UNIVERSIDADE D
COIMBRA

André Filipe de Oliveira Ribeiro

WhyWheMove

Ferramenta de Apoio à Decisão para Desenvolvimento Sustentado do Espaço Urbano

Dissertação no âmbito do Mestrado em Engenharia Informática na área Comunicações, Serviços e Infraestruturas orientadas pela Professa Doutora Ana Oliveira Alves e Professor Doutor Carlos Lisboa Bento e apresentada à Faculdade de Ciência e Tecnologia da Universidade de Coimbra ao Departamento de Engenharia Informática.

Julho de 2022

Resumo

Dispositivos que possuem a tecnologia de Posicionamento por Satélite, nomeadamente o Sistemas de Posicionamento Global (GPS) têm sido usados regularmente na recolha de dados de viagens realizadas no espaço urbano. Este uso tem beneficiado a melhoria e qualidade da informação disponível para o desenho e planeamento urbano.

Os dispositivos mais usados no quotidiano das pessoas correspondem aos "*smartphones*". Eles registam a posição, tempo e velocidade da viagem com o auxílio da tecnologia GPS. Estas características juntamente com o destino da viagem e as atividades (POIs) que lá se podem realizar permitem identificar o propósito de viagem. O propósito de viagem corresponde à identificação da realização de uma determinada atividade num determinado lugar geográfico, por exemplo, lazer, trabalho, educação, entre outras atividades quotidianas. Atualmente a inferência do propósito de viagem não pode ser registada automaticamente pelos dispositivos tecnológicos, porque estas tecnologias ainda não conseguem determinar com qualidade esta informação visto que dependem de diversos fatores: altura do dia e da semana, a sua proximidade à casa e/ou trabalho, modos de transportes usados, duração da atividade/viagem, entre outros. Desta forma, o estudo feito neste trabalho consiste em idealizar um método de aprendizagem computacional em prol desta identificação com base em dados recolhidos por *smartphones*, onde o método apresenta boas precisões na previsão desta identificação. Ao idealizarmos o modelo permitimos apoiar e impactar de forma positiva em certas estratégias promocionais, eventos, estruturações urbanas, trajetórias dos meios de transportes, entre outros, com base nos hábitos populacionais. Selecionar "*features*" e trabalhá-las também será incluído neste estudo, integrando no modelo, para proceder à obtenção da melhor previsão possível ao mesmo tempo que se verifica o impacto desta integração.

No decorrer deste trabalho foram analisados diferentes *features* do *dataset* Breadcrumbs (Moro et al., 2019), com base nas técnicas de recolha e seleção de dados de uma base de dados prosseguindo com a implementação de algoritmos que integram estas *features* e, posteriormente, determinem a melhor precisão possível na previsão da inferência do propósito de viagens.

Palavras-Chave

Mobilidade urbana; Atribuição do propósito; Inferência de atividade; Aprendizagem computacional; *Random Forest*

Abstract

Devices that have Satellite Positioning technology, namely Global Positioning Systems (GPS) have been regularly used to collect data from trips carried out in urban space. This use has benefited the improvement and quality of the information available for urban design and planning.

The most used devices in people's daily lives correspond to smartphones. They record the position, time, and speed of travel with the help of GPS technology. These characteristics, together with the destination of the trip and the activities (POIs) that can be carried out there, allow identifying the purpose of the trip. The purpose of travel corresponds to the identification of carrying out a certain activity in a certain geographical place, for example, leisure, work, education, among other daily activities. Currently, the inference of the purpose of travel cannot be automatically registered by technological devices because these technologies are still unable to determine this information with quality. In this way, the study carried out in this work consists of devising a computational learning method for this identification based on data collected by smartphones, where the method presents good accuracy in predicting this identification. When we idealized the model, we supported and positively impacted certain promotional strategies, events, urban structures, transport routes, among others, based on population habits. Selecting features and working with them will also be integrated in this study, embedded in the model, to proceed to obtain the best possible forecast while verifying the impact of this integration.

In the course of this work, different features of the Breadcrumbs dataset (Moro et al., 2019) were analysed, based on the techniques of collecting and selecting data from a database, proceeding with the implementation of algorithms that integrate these features and, later, determine the best possible accuracy in predicting travel purpose inference.

Keywords

Urban mobility; Purpose imputation; Activity inference; Machine learning; Random Forest

Agradecimentos

À minha mãe, pelo amor, carinho, apoio, por tudo.

A todos os meus familiares, que sempre se interessaram e me apoiaram, nesta caminhada.

Ao amigos e colegas do Mestrado de Engenharia Informática.

À Ana Sofia, por estar ao meu lado, apoiando-me com carinho e paciência.

À equipa AmiLab, por toda a disponibilidade e orientação dada no decorrer deste trabalho.

Em especial, à Professora Ana Alves por todo o acompanhamento, apoio, paciência e valiosa orientação ao longo deste projeto.

Índice

Lista de figuras

Lista de tabelas

Anexos

Siglas e Acrónimos

ACC Accuracy Classification Score

ANN Artificial Neural Networks

API Application Programming Interface

BN Bayesian Network

CASI Computer Assisted Self-Interviews

CATI Computer Assisted Telephone Interview

DBSCAN Density-based spatial clustering of applications with noise

DT Decision Tree

GPS Global Positioning System

GTFS General Transit Feed Specification

HAC Hierarchical Agglomerative Clustering

HG Human Geography

LBSN Location-Based Social Network

MCC Matthew's Correlation Coefficient

MTMC Mobility and Transport Microcensus

NB Naive Bayes

NL Nested Logit

PAPI Paper-And-Pencil Interview

POI Point of Interest

QGIS Quantum Geographic Information System

RF Random Forest

RGPD Regulamento Geral sobre a Proteção de Dados

SIG Sistemas de Informação Geográfica

SRID Spatial Reference System ID

SML Supervised Machine Learning

SVM Support Vector Machine

TDM Transportation Demand Management

TI/SI Trip Identification/Segment Identification

TS Transportation Science

WGS World Geodetic System

Capítulo 1 Introdução

As trajetórias das viagens realizadas pelos humanos no seu dia a dia apresentam um nível de regularidade temporal e espacial (espaço-temporal), porque a maior parte das atividades realizadas são rotineiras.

Ao longo dos anos houve uma rápida evolução dos dispositivos computacionais móveis, que hoje em dia possuem a capacidade de registar uma massiva quantidade de trajetórias dos utilizadores através do Sistema de Posicionamento Global ou *Global Positioning System* (GPS), ou seja, através de serviços baseados na localização (Geo-referenciados), redes sociais baseadas em localização, transportes e aplicativos. O agrupamento dessas trajetórias é estudado por vários investigadores onde apresentam vários métodos para a identificação dos modos de transporte usados, propósito de viagens, rotas e outros registos com base nos dados de GPS. Para realizar estes estudos é necessário possuir um conjunto de dados já trabalhado para posteriormente aplicar os vários métodos existentes, como aprendizagem computacional, métodos estatísticos e métodos baseados em regras.

O conjunto de dados "Breadcrumbs", ou seja, conjunto de dados de grande mobilidade foi disponibilizado para este estudo pelos seus autores (Moro et al., 2019). Esta campanha de recolha de dados foi feita na primavera de 2018 num período de 3 meses. Este conjunto contém dados já trabalhados de vários sensores, incluindo GPS, GSM, WiFi, Bluetooth, recolhidos pelos "*smartphones"* de 81 pessoas, na Suíça. Além dos dados do sensor do *smartphone*, o *dataset* Breadcrumbs também contém dados etiquetados, *ground truth*, sobre os locais visitados pelas pessoas, incluindo a sua identificação bem como atributos demográficos, registos de contacto, eventos, informações de estilo de vida social e relações entre os participantes. Esses atributos exclusivos tornam o *dataset* Breadcrumbs ideal para esta área de pesquisa, onde se inclui na inferência do propósito de viagem.

1.1 Contexto

A identificação dos modos de viagem que as pessoas usam nos seus deslocamentos diários com a identificação do propósito de viagem contribuem para uma melhor compreensão da mobilidade urbana. Assim, os objetivos de viagem desempenham um papel importante, uma vez que as escolhas de mobilidade (modos de viagem, rotas, horários, datas, etc.) são feitas com o objetivo de realizarem determinadas atividades específicas, por exemplo, trabalho/educação, compras, restaurantes, vida social, etc.

Modelos baseados em atividades realizadas pelo ser humano têm ganho popularidade nos últimos anos e são construídos a partir de dados obtidos da realização de um grande número de viagens com um determinado propósito. Nesse sentido, a recolha de dados de viagens é cada vez mais realizada e nos tempos de hoje é feito através das tecnologias de Sistema de Posicionamento Global ou *Global Positioning System* (GPS) incluída nos *smartphones*. Embora os dados de GPS possam fornecer informações espaço-temporal (latitude, longitude, hora, data, etc.) precisas de movimentos veiculares ou pessoais, o modo de transporte e o propósito de viagem não podem ser obtidos diretamente dos dados de GPS. Contudo, a identificação do segmento de viagem com a disponibilidade de dados contínuos de dados GPS é fundamental para determinar o propósito de

viagem e dar respostas futuras das atividades realizadas pelos viajantes num determinado local específico.

1.2 Problema

A deteção do propósito de viagem a partir de dados de sensores de GPS em *smartphones* ou em outros dispositivos dedicados surgiu como desafio de investigação nos últimos anos. No entanto, embora os sistemas de GPS façam uma boa recolha de dados é necessário identificar o erro associado a cada ponto de localização registado, completar com informações recolhidas pela falta de disponibilidade do sinal de GPS em espaços urbanos: túneis, áreas bastantes congestionadas pela edificação (edifícios altos que impedem a correta exceção do sinal GPS), interior dos edifícios, etc. Esta falta de informação nos *datasets* com a falta de disponibilidade de dados anotados faz com que exista no *dataset* uma ausência de informação precisa, para além das atividades não anotadas, porque o local onde determinado utilizador realiza a atividade pode não estar bem discriminada ou pode não estar de forma clara para atribuir um tipo de atividade à viagem. Assim, será necessário contactar com os entrevistados para obter informações de viagens não registadas como, por exemplo, pela falta de lembrança do acompanhamento dos aparelhos tecnológicos de GPS.

Quando o conjunto de dados em bruto são pré-processados torna os dados mais ricos, com um elevado número de "*features*", não apresentam os problemas que foram anteriormente referenciados. Sendo assim, é fundamental projetar e avaliar esses dados num algoritmo para facilitar a reprodutividade na identificação do propósito de viagem para a obtenção de uma melhor previsão possível.

1.3 Objetivos

O objetivo deste projeto é desenvolver um modelo que possa ser usado para identificar o propósito de viagem de diferentes utilizadores com base num conjunto de dados anonimizados de GPS, *dataset* Breadcrumbs, fornecido pelos seus autores (Moro et al., 2019). No entanto, as conclusões dos nossos resultados vão ser genéricas já que os seus resultados vão ser comparados com outro trabalho apresentado no estado da arte, autores (Montini et al., 2014), que inclusive foi realizado em outra região e com outro *dataset*, que por sua vez não está disponível publicamente, mas como grande parte do seu trabalho engloba dados que estão discriminados no *dataset* Breadcrumbs foi seguido e comparado os seus resultados. Neste sentido, as metas estabelecidas em relação a este estudo são:

- Estudar metodologias aplicadas em outros trabalhos na deteção do propósito de viagem (casa, trabalho, educação, shopping, vida social, atividades de desporto, etc.);
- Fazer a seleção e pré-processamento das *features* dos dados de GPS necessárias para determinarem os diferentes tipos de propósitos de viagem;
- Desenvolvimento de um modelo para identificar o propósito de viagem de vários utilizadores usando os dados anonimizados de GPS, aplicando métodos de aprendizagem computacional;
- Determinar a melhor precisão da inferência do propósito de viagem, ou seja, no melhor percentual de *output* da identificação do objetivo de deslocamento realizado numa viagem;
- Avaliar e validar o modelo e os seus resultados obtidos.

1.4 Estrutura

Este documento está organizado e dividido em seis capítulos que oferece uma visão geral do trabalho desenvolvido para esta tese ao longo do estágio. Este primeiro capítulo tem como objetivo fornecer uma breve introdução aos capítulos, disponibilizando a estrutura da tese. Nele apresenta o documento, descrevendo a contextualização, objetivos e problemas associados na inferência do propósito de viagem quando é utilizado no estudo dados anonimizados de GPS, estabelecendo uma ligação na evolução do estágio.

O segundo capítulo apresenta o estado da arte. Ele começa por apresentar os domínios de pesquisa necessários para o desenvolvimento deste estudo e as fontes de dados, que consistem em dados de vários utilizadores recolhidos a partir de viagens feitas num determinado período de tempo com um determinado propósito, em fontes de dados do Foursquare e outras informações que são usados em pesquisas de vários investigadores. Neste estudo ainda inclui diversos métodos utilizados por diversos investigadores na inferência do propósito de viagem, bem como na seleção e descrição das suas *features*.

O terceiro capítulo é dedicado à descrição e análise dos dados utilizados e contidos no *dataset* Breadcrumbs, posteriormente seguido do processamento dos seus dados para a integração neste estudo.

O quarto capítulo apresenta a discussão dos resultados obtidos e a sua avaliação.

O quinto e último capítulo apresenta conclusões alcançadas de acordo com o planeado e o executado, apresentando os riscos que surgiram no desenvolvimento desta dissertação e as suas mitigações.

O sexto capítulo finaliza o documento com a conclusão do trabalho desenvolvido, apresenta os desafios encarados e os trabalhos futuros de complementaridade.

Capítulo 2 Estado da Arte

O motivo pela qual as pessoas se movem tem atraído a atenção de muitos investigadores. Nesse sentido, compreender e prever o motivo da realização de uma determinada viagem está proporcionalmente relacionado com o propósito de viagem.

A revisão da literatura mostra que a previsão de destino de viagens no contexto urbano é um campo relativamente novo que está num avanço rápido devido à disponibilidade de dados precisos de pontos de movimento registados pela tecnologia GPS (*Global Positioning System)*, integrada nos *smartphones*. É importante entender o "*ground truth*", informações recolhidas para cada região, dos propósitos ou atividades das viagens que são usados para treinar os modelos.

O propósito de viagem pode ser útil para reduzir os possíveis locais alternativos que um utilizador irá visitar. Esta busca por locais rende melhores resultados que fazer uma pesquisa por hora, dia ou trajetória. Por exemplo, se um utilizador sai do local de trabalho, com o seu carro, e vai a um determinado local com diferentes pontos de interesse (POIs), e um desses corresponde a compras/restauração afastado do seu local de trabalho/casa, poderá ser útil saber que esse desvio corresponde ao "*land use*" shopping e à atividade refeição, porque esse "*land use*" poderá corresponder tanto à atividade de uma refeição como fazer compras. Um dado *land use* corresponde à classificação ou categorização de uma determinada atividade humana associado a uma parcela do território. Numa outra trajetória, se durante o deslocamento houver um desvio da rota habitual e um pai não levar consigo o filho, dificilmente estaria a ir para um destino com o tipo *land use* de educação, assumindo que o local de trabalho esteja fora das áreas educacionais, mesmo se o destino estivesse próximo dessa rota (Krause & Zhang, 2019). Estes autores usam o propósito da viagem para melhorar o modelo de previsão de qual será o destino do utilizador, determinando o objetivo depois de o utilizador já ter chegado ao destino. No nosso modelo, pretendemos determinar o propósito de viagem antes de ter chegado ao limite, muito próximo ao destino, que é conseguido de forma aproximada e atualizada com a fonte de POIs do Foursquare. Assim, a codificação do tipo de *land use* será importante para relacionar essa proximidade, que tem um efeito crítico na precisão da classificação dos propósitos de viagem e que será conseguida e atualizada com a fonte de POIs do Foursquare.

Inferir propósitos envolve buscar a atividade mais provável num tipo de local específico, ou seja, num ponto de interesse (POI), por exemplo, quando um utilizador tem como destino final de viagem ao local de trabalho, "Work". Diferentes pessoas podem envolver-se em atividades diferentes no mesmo local em casos de funcionalidades múltiplas, por exemplo, um shopping com restaurantes, compras e cinema (Nguyen et al., 2020). A inferência do propósito de viagem tem despertado o interesse tanto na comunidade científica como de empresas e decisores do espaço urbano. Em primeiro lugar, o propósito de viagem de uma determinada pessoa pode ajudar os grandes centros comerciais no que diz respeito a propagandas dos seus produtos, ajudar os meios de transportes públicos e privados nas estimativas de consumos para as viagens que irão ser feitas e assim levarem a um melhor planeamento da cidade relativamente a decisões de investimento. Estas finalidades consistem em fornecer aos clientes e aos serviços públicos e privados as melhores recomendações e os melhores serviços, antes de uma determinada pessoa realizar a sua viagem.

Inquéritos feitos aos utilizadores são usados para fazer a recolha de informações de indivíduos que, posteriormente, serão estudados para determinar o propósito de viagem. No entanto, esses

inquéritos têm várias limitações, como na recolha de dados, que com o passar dos anos e com o desenvolvimento das novas tecnologias foram sendo deduzidas a partir do contexto do utilizador.

Este capítulo aborda o estado da arte para este trabalho, compreendendo uma visão geral, existente, sobre a forma de como determinar o propósito de viagem, metodologias aplicadas, limitações, algoritmos e tecnologias usadas.

2.1 Domínios de pesquisa

Este trabalho está inserido em dois domínios de investigação:

a) Ciência em Transportes ou *Transportation Science* (TS) (Nguyen et al., 2020)
b) Geografia Humana ou *Human Geography* (HG) (Nguyen et al., 2020)

a) *Transportation Science* estuda o movimento de pessoas ou objetos de um lugar para outro através de métodos como física, pesquisa operacional e probabilidades. Na Figura 1 pode-se visualizar *features* comumente usadas em estudos de inferência do propósito de viagem. Neste domínio, os métodos de desenvolvimento são realizados após a conclusão e análise das viagens. Os estudos de TS são para criar o *ground truth* que se refere a relatórios de viagens confirmadas por meio de pesquisas de "*recall*" solicitadas, relatórios em papel (PAPI) (C. Chen et al., 2010) , websites (Shen & Stopher, 2013) e smartphones (Yazdizadeh et al., 2019). Outra forma seria rotular as atividades através da visualização e verificação do fim de viagem, por exemplo, no sistema de informação geográfica (SIG), mais conhecido como *Quantum Geographic Information System* (QGIS). Sendo assim, o foco em TS consiste no desenvolvimento e validação de métodos para derivar o propósito da viagem a partir de dados de GPS.

Feature category	Specific features	TS	HG
Geographic data	Polygon-based; POI; street map	√	√
	Working time of land use types	–	√
Activity related	Duration, time of day, day of week, start time	√	√
	Activity history	√	√
Trip related	Travel mode, mode of next trip, mode of previous trip	√	–
	Speed, distance, duration, start time, end time	√	√
Participant related	Home address, work address, occupation, age, school address, frequently visited places, working hours, gender, driving license, employment status, income, education degree, race, family structure, household information, marital status, driving frequency	√	–
Others	Social networking: Foursquare, Twitter	√	√
	Weather: temperature, precipitation, snow accumulation	√	–

Figura 1 - Estudos com dados de entrada no auxílio da inferência de propósito de viagem (Nguyen et al., 2020)

a) *Human Geography* estuda as inter-relações entre o lugar e as pessoas, o ambiente e como elas variam no espaço-tempo. Estudos de inferência de propósito baseados em GPS neste domínio realizam a análise pós-recolha com a finalidade em adquirir conhecimento geral da mobilidade de utilizadores e a sua localização de atividade. Estes estudos usam dados de grupos de utilizadores como se verifica no artigo do autor (C. Chen et al., 2018), reunidos por dispositivos de GPS dedicados. Desta forma, o HG coloca toda a atenção em elementos da atividade e organização humana como cultura, urbanização, população e transporte.

Assim, conclui-se que os domínios da Ciência em Transportes junto com o domínio da Geografia Humana são importantes na obtenção de um *ground truth.* Neste trabalho, o domínio que se enquadra corresponde ao *Transportation Science*, sendo este mais fidedigno para a inferência do propósito de viagem, porque os dados correspondem a deslocações/viagens realizadas por vários utilizadores em espaços urbanos e rurais.

2.2 Fontes de dados

Tradicionalmente, os dados relacionados com viagens eram principalmente recolhidos manualmente por meio de entrevistas a papel e lápis, *Paper-And-Pencil Interview* (PAPI), entrevistas por telefone com o auxílio do computador, *Computer Assisted Telephone Interview* (CATI), e autoentrevistas com auxílio do computador, *Computer Assisted Self-Interviews* (CASI), como mencionado no artigo (C. Chen et al., 2018). Todos estes métodos sofrem de várias limitações incluindo alto custo de pesquisa, custo espaço-tempo e coberturas de viagens imprecisas. Com evolução tecnológica surge a tecnologia GPS que melhorou estes problemas identificados anteriormente.

A tecnologia GPS é a forma confiável de obtenção de informação sobre cada etapa da viagem, substituindo a realização de inquéritos. Os dois objetivos de estudo mais importantes são os meios de transporte e o propósito de viagem de acordo com os trabalhos encontrados.

Os dispositivos GPS permitem o registo passivo, contínuo de movimento, e possíveis correções de viagens perdidas em pesquisas de "*recall*", que consistem na validação de dados com base nos inquéritos de viagens baseadas em *smartphones*. Os pesquisadores entram em contato com os entrevistados por telefone e pedem que se lembrem dos detalhes das suas viagens, o que contribui para a recolha de características de viagens precisas e completas (Xiao et al., 2016a). Apesar da tecnologia GPS ser precisa o suficiente no registo do tempo e das características posicionais da viagem, não pode registar o propósito da viagem. O propósito de viagem refere-se a uma atividade realizada em um destino de viagem. A recolha de dados por GPS é feita de forma bruta que depende do espaço urbano onde esta recolha está inserida.

Áreas urbanas congestionadas de alta densidade com prédios altos têm maior probabilidade de gerar dados com erros em comparação com dados obtidos de áreas rurais não congestionadas. Da mesma forma, as diferenças nas velocidades e acelerações dos diferentes modos de transporte serão menores em áreas urbanas congestionadas. O congestionamento do tráfego tornou-se um problema que atraiu a atenção de muitas pessoas das cidades. Para resolver este problema, os decisores de políticas de transporte em todo o mundo tentaram desenvolver modelos de demanda de viagens para prever o efeito das políticas de gestão de demanda de transporte ou mais conhecido como TDM (*Transportation Demand Management*) (Xiao et al., 2016b). TDM usa estratégias para informar e maximizar a eficiência de um sistema de transporte que leva a uma melhor mobilidade e redução do congestionamento.

Os dados de GPS podem ser registados por *smartphones* e dispositivos GPS dedicados e podem ser usados para determinar o propósito de viagem, combinando as características dos dados recolhidos pelo GPS. Os utilizadores são mais propensos a fornecer dados completos de GPS por *smartphones* do que por dispositivos GPS dedicados, porque a maioria dos utilizadores está acostumada a andar com um *smartphone* e por estes apresentarem menos falhas no acesso à rede em determinados lugares, havendo uma maior cobertura de rede para estes dispositivos através das operadoras telefónicas e estando os dispositivos constantemente atualizados.

Quando é feita a recolha de dados GPS alguns problemas são apresentados, como a perda de sinal que pode ocorrer quando os dispositivos de GPS são esquecidos em casa ou no trabalho, quando a localização de GPS não está habilitada ou quando existe ruído de sinal. Apesar destes problemas, o método de recolha de dados GPS é de baixo custo, beneficiando de forma positiva a utilização de *smartphones* (Shen & Stopher, 2014).

Nesta seção, descreve-se o uso de dados (dados GPS, *Points of Interest* (POI) e dados de redes sociais baseadas na localização ou *Location-Based Social Networks* (LBSN)), a descrição de recursos (características relacionadas com a viagem e informações de POIs relacionados com o destino da viagem) e a sua validação e avaliação.

2.2.1 Dados do Sistema de Posicionamento Global

Os dados GPS são cada vez mais usados para complementar o relatório de uma determinada viagem. Esses dados permitem a observação de alta precisão de rotas e horários de (Montini et al., 2014). Os *smartphones* são os dipositivos mais usados pelos pesquisadores para fazerem a recolha dos dados GPS. Os autores (Xiao et al., 2016a) basearam-se nesses dispositivos, através de uma aplicação multiplataforma, Android e iOS, para a recolha de dados de diversas viagens realizadas

por múltiplos utilizadores na cidade de Xangai. Esses dados incluíram a latitude, longitude, hora, rumo, velocidade instantânea e outras informações sobre a qualidade dos dados. Em geral, a maioria dos dados recolhidos pelo GPS fornecem estas informações. Os voluntários, com idades superiores aos 18 anos, instalaram a aplicação no *smartphone* onde lhes foi atribuído um ID único. Cada utilizador preencheu um formulário para fazer a recolha de dados pessoais, ou seja, endereço de email, contacto telefónico, idade, género, escolaridade, horário de trabalho, endereço residencial e de trabalho e a constituição do agregado familiar. Assim, foi possível realizar a recolha de dados de GPS das viagens feitas a 321 utilizadores desde outubro de 2013 a junho de 2015. Os propósitos selecionados foram baseados no artigo dos autores (Oliveira Marcelo et al., 2014) e corresponderam a casa, trabalho/educação, restaurantes, centro comerciais (shoppings), visitas sociais e buscar/deixar alguém. O maior número de propósitos de viagens correspondeu a casa e trabalho/educação, 2698 e 1996, respetivamente, como se pode verificar nas contagens das atividades apresentadas na Figura 2.

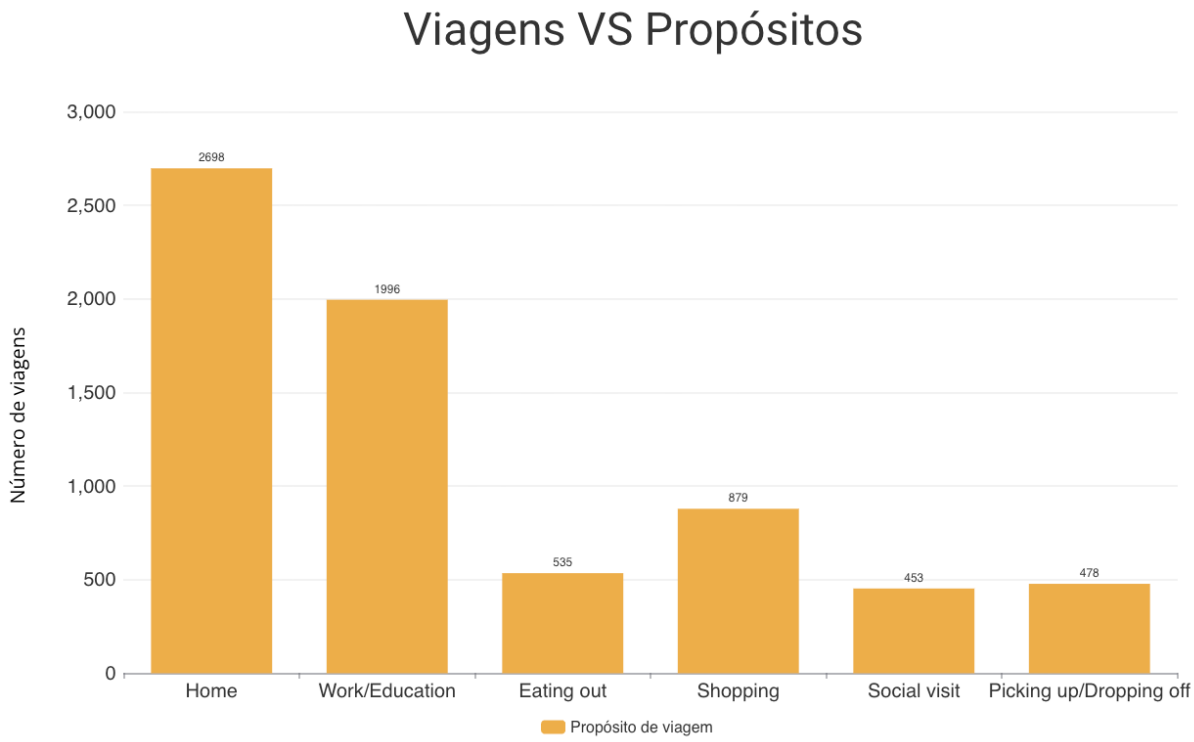


Figura 2 - Número de viagens por propósito (Xiao et al., 2016a)

Como a cidade em causa corresponde à cidade de Xangai é de prever que haja um elevado número de propósitos de viagens a centros comerciais, porque existe uma grande quantidade desses pontos de interesse na cidade. Contudo, as restantes viagens apresentaram valores semelhantes.

Outro estudo feito no mesmo país que os dados que são trabalhados neste estudo, corresponde ao estudo dos autores (Montini et al., 2014) que utilizaram um conjunto de dados GPS recolhidos na cidade de Zurique, Suíça. Os dados foram recolhidos de 156 utilizadores que usaram dispositivos GPS dedicados na vez de *smartphones*, para fazerem o registo e recolha de dados de viagens realizadas num período de uma semana. Os entrevistados ainda foram solicitados a corrigir o relatório das viagens realizadas, *recall*, incluindo o modo de transporte usado e o propósito de viagem. Os propósitos de viagens correspondentes a este trabalho de (Montini et al., 2014) consistem em casa, trabalho, compras/serviços, atividades recreativas, buscar/deixar alguém, atividades laborais fora do local de trabalho (negócios), e outras atividades. Dentro destes propósitos, alguns correspondem aos do artigo (Xiao et al., 2016a), anteriormente apresentados. As

percentagens mais elevadas condizem ao propósito casa, posteriormente seguido do propósito trabalho. Ambos os artigos aqui referenciados apresentaram o maior número de viagens realizadas para esses dois propósitos, porque são duas atividades mais realizadas por pessoas com faixas etárias superiores aos 18 anos, considerando a integração do propósito de viagem da educação no trabalho.

2.2.2 Dados de Pontos de Interesse

Os dados de pontos de interesse ou *Points of Interest* (POIs), normalmente usados para inferir o propósito da viagem, correspondem a pontos de localizações específicas que são definidas como úteis ou interessantes numa determinada região. As coordenadas GPS de POIs especificam um ponto mínimo no mapa, por exemplo no "*OpenStreetMap*" através da latitude e longitude.

Diversos investigadores estabeleceram vários critérios a usar para a identificação do propósito de viagem com base num POI. Essa identificação foi usada com base na distância entre o fim de viagem e o POI. O final de viagem sendo identificado, segundo (Lu et al., 2013), num raio de 500 metros poderá ter o propósito casa, trabalho/educação, comércio, etc., de acordo com quantidade de POIs que se encontram numa determinada parcela territorial, ou seja, nesse "*land use*". Outros autores tiveram em consideração em raios diferentes. Os autores (Bohte & Maat, 2009) empregaram um raio de 100 metros e os autores (C. Chen et al., 2010) empregaram um raio maior que este último, de 250 metros.

A previsão do propósito depende muito de fontes externas, como dos dados dos POIs que estão integrados num determinado *land use*. Alguns autores, referenciados no parágrafo anterior, classificam o tipo de *land use* nesses POIs rotulados, enquanto outros autores determinam com informações baseadas em polígonos (Shen & Stopher, 2014).

No trabalho de (Shen & Stopher, 2014) empregou-se a agregação de vários POIs com o mesmo tipo de atividade enquadrados numa área/polígono para gerarem um determinado *land use*, através de aprendizagem computacional, sendo responsável por determinar o propósito de viagem idealizado. Qualquer fim de viagem que esteja dentro desse polígono apresentará esse propósito de viagem. Desta forma, o polígono consiste numa área abrangente com um dado *land use*, que irá ser responsável por identificar o propósito de viagem. Enquanto um POI, para além de identificar também o propósito de viagem, corresponde a um ponto fixo no mapa. Para isto é preciso ter em conta que a utilização de polígonos, num determinado estudo, pode levar a erros na deteção de propósito de viagem, porque o fim de viagem de um determinado utilizador poderá estar no limiar da área do polígono e não corresponder a esse propósito, e vice-versa. Por essa razão, usar a informação de POIs será mais precisa (Xiao et al., 2016b), porque mesmo que haja mais do que um POI próximo a um destino de viagem, de acordo com os dados pessoais daquele utilizar, dos seus hábitos e da hora da atividade consegue-se determinar esse propósito. Um exemplo deste tipo corresponde ao utilizador "Professor" que regularmente às segundas-feiras vai trabalhar às 10h e entre as 9h e as 10h, na viagem para a faculdade, habitualmente toma café num estabelecimento perto de um Continente e da faculdade onde trabalha. Como existem esses três tipos de POIs, "Café", "Shopping" e "Universidade" poderá existir um conflito na determinação do propósito de viagem. Contudo através das distâncias entre os destinos e o POI, e ainda da hora desse destino é sempre possível determinar qual a atividade atribuída a esse POI.

Com isto, podemos afirmar que a inferência do propósito de viagem é geralmente detetada com base no tipo de *land use*, em dados pessoais do utilizador e ainda com base em dados de redes sociais baseadas na localização (LBSN, do termo em inglês *Location-Based Social Networks*).

2.2.3 Dados de Redes Sociais Baseadas na Localização

Estudos recentes para a deteção da atividade realizada no fim de uma viagem tendem a usar abordagens de aprendizagem computacional junto com dados de LBSN, como por exemplo, "Foursquare" e "Google Places", (Ermagun et al., 2017) e (C. Chen et al., 2018), respetivamente.

Os dados LBSN possuem um grande potencial para melhorar o conhecimento do comportamento das atividades significativas dos utilizadores e assim determinar o propósito de viagem. Alguns investigadores utilizaram dados de *check-in* do Foursquare (C. Chen et al., 2018) e outros dados do Google Places para inferir padrões de atividades humanas (Zhan et al., 2014). O check-in identifica a presença dos utilizadores num determinado local criando estatísticas. Essas estatísticas são realizadas através do número total de *check-ins* nesses locais e do total de visitas que esses utilizadores praticam.

O Google Places (Google Places, 2021) e o Foursquare (Foursquare, 2021b) permitem, por meio de uma interface de programação de aplicações ou *Application Programming Interface* (API), fazer uma pesquisa por área para identificar diversos POIs de destinos de viagem e eventos, através da sua latitude e longitude. Uma API não é mais do que um serviço que permite comunicar com outros serviços para obter determinado(s) resultado(s), sem haver a necessidade de se saber como foram implementados.

O Foursquare permite que as pessoas se registem por meio de *check-ins* aos lugares que visitaram, usando o aplicativo. Assim, é feita a categorização dos diferentes locais para determinados utilizadores (Yazdizadeh et al., 2019).

Os autores (Ermagun et al., 2017) testaram ambas as APIs. A conclusão a que se chegou foi que a API Google Places tinha uma maior área de cobertura geográfica e que utilizava um conjunto maior de informações de dados LBSN. Segundo (Garnett & Stewart, 2015), como os *smartphones* com GPS podem obter erros de registo da posição de um utilizador, variando esse erro entre os 0.56 metros e os 50 metros, com uma média de 6.5 metros, o Google Places ao permitir uma maior cobertura de área e ao fornecer um maior conjunto de POIs, é possível obter uma maior área com informações fidedignas de POIs próximos ao fim de viagem. Já os autores (C. Chen et al., 2018), para determinar com alta precisão a atividade que um determinado passageiro pretende realizar após descer do táxi, na área de Manhattan na cidade de Nova York, utilizaram dados de *check-in* do Foursquare. Algumas informações estatísticas básicas sobre esses dados de *check-in* do Foursquare podem ser visualizadas na Tabela I. Para determinar as atividades realizadas por diferentes passageiros em viagens de táxis, teve em conta o tempo de entrega do passageiro, o local de entrega e o contexto geográfico próximo, ou seja, dados das diferentes atividades que as pessoas geralmente realizam próximas ao ponto de entrega, ou seja, um dos prováveis pontos de interesse a ser visitado pelo utilizador no seu destino.

A utilização do Foursquare permitiu aos autores obterem uma descrição mais detalhada dos POIs, porque obteve uma maior precisão da atividade do passageiro ao integrar os dados das suas trajetórias com os dados deixados pelos passageiros na realização do *check-in* em POIs usando o Foursquare, na área onde o passageiro foi deixado pelo taxista. Com as informações das trajetórias dos passageiros, com as informações que os passageiros deixam ao iniciar a viagem e com os dados de *check-in* do Foursquare, não é difícil entender quais as atividades de viagem realizadas pelos passageiros, bem como a distribuição de atividades numa determinada área durante um determinado período de tempo. Assim, os passageiros têm como destino um restaurante para comer e um shopping para fazer compras, porque ambas as atividades podem ser realizadas neste

mesmo POI, "Shopping". Desta forma, com estas informações de *check-in* do Foursquare é possível ampliar a precisão da atividade realizada num determinado POI.

Tabela I - Estatísticas do conjunto de dados do artigo (C. Chen et al., 2018)

Dataset	Propriedades	Estatísticas
Redes das estradas	Número das interseções da estrada	11.999
	Número dos segmentos da estrada	15.202
Check-in do Foursquare	Número de Check-ins	>220.000
	Número de utilizadores	>38.000
	Número de POIs	>10.202
	Duração do tempo do check-in	12 meses
Trajetórias de táxis	Número de táxis	>19.000
	Número de viagens de táxi	≅ 13.000

Como resultado destas duas pesquisas, prevê-se que o estudo dos autores (Ermagun et al., 2017) no uso e na integração da API do Google Places, com a disponibilidade de informações de milhões de locais e pontos de interesse na sua base de dados, amplia o número de pontos de interesse, impactando no aumento e na precisão da previsão na inferência do propósito de viagem, porque aumenta o número de POIs e a área de abrangência. Com esta API e com a API do Foursquare é possível obter um elevado número de POIs fidedignos para dar auxílio na determinação da inferência do propósito de viagem. No entanto, devido à limitação da API do Google Places em suportar múltiplas plataformas móveis, como os sistemas usados pelos diferentes *smartphones* e por não possuir uma ampla cobertura em todo o mundo, usar o Google Places não resultaria numa melhoria considerável no nosso modelo, porque a recolha dos nossos dados é feita por dispositivos móveis, sendo relevante usar dados do Foursquare, uma vez que estes também são frequentemente mais usados pelos investigadores.

2.3 Inferência de Propósito de Viagem

O processo de inferência de propósito baseado em dados de GPS resulta na deteção do fim e do início de viagem, na seleção de características de viagem, em algoritmos e na sua validação e avaliação.

Os estudos presentes na inferência do propósito de viagem, segundo (Gong et al., 2014), estão classificados nos diferentes métodos existentes: métodos baseados em regras, métodos estatísticos e métodos de aprendizagem computacional.

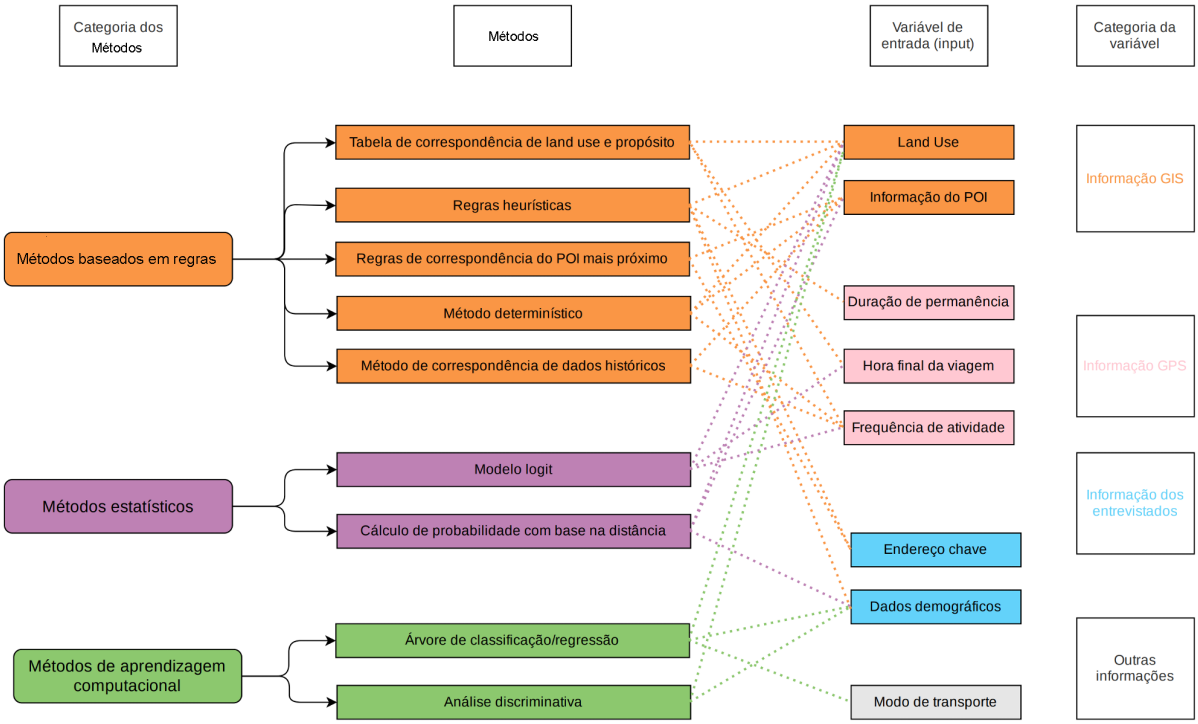


Figura 3 - Metodologias categorizadas para a inferência de propósito de viagem em pesquisas existentes e variáveis de entrada baseado em (Gong et al., 2014)

De acordo com a Figura 3 e com o artigo dos autores (Ermagun et al., 2017), os estudos mais recentes da inferência de propósito de viagem e que tendem a ser os mais promissores são baseados em métodos probabilísticos ou em métodos de aprendizagem computacional. Embora estes estudos apresentem melhores resultados quanto à precisão, por vezes os resultados se tornam artificiosos devido às diferenças nos tamanhos das amostras e na seleção da escolha das variáveis de entrada. No entanto, são as pesquisas baseadas em métodos de aprendizagem computacional que são mais abordadas e levadas em conta nos estudos. Isto se deve pelo facto de obterem as melhores precisões, como no artigo (Xiao et al., 2016a) com 96.5%, onde usaram as redes neuronais e informações baseadas em polígonos e pontos de interesse para detetar o propósito de viagem a partir dos dados de GPS. Não foi seguido este artigo, porque para além deste estudo ter sido feito numa cidade de Xangai que não corresponde à mesma cidade do *dataset* Breadcrumbs, este método requer mais pré-processamento, não lida com a falta de dados e o processo de treinamento é mais complexo, demorando mais na previsão do propósito de viagem. Sendo assim, de acordo com a Figura 3, o método considerado neste estudo corresponde ao método de árvores de classificação/regressão, ou seja, o *Random Forest* (RF) com a integração de categorias de informação GIS (informação do POI), GPS (duração de permanência, hora final da viagem e frequência da atividade) e dos entrevistados (dados demográficos).

