Mestrado em Engenharia Informática

Dissertação

Relatório Final

Murbe – Modelos de mobilidade Urbana: Inferência do modo de transporte

João Pedro Mousinho Santiago

jpsant@student.dei.uc.pt

Orientadores:

Prof. Carlos Bento

Prof.ª Ana Maria Almeida

Data: 1 de Julho de 2014

FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Resumo

Este trabalho tem como objetivo principal desenvolver um algoritmo de inferência que, com precisão, seja capaz de detetar o meio de transporte de um utilizador a partir, apenas, do seu registo de mobilidade obtido de forma ubíqua através dos dados recolhidos pelo GPS do seu *smartphone*.

A informação de localização em causa consiste em dados individuais de utilização, recolhidos usando uma aplicação *mobile* para *smartphone*, que está presentemente em desenvolvimento pelo projeto ECO-Circuitos, subprojecto do Projeto QREN Tice.Mobilidade – Sistema de serviços centrados no utilizador. Numa altura em que o meio ambiente é uma temática que ganha cada vez mais importância, é prioritário o desenvolvimento de alternativas de cariz ecológico. É com isto em mente que o ECO-Circuitos tenciona contribuir de forma a melhorar a mobilidade no dia-a-dia do utilizador. Para isso, propõe-se, através da recolha de dados do *smartphone*, traçar o perfil de mobilidade do seu proprietário e oferecer alternativas que sejam mais eficientes em geral, e mais ecológicas em particular.

Para o correto funcionamento desta aplicação *mobile*, tem de ser o menos intrusiva possível, pelo que é necessário usar um algoritmo capaz de inferir o modo de transporte utilizado em cada momento apenas utilizando os dados dos sensores do equipamento móvel. É, portanto, nesta necessidade que nasce o desenvolvimento e implementação do algoritmo de inferência que esta dissertação pretende apresentar.

Palavras-Chave: Classificação, GPS, Machine Learning, Padrão de mobilidade, Smartphone

Acrónimos

BSCR	Bus Stop Closeness Rate
FFT	Fast Fourier Transform
GPS	Global Positioning System
HCR	Heading Change Rate
QREN	Quadro de Referência Estratégico Nacional
SHCLF	Spatial Heterogeneity Constrained Levy Flight
SR	Stop Rate
VCR	Velocity Change Rate

Glossário

Dataset	Conjunto de dados
Features	Atributo que caracteriza os elementos do conjunto de dados
Smartphone	Dispositivo que possuí um sistema operacional móvel com uma capacidade computacional superior aos telemóveis tradicionais.

Lista de Figuras

Lista de Tabelas

Índice

Capítulo 1

Introdução

No contexto da mobilidade urbana, compreender qual o tipo de mobilidade do utilizador impõe-se como um papel essencial para um planeamento eficiente e sustentável. Até há algum tempo, a recolha dos dados necessários para traçar o perfil do utilizador com base no seu tipo de mobilidade só era possível através de questionários em papel [1]. Com o aparecimento dos *smartphones* e novas tecnologias como a dos sensores de monitorização do movimento e aparelhos de monitorização nos automóveis, este tipo de questionário torna-se meramente uma ferramenta de validação, uma vez que a recolha dos dados pode agora ser efetuada a partir dos sensores deste dispositivo. Este novo tipo de dados e a sua análise veio possibilitar o aparecimento de algoritmos capazes de inferir os modos de transporte do utilizador, a duração da viagem, os períodos de paragem e até mesmo os trajetos das viagens realizadas contribuindo assim para a criação do perfil de mobilidade deste, como se pode verificar no estudo feito por [2].

O QREN TICE Mobilidade[1] é um projeto ambicioso que reconhece a importância da mobilidade urbana no nosso quotidiano. Encontra-se dividido em vários subprojectos, entre os quais se encontra o projeto ECO-Circuitos, que contextualiza este trabalho de investigação. Com o objetivo de fornecer alternativas individualizadas a cada indivíduo no que toca a meios de transporte, para além da personalização, pretende oferecer alternativas mais eficientes e mais ecológicas. De forma a atingir esses objetivos, a empresa copromotora *SmartMove*[2] desenvolveu a aplicação Android, denominada *mySteps*, que recolhe todos os dados dos sensores do *smartphone*, necessários para o desenvolvimento do algoritmo que permite inferir o meio de transporte do utilizador.

O trabalho proposto encontra-se dividido em duas partes distintas. Na primeira são identificados os fatores que influenciam a mobilidade urbana e individual através do levantamento do estado de arte relativamente a modelos e projetos de aferição de mobilidade individual. A segunda parte do trabalho consiste no desenvolvimento de um algoritmo que permite inferir com precisão qual o meio de transporte que foi usado na deslocação do utilizador, utilizando os dados recolhidos pelo sensor *GPS* do seu

[1] http://tice.mobilidade.ipn.pt/index.php
[2] http://www.smartmove.pt/

smartphone. Relativamente aos dados gerados pelo acelerómetro, é apresentada uma abordagem teórica que faz uso da informação de ambos os sensores para a inferência do modo de transporte.

