2	0	18	6	15	17	14	9	16	3	8	11	7	1	13	4	5
2	12	5	0	19	1	4	10	13	18	14	16	11	7	20	2	6	3	8	9	17	15
3	7	3	5	13	4	8	16	9	0	20	17	15	14	1	6	11	10	18	2	12	19
4	5	11	14	16	13	6	3	4	1	2	8	9	15	17	7	0	20	19	18	10	12
5	14	17	15	4	13	20	12	7	10	19	1	5	11	9	6	18	8	2	0	3	16
6	2	12	13	0	4	6	16	3	15	17	5	11	9	8	18	7	14	1	19	10	20
7	13	5	9	17	15	10	4	18	8	19	12	11	1	2	3	16	7	6	14	0	20
8	19	5	17	15	12	4	8	6	20	13	9	11	10	2	1	7	18	3	16	14	0
9	20	15	17	19	12	8	5	3	1	6	18	11	10	0	7	13	4	16	9	2	14
10	14	15	17	18	8	7	2	19	5	6	13	12	3	11	9	4	16	0	1	10	20
11	16	14	4	9	13	20	5	17	15	1	10	11	18	8	0	3	2	19	12	7	6
12	1	15	17	5	8	11	20	0	4	3	2	7	10	16	14	19	9	18	12	13	6
13	8	7	0	5	9	20	2	16	6	18	15	17	4	11	19	1	14	13	3	10	12
14	7	5	13	15	14	17	0	8	18	6	4	20	10	19	9	2	1	12	16	11	3
15	3	10	14	2	20	13	17	15	8	12	0	18	7	6	19	5	16	9	4	1	11
16	0	10	12	20	14	8	6	9	13	3	16	1	7	15	17	18	19	2	11	4	5
17	12	0	6	8	10	19	18	7	14	20	16	2	11	15	3	17	1	5	13	9	4
18	9	3	12	14	20	0	10	16	1	15	17	11	4	2	6	5	13	19	18	7	8
19	3	5	10	12	0	20	6	4	14	9	15	19	2	17	11	7	16	18	1	8	13
20	12	7	3	11	10	0	14	6	20	19	18	16	15	17	8	1	9	13	2	4	5
21	13	4	0	11	8	19	18	12	10	7	20	6	3	14	2	5	9	16	1	17	15

Legenda:	
20	alta
15	média alta
10	média
5	média baixa
0	baixa

ANEXO G – ÁRVORES DE DECISÃO – CÓDIGO PYTHON

```
import numpy as np
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn import cross_validation
def importdata():
    df=pd.read_excel('/Users/xxx.xlsx',encoding='latin1')
    return df
def kfolds(df):
    Y=df.values[:,19]
    skf=cross_validation.StratifiedKFold(Y, n_folds=5, shuffle=True, random_state=None)
    return skf, Y
def split(skf,df,Y):
    for train_index, test_index in skf:
        X=df.values[:,0:19]
        X_train, X_test = X[train_index],X[test_index]
        y_train, y_test = Y[train_index],Y[test_index]
    return X,Y,X_train,X_test,y_train,y_test
def train_using_gini(X_train, X_test, y_train):
    clf=DecisionTreeClassifier(criterion="gini",random_state=None,max_depth=21)
    clf.fit(X_train, y_train)
    n_nodes = clf.tree_.node_count
    print ('\nNr nodes in the decision tree: ',n_nodes)
    return clf
def prediction(X_test,clf):
    y_pred=clf.predict(X_test)
    return y_pred
def cal_performance(y_test,y_pred):
    print('\nMatriz confusao:\n',confusion_matrix(y_test,y_pred))
    print ('\nExatidão:\n',accuracy_score(y_test,y_pred)*100)
    print('\nReport:\n',classification_report(y_test,y_pred))
def main():
    df=importdata()
    skf,Y=kfolds(df)
    X,Y,X_train,X_test,y_train,y_test=split(skf,df,Y)
    clf=train_using_gini(X_train,X_test,y_train)
    y_pred_gini = prediction(X_test, clf)
    cal_performance(y_test, y_pred_gini)
if __name__=="__main__":
    main()
```

ANEXO H – MÁQUINA DE VETORES DE SUPORTE – CÓDIGO PYTHON

```
import numpy as np
import pandas as pd
from sklearn import svm
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn import cross_validation
def importdata():
    df=pd.read_excel('/Users/xxx.xlsx',encoding='latin1')
    return df
def kfolds(df):
    Y=df.values[:,19]
    skf=cross_validation.StratifiedKFold(Y, n_folds=5, shuffle=True, random_state=None)
    return skf, Y
def split(skf,df,Y):
    for train_index, test_index in skf:
        X=df.values[:,0:19]
        X_train, X_test = X[train_index],X[test_index]
        y_train, y_test = Y[train_index],Y[test_index]
    return X,Y,X_train,X_test,y_train,y_test
def train_svm(X_train, X_test, y_train):
    clf = svm.SVC(kernel='rbf', random_state=None, gamma=1, C=1) # RBF Kernel
    clf.fit(X_train, y_train)
    return clf
def prediction(X_test,clf):
    y_pred=clf.predict(X_test)
    return y_pred
def cal_performance(y_test,y_pred):
    print('\nMatriz confusão:\n',confusion_matrix(y_test,y_pred))
    print ('\nExatidão:\n',accuracy_score(y_test,y_pred)*100)
    print('\nReport:\n',classification_report(y_test,y_pred))
def main():
    df=importdata()
    skf,Y=kfolds(df)
    X,Y,X_train,X_test,y_train,y_test=split(skf,df,Y)
    clf=train_svm(X_train,X_test,y_train)
    y_pred_svm = prediction(X_test, clf)
    cal_performance(y_test, y_pred_svm)
if __name__=="__main__":
    main()
```