A maioria dos pesquisadores usam as fontes de dados de *land use*, do Foursquare, de Especificação Geral de *Feed* de Trânsito ou *General Transit Feed Specification* (GTFS) e pesquisas de itinerários de trânsito para obterem conjuntos de dados mais ricos para serem usados nos métodos de aprendizagem computacional, como o *Random Forest*, Rede Neuronal Artificial e árvores de decisão, incluindo as variáveis de entrada dia, hora, local, duração da viagem, endereço de trabalho/residência, etc., recolhidas pelos sensores de GPS e pelas informações fornecidas pelos utilizadores nos inquéritos.

Study	Data & study area	Method detail	Input variables besides geo-coordinates	Accuracy
Rule-based methods				
Wolf et al. (2001)	156 trips from 13 participants who participated in a 3-day survey conducted in Atlanta, GA.	Matching tables developed for 11 trip purposes.	Land use, arrival time, activity duration, previous trip purpose.	79%
Wolf et al. (2004)	39 participants (28 full-time workers and 11 retirees) collected in 3 Swedish cities. Data collected over 2 years.	Heuristic rules for closest POI matching to differentiate between 10 purposes.	POI, activity duration, time of day, day of week, socio demographics, home addresses, profession and working hours.	NA
Stopher et al. (2008)	Household Travel Survey in Sydney and Travel Behavior Change Program in Canberra, Australia.	Heuristic rules used to differentiate between 10 trip purposes.	Land use, duration, occupation. Addresses for: home, school, work, frequently used stores, etc.	NA
Bohte and Maat (2009)	1104 participants from 3 municipalities in Netherlands.	Closest POI matching rules to differentiate between 7 purposes.	Home/work addresses and POI data.	43%
Wu et al. (2011)	47 participants with 131 person days from the Harbor Communities Time Location Study in 2008 and 21 person days' data in 2010 from 3 participants	Heuristic rules to differentiate between indoor, outdoor static, outdoor walking and in-vehicle travel activity.	Time and activity data merged to clusters.	>96% for indoor, >88% for in-vehicle travel, & lower for other.
Pereira et al. (2013)	1000 participants recruited from the Household Interview Travel Survey.	Historical data matching rules.	POI, socio-demographics, work hours, activity duration, travel time and historical activities and trips.	NA
Probabilistic methods				
Chen et al. (2010)	25 participant's GPS data for one weekday & 24 participant's GPS data for 5 weekdays in New York.	Multinomial logit model to differentiate between 4 purposes.	Land use, historical activity frequency, time of day and activity duration.	67–78%
Oliveira et al. (2014)	Subsample from the 2011 Atlanta Household Survey: 1352 households & 22,734 activities.	Two-level Nested Logit Model to differentiate between 12 purposes.	Land use, temporal information and socio-demographics.	60%
Machine learning and neural network methods				
Liao et al. (2007)	Data collected from 4 participants for approx. 7 days with 40,000 GPS points and 10,000 segments.	Hierarchical Conditional Random Fields to identify activities & places.	Street maps, temporal information and POI data.	85–90%
McGowen and McNally (2007)	170,000 activities data from 17,000 households in the 2000–2001 California Statewide Travel Survey. Data collected over 2-days.	Discriminant analysis & regression tree model to differentiate between 5 purposes.	Land use, socio-demographics, historical activity frequency, time of day and activity duration.	73–74%
Deng and Ji (2010)	36 participants over 3-days (226 trips) in Shanghai, China. Data collection over 2.5 months.	Decision tree models to differentiate between 7 purposes.	Land use, socio-demographics, activity duration, time of day and day of week.	87.6%
Wu et al. (2011)	47 participants with 131 person days from the Harbor Communities Time Location Study in 2008 and 21 person days' data collected in 2010 from 3 participants	Random forest model to differentiate between indoor, outdoor static, outdoor walking and in-vehicle travel activity.	Acceleration, speed and distance.	>97% for indoor, >84% for in-vehicle travel, & lower for other.
Lu and Liu (2012)	Data on 3188 trips in 2008 in the Twin Cities Metro Area, MN	Decision tree model to differentiate between 10 purposes.	Land use, temporal information, previous & next trip attributes, socio-demographics.	73.4%
Montini et al. (2014)	Data from 156 participants with 6938 activities collected in Zurich, Switzerland.	Random forests to differentiate between 8 purposes.	Land use, temporal information, home/work addresses and socio-demographic data.	80%
Oliveira et al. (2014)	Subsample from the 2011 Atlanta Household Survey: 1352 households & 22,734 activities.	Decision tree model to differentiate between 12 purposes.	Land use, temporal information, & socio-demographics.	65%
Kim et al. (2015)	Data from 793 participants with 64819 points collected in years 2012–13 in Singapore using the Future Mobility Survey application.	Random forest model to differentiate between 16 purposes.	POI	75.5%
Xiao et al. (2016)	321 participants with 2409 person days collected from October 2013 to June 2015 in Shanghai, China	Artificial neural networks to differentiate between 6 purposes	POI and polygon-based information	96.5%

Figura 4 - Estudos existentes da inferência do propósito de viagem (Ermagun et al., 2017)

Os autores (Wu et al., 2011) que desenvolveram resultados fundamentados nos métodos baseados em regras obtiveram melhores resultados, mas muito similares em viagens feitas em veículos, relativamente aos métodos baseados em aprendizagem computacional, ou seja, obtiveram 88% nos métodos baseados em regras e 84% nos métodos baseados em aprendizagem computacional para viagens realizadas de veículo, e 97% e 96% para viagens realizadas a pé, respetivamente, como se pode visualizar na Figura 4. Estes resultados foram obtidos a partir de viagens realizadas a pé ou num veículo (de um local para outro), deslocações feitas dentro de um estabelecimento (por exemplo, num shopping) e sem qualquer tipo de deslocamento, ou seja, quando um ponto de localização se encontra estático ao ar livre (por exemplo, num parque verde). Já os autores (Oliveira

Marcelo et al., 2014) obtiveram melhores resultados nos métodos de aprendizagem computacional comparativamente com os métodos probabilísticos. No seu estudo, o modelo de árvore de decisão superou o modelo logit aninhado em 5% quanto à sua precisão, ou seja, 60% no modelo logit aninhado e 65% no modelo de árvore de decisão, verificando que é mais fácil gerar modelos a partir de árvores de decisão.

As variáveis de entrada usadas pelos autores são referentes a dados de *land use*, dados de POIs e características da viagem (duração da viagem, início/fim da viagem, idade, género, ocupação, etc.). Ainda assim, alguns autores como o (Oliveira Marcelo et al., 2014) usam dados de locais populares em torno do destino, dados sociodemográficos, para obterem melhores resultados na identificação da atividade no fim de viagem (mercearias, edifícios governamentais, posto de abastecimento, etc.). No seu trabalho, (Ermagun et al., 2017) utilizam fontes de dados de redes sociais, como Foursquare ou Google Places, baseadas na localização para ampliar o número de pontos de interesse quando os dados de *land use* e os dados de localização dos destinos das viagens não são capazes de identificarem as atividades no fim de viagem. Neste artigo, empregam métodos probabilísticos e métodos de aprendizagem computacional, modelos de logit aninhado (secção 2.4.2) e *Random Forest* (RF) (secção 2.5.3), respetivamente, para demonstrarem resultados bem-sucedidos na identificação do propósito de viagem. O modelo de logit aninhado obteve piores resultados comparativamente com o modelo RF, 57.69% e 64.17%, respetivamente. Estes resultados obtiveram estas percentagens de precisão, porque as características das viagens foram aplicadas com dados de meios de comunicação social do Google Places, que são dados de *land use* ou dados de POIs, correspondentes a locais próximos de serviços de pesquisa e descobertas online como lugares próximos. Quando isto não aconteceu os valores de precisão foram inferiores.

Neste sentido, observa-se que os dados de LBSN permitem aumentar a precisão da inferência de propósito de viagem e que os modelos de árvore de decisão (secção 2.5.2) ou os modelos de RF são os modelos mais bem empregues, bastante usados e eficientes, porque para além de apresentarem uma boa precisão são mais simples de se implementar e interpretar. Os dados recolhidos pelos sensores GPS com os diários de viagem preenchidos pelos utilizadores e as pesquisas de *recall,* incluídas no *dataset*, ajudam a estes modelos de previsão identificar com uma melhor precisão na inferência do propósito de viagem.

2.3.1 Segmentação de Viagens

Uma viagem corresponde a um percurso de um local para outro, com um início e fim, com o objetivo de realizar determinada atividade. Segundo alguns pesquisadores, é necessário decorrer um determinado tempo de viagem para considerarem uma deslocação no espaço. Os autores (Stopher et al., 2008) afirmam que serão necessários 120 segundos para considerar um deslocamento como viagem, com um início e fim de viagem diferente. Outros autores discordam, (Yazdizadeh et al., 2019), defendem que são necessários 180 segundos para a ocorrência de uma viagem. Em alguns casos, como as viagens são feitas apenas de automóvel e são usados GPS dedicados para a recolha de dados, utilizam os dados do motor como desligado e ligado para fazer essa identificação (Lu et al., 2012). Porém, este método anterior como o método dos 120 segundos utilizado por (Stopher et al., 2008) não são suficientemente confiáveis, porque no caso de uma viagem ocorrer a pé a 4km/h (quilómetros por hora) serão percorridos aproximadamente 134 metros, distância pequena que poderá corresponder ao local de estacionamento próximo do local residencial do entrevistado. No entanto, se forem considerados os 180 segundos de carro a 50km/h são percorridos 2500 metros e viagens feitas a pé a 4km/h são percorridos aproximadamente 201

metros. Conclui-se assim que esses 180 segundos são suficientemente consideráveis para considerar a ocorrência de uma viagem quando são feitas a pé ou de carro.

Quando se trata de viagens feitas de táxi, pode-se substituir o tempo de deslocamento pelo ponto de entrega do passageiro feito pelo taxista, como idealizado na pesquisa de (C. Chen et al., 2018), de carro.

A deteção do modo e a deteção do propósito de viagem pode ser baseado nos resultados *Trip Identification*/*Segment Identification* (TI/SI). No caso de ocorrer a paragem permanente do veículo após os 180 segundos do início da viagem significa que este local é o fim de viagem, segundo (Shen & Stopher, 2014), porque qualquer paragem de um veículo por motivos de trânsito, paragem no semáforo ou passadeira ocorre num intervalo menor que 180 segundos. Durante a viagem poderá ocorrer perda de sinal. Se a perda de sinal estiver compreendida entre os 180 segundos e os 600 segundos numa distância inferior a 50 metros do início da viagem haverá uma atividade de curta duração, no momento da perda de sinal. Se a distância for superior a 500 metros poderá estar a ocorrer uma viagem subterrânea. A velocidade permite identificar o tipo de transporte a ser efetuado. Se a velocidade for inferior a 6km/h a viagem está a ser feita a pé, a mais de 40km poderá estar a ser feita de carro ou autocarro, conseguindo diferenciar estes dois meios de transporte através da aceleração/desaceleração, e de bicicleta se estiver compreendido entre esses dois valores de velocidade. No entanto, caso um *dataset* não possua descrição sobre os meios de transportes usados, ou apenas de forma genérica por hábitos, não se consegue distinguir as viagens realizadas de autocarro e de carro, porque as velocidades que esses carros transitam são similares, como também será difícil de distinguir as viagens realizadas de metro. Os únicos modos de transporte que possibilitam a sua identificação através das velocidades correspondem à viagem realizada a pé ou de veículo, onde esta última não é especificada (carro, autocarro, táxi, etc.).

2.3.2 Seleção de *Features*

A seleção de *features* para a inferência de propósito inclui características de dados geográficos, viagens e atividades relacionadas, participantes relacionados e meios de transportes.

Os dados geográficos são empregues na inferência do tipo de atividade utilizando recursos relacionados com Sistemas de Informação Geográfica (SIG). O limite de distância entre o final de viagem e o local de destino resulta na combinação de dados SIG e da qualidade dos dados de GPS. Dados SIG descarregados diretamente do *OpenStreetMap* existem em dois formatos geométricos, ponto e polígono, como apresentado na secção 2.2.2 pelos autores (Shen & Stopher, 2014).

Uma atividade é a principal ocupação realizada num determinado local, estando em harmonia com um determinado propósito. Uma viagem é um percurso num só sentido com a finalidade em realizar uma atividade. As viagens possuem dadas características como hora, dia, velocidade, distância, etc., e correspondem aos atributos temporais da viagem e da atividade.

As informações pessoais, endereços de casa e trabalho ajudam a impulsionar consideravelmente o desempenho geral da classificação das viagens, porque estas atividades são mais frequentes que outras finalidades, como ir almoçar/jantar a um restaurante.

Em comparação com as atividades e com os dados geográficos, as *features* relacionados com os participantes também são importantes para distinguir diferentes atividades num determinado local, porque um determinado utilizador pode ter como fim de viagem o shopping e ir ao ginásio, dado que num shopping existem ginásios, e para se conseguir identificar as atividades dentro destes espaços é preciso identificar a idade, o género e os hábitos desses utilizadores, porque maioritariamente, de acordo com a rotina de uma determinado utilizador, pode-se identificar a

atividade a realizar naquele exato momento. Sem os hábitos dos utilizadores seria pouco provável de determinar qual a atividade que foi realizada. Outras características que são úteis para a inferência de propósito são as informações de LBSN como, por exemplo, do Foursquare para dados de check-ins deixados pelos utilizadores, taxonomias/categorias desses POIs do Foursquare e pontos de interesse de locais populares fornecidos pelo Google Places, fundamentados na secção 2.2.3.

2.3.3 Taxonomia de propósitos

As categorizações das atividades abrangem um conjunto de propósitos de viagem pertencentes ao quotidiano do ser humano. Geralmente ou maioritariamente, quando realizamos qualquer tipo de deslocamento no nosso dia a dia temos um objetivo associado a essa viagem com a finalidade em realizar uma determinada atividade, ou seja, com um propósito categorizado por uma atividade. Segundo alguns autores, (Bohte & Maat, 2009), (Moro et al., 2019) e (Gao et al., 2021), a categorização de atividades de viagem estão compreendidas em educação, trabalho, lazer, vida social, casa, restauração, estações de meios de transportes e compras. Todas estas categorizações estão a abranger as taxonomias realizadas pelos diferentes utilizadores dependendo sempre do local onde será feita a recolha de dados, porque se um utilizador estiver numa região de praia vamos conter a taxonomia praia e, caso contrário, essa taxonomia não é considerada para a categorização uma vez que não existe essa atividade nesse local.

No *dataset* Breadcrumbs as taxonomias de propósitos correspondem a casa, universidade, desporto, parque/jardim, bar, clube, restaurante, shopping, paragem de metro/comboio/autocarro, trabalho, casa de amigo, cais, família, aluguer de automóveis, parque de estacionamento, associações, hotel, aeroporto, biblioteca e praia. Todas estas categorias estão a identificar atividade do local onde foi feita a sua recolha, Suíça, que podem apresentar diferentes taxonomias com base em dados de eventos mais visitados da região. A essas taxonomias ainda pode ser integrado as taxonomias /categorias existentes nos POIs do Foursquare.

2.4 Abordagens

Dado que é muito desafiador a inferência do propósito de viagem, porque o registo de dados GPS não é preciso o suficiente para o determinar, existem abordagens que aproveitam os dados de GPS, como trajetórias, POIs, localização dos utilizadores e eventos para o identificar (Meng et al., 2017). Essas abordagens dependem das abordagens baseadas em regras, métodos probabilísticos ou aprendizagem computacional, no qual, cada uma terá as suas características que impactam a inferência do propósito de viagem, como se pode visualizar na Figura 5.

	Rule-based	Probability-based	Supervised machine learning
Transferable level	Low	Medium	High
Suitable data size	Small	Medium and big	Big
Role of data selection/division	Minor	Minor	Major
Number of variables	Small	Small	Large
Ground truth	Optional	Optional	Mandatory
Mainly used in	Estimating home address, choosing location candidate	Human geography	Transportation science
Power/performance	Low	Reasonable	High
Interpretation	High	Medium	Low

Figura 5 - Abordagens de inferência de propósito de viagens e sua avaliação (Nguyen et al., 2020)

Neste artigo, (Nguyen et al., 2020), são apresentados diversos artigos científicos que apresentam estudos com diversos números de participantes, variáveis e instâncias sendo que no mínimo temos artigos com 4 participantes, 5 tipos diferentes de atividades e 3188 viagens realizadas.

2.4.1 Abordagens baseadas em Regras

Modelo baseados em regras consistem em manipular os recursos fornecidos pelos dados de GPS para determinar diferentes tipos de propósitos de viagens, como por exemplo, usufruir dos recursos do início e do fim da viagem. Combinam as informações selecionadas do local e as informações pessoais do entrevistado com uma série de regras heurísticas, técnica projetada para resolver um problema de forma mais rápida quando os métodos clássicos são mais lentos, predefinidas para inferir o propósito da viagem (métodos de regras heurísticas/regras de correspondência de múltiplos dados). São métodos simples e requerem menos fontes de dados. As abordagens baseadas em regras dependem de limiares entre o fim de viagem e o POI que são determinados com base na experiência prática, indicando que esses limiares podem depender consideravelmente do contexto onde os dados são recolhidos e são baseadas principalmente na experiência subjetiva. Contudo, a sua precisão de classificação geralmente não é alta, como se pode visualizar na Figura 5. Portanto, abordagens baseadas em regras podem não ser apropriadas para detetar determinados propósitos de viagem, porque incorporam baixa capacidade de generalização (Xiao et al., 2016b).

Regras determinísticas eram os métodos mais usados inicialmente na inferência de propósito, porque são simples e fáceis de interpretar. Com base em dados de *land use* baseados em polígonos e características temporais das atividades, os autores (Xiao et al., 2016a) atribuem um tipo de local ao fim da viagem com base no limite da distância do raio em metros, argumentado na secção 2.2.2. Porém, devido a ocasionalmente ocorrerem erros nos dados de GPS, por vezes os mais próximos não correspondem ao local visitado pelo utilizador.

2.4.2 Abordagens baseadas em Métodos Probabilísticos

Abordagens que calculam a probabilidade de cada propósito de viagem, com base nos diferentes valores das informações de localização e informações pessoais dos entrevistados, por exemplo modelos logit aninhado. Os modelos são simples e os resultados da classificação, com base na distribuição de probabilidade estatística, são relativamente altos. (Sun et al., 2021)

O modelo logit aninhado ou *Nested Logit* (NL) consiste em moderar a suposição de independências entre todas as alternativas possíveis do modelo. Modela a similaridade entre as alternativas agrupadas/aninhadas por meio da correlação de componentes, como por exemplo transportes públicos ou viagens, para estimar, especificar e interpretar as previsões (Carrasco Juan & Ortúzar Juan, 2010). A Figura 6 mostra uma estrutura do modelo logit aninhado que é completamente geral para fins de análise que podem ter três ou mais níveis.

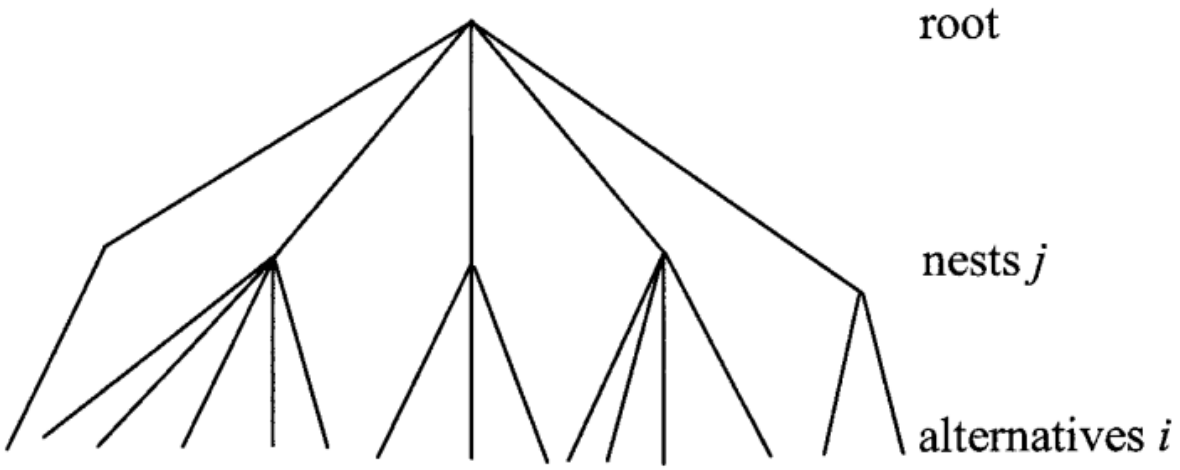


Figura 6 - Estrutura de uma árvore de modelo logit aninhada (Carrasco Juan & Ortúzar Juan, 2010)

As abordagens probabilísticas são mais flexíveis do que as regras, porque levam em consideração a relação espaço-tempo entre propósitos. Os autores (C. Chen et al., 2010) e (Oliveira Marcelo et al., 2014) ao introduzirem o modelo logit obtiveram boas precisões, verificando que os métodos estatísticos foram melhores do que os métodos baseados em regras em ambientes de alta densidade, ou seja, de acordo com as características das áreas urbanas. Os métodos probabilísticos também são mais poderosos do que os baseados em regras, menos dependentes do conhecimento subjetivo dos pesquisadores e capazes de deduzir o propósito das viagens quando o conjunto de dados é de grandes dimensões (Gong et al., 2014). No entanto, é difícil para o modelo lidar com tarefas de classificação complexas, diminuindo a sua precisão.

2.4.3 Abordagens baseadas em Métodos de Aprendizagem Computacional

Abordagens na área da inteligência artificial relacionada com a busca de um conjunto de regras e procedimentos, permitem que as máquinas possam agir e tomar decisões baseadas em dados ao invés de serem explicitamente programadas para realizar uma determinada tarefa, sendo capazes de tomar decisões com o auxílio de modelos. Correspondem a modelos de classificação computacionalmente intensivos associados às informações de localização, informações pessoais de entrevistados, informações de transporte e neste caso específico, propósitos de viagens. Usam *features* das viagens, fonte de dados, e aprendem automaticamente para se tornarem precisos no reconhecimento, apesar de serem difíceis de treinar e de otimizar. Os algoritmos mais abordados para a inferência do propósito de viagem correspondem a máquinas de vetores de suporte ou *Support Vector Machine* (SVM), árvore de decisão ou *Decision Tree* (DT), *Random Forest* (RF) e Rede Neuronal Artificial ou *Artificial Neural Networks* (ANN) (Xiao et al., 2016b).

2.5 Algoritmos de Aprendizagem Supervisionada

Os algoritmos de aprendizagem supervisionada ou *Supervised Machine Learning* (SML) são considerados as melhores escolhas do que as abordagens baseadas em regras e métodos probabilísticos. Existem muitos algoritmos de aprendizagem computacional que estão divididos em aprendizagem estatística (máquinas de vetores de suporte), baseados em lógica (árvore de decisão e *Random Forest*) e baseados em perceptrão (rede neuronal artificial), que pode ser considerado o primeiro modelo de Redes Neuronais (Silva & Ribeiro, 2018).

O SML é o mais generalizável, porque aprende com os dados para fazer previsões em vez de ser explicitamente programado por regras. Além disso, é adequado para "*big data*" como dados de GPS e SIG, bem como uma série de variáveis.

2.5.1 Máquinas de Vetores de Suporte

Algoritmo apto para lidar com problemas de classificação do propósito de viagem, através de treino e teste. Utiliza aprendizagem supervisionada quando se quer classificar dados em grupos diferentes. O SVM escolhe a reta, também chamada de hiperplano em maiores dimensões, entre dois grupos que se distancia mais de cada um.

Na Figura 7 no gráfico superior esquerdo há três hiperplanos (A, B e C), sendo o hiperplano B o certo para classificar o grupo de estrelas e círculos. No gráfico superior direito, o hiperplano B tem um erro de classificação e o A classificou tudo corretamente. Portanto, o melhor hiperplano corresponde ao A. Existem casos onde não é possível separar as duas classes usando uma linha reta, pois uma das classes está no território de outra (*outlier*), como apresentado no gráfico inferior esquerdo.

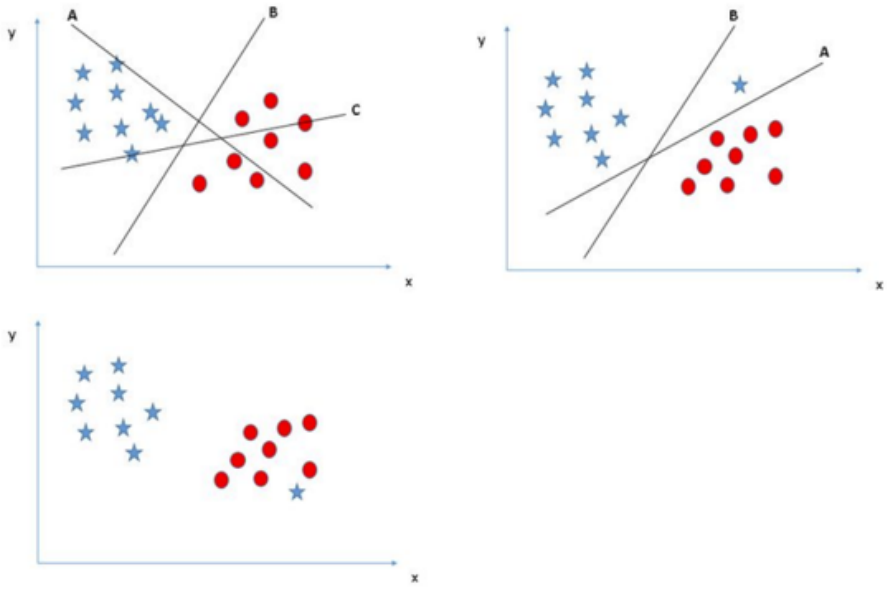


Figura 7 - Separação de classes de dados e escolha do hiperplano pelo algoritmo máquinas de vetores de suporte (Addan, 2019)

Quando se trata de inferência do propósito de viagem, apresenta a desvantagem de incapacidade para lidar com ambiguidade e apresenta performances relativamente baixas, quando aplicado a classificações multiclasse (Xiao et al., 2016b).

No seu trabalho, (Feng & Timmermans, 2016) utilizaram este algoritmo com dados recolhidos de um pequeno grupo de oito indivíduos. Eles transportavam consigo dispositivos de GPS que registavam as deslocações num período de 6 a 8 semanas. Um pequeno questionário foi feito com perguntas relacionadas com a atividade de viagem, tempo (horário de início e fim), locais de atividades e meios de transporte usados. Desta forma, os dados que foram confirmados como *ground truth* são usados para avaliar o desempenho do algoritmo SVM. A recolha de dados GPS incluiu informações como data, hora, longitude, latitude, velocidade, distância, precisão da medição, número de satélites e ainda consideraram variáveis do perfil pessoal. Algumas destas variáveis de entrada (*input*) podem ser visualizadas na Tabela II usadas para determinar os modos de transporte Figura 8, sendo estas variáveis de entrada *features* importantes na inferência do modo de transporte que também poderão ser usadas na inferência do propósito de viagem, como empregue no trabalho (Montini et al., 2014).

Tabela II - Comboio, caminhada, bicicleta, carro, autocarro, motociclo, corrida, elétrico e metro (Feng & Timmermans, 2016)

	Nome das variáveis	Especificação
Variáveis de Entrada	STDDEVSPEED	Desvio padrão da velocidade
	AVGSPEED	Velocidade média
	AVGACC	Aceleração média
	MAXSPEED	Velocidade máxima
	MAXACC	Aceleração máxima
	ACCUMDISTANCE	Distância acumulada
	RRDIST	Distância até à linha da estrada
	RTDIST	Distância até à linha do elétrico
	RMDIST	Distância até à linha do metro
	USEDSTAT	Número de satélites usados
	VIEWSAT	Número de satélites vistos
	VALID	Tipo de correção GPX
	PDOP	Precisão de posição de coordenada 3D
	HDOP	Precisão horizontal de coordenada 2D
	CAROWN	Sim - se o entrevistado tiver carro Não - caso contrário
	BIKEOWN	Sim - se o entrevistado tiver uma bicicleta Não - caso contrário
	MOTORBIKEOWN	Sim - se o entrevistado tiver um motociclo Não - caso contrário
Variáveis de Saída	MODE	

Apesar deste algoritmo obter bons resultados de precisão na previsão do modo de transporte usado, Figura 8, e ainda de ser bastante empregue, segundo os autores do algoritmo, foi o que obteve piores resultados relativamente a outros algoritmos usados, como o algoritmo da rede bayesiana ou *Bayesian Network* (BN) e o C4.5.

	A	B	C	D	E	F	G	H	I	J
BN	0.997	0.997	0.999	1	0.999	0.999	1	0.999	1	1
NB	0.848	0.969	0.934	0.799	0.836	0.926	0.949	0.98	1	0.983
LR	0.989	0.991	0.818	0.928	0.891	0.758	0.947	0.76	1	1
MP	0.998	0.974	0.916	0.926	0.965	0.743	0.989	0.985	1	1
DT	0.999	0.971	0.958	0.985	0.979	0.99	0.991	0.974	0.982	0.98
SVM	0.987	0.999	0.76	0.925	0.876	0.888	0.971	0.654	1	1
C4.5	1	0.999	0.993	0.997	0.997	0.994	0.998	0.999	0.996	0.99

Note: A-Activity episode; B-Train; C-Walking; D-Bike; E-Car; F-Bus; G-Motorbike; H-Running; I-Tram; J-Metro.

Figura 8 - Taxa de acerto no modo de transporte da viagem (Feng & Timmermans, 2016)

2.5.2 Árvores de Decisão

Algoritmo de aprendizagem supervisionada simples, mas poderoso, adequado e não paramétrico (que fazem suposições fortes sobre a forma da função de mapeamento, ou seja, procuram ajustar melhor os dados de treinamento na construção da função de mapeamento (Jason Brownlee, 2020b)) para classificação. É uma abordagem eficaz e imediata para compreender as relações entre variáveis independentes. A ideia base é resolver subproblemas mais simples partindo de um problema complexo, Figura 9. As árvores de decisão são muitos simples de compreender, com capacidades de dividir o espaço definido pelos atributos em subespaços, onde cada subespaço é associado a uma determinada classe. Um exemplo binário à ideia base, considerando a informação se está a chover no momento, podemos dividir o problema em dois mais simples: decidir não levar bicicleta se estiver a chover, e decidir não levar bicicleta se não estiver a chover. Evidentemente que o primeiro subproblema tem uma solução imediata (Silva & Ribeiro, 2018). Desta forma, a escolha do modo de transporte pode ser analisada como um problema de reconhecimento de padrões de indivíduos.

Existe um interesse crescente neste tipo de abordagens tendo sido desenvolvido várias implementações, como por exemplo o C4.5, implementado e testado pelos autores (Feng & Timmermans, 2016). Porém estes algoritmos apresentam a desvantagem para dados que incluem variáveis categóricas com diferentes números de níveis, como dados SIG, dados GPS, dados obtidos de relatórios feitos aos utilizadores e outras informações, porque poderão chegar a falsos positivos.

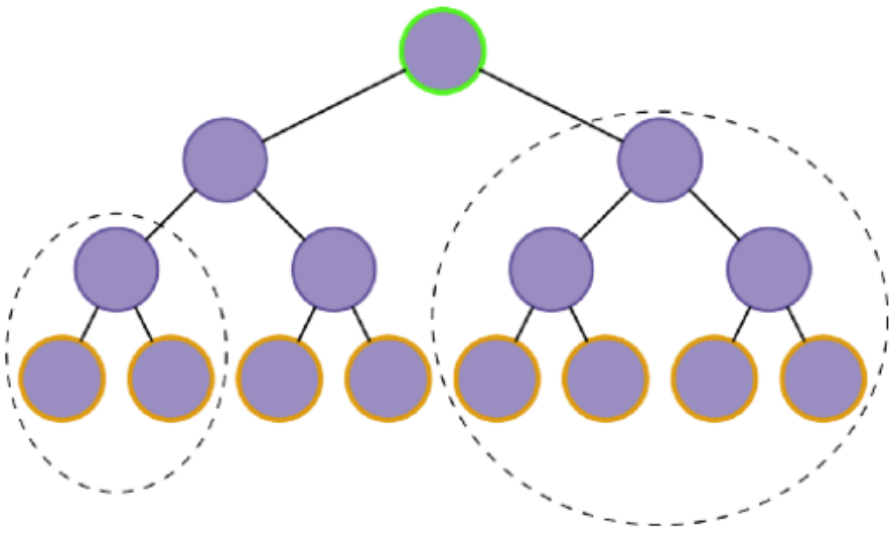


Figura 9 - Árvore de decisão dividida em subproblemas (Guilherme Fernandes, 2019)

Uma árvore de decisão que classifica corretamente todos os exemplos em um conjunto de treino, fenómeno conhecido como sobreajuste (em inglês, *overfitting*), pode não ser um classificador tão bom quanto uma árvore de menores dimensões, onde não cabem tantos dados de treino. Para evitar esse problema a maioria dos algoritmos de árvore de decisão empregam um método de "poda" (em inglês, *pruning)*, que significa que ocorre o cultivo de uma árvore grande excluindo alguma parte dela. Um método alternativo a este método de "poda" designa-se "critério de paragem" que consiste em parar de fazer crescer a árvore, subdividindo-a. Este último método é empregue pelo algoritmo ID3. Os autores (Ross Quinlan et al., 1994) empregaram ambos os métodos e chegaram à conclusão que o método de "poda" do C4.5 funciona muito melhor.

A árvore de decisão é construída e empregue para automatizar a deteção do propósito de viagem. Os atributos de entrada, de acordo com o artigo de (Lu et al., 2012), são derivados a partir dos dados de GPS e correspondem ao horário de início/fim da viagem, local de destino da viagem, duração da atividade, tipo de *land use* em SIG, bem como atributos sociodemográficos do indivíduo (idade, sexo, local de residência, nível de escolaridade, renda, etc.). O algoritmo de árvore de decisão empregue por estes autores corresponde ao algoritmo C4.5, com o método de "poda" introduzido pelos autores (Ross Quinlan et al., 1994). A amostra contém 3188 viagens para aprendizagem, ocorrendo a análise de sensibilidade para ajudar na análise do método de codificação do *land use* da viagem e assim, derivar o propósito da viagem e ajudar nas contribuições das variáveis de entrada para a sua inferência. Diferentes subcategorias de variáveis de entrada são adicionadas separadamente na árvore de decisão do propósito da viagem de forma gradual. O procedimento recursivo de inferência de propósito pode ser visualizado na Figura 10, onde ocorre a associação de vários módulos para a obtenção de uma maior precisão, neste caso, para chegar a uma precisão final de 73.37%. De acordo com esta figura, verifica-se que a variável de entrada localização do fim de viagem é a que tem maior impacto na precisão da previsão da inferência de propósito de viagem.

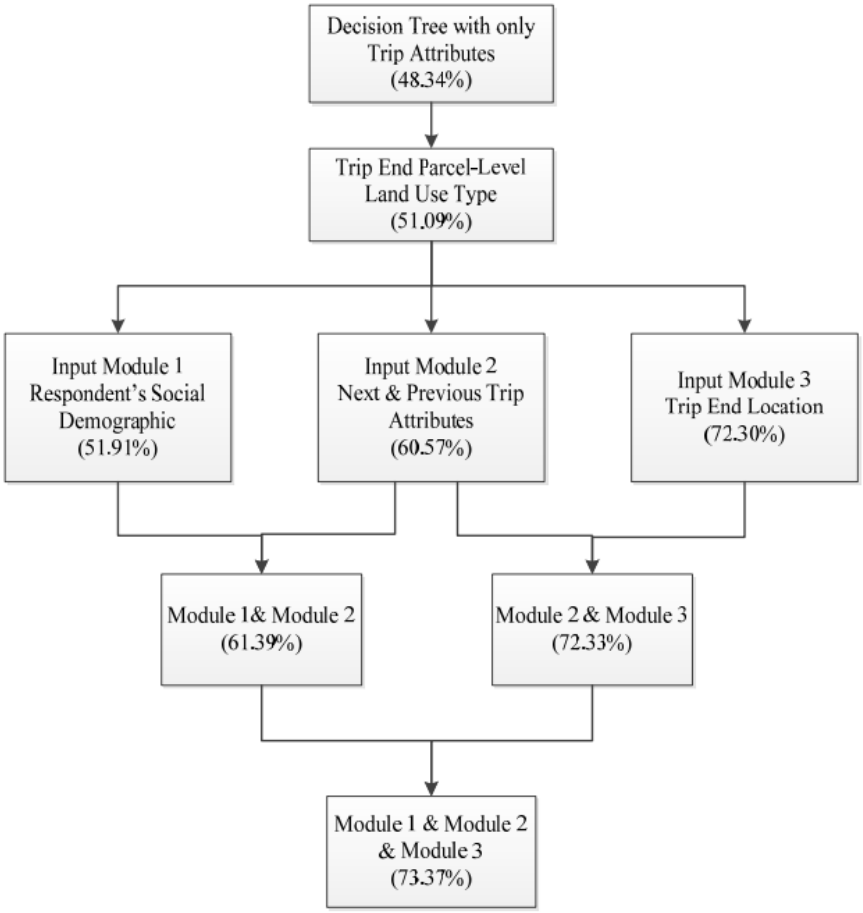


Figura 10 - Procedimento de inferência de propósito de viagem e resultados de precisão (Lu et al., 2012)

2.5.3 Random Forest

O algoritmo de floresta aleatória ou mais conhecido como *Random Forest* (RF) exibe desempenhos muito competitivos, sendo muitas vezes a escolha para muitos problemas reais de classificação. Consiste na construção de várias árvores de decisão durante o treino, Figura 11, ou seja, subconjunto de *features* com a saída do valor mais comum de uma série (moda). Este algoritmo é muito eficiente em conjuntos de dados com elevadas dimensões e elevada dimensionalidade.

As RFs foram introduzidas por (Breiman, 2001) demonstrando que bons resultados podem ser obtidos mesmo na falta de dados, uma vez que são estimadas internamente. Tecnicamente estes algoritmos possuem árvores de decisão onde cada uma conta para a classificação, através de votos. Em uma árvore de decisão regular, um conjunto de dados é dividido pelo recurso que resulta na melhor divisão. Numa RF votos diferentes são necessários para obter a melhor classificação e conseguir com que cada árvore aprenda a partir de um subconjunto diferentes de dados de treino (Montini et al., 2014).

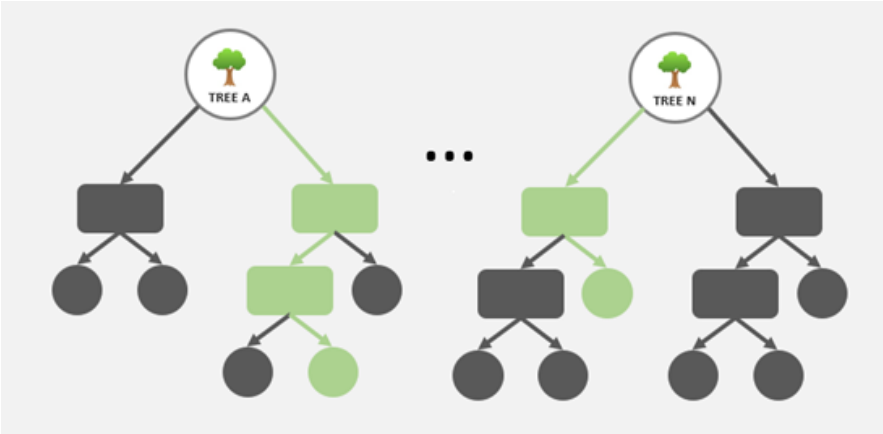


Figura 11 - Random Forest com diferentes conjuntos de árvores de decisão (Cínthia Pessanha, 2019)

No trabalho de (Montini et al., 2014) foram realizados 100 testes com o algoritmo de *Random Forest*, onde empregaram 500 árvores de decisão no algoritmo, de onde resultou uma precisão média de 82.3%. A maior precisão alcançada foi de 84.4%, tal como apresenta a Figura 12. Para esta análise foram selecionados 17 *features*/características agrupadas por 4 grupos diferentes com diferentes medidas de importância (0, 1, 2, 3, e 4, onde 0 é menos importante e 4 mais importante), apresentados na Figura 13 em diferentes cores. A cor amarela representa as *features* relativas às atividades que inclui a distância até ao local de trabalho, percentagem de caminhadas, início da atividade e dia da semana da atividade; cor laranja as *features* de dados pessoais que inclui idade, escolaridade, género e estado civil; cor vermelha a *features* de médias de conjunto de dados relativos à duração da atividade, percentagem dos dias da semana realizadas pelas atividade, número de ocorrências por dia e desvio padrão da duração da viagem; e a cor roxa relativa ao conjunto de dados gerais, como o número total de participantes nesta pesquisa.