O relatório divide-se em seis capítulos. O Capítulo 2 destina-se a apresentar os trabalhos conhecidos na área da mobilidade urbana e da deteção de meios de transporte. O Capítulo 3 são apresentados com mais detalhe os objetivos propostos por este trabalho assim como a abordagem proposta para cumprir esses objetivos. Ainda na mesma secção, é referido o trabalho que foi realizado ao longo do semestre, assim como o trabalho a realizar no segundo semestre. No Capítulo 4 encontra-se descrito o processo de recolha e tratamento de dados, da abordagem teórica desenvolvida, da implementação do algoritmo de inferência, da aplicação *Web* desenvolvida e da integração do algoritmo na aplicação móvel e na aplicação *Web*. O Capítulo 5 apresenta os testes efetuados no sentido de validar a abordagem apresentada neste trabalho, assim como a análise aos resultados obtidos. Por fim, o Capítulo 6 apresenta conclusões obtidas através do trabalho realizado e descrição do trabalho a realizar no futuro.

Capítulo 2

Estado de arte

Neste capítulo é feita uma introdução à temática da mobilidade urbana de forma a compreender melhor os conceitos envolvidos nas deslocações diárias efetuadas pelos utilizadores. Segue-se a apresentação de trabalhos efetuados na área de deteção do modo de transporte através da recolha de dados do *smartphone*.

2.1 Mobilidade Urbana

2.1.1 Tendências na aquisição de veículos motorizados

Desde 1960, tem-se vindo a verificar um aumento na aquisição de veículos motorizados privados [3]. Isto reflete-se num aumento de quilómetros automóvel viajados (*Vehicle Kilometers Travelled aka VKT,* [4]*),* na deterioração dos níveis de congestionamento e respetivo aumento do número de horas passadas a viajar tendo como consequência direta a degradação da qualidade do ar. Numa altura em que o aquecimento global se manifesta cada vez mais, esta é uma problemática a ter em conta. Ainda mais preocupante é o facto de esta tendência se verificar na maior parte dos centros urbanos tanto dos países desenvolvidos como dos países em desenvolvimento, visto que o carro é o principal responsável pela diminuição da qualidade do ar [5].

Esta problemática tem vindo a ganhar cada vez mais importância, como revela o estudo efetuado por [3], onde foram identificados os fatores que estão relacionados com a variação do *VKT*:

- Crescimento da população;
- Expansão urbana;
- Aquisição de veículos motorizados privados;
- Aproveitamento da capacidade de transporte dos veículos.

A fim de averiguar o impacto que estes tinham sobre o *VKT,* os autores do trabalho [3] realizaram um conjunto de estudos que incidiu sobre as cidades de Singapura, Hong Kong, Munich, Estocolmo, Nova Yorque, Perth e Phoenix. Podemos dividir os resultados obtidos em dois conjuntos: De um lado temos as cidades que

implementaram medidas/políticas de forma a mitigar o uso excessivo dos veículos motorizados privados. Do outro as cidades que não implementaram qualquer medida.

No conjunto das cidades que implementaram medidas encontramos as cidades asiáticas (Singapura e Hong Kong), as cidades europeias (Estocolmo e Munich) e a cidade norte americana Nova Iorque. De entre as várias medidas implementadas, apenas algumas obtiveram um efeito positivo na redução dos veículos motorizados privados. As principais medidas implementadas foram as seguintes:

- Introdução de taxas com o objetivo de desencorajar a aquisição de novos veículos;
- Restrição da circulação de veículos privados em determinadas zonas, exceto se estes tivessem mais de quatro lugares ocupados;
- Investimento na rede de transportes públicos.

Das medidas apresentadas as duas últimas foram as que se revelaram mais eficazes, pois teve como consequência direta um melhor aproveitamento da capacidade dos veículos e um incentivo para a utilização dos transportes públicos, reduzindo assim o número de quilómetros automóvel viajados. A primeira medida revelou-se mais eficaz no início da sua implementação, visto que com a melhoria da qualidade de vida das pessoas e do aumento do seu poder de compra, esta deixou de ser um obstáculo tão grande na aquisição de novos veículos motorizados [4.5.6], embora continue a ser um entrave para classes com um poder económico reduzido.

Um bom exemplo do efeito da terceira medida no *VKT* é Nova Iorque. Apesar de ser umas das cidades mais tráfego dos Estados Unidos, apresenta a média mais baixa entre as cidades americanas no que toca ao uso do automóvel privado. Dado este investimento nos transportes públicos, em 1990 o número de passageiros a frequentar os transportes públicos era 280% acima da média dos Estados Unidos da América [6] refletindo assim a influência desta medida na dinâmica urbana da cidade.