	Prediction								
Activity (truth)	Mode Transfer	Being Home	Work–Education	Shopping–Service	Recreation	Pickup–Drop-Off	Business	Other	Recall (%)
Mode transfer	490	1	1	2	0	1	0	0	99.0
Being home	5	374	4	1	2	1	0	0	96.6
Work–education	8	5	177	6	8	0	1	0	86.3
Shopping–service	13	3	3	124	19	2	1	2	74.3
Recreation	10	11	7	25	115	0	0	1	68.0
Pickup–drop-off	6	3	1	14	4	19	0	1	39.6
Business	7	2	11	9	9	0	19	0	33.3
Other	4	1	0	19	13	1	0	7	20.0
Precision (%)	90.6	93.3	85.1	62.6	69.3	82.6	82.6	69.2	

Figura 12 - Resultados obtidos de 100 testes com 500 árvores de decisão (Montini et al., 2014)

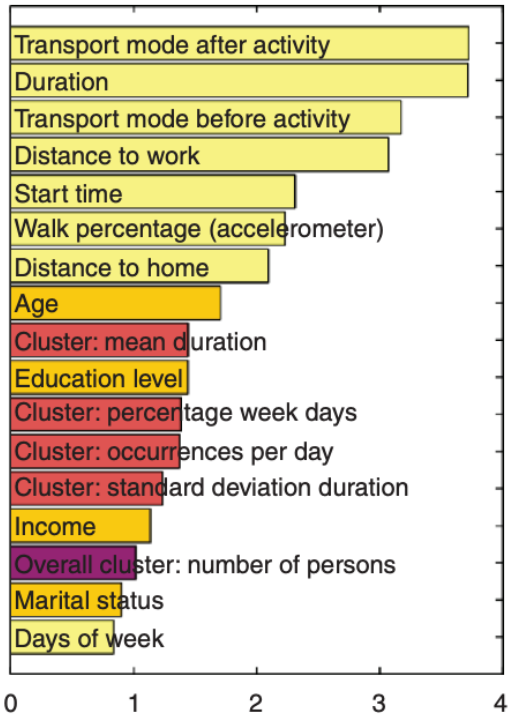


Figura 13 - Características categorizadas e usadas na avaliação da inferência do propósito de viagem (Montini et al., 2014)

As *features* desta figura têm os seguintes significados: modos de transportes antes da atividade ("Transport mode after activity"), duração da atividade ("Duration"), modos de transportes depois das atividades ("Transport mode before activity"), distância até ao local de trabalho/universidade ("Distance to work"), início da atividade ("Start time") percentagem de caminhadas ("Walk percentagem (accelerometer)"), distância até ao local residencial ("Distance to home"), idade do utilizador ("Age"), média do conjunto de dados relativos à duração da atividade ("Cluster: mean duration"), escolaridade ("Education level"), percentagem dos dias da semana realizadas pela atividade ("Cluster: percentage week days"), número de ocorrências por dia ("Cluster: occurrences per day"), desvio padrão da duração da viagem ("Cluster: standard deviation duration"), renda da casa ("Income"), identificação do número de pessoas que conhecem uma localização ("Overall cluster: number of persons"), estado civil ("Marital status") e dia da semana da atividade ("Days of week"),

A seleção desta *features*, foram realizadas com a conveniência dos autores (Montini et al., 2014) determinarem o modo de transporte, que posteriormente usaram como *feature* na inferência do propósito de viagem desta pesquisa.

2.5.4 Redes Neuronais Artificiais

Redes Neuronais Artificiais ou *Artificial Neural Netwo*rk (ANN) correspondem a um modelo matemático computacional para lidar com o problema nos métodos de previsão e tomada de decisão. Consiste no mínimo em 3 camadas: camada de entrada, camada de saída e camada oculta, como demonstrado na Figura 14. Nas camadas de entrada e saída, o número de neurónios é igual ao número de variáveis de entrada e saída, enquanto o número de neurónios na camada oculta depende do tipo de problema em causa.

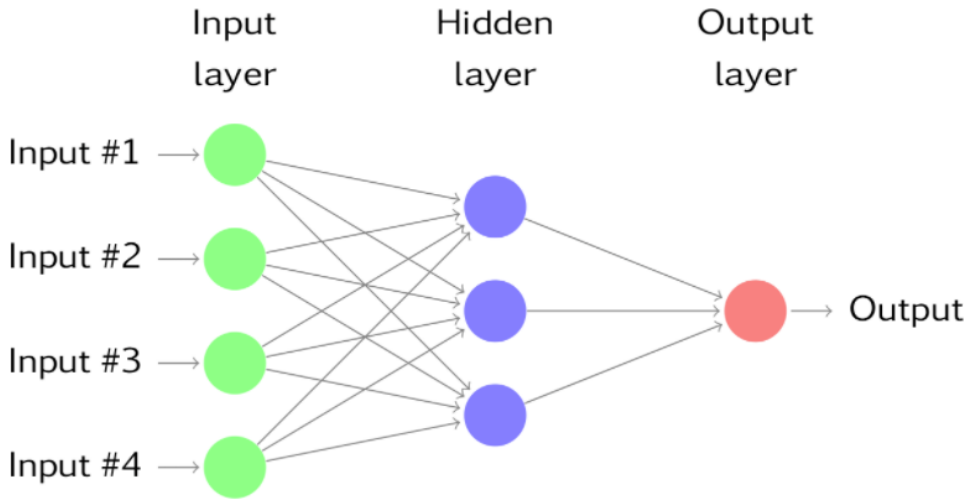


Figura 14 - Esquema de uma Rede Neuronal (Rob J Hyndman e George Athanasopoulos, 2018)

Como este algoritmo corresponde a um algoritmo não paramétrico, permite apresentar classificação de dados de GPS, porque descreve relações complexas e altamente não linear entre variáveis dependentes e independentes. Lida com dados de entrada com ruído de forma eficaz, portanto, o ANN está entre os melhores métodos para determinar o propósito de viagem, segundo o artigo (Xiao et al., 2016a).

As Redes Neuronais Artificiais são métodos de aprendizagem computacional mais abrangentes e com maiores aplicações em situações reais. São redes de neurónios à semelhança das que se encontram no cérbero humano e consistem num sistema computacional paralelo constituído por elementos de processamento muito simples, ligados entre si de forma a realizarem uma dada tarefa. Cada neurónio tem um conjunto de ligações de entrada e um conjunto de ligações de saída, Figura 14. Estas redes aprendem pela quantidade de experiências/testes e podem modificar o seu comportamento em respostas aos estímulos produzidos pelo ambiente envolvido (Xiao et al., 2016a).

Em termos de inferência de propósito de viagem, os autores (Xiao et al., 2016a) utilizam a ANN para superar a desvantagem das abordagens das máquinas de vetores de suporte, árvores de decisão e *Random Forest*, no que diz respeito ao propósito de viagem multiclasse, à recolha de dados de GPS com ruído, à incapacidade de lidar com a ambiguidade e dados que incluem variáveis categóricas com diferente número de níveis a partir das árvores de decisão. O tamanho da amostra em causa foi de 321 entrevistados, de onde resultou dados de GPS com características de 2409 dias, recolhidos de outubro de 2013 a junho de 2015. O número de dias que cada entrevistado participou desta pesquisa variou de 7 a 12 dias. As precisões das redes neuronais artificiais foram superiores a 90%, comparativamente às máquinas de vetores de suporte, no entanto, não foram apresentadas precisões para as árvores de decisão e *Random Forest* neste estudo. No entanto, quando se utiliza dados tabulares e quando se fala de interpretabilidade (Silva & Ribeiro, 2018) o algoritmo *Random Forest* é mais adequado e mais interpretável. Portanto, o RF requer menos pré-processamento, lida com a falta de dados, o processo de treino é mais simples, fazendo com que seja mais simples de usar neste estudo para a inferência do propósito de viagem.

2.6 Algoritmos de Aprendizagem não Supervisionada

Técnicas não supervisionadas, em particular *clustering*, são frequentemente usadas para análise de segmentação. O *clustering* ou agrupamento é quando os objetos são agrupados em subconjuntos

chamados *clusters*. Este método é projetado para ter grupos com as mesmas características e, em seguida, atribuí-los aos *clusters* relevantes. Um dos métodos fundamentados por (Montini et al., 2014) corresponde ao agrupamento hierárquico, técnicas baseadas em hierarquia, em que os dados de um *dataset* corresponde a um *cluster* que se vai ser dividido em diversos *clusters* cada vez menores, no caso de ser divisivo e num único *cluster* maior, ou seja, no caso de ser aglomerativo. Contudo, existem outras técnicas baseadas na partição e baseados na densidade.

2.6.1 Técnicas baseadas em hierarquia

Os métodos hierárquicos aglomerativos ou *Hierarchical Agglomerative Clustering* (HAC) a abordagem é "de baixo para cima" em que o seu próprio *cluster* e os pares de *clusters* são mesclados à medida que se sobe na hierarquia, enquanto os métodos hierárquicos divisivos correspondem ao inverso do método aglomerativo, ou seja, "de cima para baixo", inicia-se com um único grupo e termina com múltiplos *clusters*, Figura 15. Portanto, o resultado final consiste em ter diferentes *clusters* que se relacionam uns com os outros ou quão distantes eles estão.

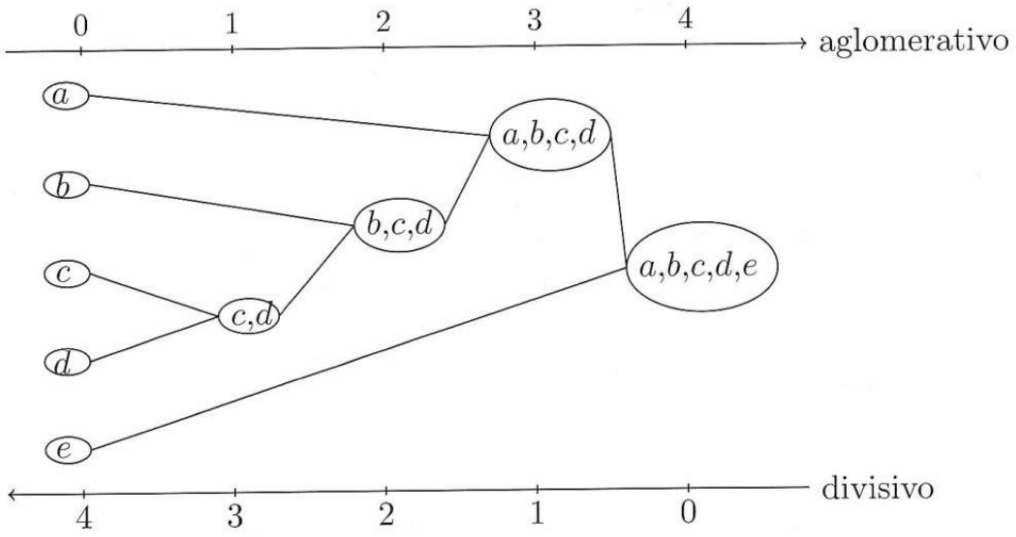


Figura 15 - Métodos hierárquicos aglomeraticos e divisivos (ProFloresta, 2022)

2.6.2 Técnicas baseadas em partição

Nas técnicas baseadas em partição, segundo os autores (Almeida Adriano et al., 2017), cada grupo ou *cluster* representa uma partição, sendo que o número de *clusters* é definido pelo utilizador. Desta forma, cada partição deve conter pelo menos um objeto e cada objeto deve pertencer somente a um grupo, o que é conhecido como separação exclusiva de grupos (em inglês, *exclusive cluster separation* ou *hard cluster*).

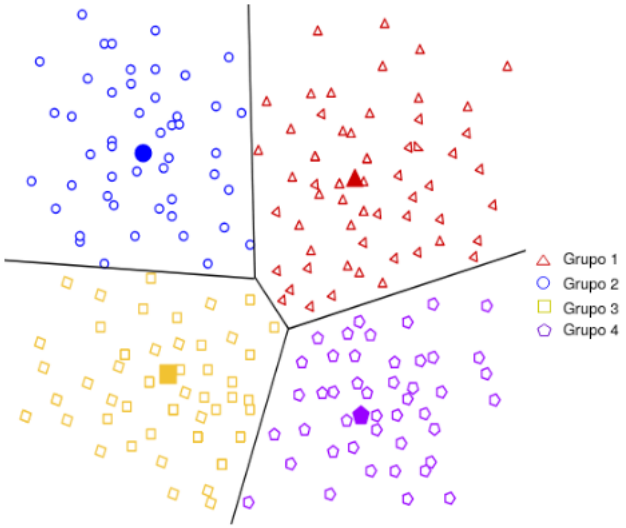


Figura 16 - Agrupamento baseado na técnica de partição (Almeida Adriano et al., 2017)

Grande parte das técnicas baseadas em partição usam medidas de distâncias para determinar o grau de similaridade de cada grupo. Outra característica importante de mencionar é a representação de cada grupo, o centroide, Figura 16, que pode ser definido pela média ou qualquer outra medida estatística. Um dos algoritmos mais utilizados baseados em partição é o *K-means* ou K-médias.

O *K-means* é uma técnica de agrupamento que usa o método de partição para dividir o conjunto de dados em *k* grupos, em que o valor de *k* é definido pelo utilizador. De forma geral, o algoritmo *K-means* cria grupos em que os objetos sejam semelhantes entre si. Então, para que esse objetivo seja atendido, em cada iteração o centróide de cada grupo é atualizado para refinar a qualidade dos grupos.

2.6.3 Técnicas baseadas em densidade

As técnicas com base na densidade segundo o livro de (Silva & Ribeiro, 2018), funcionam na deteção de áreas onde os pontos estão concentrados e onde estão separados, por áreas vazias ou esparsas. Os pontos que não fazem parte de um *cluster* são rotulados como ruído. Opcionalmente, a vizinhança dos pontos pode ser usado para encontrar grupos de pontos que se agrupam no espaço e no tempo. Estes métodos de aprendizagem são considerados não supervisionados, pois não exigem um conjunto de recursos pré-classificados para orientar ou treinar e assim encontrar *clusters* em seus dados.

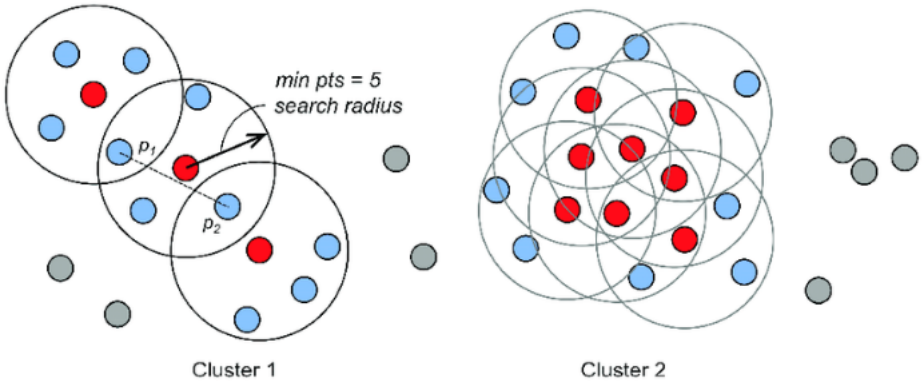


Figura 17 - Algoritmo DBSCAN e dois clusters gerados (DiFrancesco et al., 2020)

Na Figura 17, de acordo com (DiFrancesco et al., 2020), existem três tipos de pontos: pontos-chave (vermelho) são pontos que satisfazem os critérios de agrupamento (denominados pontos centrais), os pontos de fronteira (azul) que não satisfazem os critérios de agrupamento, mas estão ao alcance de um ponto-chave e os pontos de ruído (cinza) que não são nenhum dos dois tipos acima mencionados.

Um algoritmo muito utilizado para esta técnica corresponde ao DBSCAN (*Density-based spatial clustering of applications with noise*) que usa uma distância especificada para separar *clusters* densos de ruído mais esparsos. Este algoritmo é o mais rápido dos métodos de agrupamento, mas só é apropriado se houver uma distância de pesquisa muito clara a ser usada e que funcione bem para todos os agrupamentos em potencial. Isso requer que todos os *clusters* significativos tenham densidades semelhantes. Na Figura 17 o algoritmo DBSCAN usa duas regras: pontos dentro do raio de busca de um ponto-chave que fazem parte do seu *cluster* e pontos-chave que compartilham pontos de fronteira comuns que fazem parte do mesmo *cluster*, mostrado para p1 e p2 em Cluster 1, respetivamente.

2.6.4 DT Cluster

O DT Cluster (Moro et al., 2019), designado de D-Stream pelos autores (Y. Chen & Tu, 2007), está entre os algoritmos de *cluster* de fluxo mais populares e usa uma estrutura de "*grid*" fixa, onde inclui um componente online e um componente offline, Figura 18. Para um fluxo de dados, a cada passo de tempo, o componente online do D-Stream lê continuamente um novo registo de dados, coloca os dados multidimensionais em uma *grid* de densidade e atualize o vetor característico. O componente offline ajusta dinamicamente os *clusters* a cada intervalo de tempo, onde o intervalo é um parâmetro inteiro. Após a primeira lacuna, o algoritmo gera o *cluster* inicial e em seguida, o algoritmo remove periodicamente *grids* esporádicas/dispersas e regula os *clusters*.

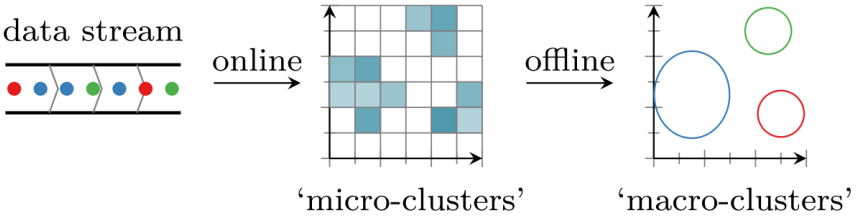


Figura 18 - Cluster de fluxo de dados para as duas fases (online e offline) usando a abordagem baseada em grid (Carnein & Trautmann, 2019)

O algoritmo distingue entre três tipos de células: células densas, células dispersas e células de transição cujo peso está entre os outros dois tipos. O algoritmo mapeia novos pontos de dados para sua respetiva célula e é inicializado atribuindo todas as células densas a *clusters* individuais. Esses aglomerados são estendidos com todas as *grids* de transição vizinhas ou são mesclados com os aglomerados de células densas vizinhas. Em intervalos regulares, o agrupamento avalia o peso de cada célula e incorpora as mudanças nos tipos de células no *cluster*. Desta forma, o algoritmo gera o *cluster* inicial e em seguida, o algoritmo remove periodicamente *grids* dispersas e regula os *clusters*.

2.7 Discussão da Seleção das *Features*

A comparação de resultados em vários cenários é essencial na determinação do melhor resultado possível, tanto para determinar o modo de transporte usado (Shafique & Hato, 2015) como para determinar o propósito de viagem (Xiao et al., 2016a). Um cenário pode incluir informações sobre vários elementos. Dado que o número de viagens para cada propósito é diferente na maioria dos casos, selecionar o conjunto de dados com o mesmo número ou a mesma proporção é importante, ou seja, tornar o *dataset* balanceado. Como resultado, determinar o melhor cenário é sempre uma etapa importante.

Os autores (Shafique & Hato, 2015) recolheram dados de 3 cidades do Japão, Niigata, Gifu e Matsuyama. Em Niigata as entrevistas foram realizadas a 12 participantes nos meses de janeiro e fevereiro de 2011, Gigu foram realizadas a 8 participantes em dezembro de 2010 e janeiro de 2011 e em Matsuyama foram realizadas entre novembro de 2010 e janeiro de 2011, a 26 participantes. Os dados recolhidos podem ser classificados como dados de localização e de viagem. Os dados de localização consistem em dados GPS e dados de acelerómetro. Os dados de acelerómetro consistem na aceleração mínima, máxima e média no movimento e transversal e vertical nas direções. Os dados de viagem abrangem as informações sobre cada viagem, ou seja, data, hora início/fim e transporte usado. Para cada cidade (Shafique & Hato, 2015) selecionaram a modalidade com menos dados e calcularam o número correspondente a 70% desses dados. Os dados iguais a esse número foram então selecionados aleatoriamente de cada transporte usado, ou seja, caminhar/andar a pé, bicicleta, carro e comboio, para formar o conjunto de dados de treinamento, deixando o resto como um conjunto de dados de teste. A Figura 19 mostra a quantidade de dados de treino usados e selecionados para cada cidade, recolhidos de segundo a segundo, onde o GPS localiza um dispositivo em qualquer lugar do mundo com precisão variável, dependendo de fatores como os satélites, ou seja, fatores da área de cobertura e de fatores do acelerómetro que mede a aceleração de um dispositivo em três direções em relação à força gravitacional (g). Isso significa que quando um dispositivo é colocado numa superfície plana, uma aceleração de 1 g é detetada na direção para baixo, enquanto a aceração zero é registada nas outras duas direções.

City	Data selection method	Mode				
		Walk	Bicycle	Car	Train	Total
Niigata	Total	164,078	3,214	61,785	425	229,502
	Equal number	298	298	298	298	1,192
	Equal proportion	114,855	2,250	43,250	298	160,653
Gifu	Total	83,645	24,559	34,678	744	143,626
	Equal number	521	521	521	521	2,084
	Equal proportion	58,552	17,191	24,275	521	100,539
Matsuyama	Total	168,687	15,404	81,653	3,631	269,375
	Equal number	2,542	2,542	2,542	2,542	10,168
	Equal proportion	118,081	10,783	57,157	2,542	188,563

Figura 19 - Tamanho da amostra de dados do modo de transporte (Shafique & Hato, 2015)

Para determinar qual algoritmo seria mais adequado, ou seja, o que apresenta o melhor desempenho, foram usados vários algoritmos: árvore de decisão, *Random Forest* e máquinas de vetores de suporte.

Random Forest teve melhor desempenho em todos os casos na Figura 20. Em particular, a sua precisão foi muita alta, com 99.8% para a média de todos os modos de transportes de 125 pontos com o método de proporção igual. Mesmo com este método a precisão é maior que 91% para o algoritmo *Random Forest*, seguindo ao algoritmo árvore de decisão e máquinas de vetores de suporte. Apesar deste estudo ser feito a determinados modos de transporte usados pelos diferentes participantes, onde o *dataset* Breadcrumbs já apresenta esse dado descrito, é relevante referenciar, porque demonstram que o modelo RF é um modelo eficiente na obtenção de bons resultados quando é trabalhado com dados de viagens.

Selection method	City	Mode	SVM		AdaBoost		Decision Tree		Random Forest	
			Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
Equal number	Niigata	Walk	92.95	92.91	97.32	93.99	100.00	87.85	100.00	94.39
		Bicycle	90.94	90.05	98.99	96.54	100.00	91.50	100.00	97.81
		Car	77.85	79.27	97.65	88.96	100.00	86.69	100.00	91.43
		Train	100.00	100.00	100.00	100.00	100.00	97.64	100.00	100.00
		All	90.44	89.21	98.49	92.67	100.00	87.59	100.00	93.64
	Gifu	Walk	85.03	85.92	95.20	90.80	100.00	85.37	100.00	93.06
		Bicycle	90.98	87.16	96.93	89.82	100.00	83.36	100.00	90.88
		Car	74.09	73.78	94.82	87.16	100.00	87.92	100.00	89.57
		Train	93.09	94.62	100.00	100.00	100.00	99.55	100.00	100.00
		All	85.80	83.22	96.74	89.77	100.00	85.67	100.00	91.85
	Matsuyama	Walk	77.50	77.07	81.83	80.91	98.47	86.26	100.00	91.81
		Bicycle	87.25	86.71	94.26	91.10	99.76	93.40	100.00	98.27
		Car	71.75	70.32	88.28	85.56	98.94	89.60	100.00	95.11
		Train	88.51	87.51	99.49	99.54	99.96	97.52	100.00	100.00
		All	81.25	75.54	90.96	82.91	99.28	87.68	100.00	93.17
Equal proportion	Niigata	Walk	97.95	97.84	98.17	97.92	99.66	99.45	100.00	99.96
		Bicycle	77.96	79.88	84.76	86.41	94.13	92.22	100.00	99.38
		Car	93.81	93.70	93.55	93.37	98.32	97.44	100.00	99.64
		Train	0.00	0.00	30.54	37.01	89.60	85.04	100.00	99.21
		All	96.37	96.30	96.61	96.42	99.20	98.78	100.00	99.86
	Gifu	Walk	97.82	97.70	97.55	97.31	99.28	98.75	100.00	99.84
		Bicycle	88.10	88.04	82.18	81.49	97.91	96.73	100.00	99.78
		Car	91.46	91.57	90.93	91.43	98.23	97.14	100.00	99.71
		Train	56.05	56.05	63.92	61.88	92.90	89.69	100.00	98.65
		All	94.41	94.36	93.15	93.00	98.76	97.97	100.00	99.79
	Matsuyama	Walk	94.10	94.09	94.10	94.13	98.72	98.14	100.00	99.88
		Bicycle	64.50	63.75	51.79	49.12	89.58	87.06	100.00	99.70
		Car	93.12	93.29	91.47	91.17	96.97	95.95	100.00	99.80
		Train	0.00	0.00	20.38	19.47	80.45	74.38	100.00	97.43
		All	90.84	90.85	89.89	89.65	97.42	96.52	100.00	99.81

Figura 20 - Resultados de precisão obtidos pelos diferentes algoritmos nas diferentes cidades (Shafique & Hato, 2015)

Já no estudo feito na Suíça por (Gao et al., 2021), mesma cidade onde foi feito o estudo de (Moro et al., 2019), para determinar o propósito de viagem foram analisadas as trajetórias de GPS de 3689 participantes desde setembro de 2019 a setembro de 2020. Consideraram apenas 91% de todas as atividades que compreendiam o limiar do país, obtendo assim 182 milhões de atividades totais num registo feito em aproximadamente 5 minutos. Apenas 43% das atividades foram rotuladas pelos participantes, ou seja, 782 600 viagens apresentaram as suas finalidades como os seus destinos. Desta forma, apenas esta percentagem foi utilizada neste estudo, embora o limite de 5 minutos de extração de dados de viagem fosse um tempo muito curto para estas atividades, como por exemplo, ir a um minimercado longe de casa.

Tabela III - Categorização das atividades realizadas pelos participantes na Suíça (Gao et al., 2021)

Categoria	Exemplos de atividades	Contagem	Percentagem
Casa	Qualquer atividade em casa	293 129	16.1
Trabalho	Qualquer atividade no local de trabalho	171 329	9.4
Lazer	Exercício, viajar	123 735	6.8
Shopping	Compras, comida	64 071	3.5
Outro		46 413	2.5
Compromisso	Viajar para trabalho	40 119	2.2
Assistência	Buscar/deixar alguém	28 189	1.5
Educação	Universidade, escola	12 694	0.7
Não rotulado	-	1 041 409	57.2
Total	-	1 821 088	100

O registo de dados através de GPS, como possuem um erro de posição entre 0.56 e 50 metros, com uma média de 6.5 metros, fundamentado na secção 2.2.3, isso é levado em consideração no estudo de (Gao et al., 2021). Este estudo conduziu o agrupamento de atividades em 8 categorias, que correspondem a casa, trabalho, lazer, shopping, outro, encargos/responsabilidades, buscar/deixar alguém e educação, Tabela III. Algumas atividades não apresentaram qualquer tipo de rotulação por parte dos participantes. Essa categorização foi baseada na "*Mobility and Transport Microcensus*" (MTMC) da Suíça de 2015. Além disso, informações de POIs da API do Google Places e informações de *land use* não foram empregues neste estudo. A razão pela qual levou a não utilização destas informações deve-se ao facto dos seus benefícios não serem comparativamente altos e por apresentarem uma precisão de inferência de propósito muito baixa. Os recursos extraídos englobaram informações pessoais (tamanho da família, emprego, idade, renda anual, trabalhador/estudante) e atividades (tipo de atividade por dia, duração, início, fim, dia da semana). Neste estudo, considerando a precisão do GPS, um limite de 30 metros é escolhido para definir a que cada atividade de viagem pertence a cada conjunto de dados de atividades agrupados por categorias. Esta categorização é feita através da mineração de dados, ou seja, através da técnica de agrupamento hierárquico, que consiste em dividir os dados em grupos/*clusters*, de forma que cada grupo tenda a ser mais semelhante possível quando comparado com outro (Bohte & Maat, 2009). Esta mineração de dados ocorreu no primeiro semestre de 2007 entre uma amostra de residentes de Amersfoort, Vee Nendaal e Zeewolde, três municípios no centro da Holanda. Este trabalho torna-se relevante, porque apresentam os mesmos modos de viagens e propósitos de viagens que os estudos anteriormente apresentam. Os entrevistados foram recrutados através de uma pesquisa realizada na Internet no final de 2005 e no total resultaram em 1104 pessoas para este estudo. Pessoas acima de 65 anos representam 4% dos entrevistados e na faixa etária entre 50 e 65 anos 35%, dos quais 57% são homens e 43% mulheres. Todos os participantes carregaram consigo um registador de dados GPS portátil com uma duração de bateria de 16 horas. Desta forma, determinaram qual o percentual de todos os modos de viagens realizados e quais as suas finalidades, Tabela IV.

Tabela IV - Modo de viagens com os respetivos percentuais e quantidades realizadas (N) (Bohte & Maat, 2009)

	Viagens corretamente derivadas	
	Percentagem (%)	Quantidade (N)
PROPÓSITO DE VIAGEM		
Trabalho	31	6199
Estudo	4	190
Compras	35	4444
Vida social	11	2120
Lazer	19	3486
Casa	74	11518
Outro	29	5729
Todos os propósitos	43	33686
MODOS DE TRANPORTE		
Carro	75	18017
Comboio	34	747
Autocarro/Elétrico/Metro	0	328
Bicicleta	72	8653
Caminhada	68	5481
Outro	7	460
Todos os modos	70	33686

O uso do carro é deduzido corretamente na maioria das vezes (75%), seguido de bicicleta (72%) e caminhar (68%). A razão apresentada para não terem atingido percentuais mais elevados deve-se à atribuição da velocidade média e máxima nas viagens, por exemplo, dirigir devagar de carro numa área em construção irá atribuir à viagem o modo de transporte bicicleta em vez de carro. O método usado na obtenção destes resultados da Tabela IV consistiu no algoritmo baseado em regras, desenvolvido pelos próprios autores, usando dados obtidos num processo de interpretação e validação dos diferentes modos de viagem e dos diferentes propósitos de viagem. Os dados que resultam do algoritmo, posteriormente, podem ser corrigidos e adicionados pelos entrevistados num aplicativo de validação e, com isto, os autores obtiveram uma precisão coerente tanto na inferência dos modos de viagens como na inferência do propósito de viagem.

Capítulo 3 Análise de dados e Processamento

Neste capítulo, apresentamos uma abordagem para identificar o propósito de viagem de um conjunto de utilizadores individuais. Esta abordagem consiste no desenvolvimento de um modelo, baseado em diferentes métodos retirados de diferentes trabalhos estudados, capítulo 2, (Feng & Timmermans, 2016), (Montini et al., 2014), (Xiao et al., 2016a), que visam demonstrar métodos desenvolvidos para detetar o objetivo de uma determinada viagem, designado de propósito de viagem.

O desenvolvimento desta abordagem proposta centra-se no estado da arte e enquadra-se na área de investigação de *Machine Learning*, com o objetivo da inferência do propósito de viagem. Além disso, acrescentar novas *features* que sejam relevantes e combinar diferentes algoritmos trará melhoramentos. Essa estratégia visa aumentar a possibilidade da obtenção de melhores resultados, tendo assim bons resultados na previsão da inferência do propósito de viagem.

O desenvolvimento será dividido em vários processos: o primeiro processo é focado em estudar as abordagens existentes e entender as diferentes metodologias e desafios que podemos enfrentar; em seguida, usamos a análise da fase anterior para construir um novo modelo que atenda aos nossos requisitos e aplicá-lo a um *dataset* Breadcrumb*s* de dados GPS; o processo final consiste na validação e avaliação dos resultados obtidos.

3.1 Dataset

O *dataset* utilizado para o desenvolvimento de um modelo inicial foi fornecido pelos autores do artigo (Moro et al., 2019), ao qual se designa de Breadcrumbs, e consiste em dados GPS anonimizados de 80 utilizadores da cidade Lausanne (Suíça), por um período de 12 semanas que se estendeu entre março e junho de 2018. A decisão de se ter um conjunto de dados com registos dos utilizadores de Lausanne, foi considerada porque o conjunto de dados é enriquecido com anotações de pontos de locais de interesse do utilizador verdadeiras (*ground truth*), atributos demográficos, relações sociais, informações de saúde, informações de mobilidade, eventos de calendário, registos de contacto e principalmente pela possibilidade de criação de novas *features* com capacidades de inferir o propósito de viagem.

Um questionário foi feito a 80 utilizadores, de acordo com o Regulamento Geral sobre a Proteção de Dados (RGPD), onde todos estes utilizadores consentiram com a utilização dos seus dados para este estudo e assim serem recolhidas as informações demográficas - género, idade, estado civil, nacionalidade, morada e morada dos pais – e informações do estilo de vida - atividades desportivas, hábitos de alimentação, área de trabalho, universidade que frequenta e meios de transportes usados. Sendo assim, o tamanho da amostra da pesquisa é de 80 utilizadores registados, dos quais deles 72 apresentam a ocorrência de eventos tendo um total de 44 mulheres e 36 homens onde a maioria dos utilizadores é estudante, apresentado na Figura 21.

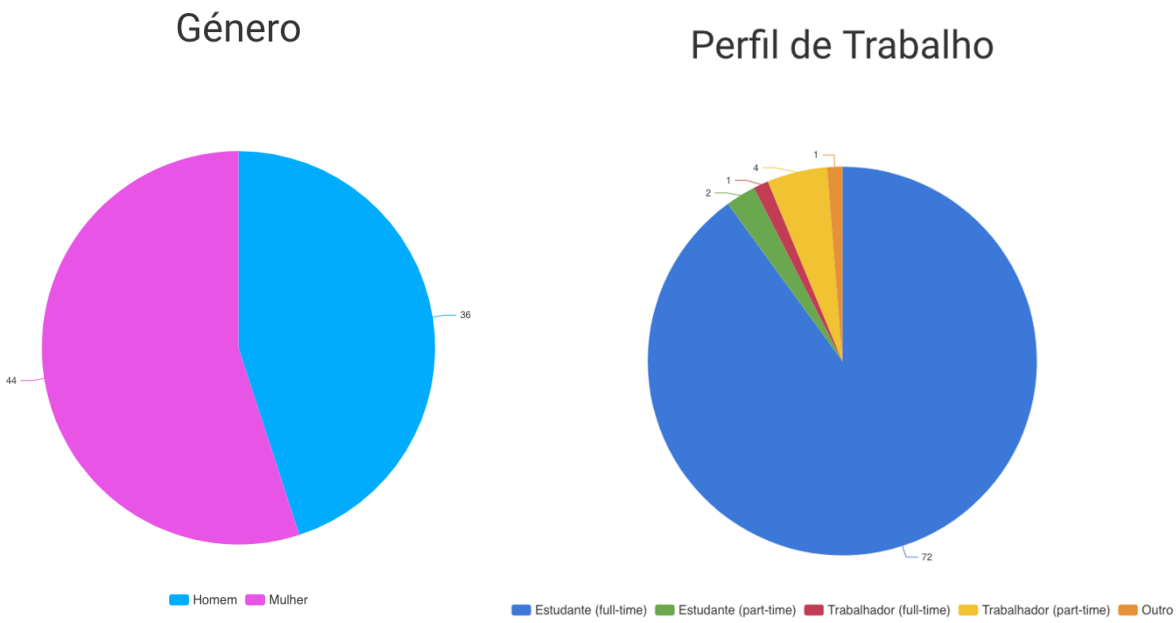


Figura 21 - Número do perfil de trabalho e do género dos utilizadores

O maior percentual de perfil de trabalho corresponde a estudantes, prevendo que as idades sejam inferiores aos 30 anos, porque grande parte das pessoas jovens que frequentam o ensino médio e superior encontra-se dentro dessa faixa etária.

De acordo com a análise prevista das idades dos utilizadores correspondeu ao expectável, tendo uma amostra maioritariamente jovem, entre os 18 e os 27 anos.

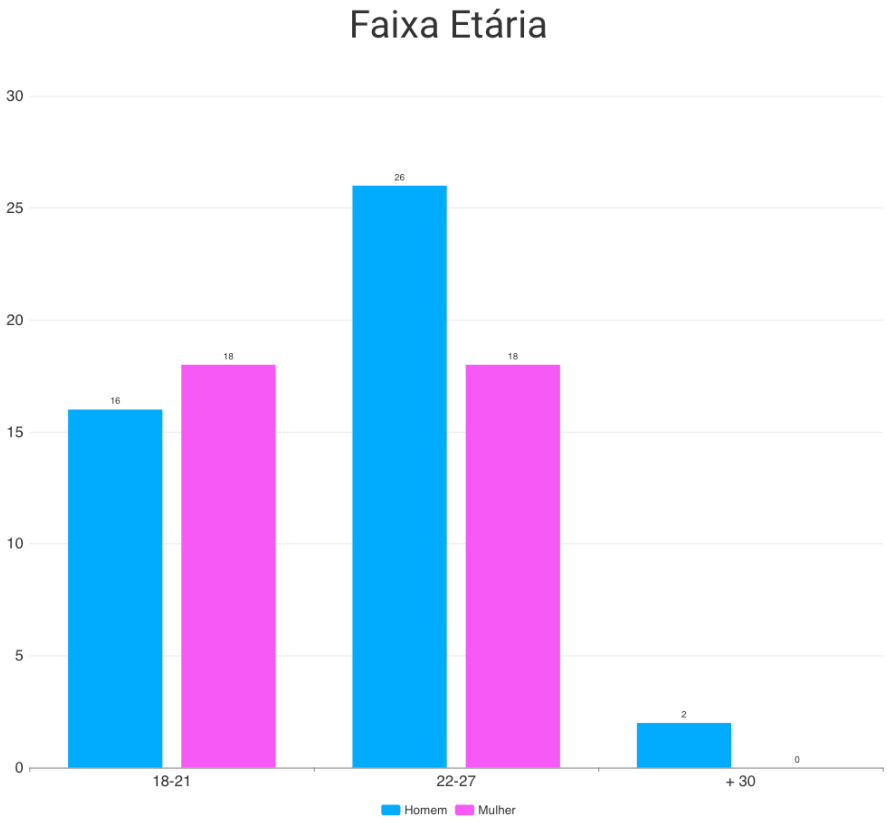


Figura 22 - Faixa etária dos utilizadores do dataset

De acordo com a Figura 22 podemos verificar que o número de utilizadores mulheres entre os 22 e os 27 anos e os 18 e 21 anos correspondem aos mesmos e que a diferença está apenas entre os homens, onde o maior número de utilizadores homens está nas idades compreendidas entre os 22 e os 27 anos. Os dois utilizadores com a idade superior aos 30 anos correspondem a uma pequena percentagem para a identificação do propósito de viagem. Apesar deste pequeno conjunto de utilizadores apresentarem a ocorrência de eventos, não se relacionam com outros utilizadores, não se relacionam com algum familiar ou amigo, quando efetuam uma determinada viagem.

Os dois grupos de nacionalidade mais importantes é suíço e francês, 44 e 21 utilizadores, respetivamente. Dos 80 utilizadores, 28 vivem com os pais e 52 vivem sozinhos, Figura 24. Os pais dos diferentes utilizadores têm a sua residência em diferentes cantões da Suíça, na França e em outros locais da Europa.

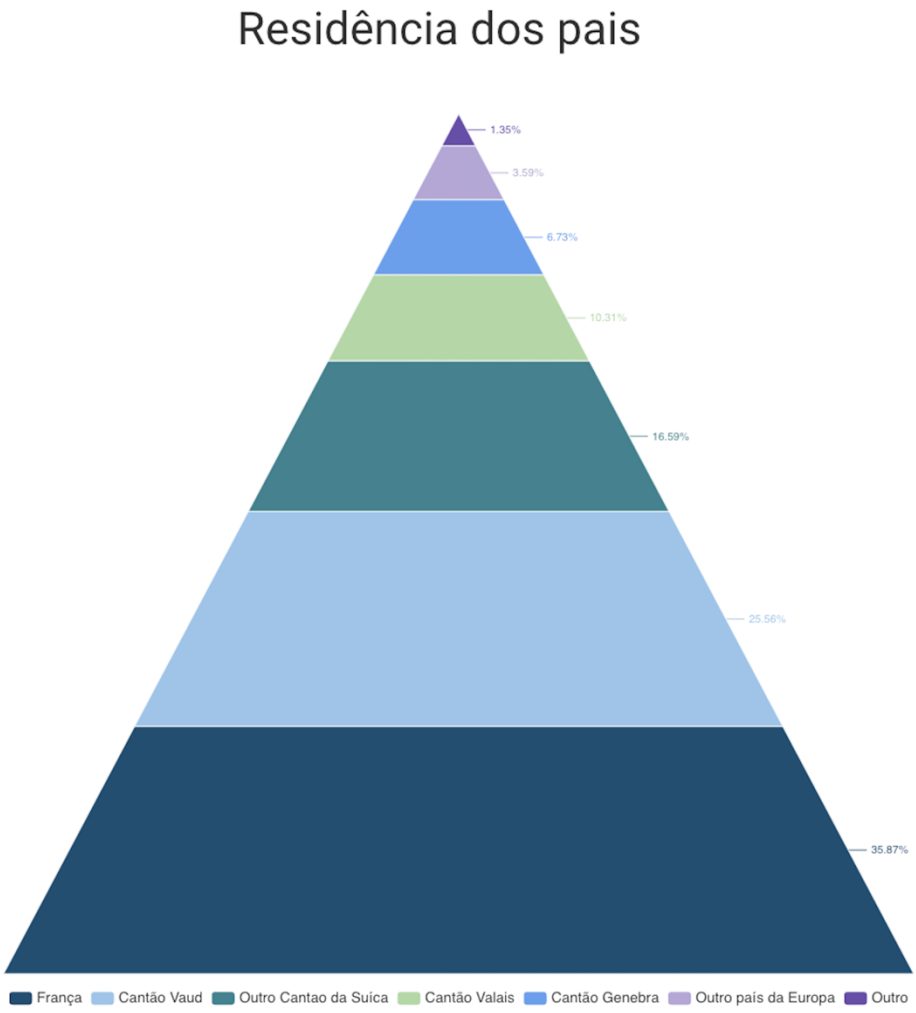


Figura 23 - Localização da casa dos pais dos 80 utilizadores

Na Figura 23 e Figura 24, podemos ver a residência dos 28 utilizadores que vivem na residência dos pais. Porém, 52 dos utilizadores vivem em residências independentes à dos pais, distribuindo 22 utilizadores na Suíça, 18 utilizadores na França, 2 utilizadores na Alemanha, 2 utilizadores na Itália e os restantes em outras regiões. Assegurando os dois grupos de nacionalidade mais importantes a Suíça e a França, como anteriormente foi referenciado.

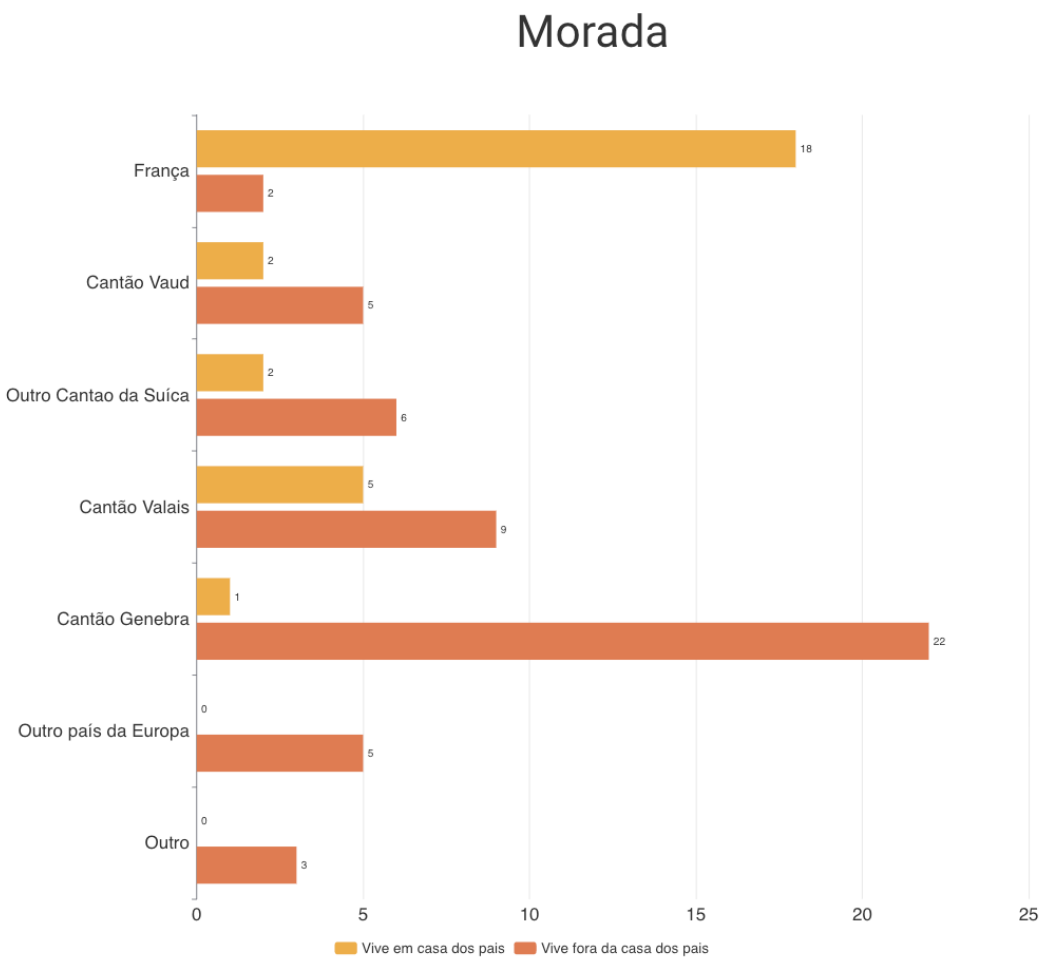


Figura 24 - Relação número de utilizadores e local de residência

Em termos de estilo de vida que os utilizadores levam, 18 utilizadores praticam desporto mais de 5 horas por semana, 40 utilizadores praticam desporto entre 1 hora e 5 horas, 21 utilizadores praticam menos de 1 hora e um dos utilizadores não respondeu a esta pergunta no inquérito. Do total, 63 utilizadores apresentam o estado civil como solteiro, dado que são maioritariamente utilizadores jovens estudantes.

Os meios de transportes utilizados durante a semana e fim de semana pelos participantes são observados através da Figura 25 e são baseados nos hábitos dos utilizadores, havendo um aumento do uso de meios de transportes próprios durante o fim de semana em relação aos dias de semana e o aumento da utilização de meios de transportes públicos nos dias de semana relativamente aos fins de semana.

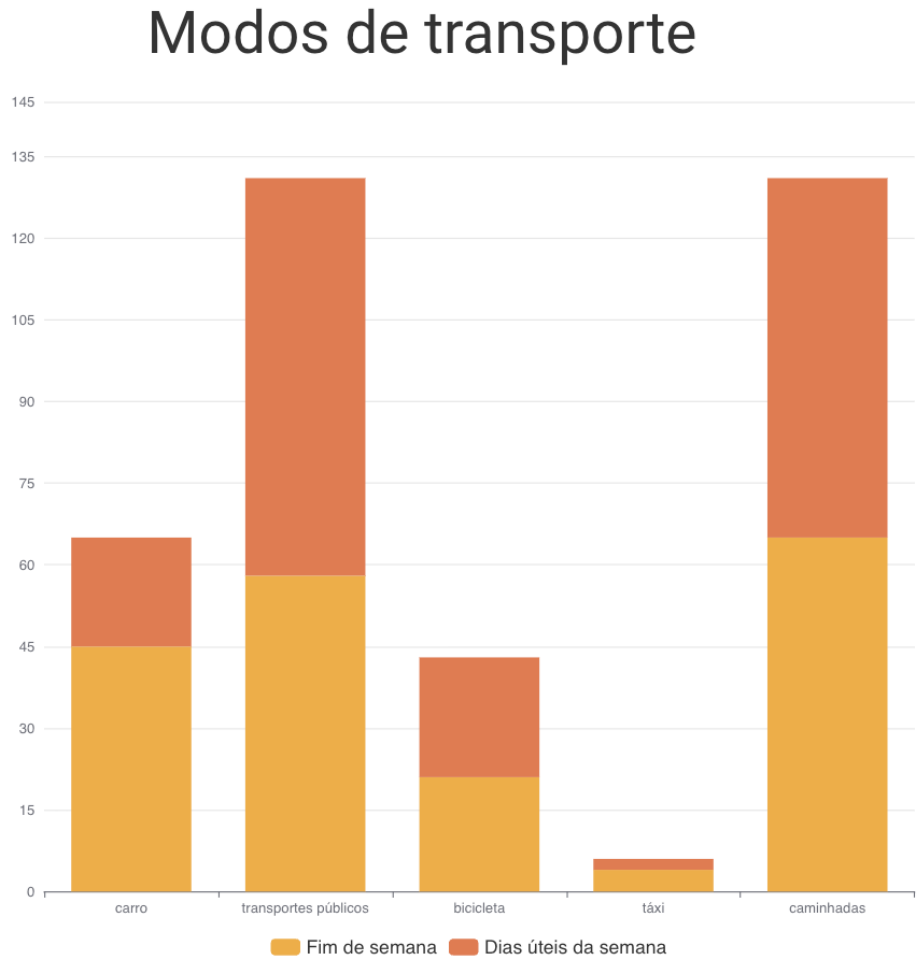


Figura 25 – Hábitos dos meios de transporte usados durante uma semana

No entanto, os hábitos de caminhada e bicicleta são semelhantes durante a semana e ao fim de semana. Apesar desta descrição no *dataset* Breadcrumbs, por serem hábitos impossibilitam o seu uso no nosso estudo, porque a sua caracterização não é específica para cada utilizador.

Através do algoritmo de *clustering* DT Cluster e critérios de seleção, os autores do *dataset* Breadcrumbs fizeram o agrupamento do conjunto de dados para processar e rotular os pontos de locais de interesse do utilizador. Os critérios de seleção incluíram o número de pontos retornados, a distância mínima entre pontos de locais de interesse do utilizador distintos e o número de parâmetros. O ponto de local de interesse do utilizador consiste num ponto no mapa com um determinado raio e está discriminado no *dataset* Breadcrumbs que correspondem à tabela "point_of_interest" da Figura 27.

O algoritmo de *clustering* DT Cluster (Moro et al., 2019), designado de D-Stream pelos autores (Y. Chen & Tu, 2007), processou os pontos de locais de interesse que representam os locais de atividades realizadas pelos utilizadores, combinando os que estavam sobrepostos e removendo esses pontos que os participantes visitaram menos de 3 vezes ao longo de toda a experiência. O algoritmo consiste no agrupamento de fluxo de dados em tempo real usando uma abordagem baseada em densidade e é essencialmente aplicado sobre os pontos demarcados pelos trajetos de GPS. Usa uma componente online que mapeia cada registo de dados de entrada numa "*grid*" e uma componente offline que calcula a densidade de cada *grid*, regiões densas que são separados por regiões de baixa densidade, que geralmente representam ruídos agrupados. Desta forma, elimina os *outliers* e reduz a complexidade do problema em causa. Ao utilizar os dados do *dataset* Breadcrumbs já estamos a tirar partido destes benefícios que o algoritmo DT Cluster arrecada.

Os rótulos possíveis dos pontos de locais de interesse do *dataset* correspondentes são: casa ("Home"), universidade ("University"), desporto ("Sport"), parque verde ("Park"), bares ("Bar"), restaurante ("Restaurant"), "Shopping", paragem de autocarro ("Bus Stop"), paragem de metro ("Metro Stop"), estação de comboios ("Train Station"), trabalho ("Work"), vida social/casa de um amigo ("Friend's Place"), estação de bicicletas ("Velo Station"), família ("Family"), carros alugados ("Car sharing"), parque de estacionamento ("Parking"), associação ("Association"), "Hotel", aeroporto ("Airport"), livraria ("Library") e praia ("Beach"), Figura 26. O conjunto de dados das colunas com todos os atributos/campos do *dataset* pode-se visualizar na Figura 27 e a sua descrição, quer das tabelas quer dos atributos, encontra-se no Anexo A: Descrição do *dataset* Breadcrumbs.

Figura 26 - Rótulos dos pontos de locais de interesse

Os pontos de locais de interesse rotulados que apresentam o maior número de frequência correspondem às Universidades, dado que o maior número de perfil de trabalho dos utilizadores é estudante. O maior número de ocupação de lazer corresponde aos Shoppings, restaurantes e locais de desporto. Os meios de transportes mais utilizados são os meios de transporte públicos, como fora indicado nesta secção, onde resultam os mais utilizados os transportes de comboio, metro e autocarro. Assim, a maioria dos pontos de locais de interesse correspondem aos rótulos semânticos de transporte, estudo e residência.

Os dados do *dataset* Breadcrumbs são etiquetados, bem como atributos demográficos, registos de contacto, eventos de calendário, informações de estilo de vida e rótulos de relacionamento social

entre os participantes. Esses atributos exclusivos tornam o *dataset* ideal para várias áreas de pesquisa, incluindo na inferência do propósito de viagem.

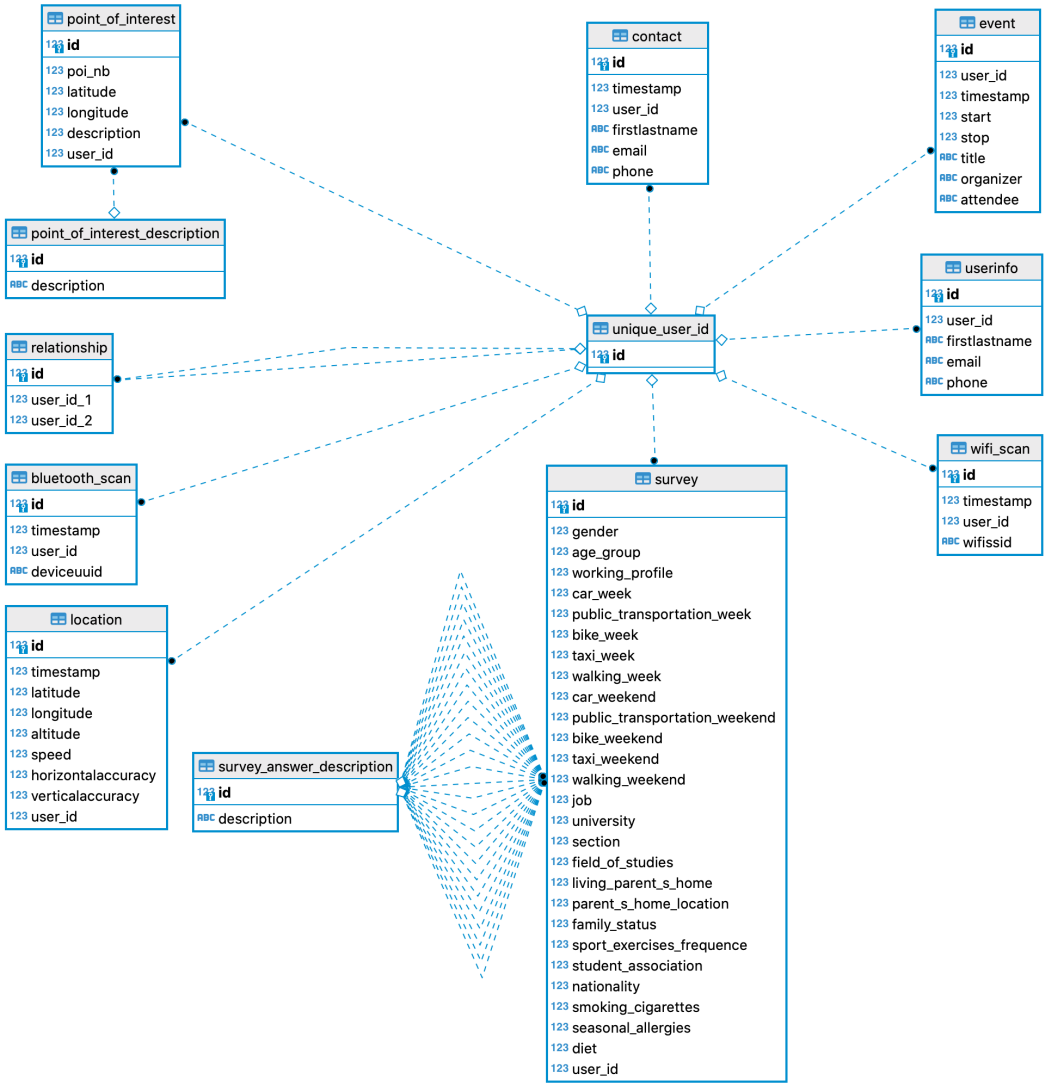


Figura 27 - Esquema do conjunto de dados do Breadcrumbs (Moro et al., 2019)

3.2 Preparação do *dataset*

A preparação do novo *dataset* que irá ser utilizado no nosso modelo para determinar com previsão a inferência do propósito de viagem será compreendido em duas fases. A primeira fase corresponde ao pré-processamento que envolve toda a parte de conversão dos dados originais do *dataset* Breadcrumbs para valores numéricos, discriminados, valores do tipo *dummy*, etc., e a segunda parte na extração de *features*, ou seja, criação e extração de variáveis finais do *dataset* propriamente dito.

3.2.1 Pré-processamento

Após a análise do conjunto de dados usando a linguagem SQL, decidiu-se usar os dados das tabelas "point_of_interest", "point_of_interest_descrition", "event" e "location" da Figura 27 para

determinar as *features* de atividade e *features* de *clustering* descritos na pesquisa de (Montini et al., 2014), secção 2.5.3. Todas as tabelas da Figura 27 estão discriminadas no Anexo A: Descrição do *dataset* Breadcrumbs e as *features* na Figura 13, utilizadas por (Montini et al., 2014), que correspondem às *features* de atividade: distância até ao local de trabalho, percentagem de caminhadas, início da atividade e dia da semana da atividade; *features* pessoais: idade, escolaridade, renda e estado civil; *features* de *clustering:* média do conjunto de dados relativos à duração da atividade, percentagem dos dias da semana realizadas pelas atividade, número de ocorrências por dia e desvio padrão da duração da viagem; e *features* gerais: número de participantes. Da Figura 27, também foram usados os dados das tabelas "survey" e "survey_answer_description" para fazer a recolha de algumas *features* pessoais dos entrevistados do *dataset* Breadcrumbs e, posteriormente, endereçar estes dados numa base de dados espacial em PostgreSQL para, posteriormente, se proceder à aplicação do SQL e importar a nossa base de dados para um ficheiro .xlsx para aplicar o algoritmo *Random Forest* (RF) na inferência do propósito de viagem a esse *dataset*. Este tipo de base de dados também permitirá transferir os dados armazenados para uma ferramenta de visualização, QGIS, para ser possível visualizar todos pontos de locais de interesse correspondentes aos destinos das viagens realizadas pelos utilizadores, Figura 37 e Figura 38. Esse destino pode ser igual ou próximo, devido ao erro associado durante a recolha de dados GPS, definido como variável "raio" no ponto um do Anexo H: Queries SQL usadas na recolha de *features*, a um ou mais pontos de locais de interesse.

Para analisar o conjunto de dados decidiu-se que inicialmente seria interessante identificar o número de paragens realizadas a um determinado ponto de local de interesse. A identificação de cada paragem de uma viagem realizada por cada utilizador/entrevistado poderá apresentar vários destinos a diferentes pontos, sendo necessário reconhecer a ambiguidade existente de cada viagem, ou seja, se existe uma mesma paragem com destino a mais de um ponto de local de interesse diferente.

Inicialmente determinamos quais as paragens realizadas pelos diferentes utilizadores através do algoritmo de deteção de paragens de trajetórias de movimento, secção 3.4.1. Através desse algoritmo obteve-se a tabela da Figura 28, onde a coluna "start_paragem" e "stop_paragem" corresponde ao início e fim da atividade, respetivamente, e "duracao" ao tempo em horas da duração da atividade. Para identificar a ambiguidade existente para os diferentes raios, 6.5, 15, 25, 35 e 50 metros, foi necessário adequar a identificação do ponto de local de interesse a cada paragem/instância da Figura 28.

user_id integer	start_paragem timestamp without time zone	stop_paragem timestamp without time zone	duracao interval
102	2018-03-26 13:29:49	2018-03-26 15:30:23	02:00:34
102	2018-03-26 17:15:04	2018-03-26 17:58:47	00:43:43
102	2018-03-26 17:58:49	2018-03-26 18:22:39	00:23:50
102	2018-03-26 18:22:43	2018-03-26 18:30:35	00:07:52
102	2018-03-26 18:32:55	2018-03-26 18:40:44	00:07:49
102	2018-03-26 18:43:14	2018-03-26 18:58:35	00:15:21
102	2018-03-26 18:59:47	2018-03-26 20:09:30	01:09:43
102	2018-03-26 20:09:40	2018-03-29 12:40:49	2 days 16:31:09
102	2018-03-29 12:41:07	2018-03-29 13:07:48	00:26:41
102	2018-03-29 13:14:29	2018-03-29 13:40:33	00:26:04

Figura 28 – Deteção de Paragens

A razão que leva à utilização deste algoritmo de deteção de paragens de trajetórias, deve-se à incompreensibilidade dos dados de identificação do início e fim da atividade no *dataset* Breadcrumbs, devido à existência de vários "*timestamps"* (data e hora) do início e fim de viagem e ainda o *timestamp* do evento, que correspondem tanto às atividades registadas pelos utilizadores como às atividades de calendarização frequentemente realizadas naquele local, respetivamente. Desta forma, será necessário determinar qual *timestamp* terá de ser considerado para determinar a duração das atividades e a distância da paragem à residência e ao local de trabalho, ou seja, determinar todas as paragens realizadas pelos diferentes utilizadores. Como o *timestamp* do evento, considerado como o momento de identificação da atividade a um determinado ponto de local de interesse de calendarização daquela região. Segundo os autores (Moro et al., 2019) do artigo Breadcrumbs, estes dados encontram-se às vezes depois do intervalo de data e hora do início e da data e hora do fim da atividade. A razão por este *timestamp* encontrar-se depois do início e fim da atividade deve-se à pós recolha desses dados, porque o utilizador em causa não o identificou no momento exato. Com isto, é necessário detetar/extrair paragens de trajetórias de movimento efetuadas pelos diferentes utilizadores para obtermos um *ground truth* da inferência do propósito para ser possível testar o nosso modelo RF. Para esta correspondência, de cada paragem a um determinado ponto de local de interesse, recorreu-se aos dados de "start_paragem" da Figura 28 para identificar a localização de cada instância a partir da associação da coluna "timestamp" da tabela "location" da Figura 27. Uma vez que os dados da coluna "timestamp" estavam no formato *Unix Timestamp* (por exemplo "1655723554"), que corresponde ao número de segundos passados, foi necessário fazer as devidas alterações à tabela convertendo os valores para o tipo *datetime* (por exemplo "2022-06-20 11:12:31"). Estas alterações foram feitas recorrendo ao segundo ponto do Anexo H: Queries SQL usadas na recolha de *features*.

Assim, obtém-se todas as localizações de paragens que possibilitam a determinação do melhor raio que comprometa a melhor ambiguidade, ou seja, fazer com que uma paragem corresponda a um único ponto de local de interesse com a identificação de uma só atividade, através da tabela "point_of_interest", como foi anteriormente argumentado. O terceiro ponto do Anexo H: Queries SQL usadas na recolha de *features* possibilitou essa identificação, através dos resultados obtidos e visualizados no capítulo 4, Figura 39, de onde resultou o raio de 6.5 metros como o raio que proporcionou uma relação de baixa ambiguidade e um número adequado de instâncias.

3.2.2 Extração de *features*

Após o pré-processamento dos dados originais para obter a nossa variável dependente de cada paragem, atividade realizada, consegue-se determinar as *features* "Distance to work" e "Distance to home" através da função SQL *st_distance(),* quarto ponto do Anexo H: Queries SQL usadas na recolha de *features*. No caso destas *features* apresentarem dois locais diferentes, porque alguns dos utilizadores do *dataset* Breadcrumbs apresentam duas residências, foi considerada a distância de paragem à residência mais próxima através da função *min()*.

Com toda a existência dos dados da Figura 29, "user_id", "start_location_latitude", "start_location_longitude", "start_paragem", "stop_paragem", "dist_metros_to_work", "dist_metros_to_home", "poi_description" e "duracao", proporciona a recolha direta das restantes *features* pessoais do *dataset* Breadcrumbs, ao mesmo tempo que se ajusta o tipo de dados das colunas.

user_id	start_location_latitude	start_location_longitude	start_paragem	stop_paragem	dist_metros_to_work	dist_metros_to_home	poi_description	duracao
102	46.511997	6.617976	2018-05-26 12:11:00	2018-05-26 12:20:42		0	Home	00:09:42
102	46.512085	6.618016	2018-05-29 03:10:50	2018-05-29 03:21:22		0	Home	00:10:32
102	46.512037	6.618031	2018-06-17 07:55:03	2018-06-17 07:59:59		0	Home	00:04:56
102	46.512099	6.618041	2018-05-07 20:56:07	2018-05-07 23:40:35		0	Home	02:44:28
102	46.512042	6.617936	2018-05-25 09:46:11	2018-05-25 09:59:43		0	Home	00:13:32
102	46.512022	6.617937	2018-04-25 12:43:59	2018-04-25 13:14:48		0	Home	00:30:49
102	46.51201	6.617936	2018-04-12 03:36:07	2018-04-12 03:41:10		0	Home	00:05:03
102	46.512018	6.618035	2018-05-17 08:12:21	2018-05-17 08:17:50		0	Home	00:05:29
102	46.512029	6.618002	2018-04-27 06:37:31	2018-04-27 07:09:15		0	Home	00:31:44
102	46.512017	6.618056	2018-05-25 06:05:12	2018-05-25 07:43:34		0	Home	01:38:22
102	46.512007	6.617982	2018-04-27 13:13:30	2018-04-27 13:42:30		0	Home	00:29:00

Figura 29 - Dataset inicial das paragens

Na coluna "stop_paragem" ao identificar o fim da realização da atividade anuncia o início da viagem daquele utilizador e a coluna "start_paragem" ao identificar o início da atividade anuncia a finalização da viagem. Assim, a duração entre o "start_paragem" e o "stop_paragem" representa a duração da atividade, ou seja, a coluna "duracao" e o contrário a duração da viagem. A coluna "duracao" foi convertida para horas atribuindo o nome da coluna como "duracao_hours". As distâncias foram consideradas com um máximo de duas casas decimais. A coluna "start_paragem" foi separada em três novas colunas, a coluna "start_time_hours_day", "day_week" e "date_year", ou seja, a coluna "start_time_hours_day" apresenta o valor da hora do dia, "day_week" o dia da semana da atividade e "date_year" a data da atividade. O "day_week" foi representado inicialmente por dados numéricos, onde:

- 0 representa o domingo
- 1 representa a segunda-feira
- 2 representa a terça-feira
- 3 representa a quarta-feira
- 4 representa a quinta-feira
- 5 representa a sexta-feira
- 6 representa o sábado

As *features* pessoais foram recolhidas diretamente para as colunas "grupo_idade", "escolaridade", "genero" e "estado_civil". Através do ponto cinco do Anexo H: Queries SQL usadas na recolha de *features*, que resultou na estruturação dos dados e recolha de *features* pessoais, obtivemos a completude do nosso *dataset* que poderá ser observado na Figura 30.

user_id	start_location_latitude	start_location_longitude	duracao_hours	dist_metros_to_work	start_time_hours_day	dist_metros_to_home	grupo_idade	escolaridade	genero	estado_civil	day_week	date_year	poi_description
102	46.511997	6.617976	0.16		12.18	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	6	2018-05-26	Home
102	46.512085	6.618016	0.17		3.18	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	2	2018-05-29	Home
102	46.512037	6.618031	0.08		7.91	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	0	2018-06-17	Home
102	46.512099	6.618041	2.74		20.93	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	1	2018-05-07	Home
102	46.512042	6.617936	0.22		9.76	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	5	2018-05-25	Home
102	46.512022	6.617937	0.51		12.73	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	3	2018-04-25	Home
102	46.51201	6.617936	0.08		3.60	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	4	2018-04-12	Home
102	46.512018	6.618035	0.09		8.20	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	4	2018-05-17	Home
102	46.512029	6.618002	0.52		6.62	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	5	2018-04-27	Home
102	46.512017	6.618056	1.63		6.08	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	5	2018-05-25	Home
102	46.512007	6.617982	0.48		13.22	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	5	2018-04-27	Home
102	46.512012	6.618032	0.24		7.94	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	4	2018-05-17	Home
102	46.512005	6.617973	0.29		5.37	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	4	2018-05-24	Home
102	46.51205	6.61798	0.38		18.26	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	3	2018-04-25	Home
102	46.512013	6.617976	0.21		21.77	0.00	Between 18 and 21 years old	Bachelor	Female	Single (Célibataire)	2	2018-05-01	Home

Figura 30 - Dataset sem as features de clustering

Todos estes procedimentos foram realizados pelo facto do modelo RF ser um modelo de aprendizagem com várias características que conferem versatilidade, podendo ser aplicado a problemas de previsão de duas classes ou multiclasses. Por esse motivo é necessário modelar as variáveis para assumir uma mistura de variáveis categóricas, contínuas ou discretas.

- Variável contínua corresponde a variáveis numéricas que têm um número infinito de valores entre dois valores quaisquer. Uma variável contínua pode ser numérica ou de data/hora, como acontece com a coluna "start_time_hours_day" e "date_year".
- Variável categórica contém um número finito de categorias ou grupos distintos. Os dados categóricos podem não ter uma ordem lógica como, por exemplo, género (colunas "grupo_idade", "escolaridade", "genero" e "estado_civil").
- Variável discreta corresponde a variáveis numéricas que têm um número contável de valores entre quaisquer valores. Uma variável discreta é sempre numérica.

O RF é uma técnica de modelação forte e muito robusta do que uma única árvore de decisão. A agregação de várias árvores limita a possibilidade de *overfitting* e erros de cálculo devido ao enviesamento, peso desproporcional.

A inexistência de dados, dados NULL, na coluna "dist_metros_to_work" no *dataset* impede com que o algoritmo RF lide com esses valores e, segundo (Montini et al., 2014), pode-se incluir a atividade "University" na atividade "Work" para obter esses valores, porque caso o utilizador não tenha referenciado o local de trabalho, mas tenha referenciado o local de estudo permitirá o cálculo da *feature* "Distance to work", caso não tenha ambos os dados esse utilizador será descartado do *dataset*. Através destes procedimentos resultou apenas 56 utilizadores no novo *dataset* dos iniciais 78. Os utilizadores que apresentem ambas as atividades foi considerada a distância à atividade "Work". Desta forma, obteve-se um *dataset* com 2365 paragens realizadas.

Todas as nossas instâncias, as 2365 paragens, foram efetuadas a uma velocidade entre o início e o fim da atividade. A média dessa velocidade permite determinar a *feature* "walk percentage" da Figura 13, que é incluída no modelo quando queremos determinar uma previsão apenas e só com *features* durante o deslocamento, ou seja, antes de o utilizador chegar ao destino. Primeiramente, segundo a secção 2.3.1 do estado da arte, todas as velocidades compreendidas até os 6km/h são consideradas como realizadas a pé/caminhar. Através da tabela "location" da Figura 27, conseguimos determinar essa média das velocidades efetuada na atividade pela função SQL *avg(),* associando o "user_id" de cada tabela e verificando se o "timestamp" está dentro do *timestamp* de início e fim de cada paragem. Todo este procedimento pode ser visualizado através da seguinte query do ponto seis do Anexo H: Queries SQL usadas na recolha de *features*.

Através dos resultados obtidos da velocidade média efetuada a pé em cada atividade conseguiu-se realizar a recolha de todas as *features* a cor amarela e laranja da Figura 13 do artigo (Montini et al., 2014), que estiveram ao nosso alcance, considerando a disponibilidade dos dados do *dataset* Breadcrumbs, excepto os modos de transportes antes e depois das atividades ("Transport mode after activity" e "Transport mode before activity") e a renda ("Income"). Essas *features* correspondem às colunas "duracao_hours", "dist_metros_to_work_university", "start_time_hours_day", "walk_percentage_trip", "dist_metros_to_home", "grupo_idade", "escolaridade", "estado_civil" e "day_week" da Figura 31. A *feature* "duracao_hours" corresponde à duração da atividade em horas; "dist_metros_to_work_university" corresponde à distância em metros daquela paragem até ao trabalho/universidade; "start_time_hours_day" corresponde ao tempo em horas do momento que iniciou a atividade; "walk_percentage_trip" corresponde à percentagem da realização a pé de uma viagem por aquele utilizador num determinado dia da semana; "dist_metros_to_home" corresponde à distância da paragem até à residência do utilizador em metros; "grupo_idade" corresponde ao grupo de idade a que pertence o utilizador, Figura 22; "escolaridade" corresponde ao nível educacional do utilizador, por exemplo Licenciatura; "estado_civil" corresponde à identificação de solteiro ou casado e "day_week" corresponde ao dia em que ocorreu a viagem/atividade.

A implicação da obtenção dos conjuntos de classes minoritárias e classes maioritárias de paragens, prejudica a aplicação do algoritmo RF. Desta forma, foi considerado todas as atividades realizadas com mais de 1% e, com isto, os POIs que foram considerados da tabela "point_of_interest" correspondentes à coluna "poi_description" são: "Home", "Friend's Place", "Sports", "University", "Metro Stop", "Family", "Bus Stop", "Shopping", "Work" e "Restaurant", Tabela IV, que especifica a atividade correspondente ao POI daquela paragem. Devido a um erro de seleção das atividades, em vez de ter sido feita a consideração da atividade "Friend's Place" fez-se a consideração da atividade "Park", não influenciando na performance do modelo uma vez que o modelo RF tem que estar preparado para qualquer tipo de atividade realizada pelos utilizadores.

Neste seguimento, resta apenas por obter as *features* de *clustering*, que se encontram a cor vermelha na Figura 13. Esta recolha e resolução é feita após os procedimentos da aplicação do algoritmo HAC da secção 3.4.2, após ser definido o número de *clusters* individualmente de cada *cluster* de paragens por utilizador e recorrer-se à seleção do "Main Cluster", através da maior quantidade de paragens do "poi_description" igual a "Home" que vai de encontro com os métodos e critérios tido na secção 3.3.

Assim, obtém-se as *features* "mean duration", "percentage week days", "occurrences per day" e "standard deviation". A *feature* "percentagem week day" corresponde à percentagem de ocorrências da atividade "Home" num determinado dia da semana no *cluster* "Main Cluster" e foi calculada após a modificação da coluna "day_week" da Figura 30, onde os valores inteiros 0, 1, 2, 3, 4, 5 e 6 foram convertidos para o tipo *string* das iniciais do dia da semana, por exemplo, 0 para "Sun" (domingo), para desta forma gerar várias colunas para cada dia da semana e termos o número de colunas igual aos dias da semana com dados do tipo *dummy*, porque variáveis do tipo *dummy* são mais adequadas ao nosso modelo, por terem apenas duas possibilidades, como as autoras (Silva & Ribeiro, 2018) mencionaram. A *feature* "occurrences per day" corresponde à quantidade de vezes cujo propósito de cada viagem corresponda a casa, ou seja, à atividade "Home", e dado que o *dataset* corresponde a mais do que uma semana fez-se uma média dessa ocorrência por dia da semana. Estas últimas duas *features* foram calculadas através da linguagem Python e os seus procedimentos podem ser visualizados no Anexo B: Recolha de *features* de *clustering*. Para os valores NULL apresentados no *dataset* foram substituídos, segundo (Alok Gupta, 2015), pelo cálculo da mediana através do seguinte *script* em Python integrado no anexo anteriormente referenciado:

Algoritmo: Cálculo da Mediana

```
Input: lista lst vazia, variável i com a entrada 1
lst = [], i = 1
while len(occurrences_per_day) > i:
    lst.append(occurrences_per_day[i]) (adicionar os valores da feature
    "ocurrences_per_day" à lista lst)
    i += 2
end
print(lst, "\n")
print("Mediana:", statistics.median(lst)) (imprimir o cálculo da mediana para
adicionar à nossa coluna e preencher os valores NULL)
```

Após todos estes procedimentos obteve-se por completo o nosso *dataset*, Figura 31, com todas as *features* referenciadas nesta secção.

user_id	duracao_hours	dist_metros_to_work_university	start_time_hours_day	walk_percentage_stop	dist_metros_to_home	grupo_idade	mean_duration	escolaridade	occurrences_per_day	percent_week_day_sun	percent_week_day_mon	percent_week_day_tue
102	1.13	44575.81	11.05	-0.23	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	2.78	44569.09	15.33	0.51	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	5.24	44568.65	11.4	0.48	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	19.13	44568.76	15.17	0.32	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	41.07	44575.93	7.88	0.27	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	46.45	44567.12	14.97	0.03	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	0.1	44566.2	12.62	-0.31	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06
102	0.16	3474.99	12.18	-1.0	0	Between 18 and 21 years old	0.92	Bachelor	1	10.94	12.5	14.06

percent_week_day_mon	percent_week_day_tue	percent_week_day_wed	percent_week_day_thu	percent_week_day_fri	percent_week_day_sat	occurrences_per_day	standart_deviation_duration	escolaridade	estado_civil	day_week
12.5	14.06	18.75	7.81	28.13	7.81	1	1.82	Bachelor	Single (Célibataire)	Mon
12.5	14.06	18.75	7.81	28.13	7.81	1	1.82	Bachelor	Single (Célibataire)	Thu
12.5	14.06	18.75	7.81	28.13	7.81	2	1.82	Bachelor	Single (Célibataire)	Thu
12.5	14.06	18.75	7.81	28.13	7.81	2	1.82	Bachelor	Single (Célibataire)	Thu
12.5	14.06	18.75	7.81	28.13	7.81	5	1.82	Bachelor	Single (Célibataire)	Wed
12.5	14.06	18.75	7.81	28.13	7.81	1	1.82	Bachelor	Single (Célibataire)	Tue
12.5	14.06	18.75	7.81	28.13	7.81	2	1.82	Bachelor	Single (Célibataire)	Tue
12.5	14.06	18.75	7.81	28.13	7.81	4	1.82	Bachelor	Single (Célibataire)	Fri

Figura 31 - Dataset com todas as features possíveis de serem obtidas

Finalmente, pode-se aplicar o algoritmo de aprendizagem computacional RF, apresentado no Anexo C: Algoritmo *Random Forest*, para determinar a previsão da inferência do propósito de viagem. Caso esta previsão inicial for de baixo valor será necessário adotar novas diligências de melhorias no modelo, ou seja, no acréscimo de novas *features* ao *dataset*.

Na Figura 30 e Figura 31, pode-se visualizar a falta de algumas colunas, "start_location_latitude" e "start_location_longitude", "user_id" e "date_year", porque não serão usados no nosso estudo, uma vez que não correspondem a *features* e sim a dados de identificação de localização da localidade no QGIS (latitude, longitude), do utilizador (user ID) e da data de ocorrência da paragem não utilizada por (Montini et al., 2014).

3.3 Métodos e critérios

Com o conjunto de dados disponibilizado, baseado em dados de GPS, e com o estudo do estado da arte começa-se a pensar num modelo que nos permita comprovar de forma eficiente que é possível identificar com elevada precisão a inferência do propósito de viagem. Para desenvolver este modelo, decidimos percorrer os trabalhos relatados sobre o modelo *Random Forest* que tomamos como referência (Breiman, 2001) e (Montini et al., 2014), combinando e adaptando os métodos mais adequados. Esta escolha foi feita, porque este modelo integra um algoritmo que é mais adequado e mais fácil que os outros algoritmos identificados na secção 2.5, ou seja, requer menos pré-processamento, poderá ocorrer a falta de alguns valores nulos ou mesmo a não existência de *features* previamente consideradas no artigo original e o processo de treino é mais simples para determinar o propósito de viagem.

Após analisar o conjunto de dados na secção 3.1, com base nos artigos (Garnett & Stewart, 2015) e (Gao et al., 2021) é aplicado a função "TrajectoryStopDetector" da biblioteca *MovingPandas* (Anita Graser, 2019) que fornece estruturas e funções de dados de trajetória para exploração e análise de dados de movimento e, assim recolher todas as paragens de viagens e permitir identificar o local de paragem transversalmente da latitude e longitude. Posteriormente, através de *queries* SQL para os diferentes raios (6.5, 15, 25, 35 e 50 metros) identifica-se a relação de ambiguidade das paragens com o número de POIs. Este critério foi tido em consideração porque, segundo o que foi mencionado na secção 2.2.3, os *smartphones* com GPS quando fazem o registo da localização de utilizadores apresentam erros relativos à sua posição real. Extraindo estas atividades e as suas localizações é necessário retificar se para cada paragem efetuada por utilizador, com a finalidade de realizar uma atividade, corresponde a um determinado ponto de local de interesse para selecionar o raio que permita obter a menor ambiguidade sem comprometer o baixo número de paragens realizadas pelos diferentes utilizadores.

Algumas das *features* da Figura 13 foram possíveis recolher de forma direta do *dataset* Breadcrumbs (dados pessoais) e calcular através de *querys* SQL ("duration", "distance home" e "distance work"). Na obtenção das outras *features* de *clustering*, que se encontram identificadas a cor vermelha na Figura 13, foi necessário recorrer ao algoritmo *Hierarchical Agglomerative Clustering* (HAC), mencionado pelos autores (Montini et al., 2014). Após realizar o *clustering* de dados e visualizar-se através de dendrogramas, para decidir o número de *clusters* que cada *cluster* de paragens por utilizador irá alocar, é necessário prosseguir com os critérios destes autores, ou seja, para cada utilizador no *dataset* efetuar os seguintes procedimentos:

1. Agrupar todas as atividades disponíveis (*clusters*)
2. Contar quantos *clusters* contêm a atividade do tipo "Home" (POI)
3. Atribuir o *cluster* com mais atividades do tipo "Home" como "Main Cluster"

Após a realização destes procedimentos irá ser calculado todas as restantes *features* de *cluster* denominado como "Main Cluster" para cada utilizador. Desta forma, consegue-se recolher e obter todas as *features* da Figura 13, exceto "Transport mode after activity" e "Transport mode before activity", pelo simples facto de existir apenas dados superficiais dos modos de transporte, ou seja, hábitos relativos aos modos de transporte usados pelos diferentes utilizadores e não o tipo de transporte usados em cada viagem. Com todas estas etapas realizadas, tendo em consideração todos os métodos e critérios referenciados no estado de arte e no artigo (Montini et al., 2014), obtém-se um *dataset* não balanceado, ou seja, há uma grande incidência de determinadas categorias dos POIs dentro do *dataset* em comparação com outras. Para não carretar problemas na construção do modelo RF em que o algoritmo não diferenciará as classes minoritárias das demais categorias, classes maioritárias, utilizam-se métricas de avaliação. Essas métricas de avaliação são aplicadas através das funções: Cálculo de Coeficiente de Correlação de Matthew's (MCC), Cálculo de Pontuação F1 (*F1-Score/F1-Measure*), Pontuação de Classificação de Precisão (ACC) e *Recall* da biblioteca *scikit-learn,* para avaliar o modelo de aprendizagem computacional RF. Dado que estas métricas são quantitativas devemos confiar nelas pois podem avaliar o nosso modelo (métricas mais simples, como precisão, que não levam em consideração dados não balanceados) e ver quais classes do modelo estão a criar conflitos entre elas e corrigir essas lacunas. Para as métricas foi considerado *average=weighted* em vez de *average=macro* porque, segundo (Bex T., 2021), *average=weighted* contabiliza a existência de dados não balanceados (desequilíbrio), das classes calculadas, multiplicando cada *score* pelo número de ocorrências de cada classe, dividindo pelo número total de amostras, enquanto *average=macro* retorna a média sem considerar a proporção de cada classe no conjunto de dados.

Para realizar a visualização de todas as paragens e locais visitados do nosso *dataset* são criadas as colunas de geometria em cada tabela com o *Spatial Reference System ID* (SRID) 4326. Um SRID é um sistema de coordenadas. Como o QGIS tem capacidade de armazenar os seus dados num único sistema de coordenadas, como *World Geodetic System* (WGS) 84, é usado o SRID 4326 que faz a combinação das *features* latitude com a longitude para a obtenção de um ponto projetado no mapa *OpenStreetMap*. Assim pode-se visualizar a correspondência das viagens aos pontos de interesse.

Com os métodos e critérios aplicados ao *dataset* Breadcrumbs aplicar-se-á o algoritmo RF para classificar o propósito de viagem e será avaliada a previsão desta determinação.

3.4 Algoritmos

Os autores (Montini et al., 2014) ao desenvolverem o seu estudo e pesquisa na classificação do modo de viagem com base no modelo RF de (Breiman, 2001), algoritmo de aprendizagem

computacional, subsequentemente usufruíram das *features* apresentadas na Figura 13 e dos resultados obtidos do modo de transporte para determinarem o propósito de viagem. Para determinar o propósito de viagem é então necessário obter as *features* a partir da recolha, seleção e manipulação do *dataset* Breadcrumbs e aplicá-las neste algoritmo RF (Datamart, 2019), utilizando num primeiro teste para obter uma previsão da inferência do propósito.

Inicialmente procede-se à identificação das paragens do *dataset* Breadcrumbs através da implementação do algoritmo de deteção de paragens de trajetórias de movimento. A obtenção destas paragens tem como objetivo permitir identificar o *ground truth* da atividade numa determinada área demarcada por um raio predefinido. A escolha destes raios, 6.5, 15, 25, 35 e 50 metros, deriva dos resultados da relação da ambiguidade com a quantidade de paragens que será apresentada e argumentada no capítulo 4. A partir deste algoritmo de deteção de paragens e após a escolha do raio de área abrangida por uma determinada atividade é possível obter as *features* "Duration", "Distance to work", "Distance to home", "Start time", "Age", "Education level", "Income", "Marital status" e "Days of week". Para as *features* de *clustering*, "mean duration", "percentage week days", "occurrences per day" e "standard deviation duration", foi necessário recorrer ao algoritmo o *Hierarchical Agglomerative Clustering* (HAC). Posteriormente, à recolha de todas as *features* é introduzido a implementação do algoritmo RF ao novo *dataset* para inferir num primeiro teste, o propósito de viagem. Assim, nesta secção será apresentado os algoritmos de deteção de paragens de trajetórias de movimento, HAC e o RF, algoritmos que fazem parte do nosso modelo.

3.4.1 Deteção de paragens de trajetórias de movimento

O algoritmo de deteção de paragens de trajetórias de movimento, seguindo o exemplo de (Anita Graser, 2022), foi implementado para detetar as diferentes paragens realizadas pelos diferentes utilizadores do *dataset* Breadcrumbs através da função "TrajectoryStopDetector" da biblioteca *MovingPandas*. A função "TrajectoryStopDetector" permite criar instâncias do movimento que permaneçam dentro de uma determinada área, definido por um determinado raio, por um determinado período de tempo. Os segmentos de paragens resultantes incluem informações espaciais e temporais sobre a localização e a duração da paragem, que correspondem à nossa *feature* "Duration". A duração definida para cada paragem foi de 180 segundos (*min_duration=timedelta(seconds=180)*) por ser um tempo suficientemente considerável para a realização de uma determinada atividade, seguindo o que foi descrito na secção 2.3.1. Para mais informações sobre a implementação do algoritmo pode ser observado no Anexo D: Algoritmo de deteção de paragens.

3.4.2 Agrupamento Hierárquico Aglomerativo

O Agrupamento Hierárquico Aglomerativo / *Hierarchical Agglomerative Clustering* (HAC) ao ser aplicado irá agrupar cada instância de paragem do nosso *dataset* em *clusters,* ou seja, constrói uma hierarquia de agrupamentos, conforme descrito na secção 2.6.1. É um método de baixo para cima em que começa em um *cluster* separado e os pares de *clusters* são mesclados à medida que se sobe na hierarquia. O número de *clusters* que são considerados no agrupamento é selecionado através da triagem feita aos dendogramas, (Saul Dobilas, 2021). Um dendrograma é um tipo específico de diagrama ou representação icónica que organiza determinados fatores e variáveis através de

análises estatísticas, ramificando de forma hierárquica permitindo a visualização do número de *clusters* agrupados. O número de *clusters* é definido através da visualização das diferentes cores de *clusters* dos diagramas e dos diferentes tipos de *linkages*, critérios de ligação. Na Figura 32 podemos observar duas cores diferentes de *clusters*, verde e vermelho, em ambos os tipos de gráficos, com diferentes tipos de *linkages* - *single*, *complete* e *average*, respetivamente - e dessa forma definimos no algoritmo, Anexo E: Algoritmo de Agrupamento Hierárquico Aglomerativo, o número de *clusters* igual a dois (*n_clusters*=2).

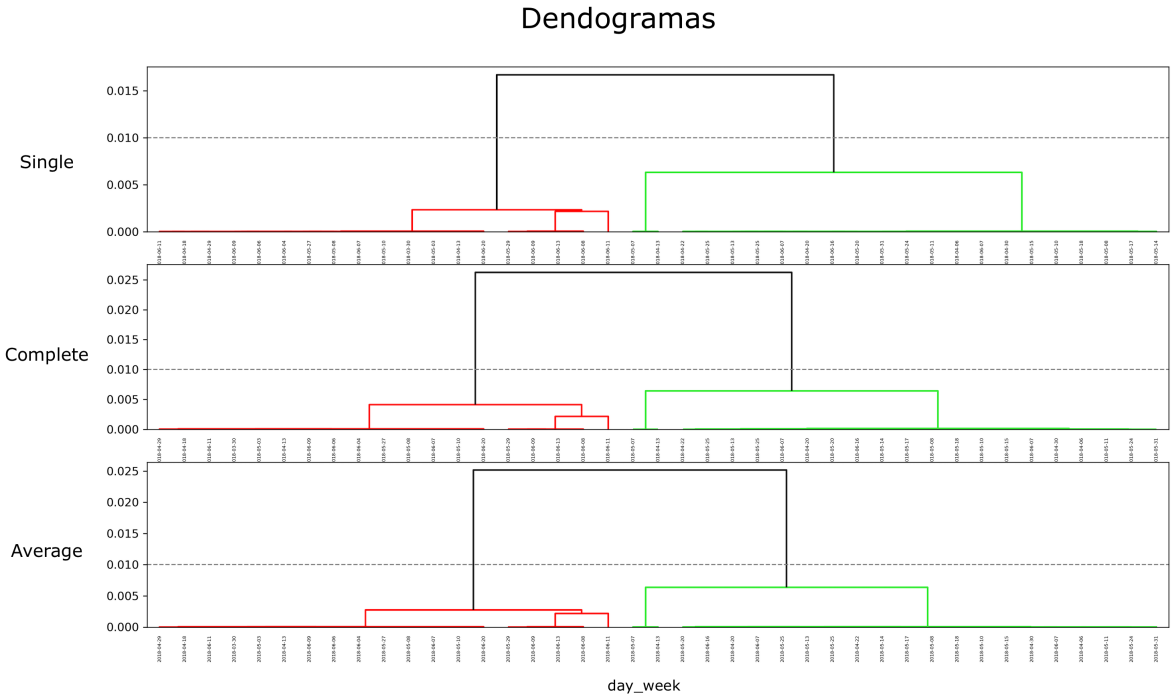


Figura 32 - Dendograma do utilizador 831 para diferentes linkages

Em cada etapa, o par de *clusters* mais próximos é mesclado em um novo *cluster*. Os *clusters* são mesclados desde que estejam mais próximos, através da identificação das distâncias (eixo do y) isto é, da *feature* "day_week" que representa o dia da paragem, poderia ser outra *feature* mas esta é a mais discriminativa para o gráfico dendograma, e através da latitude e longitude de cada paragem (eixo do x), a partir do registo de pontos de GPS. Se dois *clusters* consistem em vários locais de atividade, a distância entre os *clusters* não é direta. Assim, os *linkages*, que foram anteriormente mencionados, apresentam:

a) *Single-Linkage*: distância entre os pontos mais próximos possíveis de dois *clusters*
b) *Complete-Linkage*: distância máxima entre dois pontos de dois *clusters*
c) *Average-Linkage*: média de todas as distâncias de todos os pares de pontos possíveis de dois *clusters*

Para determinar o *cluster* ótimo, ou seja, as *features* de *clustering*, foi considerado o local de atividade "Home" como o *cluster* principal, "Main Cluster", citado no artigo (Montini et al., 2014), estando ao nosso alcance a obtenção da maioria das *features* apresentadas na Figura 13.

3.4.3 Algoritmo *Random Forest*

No novo *dataset*, após a realização da preparação dos dados, obtivemos mais colunas que variáveis, porque existem *features* que foram separadas em várias colunas para converter em amostras do tipo *dummy,* através da função *.get_dummies().* Este processo ocorreu na *feature* "percentagem week days" separando esta coluna em várias amostras para cada dia da semana, isoladamente.

Transversalmente às *features* recolhidas na preparação do *dataset*, coloca-se elas como variáveis de entrada para o algoritmo de RF (*max_features*), resultando como variável de saída na previsão da inferência do propósito de viagem. Estes hiperparâmetros do RF, o *max_features* e *n_estimators,* são configurações que não podem ser aprendidas a partir de dados regulares que se fornecem ao algoritmo, pois são embutidos no algoritmo e cada algoritmo tem o seu próprio conjunto pré-definido de hiperparâmetros. Os hiperparâmetros geralmente são ajustados para aumentar a precisão do modelo (Saurabh Gupta, 2021). No algoritmo RF irá ser definido o número de árvores de decisão igual a 500 (*n_estimators=500*), baseado e utilizado no artigo dos autores (Montini et al., 2014), que poderá variar de acordo com a previsão obtida, influenciando no impacto do seu aumentando ou na sua diminuição. Este hiperparâmetro foi definido como 500, para estar de acordo com o estado de arte e de acordo com (Montini et al., 2014). Como se pode verificar, todas estas variáveis poderão ser alternadas, com a adição ou exclusão delas, dependendo sempre dos resultados obtidos. Para isso foi realizado o ajuste dos hiperparâmetros do RF, ou seja, o número de árvores de decisão (*n_estimators*) e o número de *features* (*max_features*), sempre com o intuito de melhorar o desempenho do modelo, seguindo os exemplos dos autores (Piotr Płoński, 2019) e (Jason Brownlee, 2020a), respetivamente. Desta forma, há uma preparação de dados feito pelo nosso modelo para além da sua implementação.

Inicialmente, no algoritmo RF começa-se por selecionar as variáveis de entrada e saída mencionadas na secção 3.2.1, Figura 31, onde foi excluído as colunas do *dataset* que não entram para o estudo, através da função *.drop()*, obtendo duas variáveis "X" e "y", sendo o "X" o conjunto de variáveis de entrada e "y" a variável de saída, variável prevista. O processo de aprendizagem divide os dados em conjuntos de teste e de treino. Durante o treino, o modelo vê as respostas, ou seja, a nossa variável "y" é obtida a partir da combinação das variáveis do conjunto "X", para poder aprender a forma de prever o propósito de viagem. No entanto, como em todos os processos de aprendizagem, será estudado o conjunto de teste onde o algoritmo só tem acesso às *features* de previsão e não propriamente à sua resposta (Silva & Ribeiro, 2018). Esta divisão de teste e treino, foi feita numa proporção de 25% de teste e 75% de treino, feita de forma *default* pela função *train_test_split()* da biblioteca "Scikit-Learn" de (David Cournapeau, 2022), uma vez que não se definiu o parâmetro *test_size,* mas apenas o parâmetro *stratify=y.* O parâmetro *stratify* é responsável por fazer as divisões estratificadas dos conjuntos de dados, adequado para conjunto de dados não balanceados (Mirko Stojiljković, 2021). Após a divisão do nosso *dataset* em teste e treino, pode-se aplicar a função "RandomForestClassifier" da mesma biblioteca, através da seguinte variável:

```
forest=RandomForestClassifier(n_estimators = 500, max_features = 33, bootstrap = True)
```

A biblioteca "Scikit-Learn" é a que acompanha sempre na utilização das diversas funções de consecução de resultados de previsão neste estudo. O número de *features* (*max_features=33*) igualou-se a 33 *features*, maior que as 13 *features* das 17 da Figura 13 adquiridas a partir do *dataset* Breadcrumbs na secção 3.2, pelo simples facto de ser realizado a divisão das *features* em várias amostras com o intuito de as tornar em *features* do tipo *dummy,* como mencionado na secção 3.2.1. O hiperparâmetro *boostrap=True* permite melhorar o desempenho do modelo, porque diminui a variância do modelo sem aumentar o viés, semelhante ao processo *pruning* utilizado pelo modelo

das árvores de decisão. Isso significa que, embora as previsões de uma única árvore sejam altamente sensíveis ao ruído, a média de muitas árvores não o é, contanto que as árvores não sejam correlacionadas, ou seja, que não estabeleçam relações entre elas. Simplesmente treinar muitas árvores num único conjunto de treino forneceria árvores fortemente correlacionadas. Assim, o *bootstrap* é uma forma de descorrelacionar as árvores do modelo RF, mostrando diferentes conjuntos de treino. Apesar das funcionalidades deste hiperparâmetro, foram descartadas as atividades realizadas abaixo dos 1% pelos utilizadores, como apresentado na Tabela VI, porque a representatividade destas classes é muito baixa, não havendo um número suficiente de instâncias de uma dada categoria para treinar o modelo, como por exemplo a classe "Car Sharing". Depois destes processos é construída o modelo RF a partir do conjunto de treino pela função *forest.fit(X_train, y_train).* Agora será necessário prever a classe "X", que consiste naquela com maior probabilidade média estimada entre as árvores de decisão incluídas no nosso algoritmo, através da seguinte variável:

```
y_pred = forest.predict(X_test)
```

Neste momento estamos aptos para aplicar as diferentes métricas de previsão ao nosso algoritmo. Estas métricas consistem na previsão da precisão e do *recall*, ou da sua média harmónica *F1-Score*, apresentadas pela função *classification_report()*, e do Coeficiente de Correlação de *Matthews,* função *matthews_corrcoef()*. Existem muitas mais métricas, porém a mais adequada ao nosso *dataset* é a métrica de *Matthews* (MCC) por ser a melhor métrica para modelos treinados com dados não balanceados, como mencionado pelo autor (Igor Kuznetsov, 2019), e as restantes duas, precisão e *recall*, para comparar com os resultados obtidos por (Montini et al., 2014). De seguida, visualizou-se a matriz de confusão pela função *confusion_matrix()*, que consiste numa tabela de dados de desempenho do nosso algoritmo de classificação. Toda esta representação foi feita através dos seguintes *prints*:

```
print("MCC: ", matthews_corrcoef(y_test, y_pred))
print(classification_report(y_test, y_pred, zero_division=1))
print(“"Confusion Matrix: ", confusion_matrix(y_test, y_pred))
```

Com isto, será apresentado o processo de cálculo feito pelas diferentes métricas de classificação multiclasse:

- *Precisão*

Num problema multiclasse existem amostras positivas e negativas. Quando uma amostra positiva é falsamente classificada como negativa, chamamos isso de Falso Negativo (FN) e da mesma forma, quando uma amostra negativa é falsamente classificada como positiva, é chamada de Falso Positivo (FP) (Boaz Shmueli, 2019). Ainda assim, existem os Falsos Positivos (FP) que indicam que a amostra falsa é positivamente classificada e Verdadeiros Positivos (TP) onde as amostras corresponde ao esperado.

A precisão é calculada dividindo os TP pela soma do TP e FP. De acordo com a Figura 33, a otimização do nosso modelo é feita à medida que diminuímos o número de FP.

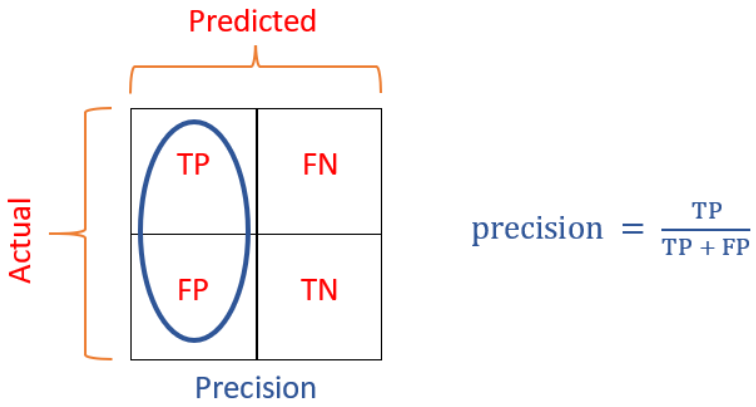


Figura 33 - Classificação multiclasse para a precisão (Bex T., 2021)

– *Recall*

O *recall* é calculado dividindo o número de TP pela soma de TP e FN. De acordo com a Figura 34, a otimização do nosso modelo para a métrica *recall* consiste na diminuição do número de FN.

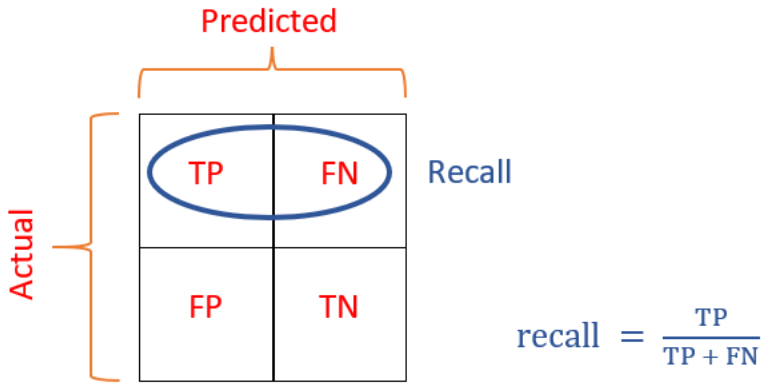


Figura 34 - Classificação multiclasse para o recall (Bex T., 2021)

– *Coeficiente de Correlação de Matthews*

O MCC é uma medida de qualidade que traduz a correlação entre os valores verdadeiros e os previstos que pode ser usada mesmo que as classes possuem tamanhos bastante diferentes. E como uma correlação pode variar entre (-1) e 1, em que um coeficiente de 1 representa uma predição perfeita, 0 representa uma predição aleatória média e (-1) uma predição inversa. O grande benefício desta métrica é fornecer uma medida balanceada para o modelo e exprimir a sua qualidade, independentemente das diferenças quantitativas entre as duas classes.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Figura 35 - Classificação multiclasse para o MCC (Bex T., 2021)

Neste sentido, quando o *recall* é alto, mas a precisão é baixa significa que a maioria dos positivos é reconhecida, mas existem muitos FP. Por sua vez quando o *recall* é baixo, mas a precisão é alta significa que o modelo erra muitos positivos, alto FN.

Segundo (Bex T., 2021) o MCC é a melhor métrica de *score* para estabelecer o desempenho de um classificador no contexto da matriz de confusão. Todo este procedimento de implementação deste algoritmo poderá ser visualizado através do Anexo C: Algoritmo *Random Forest*.

3.5 Foursquare

No âmbito de recorrer à adição de novas *features* ao nosso *dataset*, distintas às que (Montini et al., 2014) empregou, para além das *features* "genero", "working_profile" recolhidas diretamente do *dataset* Breadcrumbs, e "walk_percentage_trip" pelo cálculo da velocidade média das viagens realizadas, recorreu-se ao *dataset* Foursquare (Dingqi YANG, 2019) para realizar essa recolha. Essa recolha e adição de novas *features* consistiu na identificação dos POIs mais recentes do Foursquare num raio de 500 metros dos pontos de atividades aos pontos de locais de interesse do *dataset* Breadcrumbs, método já usado para determinar o melhor raio na identificação da atividade do POI, que posteriormente, é categorizado pela taxonomia/categoria hierárquica definida pelo (Foursquare, 2021). As taxonomias genéricas consideradas correspondem:

- *Arts and Entertainment*
- *Business and Professional Services*
- *Community and Government*
- *Dining and Drinking*
- *Event*
- *Health and Medicine*
- *Landmarks and Outdoors*
- *Retail*
- *Sports and Recreation*
- *Travel and Transportation*

O raio de 500 metros foi considerado, porque essa distância foi usada pelos autores (Lu et al., 2013) na identificação de POIs com propósitos similares aos que o *dataset* Breadcrumbs apresenta, conforme descrito no estado da arte na secção 2.2.2, sendo necessário converter esse valor de metros em graus para obtermos a geometria do ponto, feito através do ponto sete do Anexo H: Queries SQL usadas na recolha de *features*, e assim visualizamos no QGIS o local desse POI categorizado. Sabendo que 111 quilómetros correspondem a 1 grau na linha do equador (MullOverThings, 2020), 500 metros correspondem a 0.0045 graus na região de estudo, averiguando através de uma localização dos pontos de locais de interesse do *dataset* Breadcrumbs se realmente essa adição à latitude ou longitude corresponde realmente a uma distância adicional de 500 metros. Esta averiguação foi feita com o prosseguimento do ponto 8 do Anexo H: Queries SQL usadas na recolha de *features*, verificando que a distância da latitude (46.512019) e longitude (6.618055) à latitude (46.516519) e longitude (6.618055), onde esta última aditou os 0.0045 graus em relação à latitude, resultou em 497.622 metros, aproximadamente os 500 metros. Desta forma, confirmou-se a boa conversão dos metros em graus, uma vez que existe sempre um erro associado à sua conversão.

Para determinar os POIs do Foursquare que estavam num raio de 500 metros, usou-se o mesmo método usado na determinação da ambiguidade, ou seja, função ST_DWithin() do primeiro ponto do Anexo H: Queries SQL usadas na recolha de *features*. Alguns desses POIs não estavam incluídos na taxonomia/categorização hierárquica do (Foursquare, 2021), sendo necessário adicionar novas

categorizações em falta nessa taxonomia, através do ficheiro .json de (Peter Hohnson, 2015). Desta forma, uma outra categoria genérica, "Residence" é acrescentada às nossas categorias anteriores.

Os procedimentos que foram executados para adicionar as novas colunas de categorização ao nosso *dataset* correspondem à importação de todos os POIs mais recentes do Foursquare, criação de uma coluna geométrica da tabela de POIs, fazer a seleção dos POIs que estão à volta das paragens num raio de 500 metros, importar as tabelas de taxonomias/categorias do Foursquare e gerar a categorizarão do Foursquare em falta. Nestes poucos POIs identificar as suas categorias genéricas e contabilizar para cada paragem o número de categorias genéricas, anteriormente identificadas, que estão dentro do raio de 500 metros. Desta forma, é gerado as novas *features* dessas categorias genéricas onde as categorias que não apresentem nenhuma contagem igualam-se a zero. Todo este procedimento pode ser seguido e visualizado através das *queries* desenvolvidas e apresentadas no Anexo G: Integração dos POIs do Foursquare, categorização genérica e contabilização.

3.6 Eliminação de *features*

Para tornar o modelo mais preditivo foi removida do algoritmo RF as *features* calculadas após a chegada no destino, ou seja, *features* recolhidas após e durante a atividade, com o objetivo de tentar melhorar a sua performance. Essas *features* correspondem "duracao_hours", "mean_duration" e "standard_deviation_duration" da Figura 13. A *feature* "duracao_hours" corresponde ao tempo que a atividade decorreu, "mean_duration" à média que o utilizador perde na realização daquele tipo de atividade num determinado *cluster*, de acordo com os *clusters* obtidos da aplicação do HAC, e "standard_deviation_duration" ao desvio padrão da duração da atividade resultante e calculado a partir dos mesmos *clusters*.

3.7 Arquitetura

O roteiro técnico que define as fases práticas que foram seguidos ao longo do desenvolvimento deste trabalho poderá ser visualizado na Figura 36. Inicialmente recolheu-se o *dataset* fornecido pelos autores do *dataset* Breadcrumbs. De seguida, escolheu-se todos os dados necessários para a criação das *features* utilizadas pelo algoritmo *Random Forest*, de acordo com o estado da arte, exportando os dados para ficheiros .csv, após a realização da importação dos dados para a máquina pessoal, para serem mais fáceis de serem trabalhados com o auxílio da linguagem de programação *Python* junto do compilador *Jupyter Notebook*.

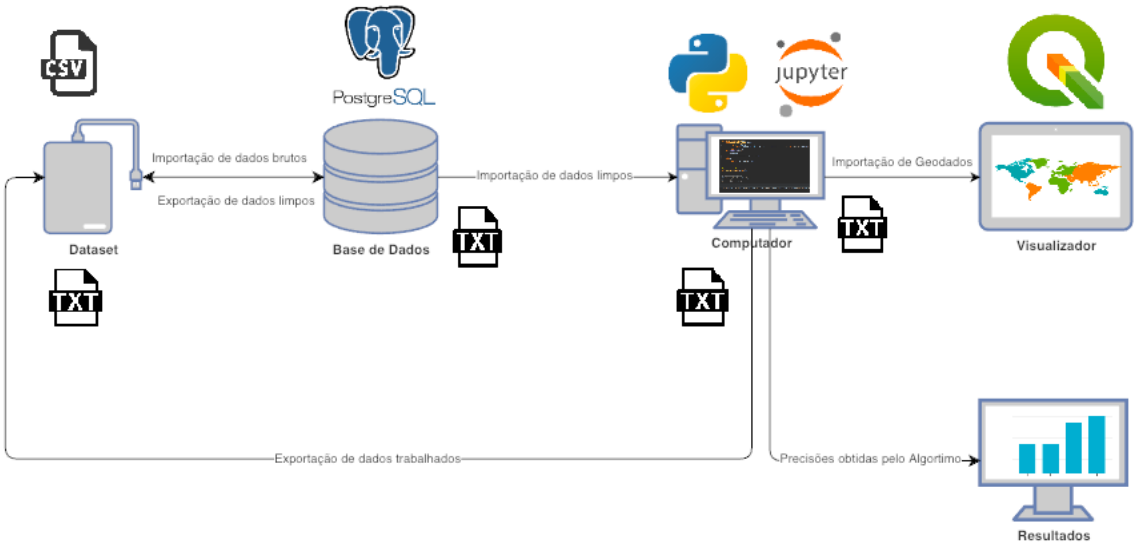


Figura 36 - Diagrama técnico de desenvolvimento na inferência do propósito de viagem

Após os dados das viagens serem trabalhados e selecionados são utilizado os atributos de latitude e longitude para visualizar todos os pontos das viagens no QGIS e assim tirar conclusões relativas aos trajetos efetuados pelos diferentes utilizadores. Após visualizar todas as viagens, com um ponto de início e um ponto de fim (POI) são recolhidas todas as *features* dessas viagens, selecionadas com base na Figura 13, para dar como *input* as *features* (variáveis independentes) e a inferência de propósito como *output* (variável dependente) no algoritmo *Random Forest*, objetivo deste trabalho. Posteriormente, são também trabalhadas e selecionadas novas *features* que influenciam na precisão do nosso modelo. Desta forma, obtemos as previsões da inferência de propósito de viagem para realizar a validação e avaliação desses resultados.

Capítulo 4 Resultados e discussão

Após a seleção do conjunto de dados, da aplicação dos critérios e dos algoritmos para obtermos um novo *dataset* com as *features* da Figura 13, as experiências foram feitas em amostras com várias viagens realizadas por diferentes utilizadores a diferentes pontos de interesse (POIs). As amostras correspondem às viagens realizadas num máximo de 78 utilizadores dos 80 existentes no *dataset* Breadcrumbs, neste caso para um raio de 6.5 metros de distância de erro registados pelo GPS, onde o ponto central corresponde à paragem efetuada por cada utilizador existindo sempre a possibilidade de uma paragem efetuada por um utilizador corresponder a mais que um POI. Neste sentido, foi necessário verificar os resultados de ambiguidade obtidos pelos diferentes raios, 6.5, 15, 25, 35 e 50 metros, para demonstrar que a seleção feita do raio de 6.5 metros corresponde à melhor relação de ambiguidade e maior número de paragens. Após a tomada de decisão da seleção do raio de erro, relativo ao ponto de registo efetuado pelo GPS ao longo da viagem e da atividade, considerou-se apenas 1% das atividades realizadas por esses utilizadores, excluindo as atividades com valores percentuais inferiores e ainda as atividades que não apresentassem de forma explícita o tipo de atividade realizada. Com isto, aplicou-se o algoritmo RF seguindo a secção 3.4.3 e verificaram-se os valores de precisão e *recall* para comparar os resultados obtidos no artigo de (Montini et al., 2014). Ainda foi utilizado a métrica de *Mattews's Correlation Coefficiente* (MCC) por ser extremamente boa para a classificação em conjuntos de dados não balanceados e para conjuntos de classes com diferentes tamanhos (Igor Kuznetsov, 2019), estando estas características incluídas no nosso *dataset* em estudo. Para melhorar a nossa precisão inicial inclui-se mais *features* ao *dataset* e dados de POIs do Foursquare.

4.1 Ambiguidade

Na Tabela V pode-se ver as médias de POIs visitados pelos entrevistados a um só ponto de local de interesse obtido pelos diferentes raios, bem como o número de paragens realizadas e respetivos utilizadores. Claramente que obtivemos poucas viagens para um raio de 6.5 metros, mas a média de visitas a um único POI aproxima-se mais do valor 1, querendo dizer que a ambiguidade é mais baixa para esse raio. Contudo, poder-se-ia continuar a baixar o raio de erro de GPS, mas isso iria impactar no número de paragens do nosso *dataset* e comprometer demais o nosso estudo, limitando a um número pequeno de instâncias e, por essa razão, consideramos um raio de 6.5 metros o raio de erro mais adequado, condizendo com o estado de arte.

Tabela V - Média e quantidade de POIs visitados durante o período de 3 meses da recolha do dataset Breadcrumbs

Raio (metros)	Número de paragens	Número de utilizadores	Média de visitas a um único POI
6.5	3386	78	1.17
15	12152	79	1.24
25	21727	79	1.27
35	30707	79	1.27
50	43802	79	1.28

A Figura 37 e Figura 38 a seguir apresentam todos os pontos de locais de interesse dos eventos/destinos de paragens realizadas pelos vários entrevistados na área de cobertura definida pelos raios indicados na Tabela V, onde o ponto central corresponde às paragens.

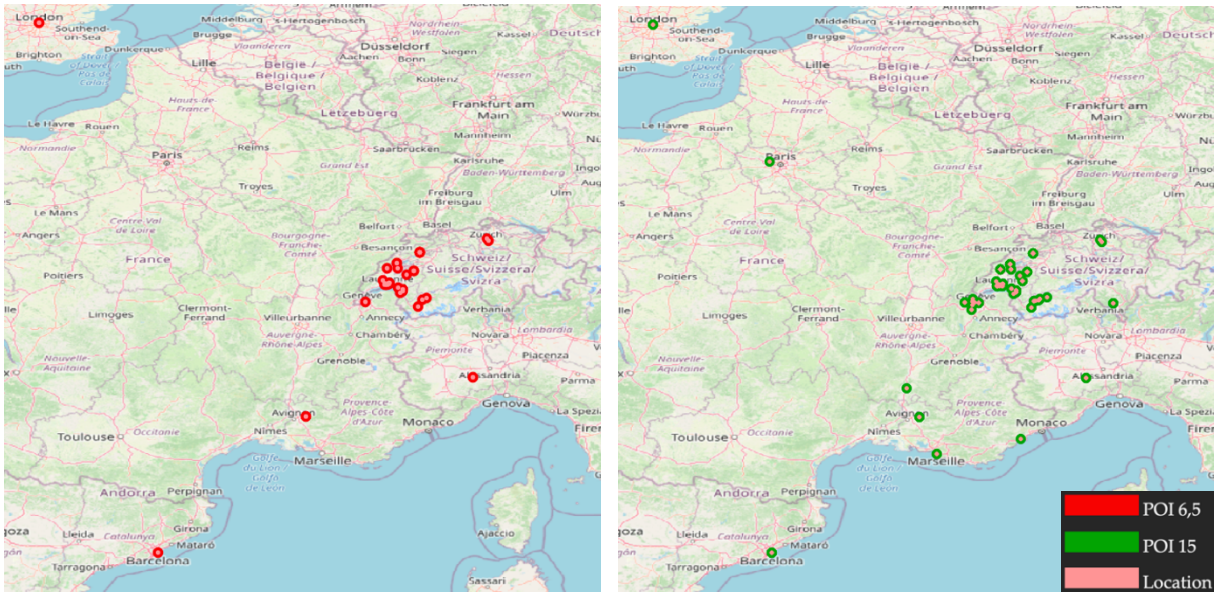


Figura 37 - Mapa à esquerda com a identificação de fim de viagem num raio de 6.5 metros e mapa à direita num raio de 15 metros

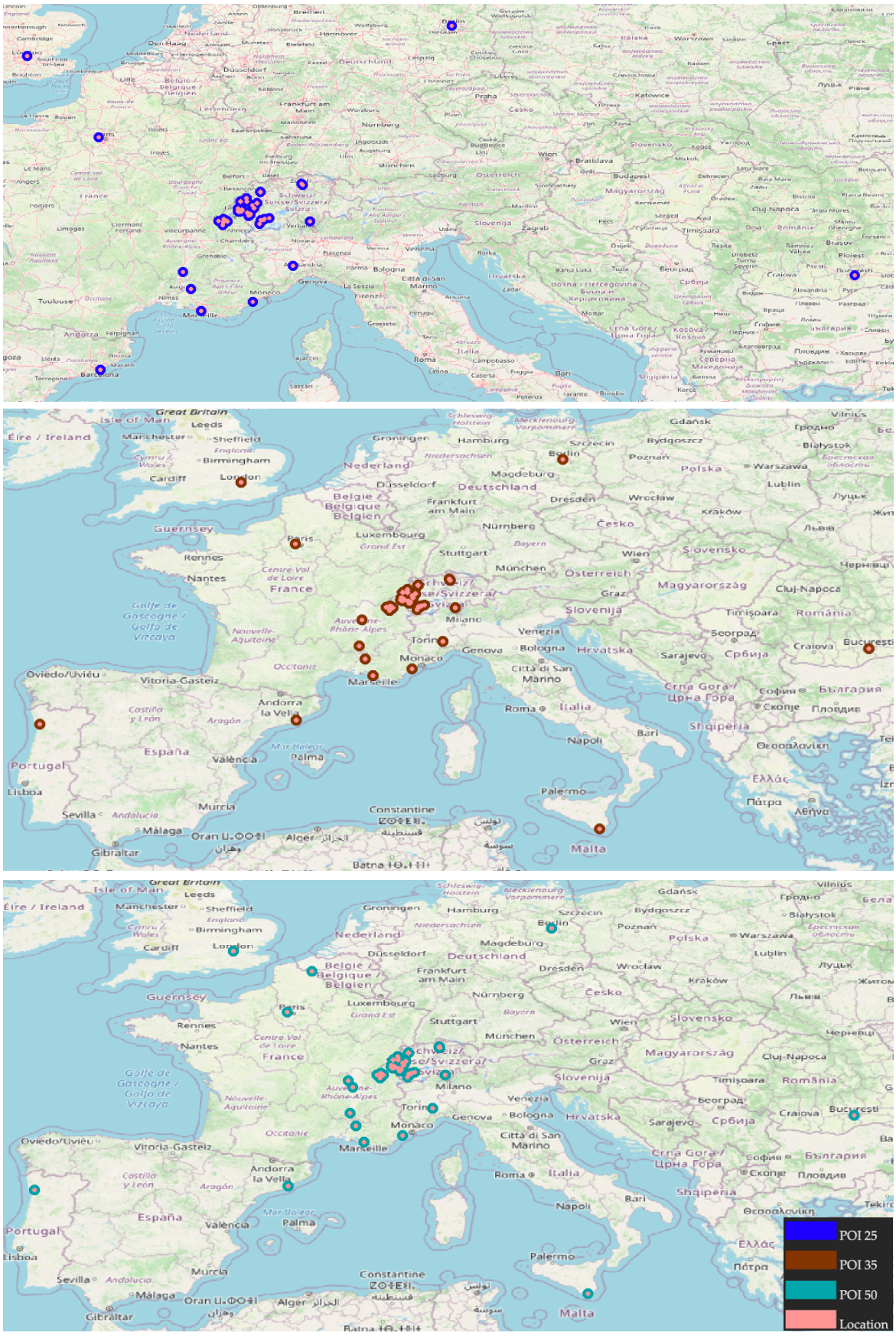


Figura 38 - Primeiro mapa com a identificação de fim de viagem num raio de 25 metros, mapa ao centro com um raio de 35 metros e último mapa num raio de 50 metros

Ao realizar a preparação do *dataset*, secção 3.2, para recolher as *features* da Figura 13 do *dataset* Breadcrumbs, obtendo todos os pontos visitados pelos entrevistados nos diferentes raios definidos, foram criadas as colunas de geometria referenciadas na secção 3.3 para ligar a base de dados espacial PostgreSQL com a ferramenta de visualização QGIS, para assim ter uma melhor perspetiva dos resultados obtidos, como se pode visualizar na Figura 37 e Figura 38. Grande parte desses pontos de locais de interesse concentram-se na Suíça, na cidade de Lausanne. Resultaram ainda alguns pontos de locais de interesse dispersos em outros locais, fora deste país, devido ao facto desses utilizadores apresentarem duas residências, uma que corresponde à casa dos pais fora da cidade e outra que corresponde ao local onde o utilizador trabalha, para além de realizarem outras atividades fora desse país. Desta forma, é possível determinar as distâncias entre os locais visitados e a sua própria residência, porque as atividades realizadas nos destinos de viagens do *dataset* Breadcrumbs encontram-se identificadas, considerando sempre a distância à residência mais próxima. A partir destas figuras, Figura 37 e Figura 38, e com a Tabela V existe uma grande variação do número de paragens quando aumentamos esse raio de erro.

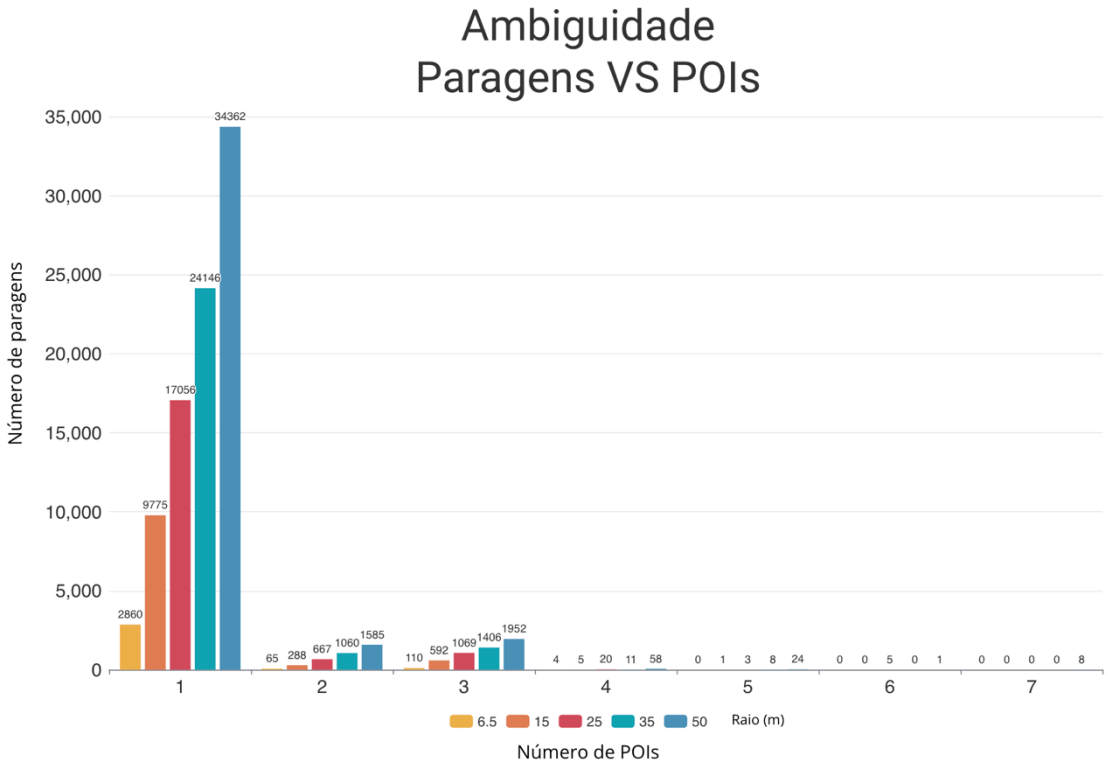


Figura 39 - Ambiguidade de paragens de viagens

Através da Figura 39, mais uma vez, confirma-se que o raio de 6.5 metros para a maioria das paragens tem apenas um ponto de local de interesse associado a cada paragem, relativamente aos outros raios, o que permite associar à paragem da viagem uma atividade mais fidedigna que determinado utilizador efetuou num determinado *timestamp*. Quanto maior o raio, maior é o número de pontos de locais de interesse com paragens associadas como já foi mencionado, porém, compromete o *ground truth* desses pontos.

4.2 Eliminação de dados

Alguns dos utilizadores não apresentavam a atividade "Home", ou seja, não especificavam a sua residência, causando a incapacidade do cálculo da *feature* "Distance to home". Este pequeno grupo

de utilizadores, 174, 321, 328 e 381, foram excluídos do *dataset* fazendo com que o *dataset* inclua apenas 74 utilizadores, havendo a diminuição do número de paragens de 3386 para 3366. O modelo *Random Forest* desconsidera os conjuntos muito pequenos de paragens em comparação com os restantes conjuntos de paragens de atividades maioritários, dando mais relevância aos conjuntos de classes maiores. Por essa razão, foi considerado a realização das paragens com mais de 1% neste estudo que se pode visualizar através dos tipos de atividades na cor verde da Tabela VI.

Tabela VI - Seleção das atividades com mais de 1% de paragens

	Raio				
POI	**6.5**	**15**	**25**	**35**	**50**
Home	2074	6657	10779	14291	18973
Friend's Place	197	578	938	1215	1606
Sports	148	419	710	1028	1397
University	219	1363	2976	4782	7904
Metro Stop	98	293	742	1117	1701
Family	144	411	738	953	1190
Bus Stop	73	232	403	570	827
Shopping	111	641	1080	1530	2171
Work	73	312	565	860	1262
Restaurant	90	571	963	1365	1983
Other	62	168	351	519	788
Train Station	27	170	564	928	1483
Bar	12	78	244	479	803
Hotel	8	37	65	92	121
Library	17	120	363	555	897
Club	7	28	99	187	305
Park	6	63	117	175	273
Beach	0	2	4	6	13
Airport	0	1	3	6	13
Association	0	2	7	14	25
Parking	0	1	3	4	13
Car Sharing	0	1	1	1	1
Velo Station	0	4	12	16	22

Total	3366	12152	21727	30707	43802
> 1%	33.66	121.52	217.27	307.07	438.02

Sendo assim, as paragens com as atividades mais relevantes correspondem a "Home", "Friend's Place", "Sports", "University", "Metro Stop", "Family", "Bus Stop", "Shopping", "Work" e "Restaurant" para todos os raios, sendo a atividade "Train Station" menos relevante apenas para o raio de 6.5 metros e "Bar" e "Library" para o raio de 6.5 e 15 metros, isto tendo em consideração as atividades realizadas com mais de 1% de paragens. A atividade "Other", mesmo não estando abaixo desse percentual foi desconsiderada pelo simples facto de não designar no *dataset* Breadcrumbs qualquer tipo de informação relativo a este ponto de local de interesse.

Figura 40 – Total de paragens por pontos de locais de interesse > 1%

Assim, obtivemos um *dataset* com 3127 paragens das 3366 paragens inicias. Observa-se na Figura 40 que no caso de considerarmos outro raio seria incluso mais atividades no nosso *dataset* e a exclusão de atividades diminuiria na Tabela VI.

Alguns utilizadores neste *dataset* não incluíam informações relativas ao local de trabalho e local de estudo, pontos de locais de interesse "Work" e "University", respetivamente. Como estes dados são necessários para obtermos a *feature* "Distance to work" da Figura 13, onde se considerou a atividade trabalho para ambos os pontos de locais de interesse "Work" e "University", seguindo a mesma ideia de (Montini et al., 2014), reduziu-se o número de utilizadores de 74 para 56 e o número de paragens de 3127 para 2365. Esta diminuição foi considerável, mas o número de instâncias ainda se manteve grande para realizarmos um bom estudo no nosso modelo.

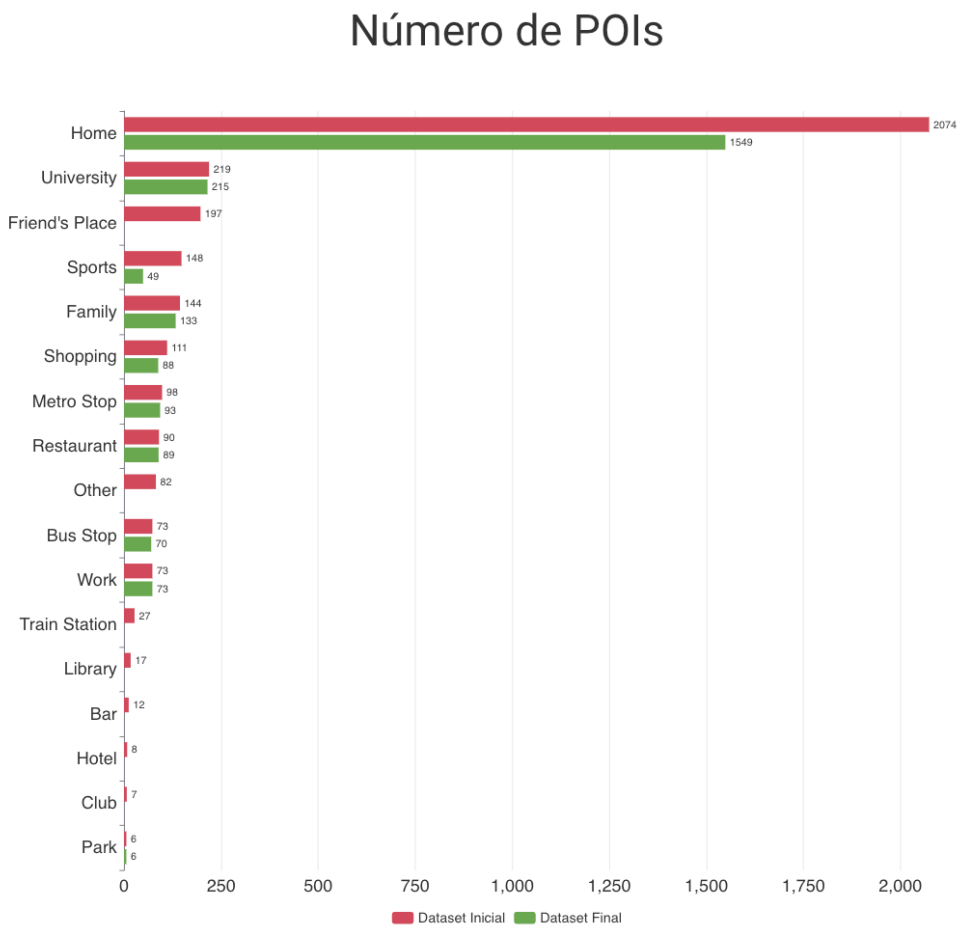


Figura 41 - Total de pontos de locais de interesse discriminados

Na Figura 41 podemos ver os POIs do *dataset* inicial com as 3386 paragens, total de paragens recolhidas pelo algoritmo de deteção de paragens de trajetórias de movimento da secção 3.4.1, e as 2365 paragens do *dataset* final após a exclusão dos utilizadores que não possuíam os dados necessários para a recolha das *features* da Figura 13. Os autores (Montini et al., 2014) integraram o perfil de trabalhador dos utilizadores trabalhadores com os utilizadores estudantes, ou seja, a atividade de ambas as categorias "Work" e "University" foram associadas numa só atividade "Work" com 288 pontos de locais de interesse.

4.3 POIs Foursquare

No decorrer da preparação do nosso *dataset* foi relevante fazer a adição de novas *features* que conseguissem identificar cada paragem com um propósito mais genérico. Este processo teve como intuito de possibilitar o aumento da previsão do nosso modelo e para isso foi necessário seguir os procedimentos da secção 3.5. Após obtermos a geometria de todos os POIs do Foursquare fez-se a sua visualização no QGIs e, posteriormente, a comparação com os POIs do *dataset* Breadcrumbs, Figura 42.

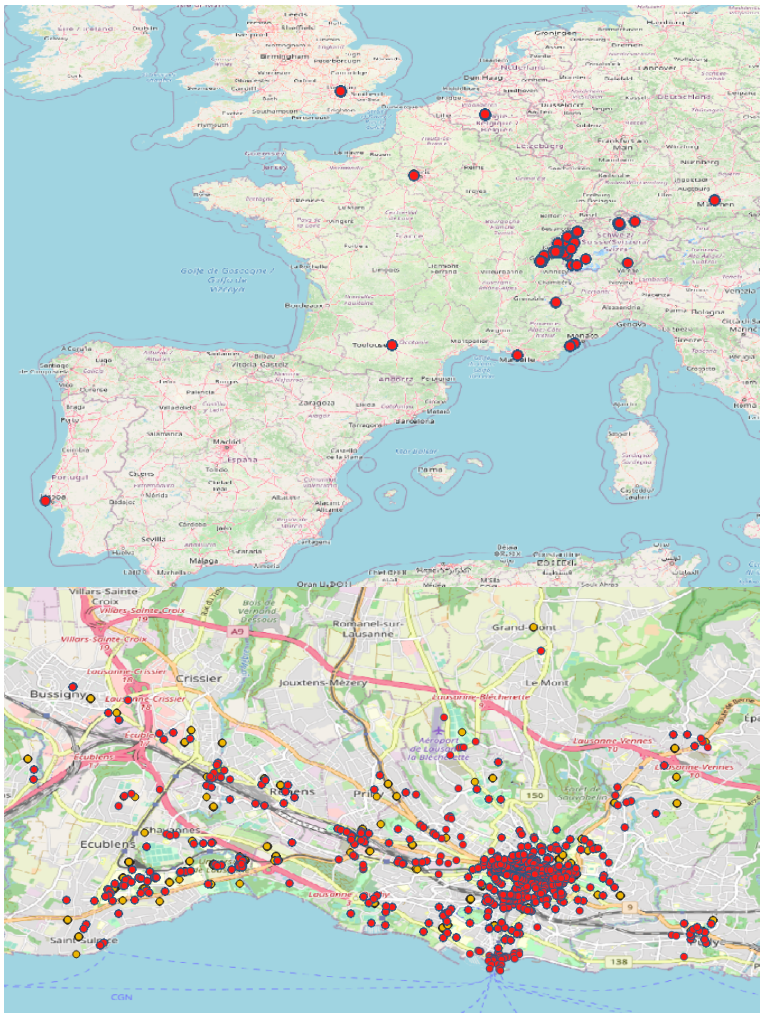


Figura 42 - POIs Foursquare a vermelho e POIs dataset Breadcrumbs a amarelo

Com base nos autores (Yazdizadeh et al., 2019), foram considerados os POIs do *dataset* Foursquare dos países Portugal, França, Alemanha, Itália e Londres para além da Suíça. Após a visualização no QGIS de todos os POIs do Foursquare recolhidos num raio de 500 metros de cada local de destino vistiado do *dataset* Breadcrumbs retificamos através de um *buffer*, Figura 43, se realmente esses POIs se encontram dentro da área demarcada.

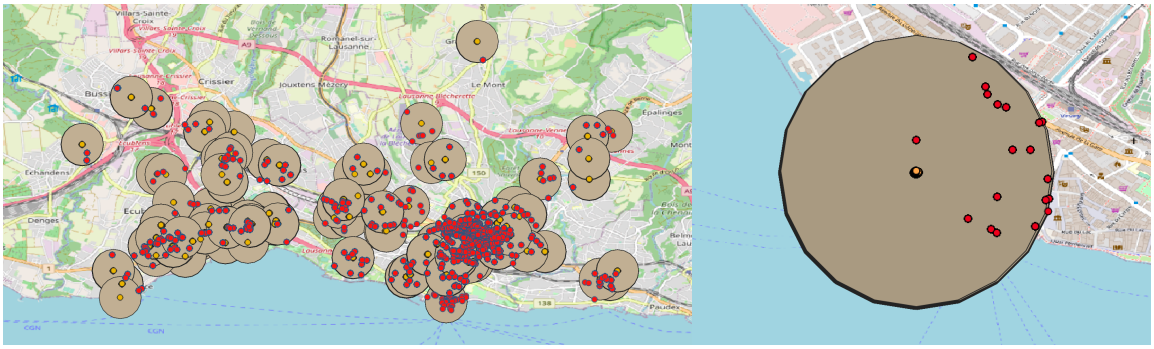


Figura 43 - POIs Foursquare num raio de 500 metros dos destinos do dataset Breadcrumbs

Desta forma, como se pode visualizar na Figura 43, figura à direita na cidade de *Vevey*, Suíça, no colégio *l'Aviron*, ponto de local de interesse representado por um ponto central a laranja como atividade "Home", contém num raio de 500 metros 1340 POIs do Foursquare, pontos a vermelho. Através desses 1340 POIs e através da categorização/taxonomia mais genérica, apresentadas na

secção 3.5, fez-se a identificação no formato *dummy* da contabilização dessas categorias genéricas, criando as *features* adicionais para cada taxonomia no nosso *dataset*, Figura 44. Todo este processo está incluso no Anexo G: Integração dos POIs do Foursquare, categorização genérica e contabilização, contabilizando o número de todas as categorias genéricas a que pertencem uma determinada paragem.

arts_and_entertainment	business_and_professional_services	community_and_government	dining_and_drinking	evento
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	6	1	1	0

health_and_medicine	landmarks_and_outdoors	residence	retail	sports_and_recreation	travel_and_transportation
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	1	0	1	1	4

Figura 44 - Features da contagem da taxonomia/categoria do Forsquare de cada paragem

A disponibilidade destes dados de redes sociais baseadas na localização ou LBSN permitem melhorar o conhecimento do comportamento das atividades significativas dos utilizadores e assim, fazer com que o nosso modelo determine o propósito de viagem com maior capacidade.

4.4 Previsão

Inicialmente calculamos a previsão da precisão e *recall* através do modelo de RF, Anexo C: Algoritmo *Random Forest*, para as 13 *features* conseguidas das 17 de (Montini et al., 2014). Essas *features* correspondem às *features* de atividade: duração da atividade ("Duration"), distância até ao local de trabalho/universidade ("Distance to work"), considerando a educação como trabalho, percentagem de caminhadas ("Walk percentage"), distância até ao local de residência ("Distance to home"), início da atividade ("Start time") e dia da semana da atividade ("Days of week"); *features* pessoais: idade ("Age"), escolaridade ("Education level") e estado civil ("Marital status"); *features* de *clustering:* média do conjunto de dados relativos à duração da atividade ("mean duration"), percentagem dos dias da semana realizadas pela atividade ("percentage week days"), número de ocorrências por dia ("occurrences per day") e desvio padrão da duração da viagem ("standard deviation duration"). Destas 17 *features* não se conseguiram recolher ou determinar as *features* relativas ao modo de transporte ("Transport mode after activity" e "Transport mode before activity") e a *feature* "Income", porque o *dataset* Breadcrumbs não inclui dados suficientes para realizar a resolução dos modos de transporte e nenhuma informação da renda paga pelos utilizadores. A outra *feature* corresponde a "Overall cluster: number of person" e não foi considerada, porque consiste na identificação do número de pessoas que conhecem uma localização. Como este recurso deve ser tratado com cuidado porque é dependente do conjunto de dados, ou seja, para um grupo muito homogêneo que trabalha na mesma universidade, o trabalho

é um lugar que todos conhecem ou para um conjunto de dados mais diversificado de uma região, as estações de metro são mais propensas a serem conhecidas por várias pessoas, esta identificação de conhecimento dos diferentes pontos de locais de interesse pelos utilizados não está discriminada no *dataset* disponibilizado, impossibilitando a sua integração no modelo.

Numa primeira versão utilizamos essas 13 *features*, Tabela VII, seguidamente numa outra versão sem as *features* após e durante a realização da atividade, ou seja, sem as *features* "Duration", "mean duration" e "standard_deviation_duration", Tabela VIII. Finalmente, numa última versão utilizou-se as *features* da versão anterior mais as *features* adicionais "genero", "working_profile", "walk_percentage_trip", e taxonomia de POIs do Foursquare, Tabela IX. A *feature* "genero" corresponde ao género do utilizador, "working_profile" indica se o utilizador é trabalhador, trabalhador/estudante ou estudante, "walk_percentage_trip" corresponde à percentagem de viagens realizadas a pé pelo utilizador e *features* de taxonomia de POIs do Foursquare correspondentes à Figura 44. Em todas as versões foram consideradas as médias ponderes, ou seja, *average=weighted* onde a média considera a proporção de cada classe no conjunto de dados adequados a um *dataset* não balanceado.

Desta forma, conseguimos visualizar a nossa primeira versão, resultado de uma previsão padrão, Tabela VII.

Tabela VII - Confusion Matrix: Random Forest com 500 árvores (n_estimators=500) e 13 features (max_features=33)

	Prediction									
Activity (truth)	1	3	4	7	8	9	10	13	16	Recall (%)
1	388	0	0	0	0	0	0	0	0	100
3	0	8	0	2	1	0	0	1	0	67
4	0	0	2	0	0	0	0	0	0	100
7	0	0	0	4	14	0	0	3	1	18
8	0	1	0	8	9	1	3	0	0	41
9	0	1	0	0	0	4	12	1	0	22
10	0	0	0	1	0	13	5	4	0	22
13	0	0	0	9	0	0	5	58	0	81
16	0	0	0	0	1	0	0	0	32	97
Precision (%)	100	80	100	17	36	22	20	87	97	

1 – Home; 3 – Sports; 4 – Park; 7 – Restaurant; 8 – Shopping; 9 – Bus Stop; 10 – Metro Stop; 13 – Work/University; 16 – Family

Nota: A média ponderada (*average=weighted*) de precisão e de *recall* para 100 testes é de 86% e de MCC é de 75%.

	Prediction								
Activity (truth)	Mode Transfer	Being Home	Work–Education	Shopping–Service	Recreation	Pickup–Drop-Off	Business	Other	Recall (%)
Mode transfer	490	1	1	2	0	1	0	0	99.0
Being home	5	374	4	1	2	1	0	0	96.6
Work–education	8	5	177	6	8	0	1	0	86.3
Shopping–service	13	3	3	124	19	2	1	2	74.3
Recreation	10	11	7	25	115	0	0	1	68.0
Pickup–drop-off	6	3	1	14	4	19	0	1	39.6
Business	7	2	11	9	9	0	19	0	33.3
Other	4	1	0	19	13	1	0	7	20.0
Precision (%)	90.6	93.3	85.1	62.6	69.3	82.6	82.6	69.2	

Figura 45 - Resultados obtidos para 500 árvores de decisão (Montini et al., 2014)

A média de previsão obtida pelos autores (Montini et al., 2014) em 100 testes foi de 84.4% e neste primeiro estudo foi de 86%, ou seja, superior ao resultado espectável. Porém, apenas existem três atividades análogas ao nosso *dataset*, "Home", "Work/University" e "Shopping" correspondentes a "Being Home", "Work-Education" e "Shopping-Service", respetivamente. Sendo que apenas a atividade "Shopping", isoladamente apresentou uma previsão mais baixa que o estudo de (Montini et al., 2014).

Este bom resultado geral da média de previsão resultou do tratamento de dados das *features*, por haver o cuidado de respeitar todos os problemas que o nosso *dataset* apresentou, como ambiguidade, desequilíbrio entre classes e conversão do tipo de classes para variáveis que fossem mais favoráveis na aplicação do modelo *Random Forest.* Este bom resultado também advém da média entre classes e da integração de *features* recolhidas durante e após a realização das atividades. Por esse motivo, após o estudo dos hiperparâmetros realizou-se outro teste de previsão para o *dataset* sem essas *features,* para que o nosso modelo esteja apto a determinar o propósito sem qualquer informação relativa à atividade. Estes resultados podem ser visualizados na Tabela VIII.

Dado que os hiperparâmetros podem influenciar na previsão do nosso modelo, uma vez que foram considerados inicialmente os valores padrão, numa primeira análise houve o estudo do número de *features* (*max_features*) a considerar no modelo, segundo (Jason Brownlee, 2020a), com o intuito de tornar o nosso modelo mais rentável e eficiente quando aplicado.

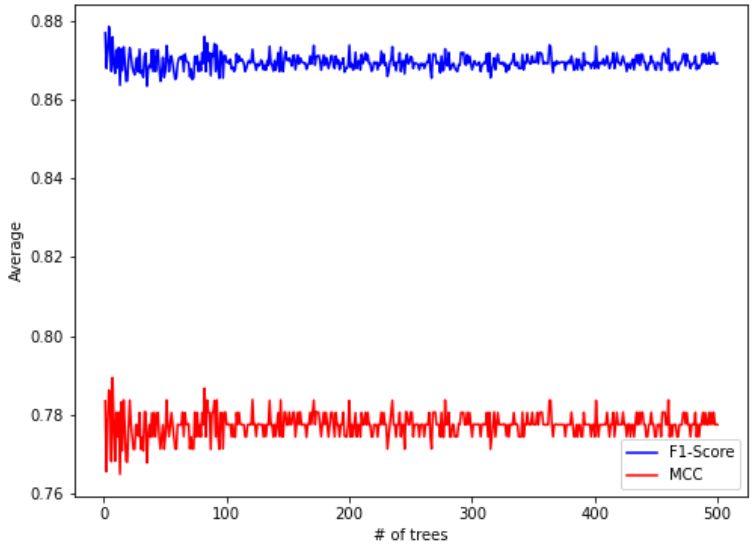


Figura 46 - Oscilações de previsão para a métrica MCC e F1-Score

A partir da Figura 46 podemos ver que independente do número de árvores de decisão, existem pequenas oscilações num intervalo de 0 e 1%, ocorrendo maiores oscilações até às 100 árvores de decisão, sendo menor esta oscilação para a métrica MCC do que para a métrica *F1-Score*, que envolve a métrica de precisão e *recall*. O MCC tem um valor médio mais baixo dado que a mesma é mais adequada e precisa ao nosso modelo, porque temos um *dataset* não balanceado como já foi referenciado na secção 3.4.3. Desta forma, conclui-se que a partir das 100 árvores de decisão não será relevante fazer experiências para o nosso *dataset*, não impactando na previsão de forma negativa ou positiva. O outro hiperparâmetro estudado, como mencionado mais uma vez na secção 3.4.3, corresponde ao número de *features* máximas (*max_features*), idealizando a sua análise de acordo com o autor (Piotr Płoński, 2019).

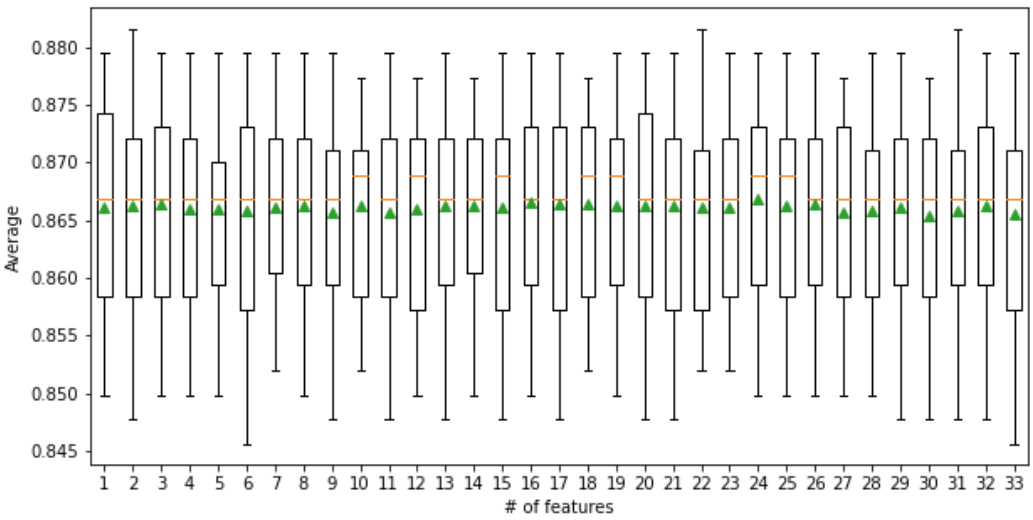


Figura 47 - Boxplot da influência do max_features na previsão

Após verificarmos, a partir do *boxplot* da Figura 47, o número de *features* não será relevante no nosso modelo, porque a sua influência é pouco relevante para a previsão, considerando desta forma o número total de *features* do nosso *dataset*. Este estudo dos hiperparâmetros já foi considerado na obtenção da previsão obtida na Tabela VIII.

Tabela VIII - Confusion Matrix: Random Forest com 100 árvores (n_estimators=100) e 9 features (max_features=29)

	Prediction									
Activity (truth)	1	3	4	7	8	9	10	13	16	Recall (%)
1	388	0	0	0	0	0	0	0	0	100
3	0	11	0	0	0	0	0	1	0	92
4	0	0	0	0	0	1	0	1	0	0
7	0	1	0	1	12	0	0	8	0	5
8	0	0	0	16	2	1	0	2	1	9
9	0	0	0	0	0	3	14	1	0	17
10	0	0	0	2	0	7	9	5	0	39
13	0	0	0	2	0	0	4	66	0	92
16	0	1	0	0	0	0	0	0	32	97
Precision (%)	100	85	100	5	14	25	33	79	97	

1 – Home; 3 – Sports; 4 – Park; 7 – Restaurant; 8 – Shopping; 9 – Bus Stop; 10 – Metro Stop; 13 – Work/University; 16 – Family

Nota: A média ponderada (*average=weighted*) de precisão e de *recall* é de 87% e de MCC uma média de 75%.

Após desconsiderarmos as *features* recolhidas durante e após a realização das atividades e estudar os valores a considerar nos hiperparâmetros a nossa média de precisão e *recall* não se alterou significativamente, contudo, a previsão individual no geral diminui para grande parte das atividades. Algumas das atividades como "Restaurant", "Shopping", "Bus Stop" e "Metro Stop" apresentam uma precisão e *recall* muito baixo em ambos os casos, Tabela VII como na Tabela VIII, havendo a necessidade de recorrer à adição de novas *features* de categorização genérica do Foursquare, Figura 44, para impactar no aumento dessa previsão, seguindo os procedimentos da secção 3.5, em sincronia com a adição das *features* já recolhidas "genero", "working_profile" e "walk_percentage_trip".

Tabela IX - Confusion Matrix: Random Forest com 100 árvores (n_estimators=100) e 27 features (max_features=51)

	Prediction									
Activity (truth)	1	3	4	7	8	9	10	13	16	Recall (%)
1	388	0	0	0	0	0	0	0	0	100
3	0	9	0	0	0	0	0	1	2	75
4	0	0	1	0	0	0	1	0	0	50
7	0	0	0	3	13	0	1	4	1	14
8	0	0	0	11	6	0	1	0	4	27
9	0	0	0	0	0	6	12	0	0	33
10	0	0	0	1	1	13	7	0	1	30
13	0	0	0	4	0	0	8	60	0	83
16	0	1	0	0	0	0	0	0	33	100
Precision (%)	100	100	100	16	30	32	23	92	80	

1 – Home; 3 – Sports; 4 – Park; 7 – Restaurant; 8 – Shopping; 9 – Bus Stop; 10 – Metro Stop; 13 – Work/University; 16 – Family

Nota: A média ponderada (*average=weighted*) de precisão e de *recall* é de 87% e de MCC uma média de 75%.

Na Tabela IX, a previsão individual de grande parte das atividades aumentou consideravelmente como "Park", "Restaurant", "Shopping" e "Bus stop", onde inevitavelmente "Metro Stop" e "Wrok/University" diminuíram.

Como em alguns dos casos os resultados obtidos de cada atividade individual ainda não corresponderam ao expectável, ou seja, estiveram abaixo de uma previsão de 50% e como a atividade "Shopping" pode incluir a atividade "Restaurant" e as atividades "Bus Stop" e "Metro Stop" podem estar associadas na mesma atividade a paragens de transporte, fez-se a junção destas atividades numa só. Desta forma, a atividade 7 e 8 passa a ser a atividade 9, "Shopping/Restaurant", e a atividade 9 e 10 como "Metro/Bus Stop", atividade 10.

Tabela X - Confusion Matrix: Random Forest com 100 árvores (n_estimators=100), 27 features (max_features=51) e união da atividade 7/8 e 9/10

	Prediction							
Activity (truth)	1	3	4	8	10	13	16	Recall (%)
1	388	0	0	0	0	0	0	100
3	0	9	0	0	0	2	1	75
4	0	0	2	0	0	0	0	100
8	0	0	0	33	1	10	0	75
10	0	0	0	3	35	3	0	85
13	0	3	0	8	2	59	0	82
16	0	0	0	0	0	0	33	100
Precision (%)	100	75	100	75	92	80	97	

1 – Home; 3 – Sports; 4 – Park; 8 – Shopping/Restaurant; 10 – Metro/Bus Stop; 13 – Work/University; 16 – Family

Nota: A média ponderada (*average=weighted*) de precisão e de *recall* é de 94% e de MCC uma média de 90%.

Com isto, o nosso modelo RF conseguiu obter boas prestações para estes tipos de atividades, como se pode observar na Tabela X, alcançando os objetivos pretendidos para o estudo em causa. No entanto, para as diferentes tabelas, apercebemos que existem valores foram da linha diagonal da matriz de confusão (*confusion matrix*) e isto se deve à imperfeição do nosso modelo. Apesar de obtermos uma média de previsão, *recall* e MCC consideravelmente boa, o modelo ainda confunde algumas classes com outras, mesmo depois de agregar a atividade "Shopping" com "Restaurant" e "Metro Stop" com "Bus Stop". Contudo, como sabemos, em todos os modelos existem sempre melhorias a realizar o aumento da previsão da identificação das diferentes atividades, e isto poderá resultar na adição de novas *features* mais específicas a cada atividade no nosso *dataset*, individualmente, para aumentar a previsão da inferência do propósito do nosso modelo. Estas etapas terão como objetivos futuros na secção 6.3.

Capítulo 5 Planeamento

Este projeto foi planeado desde o início do primeiro semestre pela equipa AmILab, no qual me integro. A equipa realiza reuniões semanais onde todos os membros, incluindo os orientadores, discutem o progresso que foi realizado durante essa semana, quais as metas concluídas, dificuldades e quais os próximos objetivos a concluir. Qualquer tipo de questão discutida que resultasse em dúvida, estabeleceu-se contacto com os autores do *dataset* Breadcrumbs, (Moro et al., 2019), para prosseguir com o decorrer do trabalho e mitigar os riscos que foram encontrados, retirando partido dos esclarecimentos ajudando na melhoria da inferência de propósito de viagem. No segundo semestre procedeu-se à recolha e desenvolvimento de *features*, de acordo com o estado de arte, para que o modelo tenha uma boa previsão. Mais de metade deste trabalho esteve em volta da manipulação e recolha de dados, porque é necessário que o *dataset* tenha as características necessárias na aplicabilidade da inferência do propósito de viagem. Desta forma, neste capítulo é apresentado o cronograma e riscos associados ao desenvolver deste trabalho.

5.1 Cronograma

O projeto foi dividido em várias fases, sendo a primeira integrada na pesquisa do estado da arte, realizada no primeiro semestre, e no segundo semestre resultou na recolha, processamento dos dados, treino e teste do classificador, e finalmente, aplicação do classificador treinado para determinar a previsão da inferência do propósito de viagem. Durante a pesquisa, começou-se por analisar os dados do *dataset* para entender as suas características e de seguida processá-las, seguidamente da sua recolha, do processamento e extração de novas *features* dos dados do *dataset,* disponibilizado por (Moro et al., 2019), para ser adquirido um *dataset* com as condições necessárias para a aprendizagem do nosso modelo de classificação.

5.1.1 Primeiro semestre

Na Figura 48 é apresentado o gráfico com o cronograma GANTT para o início do primeiro semestre. O objetivo consiste em iniciar pesquisas de artigos relevantes onde os autores apresentam abordagens e metodologias adequadas para determinar diversos modos ou propósitos de viagens realizados por um grupo de utilizadores, numa determinada região.

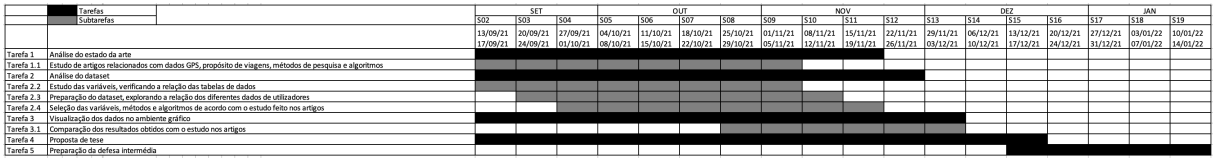


Figura 48 - Gráfico de GANTT do primeiro semestre com as tarefas previstas

Uma vez que tínhamos o *dataset* Breadcrumbs, começou-se por realizar a análise do conjunto de dados e metodologias de experimentos estudados no estado da arte. Também se decidiu que se deveria dividir o desenvolvimento da proposta de tese em quatro etapas: na primeira etapa documentar o estado da arte, analisar o *dataset* fornecido, visualizar os dados do *dataset* já

minimamente trabalhados no QGIS e, posteriormente, preparar a proposta da tese, feita ao longo de todo o semestre, finalizando com a preparação da defesa intermédia.

No entanto, ao longo do semestre tivemos que mudar o percurso do nosso cronograma, Figura 49. Estas mudanças foram efetuadas, porque foi necessário realizar pesquisas mais exaustivas de artigos relacionados com os dados de GPS, propósitos de viagens e métodos usados para estes fins. Após esse estudo, iniciou-se o estudo e análise do *dataset* Breadcrumbs sendo necessário fazer a escolha e recolha das *features* de acordo com o estudo anteriormente fundamentado. Ainda foi necessário recorrer ao desenvolvimento de *queries* à base de dados espacial onde estava importado o *dataset*, para posterior processamento.

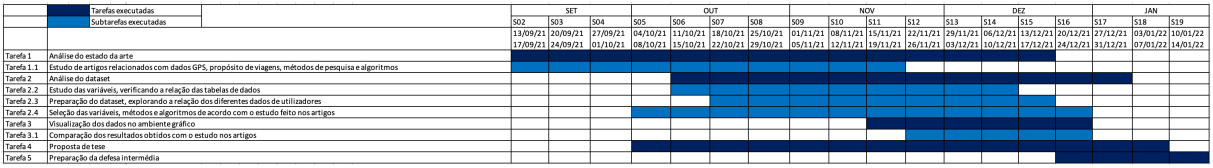


Figura 49 - Gráfico de GANTT do primeiro semestre com as tarefas executadas

Nesta primeira fase não foi possível determinar todas as *features* possíveis da Figura 13, para aplicar no modelo e obter uma previsão inicial para a inferência do propósito de viagem.

5.1.2 Segundo semestre

O planeamento do segundo semestre é apresentado no cronograma GANTT da Figura 50. Levando em consideração com a contínua obtenção das *features* selecionadas no primeiro semestre. Após a obtenção destas *features*, dá-se prosseguimento ao desenvolvimento de *queries* para realizarem a determinação e recolha de algumas *features* que não podem ser recolhidas diretamente do *dataset* Breadcrumbs e que serão necessárias para o algoritmo *Random Forest*, seguindo o estudo de (Montini et al., 2014), para determinar a previsão do propósito de viagem. Posteriormente, de acordo com os resultados obtidos, calcular uma previsão próxima aos resultados obtidos por estes autores. No entanto, ao longo deste processo, poderá ser necessário aprimorar o estado da arte, caso algumas das *features* não sejam possíveis de obter. Essa melhoria dependerá também de um estudo mais exaustivo do estado da arte, porque apesar de se ter escolhido as *features* e os critérios do capítulo 3, não significa que não seja necessário realizar algumas alterações. Todo este desenvolvimento será um processo que poderá sofrer algum tipo de evolução, com o objetivo de melhorar os seus resultados.

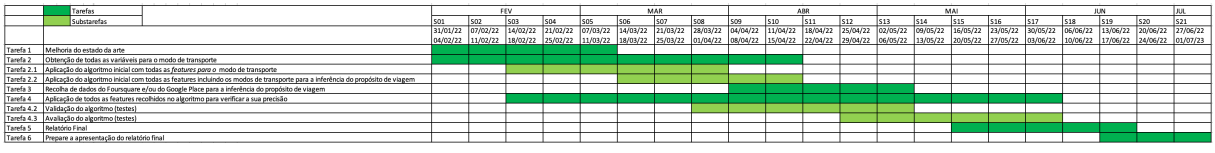


Figura 50 - Gráfico de GANTT do segundo semestre com as tarefas previstas

Ao longo do segundo semestre foi necessário recorrer a outro planeamento, Figura 51, porque durante a recolha de *features* não houve a possibilidade de obter dados relativos aos modos de transportes efetuados pelos diferentes utilizadores, não havia dados que possibilitassem essa determinação ou recolha de forma explícita. No entanto, através da velocidade registada pelos dispositivos de GPS, conseguiu-se identificar se a viagem era realizada de veículo ou simplesmente a pé. Assim, integrou-se ao nosso *dataset* todas as *features* inicialmente recolhidas, incluindo esta última que se considerou como meio de transporte, seguida do género e perfil de trabalho que não se encontram no estudo de (Montini et al., 2014) e, posteriormente, obter uma previsão inicial.

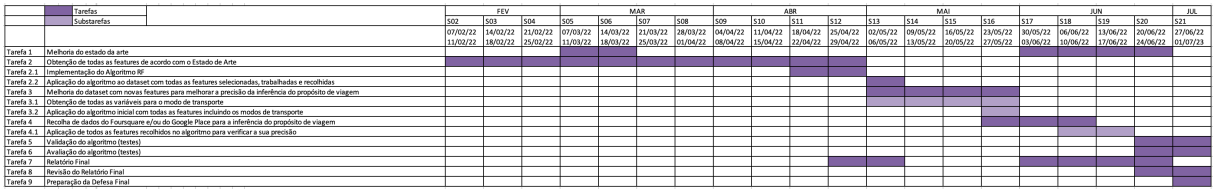


Figura 51 - Gráfico de GANTT do segundo semestre com as tarefas executadas

Houve ainda a necessidade de utilizar alguns algoritmos para obtermos um *ground truth* dos pontos de locais de interesse e ainda fazer algumas alterações no tipo de dados das *features* já recolhidas no primeiro semestre, para um tipo de dados que pudesse ser usado no nosso modelo como o preenchimento dos valores ausentes. Estas alterações causaram algumas modificações no cronograma, não influenciando no progresso e objetivos do projeto. Após obtermos uma previsão inicial, fez-se a recolha de mais dados do Foursquare para adicionar novos dados ao conjunto de *features* no *dataset*, sempre com o intuito de melhorar a nossa previsão, ao mesmo tempo que avaliamos o modelo e preparamos o relatório final deste projeto e, assim termos um modelo apto a identificar o propósito de viagem com uma boa previsão.

5.2 Riscos e Mitigação

Alguns riscos poderão afetar o desenvolvimento desta dissertação ao longo do segundo semestre. Para isso, será necessário apresentar o impacto desses riscos no desenvolver deste projeto e que soluções possíveis podem ser aplicadas para mitigar estes riscos.

Tabela XI - Escala e avaliação de riscos

	Probabilidade	**Impacto**
1	Improvável	Baixo
2	Provável	Médio
3	Muito provável	Alto

		Probabilidade		
		1	2	3
Impacto	1			
	2			
	3			

Na Tabela XI são apresentados os riscos a cor verde, amarelo e vermelho em que o verde apresenta o risco que pode ser mitigado rapidamente, a cor amarela representa um risco que pode ter algum impacto no decorrer do projeto e também poderá resultar em algum impacto no seu desenvolvimento, porque poderá ser mais difícil de mitigar. A cor vermelha significa que o risco irá interferir no desenvolver do projeto e a sua mitigação poderá ser difícil ou impossível de resolver.

Tabela XII - Classificação dos riscos

	Risco	Probabilidade	Impacto	Avaliação
1	Artigos com f*eatures* incompatíveis com o *dataset* Breadcrumbs	2	3	
2	Atraso na recolha de *features*	2	1	
3	Falta de um *ground truth* dos pontos de locais de interesse do *dataset* Breadcrumbs	2	2	
4	Existência de uma alta ambiguidade na correspondência de uma localização de atividade para um determinado ponto de local de interesse	2	3	
5	Precisão muito baixa	3	1	
6	Incompatibilidade das *features* do dataset Breadcrumbs com as features de (Montini et al., 2014)	1	3	
7	Lapso na seleção das atividades quando se realiza a implementação dos diferentes algoritmos	2	2	

Para mitigar os riscos identificados na Tabela XII são apresentados os seguintes planos:

Risco 1: No caso de nenhum artigo apresentar um estudo que integre *features* que estejam diretamente relacionadas com os dados do *dataset* Breadcrumbs será necessário recorrer a outro tipo de características para conseguir fazer determinar *features* similares, ou seja, que estejam próximas à sua categorização. Um exemplo que ocorreu neste estudo corresponde à falta de dados relacionados com o modo de transporte usado durante a viagem, estando apenas apresentado como hábitos dos utilizadores. Porém, através da velocidade descrita no *dataset,* ao longo das viagens, conseguiu-se determinar se a viagem foi realizada num veículo ou simplesmente a pé. Esta *feature,* não sendo totalmente explícita para os meios de transporte usados, está inclusa nesta característica categórica. Caso não exista nenhuma característica no *dataset* que se relacione com as *features* de um determinado artigo, será necessário recorrer a outros estudos implicando um grande atraso no desenvolvimento e obtenção das *features* para serem aplicadas no modelo, e por essa razão é necessário realizar um bom estudo no estado de arte ao mesmo tempo que se verifica a viabilidade da extração de características do *dataset* disponibilizado.

Risco 2: No caso de ocorrer atraso na recolha das *features*, o que é um risco habitual de acontecer, dado que se perde muito tempo de trabalho na recolha, manipulação e extração de *features* dos dados, teremos que adiar a entrega do documento final para a época especial. No entanto, não é algo que implique um mau estudo e trabalho. Por isso, com uma boa pesquisa realizada no estado da arte e com a possível aplicação de *queries* em *Python* e SQL que façam com que o processo de recolha de *features* seja mais rápido e eficiente, conseguindo mitigar este risco e conduzir o trabalho numa boa conduta. Felizmente, este risco não ocorreu no nosso trabalho.

Risco 3: Como é necessário realizar a correspondência de cada ponto de local de interesse do *dataset* Breadcrumbs à latitude e longitude da localização das viagens, com o objetivo de obter um *ground truth*, houve a necessidade de entrar em contacto com os autores do *dataset* Breadcrumbs e tentar perceber se a coluna "event" corresponde à atividade ou a possíveis eventos realizados naquela região. Caso os autores do *dataset* demorassem a dar resposta, iria impactar de forma negativa no avanço do nosso estudo, uma vez que precisávamos dos dados de localização para obter um *ground*

truth dos locais de interesse. Felizmente os autores foram rápidos a dar resposta aos emails e inevitavelmente a coluna "event" corresponde a eventos calendarizados naquela região. Este risco levou com que houvesse uma alternativa de mitigação com a implementação do algoritmo de deteção de paragens de trajetórias de movimento para se conseguir determinar com previsão o local de início e fim da atividade realizada por um determinado utilizador. Caso a coluna "event" corresponde-se à realização das atividades esse processo de deteção de paragens não tinha de ser implementado no modelo, apoiando num maior avanço do nosso trabalho.

Risco 4: No caso de ocorrer uma elevada ambiguidade quando é realizada a seleção do melhor raio de GPS que comprometa a identificação do tipo de atividades a uma determinada paragem, no caso de uma paragem possuir várias atividades associadas não se consegue um *ground truth* coerente. Para mitigar essa incoerência pode-se diminuir o raio de GPS até obter um tipo de atividade a uma só determinada paragem, mas poderá comprometer um baixo número de ocorrências de instâncias de paragens no *dataset*, como foi mostrado na seleção do raio, secção 4.1, com o intuito de diminuir essa ambiguidade. No entanto, o número de paragens para o nosso estudo foi suficientemente adequado para aplicar o nosso modelo o que não implicou o baixo número de instâncias, porque este risco não está dependente do estudo, mas sim da recolha fidedigna dos dados das viagens.

Risco 5: Se a previsão obtida pelo modelo for inferior ao pretendido, ou seja, pior que a previsão obtida pelos autores identificados no estado da arte, teríamos de aplicar mais informações de dados naquela região, dados de redes sociais baseadas, para ajudar no aumento da previsão da inferência do propósito de viagem. Inicialmente a previsão obtida pelo nosso modelo, no seu global, foi melhor que a dos autores (Montini et al., 2014), mas algumas das atividades apresentaram médias de previsão muito baixas. Este tipo de risco é muito frequente e por essa razão já existia a forma de o mitigar, ou seja, com a adição de taxonomias/categorias genéricas do Foursquare mais abrangentes do que as apresentadas no *dataset* Breadcrumbs, e caso esta previsão insista realizar a junção de atividades que estejam relacionadas ou a implementação de *features* que estão diretamente associadas aquela atividade.

Risco 6: Após o desenvolvimento do estado de arte, estudo e seleção de vários artigos que servem como apoio no nosso trabalho futuro, ao induzir o trabalho em certas pesquisas poderá não existir nenhuma pesquisa compatível com o nosso estudo em causa. Apesar de o estado de arte ser realizado com base nos nossos objetivos em estudo, pesquisas que tenham processos de desenvolvimentos similares ao nosso *dataset*, poderá não existir dados iguais ou idênticos ao estudo em causa. A probabilidade deste risco acontecer é muito baixa, no entanto, se isto acontecer implicaria uma obtenção de resultados que não poderiam ser comparados com nenhuma pesquisa, nunca prevendo se realmente o nosso modelo estava a seguir um caminho certo ou se os resultados corresponderam ao expectável. Não existe forma de mitigar este risco.

Risco 7: Na implementação dos vários algoritmos, é preciso ter o cuidado de escolher as variáveis corretas, *features* e suas atividades. No nosso modelo, quando se fez a seleção das atividades com mais de 1% de paragens efetuadas no *dataset* houve a troca de duas atividades, ou seja, da atividade "Friend's Place" com a atividade "Park", onde esta última atividade é a que estava abaixo dos 1% de paragens realizadas pelos utilizadores e que não deveria ser considerada. Este erro não influenciou na nossa previsão, dada que a atividade não estava a ser muito considerada pelo modelo devido à capacidade de *boostraping* do algoritmo RF. No entanto, caso implicasse um risco enorme no nosso modelo a forma de o mitigar consistia no regressar do início da seleção das atividades e corrigir esse erro, obviamente que seria necessário percorrer vários processos de implementação como de estudo, causando uma perda de tempo no nosso trabalho, mas seria sempre possível de o resolver.

Capítulo 6 Conclusão

No decorrer desta dissertação, junto com a equipa AmILab, foram concluídos vários objetivos com o estudo feito no estado da arte, ou seja, estudos e pesquisas realizados por vários autores que aplicaram modelos na inferência de modos de transporte como propósito de viagens com base em *features*, dados de GPS, recolhidos por diferentes dispositivos que possuíam tal tecnologia.

Analisando qual o modelo que apresentou melhores previsões, ou seja, o melhor percentual na identificação do *output* do estudo em causa pelos diferentes investigadores, selecionou-se o algoritmo *Random Forest* e as *features* relevantes empregues nesse algoritmo. As análises destes resultados permitiram compreender quais eram as *features* mais usadas e quais os modelos que, geralmente, apresentavam melhores resultados. Desta forma, prosseguiu-se para o estudo do *dataset* Breadcrumbs, disponibilizado pelos autores (Moro et al., 2019), Anexo F: "Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations", para identificar e comparar quais as *features* que poderiam ser recolhidas diretamente do *dataset* Breadcrumbs para o algoritmo já selecionado.

Embora o algoritmo tivesse sido escolhido juntamente com as *features*, de acordo com as melhores previsões, algumas delas não foram possíveis de obter diretamente do *dataset*, porque alguns dados estavam em falta ou porque alguns dados das *features* eram hábitos do próprio utilizador, como por exemplo, a frequência de um entrevistado andar de carro durante o fim de semana. Como algumas destas *features* não confirmavam tais informações selecionadas, levou à necessidade de ocorrer mais uma análise e estudo, com o objetivo de obter tais informações através de outras *features* que o *dataset* Breadcrumbs permite extrair. Desta forma e com o auxílio do desenvolvimento e implementação de algoritmos para fazer a recolha dos *timestamps* de *start* e *stop* de viagem, conseguiu-se determinar as paragens, identificar as taxonomias genéricas das atividades e a duração das mesmas, entre outros processos que foram efetuados pelos diferentes utilizadores durante a realização de uma viagem. Apesar de se tentar ultrapassar estes obstáculos, não se pode afirmar que o nosso modelo futuramente poderá proporcionar uma boa previsão para outros *dataset*, porque a recolha e seleção de *features* impacta no tipo de algoritmos ou métodos usados, e por isso existe sempre uma preparação e estudo inicial do *dataset* específica para cada estudo. No entanto, após feita toda essa recolha de *features* e sabendo que o nosso *dataset* não é balanceado, o método direcionado para este estudo correspondeu ao *Random Forest* (RF), porque proporciona uma melhor previsão da inferência do propósito de viagem, de acordo com o estado de arte.

Finalmente, após todo o desenvolvimento do modelo, ou seja, processo de recolha de *features*, estudo do melhor método de aprendizagem computacional e desenvolvimento, permitiu o desencadear de um estudo com uma boa previsão da inferência do propósito de viagem, concluindo assim com sucesso o objetivo final desta dissertação.

6.1 Principais Contribuições

Numa fase inicial deste trabalho, ao rever o estado da arte para identificar trabalhos que permitissem identificar o propósito de viagem com base na aprendizagem computacional, verificou-se que grande parte desses trabalhos utilizavam algumas *features* após a realização das atividades, como o caso dos autores (Montini et al., 2014). Estas *features* estão expressas em (Montini et al., 2014) e podem ser visualizadas na Figura 13, correspondentes a "duration", "Walk

percentage", "Cluster: mean duration", "Cluster: standard deviation duration" que condizem às colunas do nosso *dataset,* transversalmente a "duracao_hours", "walk_average_stop", "mean_duration", "standard_deviation_duration", respetivamente. Comparando este trabalho com alguns artigos publicados, se estas *features* forem removidas do nosso modelo, a sua aplicação é escalável, ou seja, permite determinar os locais importantes com base no perfil do utilizador e recomendar, de forma publicitária, vários locais à volta do seu objetivo, dado que se previu antecipadamente o objetivo do utilizador. Para isso, foi necessário remover essas *features* do nosso modelo e verificar se a previsão é aceitável.

Tabela XIII - Confusion Matrix: Random Forest sem as features após e durante a atividade

	Prediction							
Activity (truth)	1	3	4	8	10	13	16	Recall (%)
1	388	0	0	0	0	0	0	100
3	0	10	0	0	0	0	2	83
4	0	1	1	0	0	0	0	50
8	0	2	0	31	3	6	2	70
10	0	1	0	1	35	4	0	85
13	0	0	0	7	3	62	0	86
16	0	0	0	0	0	0	33	100
Precision (%)	100	71	100	79	85	86	89	

1 – Home; 3 – Sports; 4 – Park; 8 – Shopping/Restaurant; 10 – Metro/Bus Stop; 13 – Work/University; 16 – Family

Nota: A média ponderada (*average=weighted*) de precisão e de *recall* é de 95% e de MCC uma média de 90%.

Como se pode visualizar na Tabela XIII a nossa média previsiva é adequada, onde grande parte dos resultados individuais de cada atividade foram bons e por esse motivo o nosso modelo fica apto para proporcionar com o decorrer da viagem, antes da realização da atividade, indicações ao utilizador de determinados eventos e envio de coupons promocionais em certas lojas/restaurantes que se encontram ao redor do local dessa atividade a realizar. Um exemplo corresponde à realização da atividade "Shopping" efetuado por um utilizador com o intuito de fazer compras no Continente ou noutro supermercado. Antes de o utilizador chegar ao seu destino irá receber coupons promocionais não só do Continente, mas também do Pingo Doce, porque se encontra dentro de um determinado raio correspondente à atividade "Shopping" e o nosso modelo ao prever essa atividade vai gerir esses coupons promocionais e enviar ao utilizador.

Em última análise, este trabalho comprova que, além de conseguir determinar com previsão o local das atividades dos utilizadores poderá apoiar na gestão de marketing de determinadas entidades de venda de produtos.

6.2 Desafios

Desde o início do projeto, durante o seu desenvolvimento, o grande desafio consistiu na recolha de todas as *features* que os autores (Montini et al., 2014) documentaram. Não foi possível recolher

todas as 17 *features* da Figura 13, mas através do processamento de dados disponibilizados no *dataset* Breadcrumbs, foi possível recolher 13 delas. As *features* que não se conseguiram recolher ou determinar corresponderam às *features* relativas ao modo de transporte, porque o *dataset* Breadcrumbs não inclui dados suficientes para realizar essa resolução. No entanto, antes de realizar a recolha das *features* foi necessário validar as nossas instâncias de paragens do *dataset* para identificar a ambiguidade existente no *dataset* Breadcrumbs, sendo um desafio repetitivo de realizar e interpretar, porque esta etapa teve de ser feita para os vários raios de GPS mencionados no estado de arte. Caso a seleção do raio seja feita de forma errada irá prejudicar em todo o trabalho, fazendo com que o processo de obtenção das *features* tenha de ser idealizado de novo, por causa das *features* de *clustering*.

6.3 Trabalho Futuro

Num trabalho futuro será importante integrar um *dataset* no nosso modelo que contenha dados explícitos do uso dos meios de transportes ao longo da viagem, para desta forma conseguirmos replicar de forma completa o trabalho dos autores (Montini et al., 2014). Infelizmente, o *dataset* Breadcrumbs só apresenta dados de hábitos relativos aos meios de transportes usados pelos utilizadores, impossibilitando a recolha e integração dessas *features* no nosso estudo, ou seja, das *features* "Transport mode after activity" e "Transport mode before activity" incluídas nas 17 *features* da Figura 13. No entanto, como o *dataset* Breadcrumbs apresenta dados das velocidades das viagens realizadas, segundo (Feng & Timmermans, 2016), através do cálculo das variáveis da Tabela II possibilita a determinação dos modos de transportes que, posteriormente, são usados futuramente na aplicação do nosso modelo. Contudo, no caso de existir congestionamento numa determinada região poderá ocorrer conflito de identificação do modo de transporte de carro do modo de transporte de bicicleta, porque os carros durante um determinado tempo irão transitar a uma velocidade baixa. Para além desta *features* não estarem presentes no nosso *dataset* Breadcrumbs, também não tínhamos a informação relativa à *feature* "Income".

No nosso modelo, como não foram adicionados dados de check-ins deixados pelos utilizadores do Foursquare naquela região, Suíça, e em outros locais onde várias atividades foram realizadas pelos diferentes utilizadores, será relevante num trabalho futuro integrar os dados de popularidade no nosso estudo e identificar se houve melhoria do nosso modelo, uma vez que apenas foi contabilizado qual a oferta de serviços à volta de um determinado destino de viagem de forma genérica (através das categorias de topo da taxonomia do Foursquare). A adição de dados sociodemográficos dessas regiões também terá um impacto positivo para identificar mercearias, edifícios governamentais, postos de abastecimento, etc., para identificar melhor as paragens de curto prazo a meio de uma determinada viagem longa. Finalmente, através dos dados de desporto do *dataset* Breadcrumbs aproveitar essa *feature,* mesmo que não tenha sido considerada por (Montini et al., 2014), para acrescentar novas *features* de atividade desportiva de cada utilizador, com o intuito de determinar a identificação dessa atividade em locais como shoppings, dado que este tipo de ponto de local de interesse pode englobar vários tipos de atividades, como compras, almoço/jantar e desporto ou em locais ao ar livre, como em parques.

Quando se fez a união da classe "Shopping" com "Restaurant" e da classe "Metro Stop" com "Bus Stop", será necessário a implementação de um novo subclassificador para ser treinado apenas com paragens a estes pontos de locais de interesse e, posteriormente, verificar se existe a distinção entre estas classes. Ou seja, verificar se o subclassificador consegue distinguir "Metro Stop" de "Bus stop" e "Shopping" de "Restaurant" onde apenas temos instâncias das paragens apenas a esses destinos. Assim, quando o nosso primeiro classificador indicar a classe composta

Restaurant/Shopping, ou Metro/Bus stop, incluirá nas instâncias a estas paragens a previsão do resultado do subclassificador, respetivamente.

Futuramente, com a aplicabilidade de um *dataset* com todos estes dados e com o acréscimo destas *features,* poderá possibilitar uma melhor previsão e um novo estudo do nosso modelo na inferência dos meios de transportes usados pelos diferentes utilizadores, antes de serem realizadas as viagens, avaliando e validando o modelo desenvolvido para trabalho.

Referências

Addan, D. (2019). *Support Vector Machine*.

Almeida Adriano, Carvalho Felipe, & Menino Felipe. (2017). *Introdução ao Machine Learning*. Github. https://dataat.github.io/introducao-ao-machine-learning/index.html

Alok Gupta. (2015, April 7). *Overcoming Missing Values In A Random Forest Classifier*. The Airbnb Tech Blog.

Anita Graser. (2019). *MovingPandas*. GitHub. https://github.com/anitagraser/movingpandas

Anita Graser. (2022, January 12). *Detecting stops*. Github.

ArcGis Pro. (2022). *How Density-based Clustering works*. Esri.

Bex T. (2021, June 9). *Comprehensive Guide to Multiclass Classification Metrics*. Towards Data Science. https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd

Boaz Shmueli. (2019, July 2). *Multi-Class Metrics Made Simple, Part I: Precision and Recall*. Towards Data Science. https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285–297. https://doi.org/10.1016/j.trc.2008.11.004

Breiman, L. (2001). *Random Forests* (Vol. 45).

Carnein, M., & Trautmann, H. (2019). Optimizing Data Stream Representation: An Extensive Survey on Stream Clustering Algorithms. *Business and Information Systems Engineering*, *61*(3), 277–297. https://doi.org/10.1007/s12599-019-00576-5

Carrasco Juan, & Ortúzar Juan. (2010). Review and assessment of the nested logit model. *Transport Reviews*.

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice*, *44*(10), 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Chen, C., Jiao, S., Zhang, S., Liu, W., Feng, L., & Wang, Y. (2018). TripImputor: Real-Time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Transactions on Intelligent Transportation Systems*, *19*(10), 3292–3304. https://doi.org/10.1109/TITS.2017.2771231

Chen, Y., & Tu, L. (2007). *Density-Based Clustering for Real-Time Stream Data*.

Cínthia Pessanha. (2019, November 20). *Random Forest: como funciona um dos algoritmos mais populares de ML*.

Datamart. (2019, September 27). *Logistic Regression in Python with the Titanic Dataset*. https://www.datarmatics.com/data-science/logistic-regression-in-python-with-the-titanic-dataset/

David Cournapeau. (2022, May). *Scikit-learn*. https://scikit-learn.org/stable/about.html#people

DiFrancesco, P. M., Bonneau, D., & Hutchinson, D. J. (2020). The implications of M3C2 projection diameter on 3D semi-automated rockfall extraction from sequential terrestrial laser scanning point clouds. *Remote Sensing*, *12*(11). https://doi.org/10.3390/rs12111885

Dingqi YANG. (2019, September). *Foursquare Dataset*. Dingqi YANG's Homepage. https://sites.google.com/site/yangdingqi/home/foursquare-dataset?authuser=0

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., & Das, K. (2017). Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, *77*, 96–112. https://doi.org/10.1016/j.trc.2017.01.020

Feng, T., & Timmermans, H. J. P. (2016). Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *Transportation Planning and Technology*, *39*(2), 180–194. https://doi.org/10.1080/03081060.2015.1127540

Foursquare. (2021a). *Foursquare Categories*. Foursquare. https://developer.foursquare.com/docs/categories

Foursquare. (2021b, November 28). *FOURSQUARE/developers*. https://developer.foursquare.com/

Gao, Q., Molloy, J., & Axhausen, K. W. (2021). Trip Purpose Imputation Using GPS Trajectories with Machine Learning. *ISPRS International Journal of Geo-Information*, *10*(11), 775. https://doi.org/10.3390/ijgi10110775

Garnett, R., & Stewart, R. (2015). Comparison of GPS units and mobile Apple GPS capabilities in an urban landscape. *Cartography and Geographic Information Science*, *42*(1), 1–8. https://doi.org/10.1080/15230406.2014.974074

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia - Social and Behavioral Sciences*, *138*, 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239

Google Places. (2021, November 28). *Places API Usage and Billing*. https://developers.google.com/maps/documentation/places/web-service/usage-and-billing

Guilherme Fernandes, F. M. e B. C. (2019, September 1). *Modelos de Predição | Decision Tree*.

Igor Kuznetsov. (2019, May 9). *Metrics for Imbalanced Classification*. Towards Data Science. https://towardsdatascience.com/metrics-for-imbalanced-classification-41c71549bbb5

Jason Brownlee. (2020a, April 27). *How to Develop a Random Forest Ensemble in Python*. Machine Learning Mastery. https://machinelearningmastery.com/random-forest-ensemble-in-python/

Jason Brownlee. (2020b, August 15). *Parametric and Nonparametric Machine Learning Algorithms*. Machine Learning Algorithms.

Krause, C. M., & Zhang, L. (2019). Short-term travel behavior prediction with GPS, land use, and point of interest data. *Transportation Research Part B: Methodological*, *123*, 349–361. https://doi.org/10.1016/j.trb.2018.06.012

Lu, Y., Zhu, S., & Zhang, L. (2012). *A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys*.

Lu, Y., Zhu, S., & Zhang, L. (2013, January 13). *Imputing Trip Purpose Based on GPS Travel Survey Data and Machine Learning Methods*. 1–18.

Meng, C., Cui, Y., He, Q., Su, L., & Gao, J. (2017). Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-January*, 1319–1324. https://doi.org/10.1109/BigData.2017.8258062

Mirko Stojiljković. (2021). *Split Your Dataset With scikit-learn's train_test_split()*. Real Python. https://realpython.com/train-test-split-python-data/#reader-comments

Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. (2014). Trip purpose identification from GPS tracks. *Transportation Research Record*, *2405*, 16–23. https://doi.org/10.3141/2405-03

Moro, A., Kulkarni, V., Ghiringhelli, P. A., Chapuis, B., Huguenin, K., & Garbinato, B. (2019). Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 508–511. https://doi.org/10.1145/3347146.3359341

MullOverThings. (2020, December 10). *How do you change units to meters in Qgis?* MullOverThings. https://mull-over-things.com/how-do-you-change-units-to-meters-in-qgis/

Nguyen, M. H., Armoogum, J., Madre, J. L., & Garcia, C. (2020). Reviewing trip purpose imputation in GPS-based travel surveys. In *Journal of Traffic and Transportation Engineering (English Edition)* (Vol. 7, Issue 4, pp. 395–412). Chang'an University. https://doi.org/10.1016/j.jtte.2020.05.004

Oliveira Marcelo, Vovsha Peter, Mitchell Michael, & Wolf Jean. (2014, January 1). Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys. *First Published*.

Peter Hohnson. (2015). *MissingLink*. GitHub. https://gist.github.com/missinglink/a0344050d3e2b52256d7/

Piotr Płoński. (2019, April 5). *Does Random Forest overfit?* Mljar.

ProFloresta. (2022, April 18). *Análise de agrupamento*. https://files.cercomp.ufg.br/weby/up/417/o/Aula_3_An%C3%A1lise_de_agrupamento.pdf

Rob J Hyndman e George Athanasopoulos. (2018). *Previsão: Princípios e Prática* ((2ª ed)).

Ross Quinlan, by J., Kaufmann Publishers, M., & Salzberg, S. L. (1994). *Programs for Machine Learning* (Vol. 16).

Saul Dobilas. (2021, May 9). *HAC: Hierarchical Agglomerative Clustering — Is It Better Than K-Means?* Towards Data Science.

Saurabh Gupta. (2021, January 22). *Hyperparameters of Random Forest Classifier*. GeeksforGeeks.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, *42*(1), 163–188. https://doi.org/10.1007/s11116-014-9541-6

Shen, L., & Stopher, P. R. (2013). A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies*, *36*, 261–267. https://doi.org/10.1016/j.trc.2013.09.004

Shen, L., & Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. In *Transport Reviews* (Vol. 34, Issue 3, pp. 316–334). Routledge. https://doi.org/10.1080/01441647.2014.903530

Silva, C., & Ribeiro, B. (2018). *CATARINA SILVA BERNARDETE RIBEIRO COMPUTACIONAL APRENDIZAGEM EM ENGENHARIA*.

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, *16*(3), 350–369. https://doi.org/10.1016/j.trc.2007.10.002

Sun, H., Chen, Y., Wang, Y., & Liu, X. (2021). Trip purpose inference for tourists by machine learning approaches based on mobile signaling data. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03346-y

Wu, J., Jiang, C., Houston, D., Baker, D., & Delfino, R. (2011). Automated time activity classification based on global positioning system (GPS) tracking data. *Environmental Health: A Global Access Science Source*, *10*(1). https://doi.org/10.1186/1476-069X-10-101

Xiao, G., Juan, Z., & Zhang, C. (2016a). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, *71*, 447–463. https://doi.org/10.1016/j.trc.2016.08.008

Xiao, G., Juan, Z., & Zhang, C. (2016b). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, *71*, 447–463. https://doi.org/10.1016/j.trc.2016.08.008

Yazdizadeh, A., Patterson, Z., & Farooq, B. (2019). An automated approach from GPS traces to complete trip information. *International Journal of Transportation Science and Technology*, *8*(1), 82–100. https://doi.org/10.1016/j.ijtst.2018.08.003

Zhan, X., Ukkusuri, S. v., & Zhu, F. (2014). Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Networks and Spatial Economics*, *14*(3–4), 647–667. https://doi.org/10.1007/s11067-014-9264-4

Anexo A: Descrição do *dataset* Breadcrumbs

Tabelas e Atributos

unique_user_id

- id (BIGINT): identificador de um usuário (*primary key*)

userinfo

- id (BIGINT): identificador de um userinfo (*primary key*)
- firstlastname (VARCHAR): nome e sobrenome do usuário com *hash*
- email (VARCHAR): e-mail do utilizador com *hash*
- phone (VARCHAR): número de telefone do utilizador com *hash*
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

location

- id (BIGINT): identificador do local (*primary key*)
- timestamp (BIGINT): etiqueta de data/hora, indicando o local
- latitude (DOUBLE): latitude da localização
- longitude (DOUBLE): longitude da localização
- altitude (DOUBLE): altitude do local
- speed (DOUBLE): velocidade registado no momento da localização
- horizontalAccuracy (DOUBLE): precisão horizontal da localização
- verticalAcuracy (DOUBLE): precisão vertical da localização do local
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

contact

- id (BIGINT): identificador do registo de contato (*primary key*)
- timestamp (BIGINT): etiqueta de data/hora, indicando o registo de contacto
- firstlastname (VARCHAR): nome e sobrenome do registo de contacto com *hash*
- email (VARCHAR): e-mail do registro do contacto com *hash*
- phone (VARCHAR): número de telefone do registo do contacto com *hash*

- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

Nota. Se vários e-mails ou telefones forem indicados para um contato, isso significa que pode haver vários registos semelhantes com um único e-mail diferente ou atributo de telefone para cada um. 0 é indicado se não houver e-mail ou telefone.

event

- id (BIGINT): identificador do registo de evento (*primary key*)
- timestamp (BIGINT): etiqueta de data/hora, indicando o registo do evento
- title (VARCHAR): título do registo do evento com *hash*
- start (BIGINT): etiqueta de data/hora, indicando o registo do inicio do evento
- stop (BIGINT): etiqueta de data/hora, indicando o registo do fim do evento
- organizer (VARCHAR): organizador do registo do evento com *hash*
- attendee (VARCHAR): participante do registo do evento com *hash*
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

Se vários participantes forem indicados para um evento, isso significa que pode haver vários registos semelhantes com um único atributo de participante diferente para cada um. 0 é indicado que não há nenhum participante ou telefone.

bluetooth_scan

- id (BIGINT): identificador do registo de evento (*primary key*)
- timestamp (BIGINT): etiqueta de data/hora, indicando a verificação de Bluetooth realizada
- deviceuuid (VARCHAR): UUID do dispositivo periférico remoto detetado por bluetooth com *hash*
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

Se vários deviceuuids forem capturados para uma ferramenta de *scan* de bluetooth, isso significa que pode haver vários registos semelhantes com um único deviceuuid diferente para cada um. 0 é indicado se não houver nenhum deviceuuid capturado.

wifi_scan

- id (BIGINT): identificador do registo de evento (*primary key*)
- timestamp (BIGINT): etiqueta de data/hora, indicando o momento a verificação de wifi realizada
- wifissid (VARCHAR):
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

Se vários wifissids forem capturados para u ma ferramenta de *scan* de wi-fi, isso significa que pode haver vários registos semelhantes com um único atributo wifissid diferente para cada um. 0 é indicado se não houver wifissid capturado.

relationship

- id (BIGINT): identificador do registo do evento (*primary key*)
- user_id_1 (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)
- user_id_2 (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

point_of_interest

- id (BIGINT): identificador do registo do evento (*primary key*)
- poi_nb (BIGINT): número do ponto de interesse de um utilizador
- latitude (DOUBLE): latitude que descreve o ponto de interesse
- longitude (DOUBLE): longitude descrevendo o ponto de interesse
- description_id (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key*)
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

O description_id de um ponto de interesse pode ser encontrado na tabela «point_of_interest_description» abaixo (usando o id da tabela abaixo).

point_of_interest_description

- id (BIGINT): identificador do registo do evento (*primary key*)
- description (VARCHAR): descrição do ponto de interesse (por exemplo, Home, Work)

survey

- id (BIGINT): identificador do registo do evento (*primary key*)
- gender (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key*)
- age_group (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- working_profile (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- car_week (BIGINT):identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- public_transportation_week (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- bike_week (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- taxi_week (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- walking_week (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- car_weekend (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- public_transportation_weekend (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- bike_weekend (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- taxi_weekend (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- walking_weekend (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- job (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*

- university (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- section (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- field_of_studies (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- living_parent_s_home (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- parent_s_home (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- family_status (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- sport_exercises_frequence (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- student_association (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- nationality (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- smoking_cigarettes -(BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- seasonal_allergies (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- diet (BIGINT): identificador de uma descrição contida na tabela «point_of_interest_description» (*foreign key)*
- user_id (BIGINT): identificador único de um utilizador (id) contido na tabela «unique_user_id» (*foreign key*)

survey_answer_description

- id (BIGINT): identificador do registo do evento (*primary key*)
- description (VARCHAR): resposta da pesquisa (por exemplo, Bachelor, Master)

Anexo B: Recolha de *features* de *clustering*

1) Determinar as *features* "percentage week day" e "occurrences per day"

```
#Para a manipulação de dados
import pandas as pd

# Ler o ficheiro csv
df=pd.read_excel('/Users/ribeiro/Desktop/clusters_6_5/hac_362_6_5.xlsx')

df = df[(df['poi_description'] == 'Home') & (df['Clust_a'] == 0)]

############ percentage week day ############
df_sun = df.loc[df['day_week'] == 'Sun']
df_mon = df.loc[df['day_week'] == 'Mon']
df_tue = df.loc[df['day_week'] == 'Tue']
df_wed = df.loc[df['day_week'] == 'Wed']
df_thu = df.loc[df['day_week'] == 'Thu']
df_fri = df.loc[df['day_week'] == 'Fri']
df_sat = df.loc[df['day_week'] == 'Sat']

percent_week_day_sun = (len(df_sun)*100)/len(df)
percent_week_day_mon = (len(df_mon)*100)/len(df)
percent_week_day_tue = (len(df_tue)*100)/len(df)
percent_week_day_wed = (len(df_wed)*100)/len(df)
percent_week_day_thu = (len(df_thu)*100)/len(df)
percent_week_day_fri = (len(df_fri)*100)/len(df)
percent_week_day_sat = (len(df_sat)*100)/len(df)

percentage_week_days = []
print(percent_week_day_sun)
percentage_week_days.append("percent_week_day_sun")
percentage_week_days.append(percent_week_day_sun)
print(percent_week_day_mon)
percentage_week_days.append("percent_week_day_mon")
percentage_week_days.append(percent_week_day_mon)
print(percent_week_day_tue)
percentage_week_days.append("percent_week_day_tue")
percentage_week_days.append(percent_week_day_tue)
print(percent_week_day_wed)
percentage_week_days.append("percent_week_day_wed")
percentage_week_days.append(percent_week_day_wed)
print(percent_week_day_thu)
percentage_week_days.append("percent_week_day_thu")
percentage_week_days.append(percent_week_day_thu)
print(percent_week_day_fri)
```

```
percentage_week_days.append("percent_week_day_fri")
percentage_week_days.append(percent_week_day_fri)
print(percent_week_day_sat)
percentage_week_days.append("percent_week_day_sat")
percentage_week_days.append(percent_week_day_sat)

sum([percent_week_day_sun, percent_week_day_mon, percent_week_day_tue,
percent_week_day_wed, percent_week_day_thu, percent_week_day_fri,
percent_week_day_sat]) #verificar se a soma corresponde a 100%

print(percentage_week_days)

############ occurrences per day ############
occurrences_per_day = []

i = 0
while len(df['date_year'].unique()) > i:
    date = df[(df['date_year'] == df['date_year'].unique()[i])]
    print(df['date_year'].unique()[i])
    print(len(date))
    occurrences_per_day.append(df['date_year'].unique()[i])
    occurrences_per_day.append(len(date))
    i += 1
```

2) Para todos os valores NULL de cada utilizador, se existirem outros dados, calcular a mediana desses dados para substituir os valores ausentes

```
#### mediana para os valores NULL ####
import statistics

print(occurrences_per_day, "\n")

lst = []
i = 1
while len(occurrences_per_day) > i:
    lst.append(occurrences_per_day[i])
    i += 2

print(lst, "\n")

print("Mediana:", statistics.median(lst))
```

Anexo C: Algoritmo *Random Forest*

1) Importar o nosso *dataset* para o Jupyter Notebook

```
# Pandas é usado para manipulação de dados
import pandas as pd

# Ler os dados e exibir as primeiras 5 linhas
dataset = pd.read_csv('/Users/ribeiro/Desktop/dataset_all_6_5.csv', encoding="utf-8")
dataset.head(5)
```

2) Transformar os dados das tabelas com variáveis categóricas em variáveis do tipo *dummy* (0 ou 1)

```
dataset = pd.get_dummies(dataset, prefix=['grupo_idade', 'escolaridade', 'estado_civil',
'day_week'], columns = ['grupo_idade', 'escolaridade', 'estado_civil', 'day_week'])
```

3) Selecionar as minhas variáveis de entrada, "X", e a variável de saída, "y"

```
y = dataset.loc[:, ['poi_description']]

# Remover as features que não irão fazer parte do estudo
dataset = dataset.drop('user_id', axis = 1)
dataset = dataset.drop('start_location_latitude', axis = 1)
dataset = dataset.drop('start_location_longitude', axis = 1)
dataset = dataset.drop('date_year', axis = 1)
dataset = dataset.drop('poi_description', axis = 1)
dataset = dataset.drop('walk_average_stop', axis = 1)


# features após e durante a atividade
dataset = dataset.drop('duracao_hours', axis = 1)
dataset = dataset.drop('mean_duration', axis = 1)
dataset = dataset.drop('standart_deviation_duration', axis = 1)

# features adicionais
dataset = dataset.drop('genero', axis = 1)
dataset = dataset.drop('working_profile', axis = 1)
dataset = dataset.drop('walk_percentage_trip', axis = 1)

X = dataset
```

4) Dividir o *dataset* entre treino e teste, 75% e 25%, através do parâmetro padrão do *stratify* e aplicar o modelo de Random Forest de acordo com o número máximo de *features* e 500 árvores de decisão

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y.values.ravel(), stratify=y, shuffle=True)

# Importar o modelo Random Forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import matthews_corrcoef, recall_score, precision_score, confusion_matrix

'''
n_estimators — o número de árvores de decisão que executaremos no modelo
max_depth — define a profundidade máxima possível de cada árvore
max_features — o número máximo de recursos que o modelo considerará ao determinar uma divisão
bootstrapping — o valor padrão para isso é True, o que significa que o modelo segue os princípios de bootstrapping
max_samples — Este parâmetro assume que o bootstrapping está definido como True, caso contrário, este parâmetro não se aplica. No caso de True, esse valor define o maior tamanho de cada amostra para cada árvore.
random_state - Controla tanto a aleatoriedade do bootstrap das amostras usadas na construção de árvores (if bootstrap=True) quanto a amostragem dos recursos a serem considerados ao procurar a melhor divisão em cada nó (if).
Sempre que a randomização faz parte de um algoritmo Scikit-learn, um random_state parâmetro pode ser fornecido para controlar o gerador de números aleatórios usado. Observe que a mera presença de random_statenão significa que a randomização seja sempre usada, pois pode depender de outro parâmetro, por exemplo shuffle, ser definido.

average=micro diz à função para calcular f1 considerando o total de verdadeiros positivos, falsos negativos e falsos positivos (independentemente da previsão para cada rótulo no conjunto de dados)
average=macro diz à função para calcular f1 para cada rótulo e retorna a média sem considerar a proporção de cada rótulo no conjunto de dados.
average=weighted diz à função para calcular f1 para cada rótulo e retorna a média considerando a proporção de cada rótulo no conjunto de dados.
average=samples diz à função para calcular f1 para cada instância e retorna a média. Use-o para classificação multirrótulo.
'''

#Verificar o número de instâncias e features
print('Formato dos recursos de teste:', X_test.shape)

forest = RandomForestClassifier(n_estimators = 500, max_features = 33, bootstrap = True)
forest.fit(X_train, y_train)
y_pred = forest.predict(X_test)

print("MCC: ", matthews_corrcoef(y_test, y_pred)) # Métrica boa para avaliar modelos em dados #não balanceados
print(classification_report(y_test, y_pred, zero_division=1))
print("Confusion Matrix\n", confusion_matrix(y_test, y_pred))
```

5) Calcular a media de previsão para 100 testes

```
from sklearn.metrics import matthews_corrcoef, f1_score

j = 0
mcc = 0
f1 = 0
while j < 100:
   mcc += matthews_corrcoef(y_test, y_pred)
   f1 += f1_score(y_test, y_pred, average="weighted")
   j += 1


print("Average MCC to 100 test: ", mcc/j)
print("Average F1-Score to 100 test: ", f1/j)
```

6) Examinar os hiperparâmetros

```
########## Contabilizar o número de árvores de decisão ideais para o nosso modelo

%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.metrics import matthews_corrcoef, f1_score

forest = RandomForestClassifier(n_estimators = 1, max_features = 29, bootstrap = True)
trees, f1_test_loss, mcc_test_loss = [], [], []
for iter in range(500):
   forest.fit(X_train, y_train)
   y_test_predicted = forest.predict(X_test)

   mcc_test = matthews_corrcoef(y_test, y_test_predicted)

   f1_test = f1_score(y_test, y_test_predicted, average="weighted")



   print("Iteration: {} Test F1-Score: {} Test MCC: {}".format(iter, f1_test, mcc_test))
   trees += [forest.n_estimators]
   f1_test_loss += [f1_test]
   mcc_test_loss += [mcc_test]

   forest.n_estimators += 1
plt.figure(figsize=(8,6))
plt.plot(trees, f1_test_loss, color="blue", label="F1-Score")
plt.plot(trees, mcc_test_loss, color="red", label="MCC")
plt.xlabel("# of trees")
plt.ylabel("Average");
plt.legend()
plt.savefig('/Users/ribeiro/Desktop/MCC.png')

########## Verificar o max_features adequado
```

```
from numpy import mean
from numpy import std
from numpy import arange
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.ensemble import RandomForestClassifier
from matplotlib import pyplot

model = forest.fit(X_train, y_train.ravel())

# get a list of models to evaluate
# get a list of models to evaluate
def get_models():
    models = dict()
    # explore number of features from 1 to 7
    for i in range(1,36):
        models[str(i)] = RandomForestClassifier(max_features=i)
    return models

# evaluate a given model using cross-validation
def evaluate_model(model, X, y):
    # define the evaluation procedure
    cv = RepeatedStratifiedKFold(n_splits=5, n_repeats=3, random_state=1)
    # evaluate the model and collect the results
    scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
    return scores

# get the models to evaluate
models = get_models()
# evaluate the models and store results
results, names = list(), list()
for name, forest in models.items():
    # evaluate the model
    scores = evaluate_model(model, X, y)
    # store the results
    results.append(scores)
    names.append(name)
    # summarize the performance along the way
    print('>%s %.3f (%.3f)' % (name, mean(scores), std(scores)))
```

Anexo D: Algoritmo de deteção de paragens

1) Obter todas as paragens da tabela "location" do *dataset* Breadcrumbs

```
import pandas as pd
import geopandas as gpd
from geopandas import GeoDataFrame, read_file
from shapely.geometry import Point, LineString, Polygon
from datetime import datetime, timedelta
import movingpandas as mpd
import warnings

warnings.filterwarnings('ignore')

mpd.show_versions

df = pd.read_csv('/Users/ribeiro/Desktop/location.txt', delimiter = ',')

traj_collection = mpd.TrajectoryCollection(df, 'user_id', t='location_datetime',
x='location_longitude', y='location_latitude')

#Gravar dados das paragens num ficheiro (.txt) para cada utilizador
i = 0
while len(traj_collection) > i:
   #Stop detection with a single trajectory
   my_traj = traj_collection.trajectories[i]
   detector = mpd.TrajectoryStopDetector(my_traj)

   #Duração de paragem
   with open('/Users/ribeiro/Desktop/stop_all/readme{}.txt'.format(i), 'a') as f:
      stop_time_ranges = detector.get_stop_time_ranges(min_duration=timedelta(seconds=180),
max_diameter=100)
      for x in stop_time_ranges:
         f.write(str(x))
         f.write('\n')
         print(str(x))
      i += 1
```

2) Selecionar os dados das paragens relevantes, ou seja, "user_id", "start_paragem", "stop_paragem" e "duração" e adicionar estes dados num ficheiro de texto (.txt) para importar para a nossa base de dados SQL

```
import re

def read_file(j, i):
   with open('/Users/ribeiro/Desktop/stop_all/readme{}.txt'.format(j), "r") as f:
      lines = f.read()
```

```
        read_file = re.split(': | - | |\n', lines)
        while len(read_file) > i:
            data = read_file[i] + ';' + read_file[i+1] + ' ' + read_file[i+2] + ';' + read_file[i+3] + ' ' +
read_file[i+4] + ';' + read_file[i+6] + ' ' + read_file[i+7] + ' ' + read_file[i+8][:-1]
            file.write(data + '\n')
            i += 10
j = 0
i = 1

file = open("/Users/ribeiro/Desktop/stop_all/paragens_all.txt","a+")
file.write("user_id; start_paragem; stop_paragem; duracao\n")
while 79 > j: #80 nº máximo de utilizadores
    read_file(j, i) # Ler os dados do ficheiro
    j += 1

file.close()
```

Anexo E: Algoritmo de Agrupamento Hierárquico Aglomerativo

1) Numa primeira fase explora-se todos os dendogramas para os diferentes tipos de *linkages* para cada utilizador

```
from sklearn.cluster import AgglomerativeClustering # Para cluster HAC
import scipy.cluster.hierarchy as sch # Para o Dendograma HAC
import matplotlib.pyplot as plt # Para visualizar os dados do dataset
import pandas as pd # Para manipular os dados do dataset

# Ler os dados do dataset do ficheiro csv
df = pd.read_csv('/Users/ribeiro/Desktop/dataset_all_6_5_most1.csv', encoding='utf-8')

# Visualizar os dados do dataset
df.head(6042) #6041 linhas, ou seja, instâncias de paragem

# Verificar no dataset os utilizadores que têm mais que uma paragem
user_id_2 = []

i = 0
while len(user_id) > i:
    dados_utilizador = df.loc[df['user_id'] == user_id[i]]

    while True:
        if len(dados_utilizador) == 1:
            i += 1
            dados_utilizador = df.loc[df['user_id'] == user_id[i]]
        else:
            user_id_2.append(user_id[i])
            break
    i += 1

print(user_id_2) # Visualizar todos os utilizadores com mais que uma paragem
len(user_id_2) # Contar o número de utilizadores

# Gerar todos os Dendogramas para os 3 tipos de linkages para todos os utilizadores
# (user_id_2) e guardar no Desktop pessoal
i = 0
while len(user_id_2) > i:
    dados_utilizador = df.loc[df['user_id'] == user_id_2[i]]

    # Selecionar os atributos para realizar o clustering
    X1 = dados_utilizador.loc[:,['start_location_latitude', 'start_location_longitude']]
```

```
    # Iniciar a função subplot usando o número de linhas e colunas
    figure, axes = plt.subplots(3, figsize=(16,9), dpi=300)

# Criar os diferentes linkages
    Z1 = sch.linkage(X1, method='single', optimal_ordering=True) # have method='single' | 
method='complete' | method='average'
    Z2 = sch.linkage(X1, method='complete', optimal_ordering=True) # have method='single' | 
method='complete' | method='average'
    Z3 = sch.linkage(X1, method='average', optimal_ordering=True) # have method='single' | 
method='complete' | method='average'

    # Definir as cores para diferentes clusters
    sch.set_link_color_palette(['red', '#34eb34', 'blue', '#ae34eb'])

    # Desenhar os Dendrogramas
    sch.dendrogram(Z1, leaf_rotation=90, leaf_font_size=4, 
labels=list(dados_utilizador['date_year']),
                color_threshold=0.1, above_threshold_color='black', ax = axes[0])
    sch.dendrogram(Z2, leaf_rotation=90, leaf_font_size=4, 
labels=list(dados_utilizador['date_year']),
                color_threshold=0.1, above_threshold_color='black', ax = axes[1])
    sch.dendrogram(Z3, leaf_rotation=90, leaf_font_size=4, 
labels=list(dados_utilizador['date_year']),
                color_threshold=0.1, above_threshold_color='black', ax = axes[2])

    # Adicionar a linha horizontal
    axes[0].axhline(y=0.1, c='grey', lw=1, linestyle='dashed')
    axes[1].axhline(y=0.1, c='grey', lw=1, linestyle='dashed')
    axes[2].axhline(y=0.1, c='grey', lw=1, linestyle='dashed')

    # Guardar os plots dos Dendogramas
    plt.savefig('/Users/ribeiro/Desktop/teste/dendograma_{}_6_5.png'.format(user_id_2[i]))
    plt.close()

    i += 1
```

2) Numa segunda fase seleciona-se os dados específicos de um determinado utilizador para agrupar os dados por distâncias definindo o número de cluster (*n_clusters*) de acordo com a triagem do dendograma do *cluster* por utilizador

```
# Seleção dos dados por utilizador (i)
dados_utilizador = df.loc[df['user_id'] == user_id_2[0]] # user_id_2[0]] corresponde ao
# primeiro utilizador da nossa lista de utilizadores (user_id_2 = [])

X = dados_utilizador.loc[:,['start_location_latitude', 'start_location_longitude']]

# Definir o modelo e os parâmetros
# Tipos de linkage: { ‘complete’, ‘average’, ‘single’}
# n_clusters corresponde ao número de clusters a definir
```

```
model_a = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='average')
model_c = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='complete')
model_s = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='single')

# Aplicação dos dados no HAC
clust_a = model_a.fit(X)
clust_c = model_c.fit(X)
clust_s = model_s.fit(X)

# Anexe rótulos de cluster de volta ao conjunto de dados de localização
dados_utilizador.loc[:,['Clust_a']]=clust_a.labels_
dados_utilizador.loc[:,['Clust_c']]=clust_c.labels_
dados_utilizador.loc[:,['Clust_s']]=clust_s.labels_

# Guardar os dados num ficheiro excel
datatoexcel =
pd.ExcelWriter("/Users/ribeiro/Desktop/clusters_6_5/hac_{}_6_5.xlsx".format(user_id_2[0]))

dados_utilizador.to_excel(datatoexcel)

datatoexcel.save()
```

3) Identificar o cluster que apresentas mais atividades (POI) do tipo “Home” para definir esse cluster como “Main Cluster”

```
# Contar a quantidade de POIs do tipo “HOME” nos clusters
def check_home(id_user):
    i = 0
    count = 0

    while len(df_loc) > i:
        #print(df_loc.iloc[i][13])
        if df_loc.iloc[i][13] == "Home":
            count += 1
        i += 1
    print("User ID: {} || Home: {}".format(id_user, count))
    return count

l = 0
j = 0
while len(user_id) > j:
    df_loc = df.loc[df['user_id'] == user_id[j]]
    #print(user_id[j])
    count = check_home(user_id[j])

    if count > 0:
        l += 1

    j += 1
```

Anexo F: "Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations"

Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations

Arielle Moro
University of Lausanne, Switzerland
Arielle.Moro@unil.ch

Vaibhav Kulkarni
University of Lausanne, Switzerland
Vaibhav.Kulkarni@unil.ch

Pierre-Adrien Ghiringhelli
University of Lausanne, Switzerland
Pierre-Adrien.Ghiringhelli@unil.ch

Bertil Chapuis
University of Lausanne, Switzerland
Bertil.Chapuis@unil.ch

Kévin Huguenin
University of Lausanne, Switzerland
Kevin.Huguenin@unil.ch

Benoît Garbinato
University of Lausanne, Switzerland
Benoit.Garbinato@unil.ch

ABSTRACT

Rich human mobility datasets are fundamental for evaluating algorithms pertaining to geographic information systems. Unfortunately, existing mobility datasets–that are available to the research community–are restricted to location data captured through a single sensor (typically GPS) and have a low spatiotemporal granularity. They also lack ground-truth data regarding points of interest and the associated semantic labels (e.g., "home", "work", etc.). In this paper, we present *Breadcrumbs*, a rich mobility dataset collected from multiple sensors (incl. GPS, GSM, WiFi, Bluetooth) on the smartphones of 81 individuals. In addition to sensor data, Breadcrumbs contains ground-truth data regarding people points of interest (incl. semantic labels) as well as demographic attributes, contact records, calendar events, lifestyle information, and social relationship labels between the participants of the study. We describe the data collection methodology and present a preliminary quantitative analysis of the dataset. A sanitized version of the dataset as well as the source code will be made available to the research community.

CCS CONCEPTS

• **Information systems** Spatial-temporal systems.

KEYWORDS

mobility dataset; point of interest annotations

ACM Reference Format:
Arielle Moro, Vaibhav Kulkarni, Pierre-Adrien Ghiringhelli, Bertil Chapuis, Kévin Huguenin, and Benoît Garbinato. 2019. Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359341

1 INTRODUCTION

Modeling human mobility is gaining importance as cities are experiencing growth and rapid transformations; this modeling demands a good understanding of individual mobility behaviors. Therefore, rich mobility datasets are fundamental for designing and evaluating algorithms pertaining to human-related geographic information systems (GIS) and for facilitating experimental reproducibility. Their availability have spurred different complex problems around the mobility domain, such as predictive queries [9], object tracking [21], trajectory indexing [4], mobility modeling [1], and location privacy [19].

As detailed in Table 1, many mobility datasets have already been made available to the research community (e.g., [13, 16, 17, 22, 23, 25]). Unfortunately, these datasets have several limitations, which include: (1) the lack of location data and related information captured from multiple sensors; (2) the unavailability of location data at a high spatiotemporal granularity throughout the data collection; (3) the lack of ground-truth information regarding participant points of interests (POI); (4) the unavailability of semantic information regarding POIs. For example, despite the proliferation of smartphones equipped with multiple sensors, datasets such as [17, 23, 24] are restricted to location data derived from either GPS, GSM, WiFi or Bluetooth. Gaining access to high granularity multi-sensor location data can lead to richer comparative and compositional studies [16]. Another example relates to the lack of ground-truth and semantic information in existing datasets. This information is crucial for research domains such as social network pattern mining [7, 10], as it is the only credible way to validate certain semantical results.

In this paper, we introduce Breadcrumbs, a rich mobility dataset that contains high-granularity data from GPS, WiFi, Bluetooth and accelerometer sensors from 81 individuals in Lausanne (Switzerland) for a period of 12 weeks that spanned between March and June 2018. This novel dataset addresses the limitations of the aforementioned datasets: it is enriched with POIs ground-truth annotations (incl. semantic labels), demographic attributes, social relationships, health information, mobility information, calendar events and contact records. This information is especially important given that, in the last decade, there has been an increasing demand to understand the behavior of individuals in multiple domains [15]. In the following sections, we describe the data collection methodology and present a preliminary quantitative analysis of the dataset. A sanitized version of the dataset and the source code will be made available to the research community at https://bread-crumb.github.io.

2 DATA COLLECTION METHODOLOGY

In order to build the Breadcrumbs dataset, we organized a data collection campaign in Lausanne in the spring of 2018. We recruited participants through a specialized unit called Labex at the University of Lausanne, which manages a pool of around 8,000 individuals

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359341

Dataset	Collection / Publication	#Participants	Duration	#Events	Sampling	Location	GPS	Check-ins	GSM	Wifi	Bluetooth	Annotation
GeoLife (Zheng et al. [25])	2007-2012 / 2012	182	5.5 years	25M	5 sec	Beijing, CN	✓	✗	✗	✗	✗	✗
MDC (Kiukkonen et al. [13])	2009-2011 / 2012	185	3 years	11M	-	Lausanne, CH	✓	✗	✓	✓	✓	relationships
Privamov (Mokhtar et al. [16])	2014-2016 / 2017	100	15 months	15M	-	Lyon, FR	✓	✗	✓	✓	✗	✗
Reality Mining (Pentland [17])	2004 / 2009	100	9 months	5M	-	Boston, US	✗	✗	✗	✗	✓	relationships
FourSquare (Yang et al. [23])	2011-2012 / 2013	3112	10 months	9M	-	New York, US	✗	✓	✗	✗	✗	relationships
blebeacon (Sikeridis et al. [20])	2016 / 2018	46	1 month	5M	-	California, US	✗	✗	✗	✗	✓	✗
hyccups (Ciobanu and Dobre [8])	2012 / 2016	72	63 days	-	-	Bucharest, RO	✗	✗	✗	✓	✗	relationships
sigcomm2009 (Pietilainen and Diot [18])	2009 / 2012	76	2 days	-	120 sec	Barcelona, ES	✗	✗	✗	✓	✓	✗
telefonica (Bogomolov et al. [3])	2013 / 2014	342	4 weeks	-	-	ES	✗	✗	✓	✗	✗	✗
ParticipAct (Chessa et al. [6])	2013-2015 / 2017	300	1 year	-	-	Bologna, IT	✓	✗	✗	✓	✓	✗
Nodobo (Bell et al. [2])	? / 2011	27	4 months	5M	-	Glasgow, GB	✗	✗	✓	✓	✗	✗
d4d challenge (Furletti et al. [11])	2016 / 2016	9M	1 year	-	-	SN	✗	✗	✓	✗	✗	✗
Gowalla (Cho et al. [7])	2008-2010 / 2011	196,591	1.5 years	6M	-	Worldwide	✗	✓	✗	✗	✗	relationships
Brightkite (Chessa et al. [6])	2008-2010 / 2010	58,228	1.5 years	4M	-	Worldwide	✗	✓	✗	✗	✗	relationships
Breadcrumbs	2018 / 2019	81	12 weeks	14M	50 sec	Lausanne, CH	✓	✗	✗	✓	✓	ground-truth semantic labels relationships

Table 1: Comparative summary of popular mobility datasets available to the community (GPS / Check-ins / GSM / Wifi / Bluetooth).

(mostly students) who registered for behavioral experiments. We contacted them by e-mail; those who were interested had to fill a short questionnaire (i.e., a screener) in order to verify their eligibility for the experiment. The main criterion was to have an iPhone with a recent version of iOS ($\geq$11.2.6) and to use it as their main phone. Eligible participants had to sign a consent form. Then, they had to install a mobile application (developed by us) on their smartphones and to keep it installed and running during the whole experiment.

The system architecture for collecting the data is presented in Figure 1. The sampling (periodic vs. motion-based) and upload (e.g., GSM vs. WiFi) strategies were carefully calibrated so that the impact on the battery life was acceptable, i.e., the battery life of the phone should be at least one day for a normal usage in the best case scenario with a recent model of iPhone. We put in place a number of mechanisms (e.g., backup, replication, notifications) to ensure a reliable and steady collection of data. The mobile application collected data from various sensors: GPS location, WiFi scans (i.e., neighboring SSIDs) and Bluetooth scans (i.e., neighboring UUIDs), and acceleration. The collected data was pre-processed directly on smartphones, for privacy reasons, and then uploaded to our backend where it was stored in a persistent database (see Figure 2 for the complete schema).

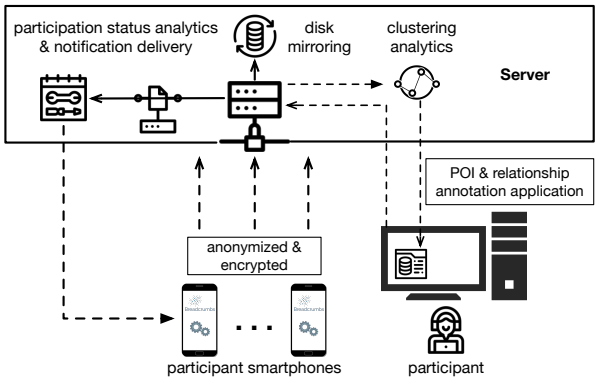


Figure 1: System architecture of the Breadcrumbs data collection.

In the middle of the experiment, we sent a questionnaire to each participant of the study in order to collect demographic (gender, age, etc.) and lifestyle (sport activities, smocking habits, transportation mode preferences, etc.) information. At the end of the experiment, participants had to fill an exit questionnaire in order for us to collect ground-truth data regarding their POIs (incl. semantic labels) and relationship information (e.g., friendship with other participants). To collect the ground-truth, we first extracted points of interest from their full mobility traces (i.e., over the whole experiment). We tested and compared four different clustering algorithms based on the MDC [13] dataset (same region as Breadcrumbs) and on the Geolife [25] dataset: (1) DJ Cluster [26], (2) DT Cluster [5], (3) TD Cluster [12] and (4) Capstone [14], which operates without parameters. Our selection criteria included the number of returned POIs, the minimum distance between distinct POIs, and the number of parameters. We selected DT Cluster [5] and further processed the returned POIs by merging overlapping POIs (a POI consists of a point on the map and a radius) and removing those that the participants visited less than 3 times over the course of the whole experiment. Each participant was shown the POIs resulting from the analysis of her/his mobility trace, then had to validate/invalidate each of them and to annotate each valid one with a semantic label. The set of possible labels was predefined; it contained the following nine categories: transport, study, residency, work, sustenance, shopping, sports, leisure and other (free-text).

The participants were compensated for their participation with CHF 100 ($\sim$USD 100) in cash, which they received at the very end of the experiment. The experiment was approved by the ethical committee of our institution.

3 QUANTITATIVE ANALYSIS

In this section, we report on our preliminary quantitative analysis of the Breadcrumbs dataset and present the different feature sets, alongside with the associated descriptive statistics. The Breadcrumbs dataset contains 34,080,964 records of GPS, WiFi and Bluetooth data points. The aggregate distance travelled by the participants amounts to 548,210 km, and the average distance travelled per participant is 6768 ± 4336 km. We collected the geospatial coordinates at an average of 79 ± 36 points per hour for each participant. The WiFi scans amount to 105 ± 49 SSIDs per hour per participant and the Bluetooth scans result in 7 ± 12 device UUIDs per hour for each participant. Additionally, each participant had an average of 280 ± 183 unique contacts in their contact list.

Table 2 shows the total number of records collected by the different sensors as well as the minimum, the median, the average, the standard deviation and the maximum of records per user. The

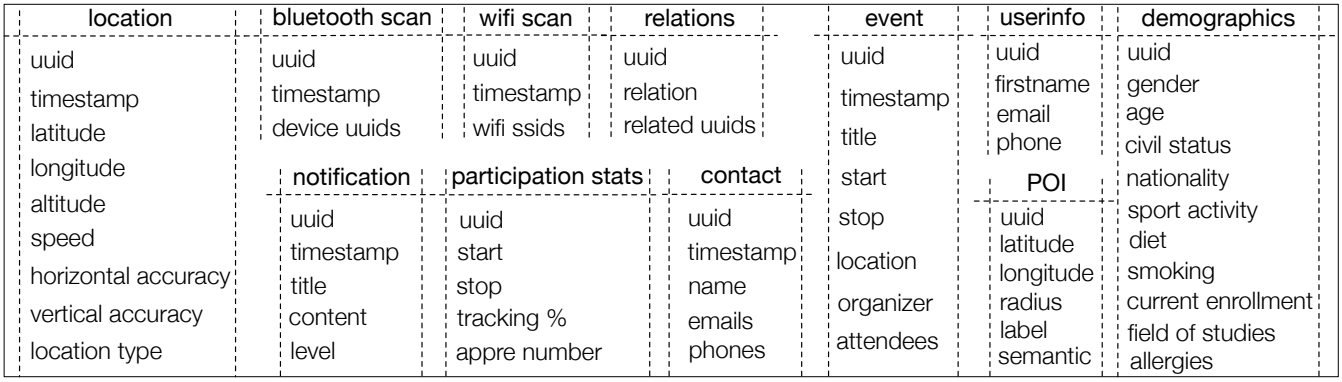
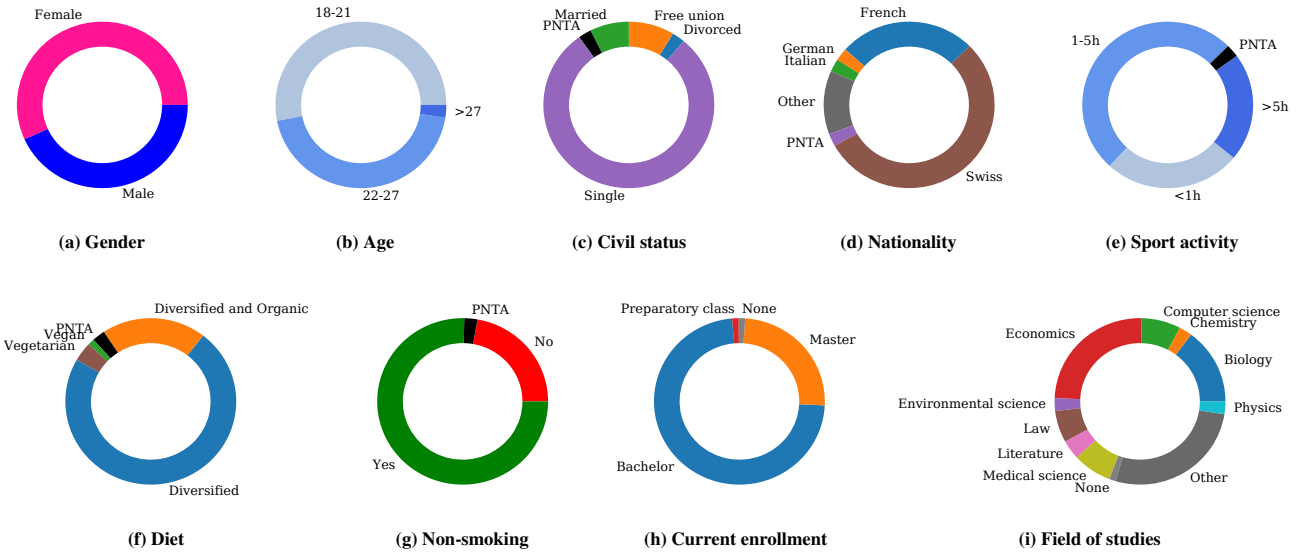


Figure 2: Database schema of the Breadcrumbs dataset.



(a) Gender **(b) Age** **(c) Civil status** **(d) Nationality** **(e) Sport activity**

(f) Diet **(g) Non-smoking** **(h) Current enrollment** **(i) Field of studies**

Figure 3: Demographics of the Breadcrumbs dataset (PNTA means "prefer not to answer").

summary of the GPS location data is presented in Table 3. The horizontal and the vertical accuracy is reported by the *Core Location API* provided by Apple.

Regarding the demographics, 56.79% of the participants identified as females, as shown in Figure 3a. The largest age groups present in the campaign are 18-21 and 22-27, with 53.09% and 44.44% respectively, as depicted in Figure 3b. In Figure 3c, the most represented civil status group is the "Single" category, i.e., 79.01%. The two most important nationality groups are "Swiss" and "French", 54.32% and 25.93% respectively, as indicated in Figure 3d. In terms of sport activities, 25.93% of the participants do sport exercises less than one hour per week, 50.62% between one and five hours per week and 20.99% more than 5 hours (see Figure 3e). Figure 3f and Figure 3g show that 72.84% of the participants have a diversified diet and 75.31% are not smoking. Figure 3h indicates that 72.84% participants were enrolled in a bachelor's degree program and 24.69% in a master's degree program. Finally, we observe that most of the participants are studying economics and biology, 24.69% and 14.81% respectively, as seen in Figure 3i.

Type	#Records	Min/usr	Median/usr	Avg./usr	STD/usr	Max/usr
GPS	13,903,934	22,418	168,050	171,654	7820	469,298
WiFi	18,669,063	15,888	234,550	230,482	107,482	426,885
Bluetooth	51,424	0	93	704	1063	5803
Accelerometer	11,661,738	17,759	131,177	143,972	71,364	415,666

Table 2: Number of data points and ratio per user.

Variable	Q05	Median	Avg.	STD	Q95
Longitude	3.962	6.589	6.618	4.509	8.465
Latitude	44.040	46.520	46.238	1.997	47.407
Altitude	64.583	415.500	465.858	557.575	753.903
Speed	0.001	9.690	13.455	16.965	35.390
Horizontal accuracy	5.000	12.000	70.792	1210.320	200.000
Vertical accuracy	3.000	6.000	14.842	111.470	29.714

Table 3: Descriptive statistics of the GPS data points.

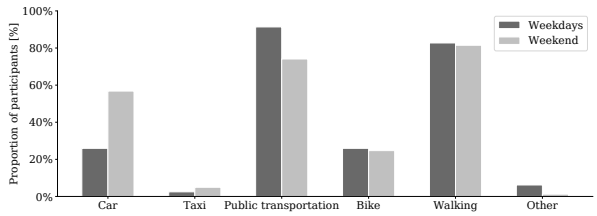


Figure 4: Transportation mode preferences for weekdays and weekend.

Figure 4 shows the transportation modes utilized during weekdays and weekend by the participants. We observe an increase in the usage of private transportation modes (cars) during the weekend as compared to the weekdays. However, walking and biking habits look similar during the weekdays and the weekend. As shown in Figure 5, the majority of the POIs correspond to the transport, study and residency semantic labels (top-level categories).

4 CONCLUSION

In this paper, we have introduced Breadcrumbs, a rich mobility dataset. In addition to demographic attributes, contacts, calendar records and social relationships, we have provided the semantic labels and the ground-truth for the points of interest. We have described the complete data-collection process and our methodology to collect ground-truth information. Our qualitative analysis sheds light on several aspects of this dataset, including the POI distribution. A sanitized version of the dataset as well as the source code will be made available to the research community at https://bread-crumb.github.io to facilitate and advance GIS research. This new dataset opens plenty of promising research avenues, such as the combination of sensor data (GPS, Wifi, Bluetooth, etc.) with demographic data, and the possibility to validate research results with a ground-truth.

ACKNOWLEDGMENTS

We thank the HEC-Labex team for their help during all the steps of the data-collection campaign. This research work was partially supported by the Business Information Systems and Architecture (BISA) research laboratory and the Faculty of Business and Economics (HEC Lausanne) at the University of Lausanne and by the Swiss National Science Foundation with grant #157160.

REFERENCES

[1] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207.
[2] Stephen Bell, Alisdair McDiarmid, and James Irvine. 2011. Nodobo: Mobile phone as a software sensor for social network research. In *Proc. of VTC*.
[3] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proc. of ICMI*.
[4] V. Prasad Chakka, Adam Everspaugh, and Jignesh M. Patel. 2003. Indexing Large Trajectory Data Sets With SETI. In *Proc. of CIDR*.
[5] Yixin Chen and Li Tu. 2007. Density-based clustering for real-time stream data. In *Proc. of KDD*.
[6] Stefano Chessa, Michele Girolami, Luca Foschini, Raffaele Ianniello, Antonio Corradi, and Paolo Bellavista. 2017. Mobile crowd sensing management with the ParticipAct living lab. *Pervasive and Mobile Computing* 38 (2017).
[7] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proc. of KDD*.
[8] Radu I. Ciobanu and Ciprian Dobre. 2016. CRAWDAD dataset upb/hyccups (v. 2016-10-17). Downloaded from https://crawdad.org/upb/hyccups/20161017. https://doi.org/10.15783/C7TG7K

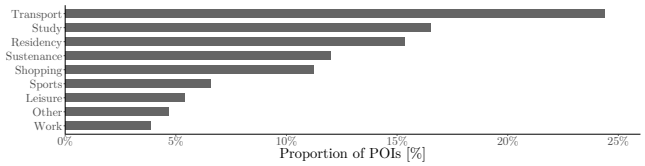


Figure 5: Distribution of POIs according to their semantic labels.

[9] Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proc. of SIGIR*.
[10] Nathan Eagle and Alex Pentland. 2005. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10 (2005), 255–268.
[11] Barbara Furletti, Roberto Trasarti, Paolo Cintia, and Lorenzo Gabrielli. 2017. Discovering and understanding city events with big data: the case of rome. *Information* 8, 3 (2017), 74.
[12] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show me how you move and I will tell you who you are. In *Proc. of SIGSPATIAL Workshop SPRINGL*.
[13] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha K. Laurila. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign.
[14] Vaibhav Kulkarni, Arielle Moro, Bertil Chapuis, and Benoît Garbinato. 2017. Extracting Hotspots Without A-priori by Enabling Signal Processing over Geospatial Data. In *Proc. of SIGSPATIAL*.
[15] Zhenhui Li, Bolin Ding, Jiawei Han, Roland Kays, and Peter Nye. 2010. Mining periodic behaviors for moving objects. In *Proc. of KDD*.
[16] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D 'Alu, Vincent Primault, Patrice Raveneau, Hervé Rivano, and Razvan Stanica. 2017. PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets.
[17] Alex Pentland. 2009. Reality mining of mobile communications: Toward a new deal on data. *The Global Information Technology Report 2008–2009* 1981 (2009).
[18] Anna-Kaisa Pietilainen and Christophe Diot. 2012. CRAWDAD dataset thlab/sigcomm2009 (v. 2012-07-15). Downloaded from https://crawdad.org/thlab/sigcomm2009/20120715. https://doi.org/10.15783/C70P42
[19] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 23. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys & Tutorials* 21, 3 (23), 2772–2793. https://doi.org/10.1109/COMST.2018.2873950
[20] Dimitrios Sikeridis, Ioannis Papapanagiotou, and Michael Devetsikiotis. 2019. CRAWDAD dataset unm/blebeacon (v. 2019-03-12). Downloaded from https://crawdad.org/unm/blebeacon/20190312.
[21] Chieh-Chih Wang, Charles E. Thorpe, Sebastian Thrun, Martial Hebert, and Hugh F. Durrant-Whyte. 2007. Simultaneous Localization, Mapping and Moving Object Tracking. *I. J. Robotics Res.* 26 (2007).
[22] Xiao-Yong Yan, Xiao-Pu Han, Bing-Hong Wang, and Tao Zhou. 2013. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Nature Scientific reports* 3 (2013).
[23] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. 2013. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs. In *Proc. of UbiComp*.
[24] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2015).
[25] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.* 33 (2010).
[26] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. 2004. Discovering personal gazetteers: an interactive clustering approach. In *Proc. of GIS Workshops*.

Anexo G: Integração dos POIs do Foursquare, categorização genérica e contabilização

1) Importar todos os POIs (mais recentes) do Foursquare

```
create table foursquare_pois (venue_id varchar(255), latitude float8, longitude float8,
venue_category varchar(255), country_code varchar(255))
```

2) Criar uma coluna geométrica da tabela de POIs

```
select addgeometrycolumn ('foursquare_pois', 'geom', 4326, 'POINT', 2);
update foursquare_pois set geom = st_setsrid(ST_MakePoint(longitude, latitude), 4326);
```

3) Criar um índice espacial sobre esta coluna geométrica

```
CREATE INDEX geom_index_foursquare ON foursquare_pois USING GIST (geom);
```

4) Fazer a seleção dos POIs (e apenas estes) que estão à volta das paragens num raio de 500 metros (criar uma tabela só com estes)

```
copy (select distinct (latitude, longitude, venue_category, country_code), start_location_latitude,
start_location_longitude, latitude, longitude, venue_category, country_code
from dataset_all_6_5, foursquare_pois
where ST_DWithin(dataset_all_6_5.geom, foursquare_pois.geom, 0.0045)
order by start_location_latitude, start_location_longitude) to
'/Users/ribeiro/Desktop/dataset_foursquare_all_6_5.csv' delimiter ';' header csv
```

```
create table dataset_foursquare_all_6_5 (one_stop varchar(294967), start_location_latitude float8,
start_location_longitude float8, latitude float8, longitude float8, venue_category varchar(255),
country_code varchar(255))
```

5) Importar as tabelas de taxonomy/categorias do Foursquare e gerar a categorizarão do foursquare em falta (missing)

```
create table foursquare_taxonomy_missing (id_taxonomy varchar(255), category_1 varchar(255),
pluralName varchar(255), shortName varchar(255), categories_id varchar(255), category_2
varchar(255), categories_pluralName varchar(255), categories_shortName varchar(255),
categories_categories_id varchar(255), category_3 varchar(255),
categories_categories_pluralName varchar(255), categories_categories_shortName varchar(255))
```

```
copy (select distinct(foursquare_taxonomy_missing_all.category_3),
foursquare_taxonomy_missing_all.category_1, foursquare_taxonomy_missing_all.category_2
from foursquare_taxonomy_missing_all, foursquare_taxonomy
where foursquare_taxonomy_missing_all.category_2 <> foursquare_taxonomy.category_2 and
foursquare_taxonomy_missing_all.category_3 <> foursquare_taxonomy.category_3
```

```
order by foursquare_taxonomy_missing_all.category_1,
foursquare_taxonomy_missing_all.category_2, foursquare_taxonomy_missing_all.category_3) to
'/Users/ribeiro/Desktop/foursquare_taxonomy_missing.csv' delimiter ';' header csv

create table foursquare_taxonomy_missing (category_1 varchar(255), category_2 varchar(255),
category_3 varchar(255))
```

6) Nestes poucos POIs ver as suas categorias genéricas e contar para cada paragem o nº de categorias genéricas (10) que estão à volta e assim gerar 10 features no dataset (taxonomy e taxonomy_missing)

```
copy (select distinct (start_location_latitude, start_location_longitude, latitude, longitude,
venue_category, country_code),
start_location_latitude, start_location_longitude, latitude, longitude, venue_category,
country_code, category_1
from dataset_foursquare_all_6_5, foursquare_category_all
where venue_category = category_1 or
venue_category = category_2 or
venue_category = category_3 or
venue_category = category_4
order by start_location_latitude, start_location_longitude) to
'/Users/ribeiro/Desktop/dataset_foursquare_category_all_6_5.csv' delimiter ';' header csv

create table dataset_foursquare_category_all_6_5 (one_stop varchar(294967),
start_location_latitude float8, start_location_longitude float8, latitude float8, longitude float8,
venue_category varchar(255), country_code varchar(255), category_1 varchar(255))

copy (select start_location_latitude, start_location_longitude, category_1, count(category_1) as
count_category_1
from dataset_foursquare_category_all_6_5
group by start_location_latitude, start_location_longitude, category_1) to
'/Users/ribeiro/Desktop/count_category_1.csv' delimiter ';' header csv

create table count_category_1 (start_location_latitude float8, start_location_longitude float8,
category_1 varchar(255), count_category_1 integer)
```

7) De seguida gerar as várias tabelas com cada coluna da categoria_1 (genérica) e a sua contabilização em cada start_location. As categorias que não existirem a contagem é 0

```
copy (select distinct (dataset_all_6_5.start_location_latitude,
dataset_all_6_5.start_location_longitude), dataset_all_6_5.start_location_latitude,
dataset_all_6_5.start_location_longitude,
arts_and_entertainment, business_and_professional_services, community_and_government,
dining_and_drinking, evento,
health_and_medicine, landmarks_and_outdoors, residence, retail, sports_and_recreation,
travel_and_transportation
from dataset_all_6_5
left join count_category_arts_and_entertainment
```

```
on dataset_all_6_5.start_location_latitude =
count_category_arts_and_entertainment.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_arts_and_entertainment.start_location_longitude
left join count_category_business_and_professional_services
on dataset_all_6_5.start_location_latitude =
count_category_business_and_professional_services.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_business_and_professional_services.start_location_longitude
left join count_category_community_and_government
on dataset_all_6_5.start_location_latitude =
count_category_community_and_government.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_community_and_government.start_location_longitude
left join count_category_dining_and_drinking
on dataset_all_6_5.start_location_latitude =
count_category_dining_and_drinking.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_dining_and_drinking.start_location_longitude
left join count_category_event
on dataset_all_6_5.start_location_latitude = count_category_event.start_location_latitude and
dataset_all_6_5.start_location_longitude = count_category_event.start_location_longitude
left join count_category_health_and_medicine
on dataset_all_6_5.start_location_latitude =
count_category_health_and_medicine.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_health_and_medicine.start_location_longitude
left join count_category_landmarks_and_outdoors
on dataset_all_6_5.start_location_latitude =
count_category_landmarks_and_outdoors.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_landmarks_and_outdoors.start_location_longitude
left join count_category_residence
on dataset_all_6_5.start_location_latitude = count_category_residence.start_location_latitude and
dataset_all_6_5.start_location_longitude = count_category_residence.start_location_longitude
left join count_category_retail
on dataset_all_6_5.start_location_latitude = count_category_retail.start_location_latitude and
dataset_all_6_5.start_location_longitude = count_category_retail.start_location_longitude
left join count_category_sports_and_recreation
on dataset_all_6_5.start_location_latitude =
count_category_sports_and_recreation.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_sports_and_recreation.start_location_longitude
left join count_category_travel_and_transportation
on dataset_all_6_5.start_location_latitude =
count_category_travel_and_transportation.start_location_latitude and
dataset_all_6_5.start_location_longitude =
count_category_travel_and_transportation.start_location_longitude) to
'/Users/ribeiro/Desktop/start_location_category_generic.csv' delimiter ';' header csv
```

```
create table start_location_category_generic (one_stop varchar(294967), start_location_latitude
float8, start_location_longitude float8, arts_and_entertainment varchar(255),
business_and_professional_services varchar(255), community_and_government varchar(255),
dining_and_drinking varchar(255), evento varchar(255),  health_and_medicine varchar(255),
landmarks_and_outdoors varchar(255), residence varchar(255), retail varchar(255),
sports_and_recreation varchar(255), travel_and_transportation varchar(255))

ALTER TABLE start_location_category_generic ALTER COLUMN arts_and_entertainment
TYPE integer USING (arts_and_entertainment::integer);

update start_location_category_generic set arts_and_entertainment = 0
where arts_and_entertainment is NULL
```

8) Adicionar as colunas das categorias principais

```
copy (select user_id, dataset_foursquare_all_6_5.start_location_latitude,
dataset_foursquare_all_6_5.start_location_longitude, duracao_hours,
dist_metros_to_work_university, start_time_hours_day, walk_percentage_stop,
dist_metros_to_home, grupo_idade, mean_duration, escolaridade,
occurrences_per_day, percent_week_day_sun, percent_week_day_mon, percent_week_day_tue,
percent_week_day_wed, percent_week_day_thu,
percent_week_day_fri, percent_week_day_sat, standart_deviation_duration, genero,
estado_civil, day_week, working_profile, walk_percentage_trip,
date_year, poi_description, latitude, longitude, venue_category, country_code,
arts_and_entertainment, business_and_professional_services, community_and_government,
dining_and_drinking, evento,
health_and_medicine, landmarks_and_outdoors, residence, retail, sports_and_recreation,
travel_and_transportation
from start_location_category_generic, dataset_all_6_5, dataset_foursquare_all_6_5
where start_location_category_generic.start_location_latitude =
dataset_all_6_5.start_location_latitude and
start_location_category_generic.start_location_longitude =
dataset_all_6_5.start_location_longitude and
start_location_category_generic.start_location_latitude =
dataset_foursquare_all_6_5.start_location_latitude and
start_location_category_generic.start_location_longitude =
dataset_foursquare_all_6_5.start_location_longitude and
dataset_all_6_5.start_location_latitude = dataset_foursquare_all_6_5.start_location_latitude and
dataset_all_6_5.start_location_longitude = dataset_foursquare_all_6_5.start_location_longitude
order by user_id, dataset_foursquare_all_6_5.start_location_latitude,
dataset_foursquare_all_6_5.start_location_longitude) to
'/Users/ribeiro/Desktop/dataset_location_foursquare_category_all_6_5.csv' delimiter ';' header
csv

create table dataset_location_foursquare_category_all_6_5 (user_id integer,
start_location_latitude float8, start_location_longitude float8, duracao_hours double precision,
```

```
dist_metros_to_work_university double precision, start_time_hours_day double precision,
walk_percentage_stop double precision, dist_metros_to_home double precision, grupo_idade
varchar(255), mean_duration double precision, escolaridade  varchar(255),
occurrences_per_day double precision, percent_week_day_sun double precision,
percent_week_day_mon double precision, percent_week_day_tue double precision,
percent_week_day_wed double precision, percent_week_day_thu double precision,
percent_week_day_fri double precision, percent_week_day_sat double precision,
standart_deviation_duration double precision, genero varchar(255), estado_civil varchar(255),
day_week varchar(255), working_profile varchar(255), walk_percentage_trip double precision,
date_year date, poi_description integer, latitude float8, longitude float8, venue_category
varchar(255), country_code varchar(255), arts_and_entertainment integer,
business_and_professional_services integer, community_and_government integer,
dining_and_drinking integer, evento integer,
health_and_medicine integer, landmarks_and_outdoors integer, residence integer, retail integer,
sports_and_recreation integer, travel_and_transportation integer)

copy (select user_id, dataset_all_6_5.start_location_latitude,
dataset_all_6_5.start_location_longitude, duracao_hours,
dist_metros_to_work_university, start_time_hours_day, walk_percentage_stop,
dist_metros_to_home, grupo_idade, mean_duration, escolaridade,
occurrences_per_day, percent_week_day_sun, percent_week_day_mon, percent_week_day_tue,
percent_week_day_wed, percent_week_day_thu,
percent_week_day_fri, percent_week_day_sat, standart_deviation_duration, genero,
estado_civil, day_week, working_profile, walk_percentage_trip,
date_year, poi_description, arts_and_entertainment, business_and_professional_services,
community_and_government, dining_and_drinking, evento,
health_and_medicine, landmarks_and_outdoors, residence, retail, sports_and_recreation,
travel_and_transportation
from start_location_category_generic, dataset_all_6_5
where start_location_category_generic.start_location_latitude =
dataset_all_6_5.start_location_latitude and
start_location_category_generic.start_location_longitude =
dataset_all_6_5.start_location_longitude
order by user_id, dataset_all_6_5.start_location_latitude,
dataset_all_6_5.start_location_longitude) to
'/Users/ribeiro/Desktop/dataset_location_category_all_6_5.csv' delimiter ';' header csv

create table dataset_location_category_all_6_5 (user_id integer, start_location_latitude float8,
start_location_longitude float8, duracao_hours double precision,
dist_metros_to_work_university double precision, start_time_hours_day double precision,
walk_percentage_stop double precision, dist_metros_to_home double precision, grupo_idade
varchar(255), mean_duration double precision, escolaridade  varchar(255),
occurrences_per_day double precision, percent_week_day_sun double precision,
percent_week_day_mon double precision, percent_week_day_tue double precision,
percent_week_day_wed double precision, percent_week_day_thu double precision,
percent_week_day_fri double precision, percent_week_day_sat double precision,
standart_deviation_duration double precision, genero varchar(255), estado_civil varchar(255),
day_week varchar(255), working_profile varchar(255), walk_percentage_trip double precision,
```

date_year date, poi_description integer, arts_and_entertainment integer,
business_and_professional_services integer, community_and_government integer,
dining_and_drinking integer, evento integer,
health_and_medicine integer, landmarks_and_outdoors integer, residence integer, retail integer,
sports_and_recreation integer, travel_and_transportation integer)

Anexo H: Queries SQL usadas na recolha de *features*

1) Identificação dos pontos de locais de interesse de cada paragem num determinado raio

```
select poi_description_all.user_id, poi_latitude, poi_longitude, start_location_latitude,
start_location_longitude, poi_description_all.description, start_paragem, stop_paragem, duracao
from poi_description_all, start_stop_paragem_location
where ST_DWithin(ST_MakePoint(start_stop_paragem_location.start_location_longitude,
start_stop_paragem_location.start_location_latitude)::geography, ST_MakePoint(poi_longitude,
poi_latitude)::geography, raio) and poi_description_all.user_id =
start_stop_paragem_location.user_id
```

Tabela "poi_description_all": incluí os dados da tabela "point_of_interest" e "point_of_interest_description" do *dataset* Breadcrumbs com o intuito de identificar o nome da atividade dos pontos de locais de interesse.
Tabela "start_stop_paragem_location": corresponde à identificação do início e fim das paragens efetuadas pelos utilizadores do *dataset* Breadcrumbs, ou seja, ao momento de realização da atividade.

2) Conversão do timestamp Unix Time Stamp para Date time

```
alter table nome_tabela add nome_coluna timestamp not null default '2021-01-01 00:00:00';
update nome_tabela set nome_coluna = to_timestamp(location.timestamp);
```

3) Contagem do número de POIs do *dataset* Breadcrumbs a cada paragem

```
select user_id, poi_latitude, poi_longitude, start_location_latitude, start_location_longitude,
duracao, count(poi_description_all) as count_poi_description
from poi_paragem_description_all_6_5
group by user_id, poi_latitude, poi_longitude, start_location_latitude, start_location_longitude,
duracao
order by count_poi_description, user_id
```

Tabela "poi_paragem_description_all_6_5": inclui os POIs do *dataset* Breadcrumbs e as respetivas paragens efetuadas pelos diferentes utilizadores como a duração da sua realização

4) Cálculo da distância de cada paragem ao trabalho e à residência mais próxima para cada utilizador

```
min(st_distance(ST_GeogFromText('SRID=4326;POINT(' ||
poi_paragem_description_all_6_5.start_location_longitude || ' ' ||
poi_paragem_description_all_6_5.start_location_latitude || ')'),
ST_GeogFromText('SRID=4326;POINT(' ||  poi_longitude_work || ' ' || poi_latitude_work ||
')')))
min(st_distance(ST_GeogFromText('SRID=4326;POINT(' ||
poi_paragem_description_all_6_5.start_location_longitude || ' ' ||
poi_paragem_description_all_6_5.start_location_latitude || ')'),
```

ST_GeogFromText('SRID=4326;POINT(' || poi_longitude_home || ' ' || poi_latitude_home || ')')))

5) Query para formalizar o *dataset* com todas as *features* das atividades e *features* pessoais, exceto as *features* de *clustering*

```
select distwork_disthome_6_5.user_id, distwork_disthome_6_5.start_location_latitude,
distwork_disthome_6_5.start_location_longitude, trunc(extract(epoch from
poi_paragem_description_all_6_5.duracao)/3600, 2) AS duracao_hours,
trunc(distwork_disthome_6_5.dist_metros_to_work::numeric, 2) AS dist_metros_to_work,
trunc(extract(epoch from distwork_disthome_6_5.start_paragem::time)/3600, 2) AS
start_time_hours_day, trunc(distwork_disthome_6_5.dist_metros_to_home::numeric, 2) AS
dist_metros_to_home, info_user_id.grupo_idade, info_user_id.escolaridade,
info_user_id.genero, info_user_id.estado_civil, extract(dow from
date(distwork_disthome_6_5.start_paragem)) AS day_week,
date(distwork_disthome_6_5.start_paragem) AS date_year,
distwork_disthome_6_5.poi_description
from distwork_disthome_6_5, poi_paragem_description_all_6_5, info_user_id
where distwork_disthome_6_5.user_id = poi_paragem_description_all_6_5.user_id and
distwork_disthome_6_5.start_location_latitude =
poi_paragem_description_all_6_5.start_location_latitude and
distwork_disthome_6_5.start_location_longitude =
poi_paragem_description_all_6_5.start_location_longitude and
distwork_disthome_6_5.start_paragem = poi_paragem_description_all_6_5.start_paragem and
distwork_disthome_6_5.stop_paragem = poi_paragem_description_all_6_5.stop_paragem and
distwork_disthome_6_5.poi_description = poi_paragem_description_all_6_5.poi_description_all
and distwork_disthome_6_5.user_id = info_user_id.user_id
```

Tabela "distwork_disthome_6_5": incluí os dados das distâncias das paragens dos utilizadores ao local de trabalho e local residencial.

6) Cálculo da velocidade média feita durante as atividades

```
select start_stop_paragem_location.user_id, avg(speed), start_location_latitude,
start_location_longitude
from start_stop_paragem_location, localizacao
where start_stop_paragem_location.user_id = localizacao.user_id and
location_datetime between start_paragem and stop_paragem
group by start_stop_paragem_location.user_id, start_location_latitude, start_location_longitude
order by start_stop_paragem_location.user_id
```

A tablea localizacao corresponde à tabela "location" do *dataset* Breadrumbs

7) Conceção da geometria dos POIs do Foursquare

```
select addgeometrycolumn ('foursquare_pois', 'geom', 4326, 'POINT', 2);
update foursquare_pois set geom = st_setsrid(ST_MakePoint(longitude, latitude), 4326);
```

8) Confirmar se a distância de 0.0045 graus correspondentes a uma distância de 500 metros de um dos pontos de locais de interesse do *dataset* Breadcrumbs

```
SELECT ST_Distance(
        'SRID=4326;POINT(46.512019 6.618055)'::geography,
        'SRID=4326;POINT(46.517519 6.618055)'::geography
        );

SELECT ST_Distance(
        'SRID=4326;POINT(46.512019 6.618055)'::geography,
        'SRID=4326;POINT(46.512019 6.622555)'::geography
        );
```