Os resultados obtidos em Perth e Phoenix traduzem o que teria acontecido caso as medidas mencionadas anteriormente não tivessem sido tomadas [6]. Não havendo nenhuma restrição, nenhum imposto extra nem qualquer tipo de investimento significativo na rede de transportes públicos, o resultado observado foi o aumento acentuado de aquisição de veículos motorizados privados e menor aproveitamento da

capacidade de transporte dos mesmos. Isto levou a um aumento no número de quilómetros viajados e por conseguinte da degradação da qualidade do ar.

2.1.2 Influência da morfologia urbana nos padrões de mobilidade

Conhecer a influência que a morfologia de diferentes cidades tem na mobilidade dos seus habitantes, é um ponto importante no projeto em que este trabalho se insere. Como tal, foi analisado o estudo apresentado por [9] para averiguar até que ponto a morfologia urbana afeta o padrão de mobilidade dos humanos. Neste trabalho foi recolhida informação das chamadas a partir de telemóveis de sete cidades do nordeste da China. A atividade individual num determinado território é medida através de um método chamado raio de giração (*radius of gyration*, [9]). Esta medida é determinada pelos lugares mais visitados pelo utilizador o que confere uma maior robustez em relação a anormalidades registadas nas viagens dos utilizadores, isto é, caso as atividades se desviem da rotina o resultado obtido pelo raio de giração não será afetado. Ainda relativamente a esta medida quanto mais elevado for o seu valor, maior será a atividade registada pelo utilizador. De seguida é aplicada uma distribuição parcial cuja aproximação é efetuada através da lei exponencial, visto que reflete melhor o decaimento da distância da mobilidade humana.

Como era de esperar, cidades maiores ou com forma irregular levam a que os seus habitantes percorram maiores distâncias, enquanto que os habitantes de cidades pequenas apresentavam menos atividade no seu dia-a-dia. Para confirmar até que ponto as conclusões estavam corretas e visto que não existem duas cidades com morfologia igual, foi efetuada uma simulação onde foram criados uma série de polígonos: De um lado polígonos com forma semelhante mas tamanhos distintos, do outro polígonos com formas distintas mas tamanho semelhante. A esta simulação deu-se o nome de *Monte Carlo* e para simular o movimento humano de forma realista foi aplicado o modelo *Spatial Heterogeneity Constrained Levy Flight*, que veio comprovar os resultados esperados e que o tamanho e a forma da cidade influenciam a mobilidade dos seus habitantes.

2.1.3 Mobilidade e os diferentes segmentos da população

Agora que se tornou clara a influência da morfologia das cidades na mobilidade é necessário verificar de que forma os diferentes segmentos da população são afetados nas suas deslocações. Com o objetivo de investigar os fatores que influenciam a distância viajada pelos habitantes das áreas urbanas do Canada, foi efetuado um estudo por [10] que teve como foco principal três cidades canadianas: Hamilton, Toronto e Montreal onde os segmentos da população avaliados foram: Idosos (adultos com idade superior a 65 anos), agregados constituídos por pais solteiros e agregados com um rendimento baixos (no caso de Montreal).

O método escolhido para efetuar a análise da distância viajada pelos participantes foi baseada numa regressão multivariada, uma técnica de análise padrão que é reforçada através de uma expansão espacial dos coeficientes tirando assim melhor partido da informação geográfica detalhada disponível. No entanto a diferença registada entre os três grupos não foi tão acentuada como o esperado nem igual em todas as cidades. No caso dos idosos estes revelam padrões de mobilidade mais limitados em Hamilton, no entanto os agregados constituídos por pais solteiros em Toronto e Montreal são os que apresentam um menor nível de mobilidade. Este estudo permite-nos concluir que segmentos mais vulneráveis da população possuem condições de mobilidade mais restringidas, em particular estes três grupos apresentam uma distância viajada menor relativamente aos outros segmentos da população, contribuindo assim para o projeto com um conhecimento mais aprofundado relativamente à mobilidade praticada pelos diferentes segmentos da população.

2.2 Mobilidade Individual

Até ao momento foram analisados trabalhos de referência na mobilidade no contexto urbano, assim como os fatores que influenciam a distância viajada por um individuo. Esta presente secção destina-se a avaliar a mobilidade individual e os fatores que a influenciam.

2.2.1 Padrão de mobilidade

É importante compreender a mobilidade dos utilizadores e os padrões que estes desenvolvem ao longo do tempo. Torna-se portanto essencial saber quais os fatores que os levam a percorrer maiores distâncias assim como os motivos dessas mesmas deslocações. A relação que existe com os colegas de trabalho, familiares, amigos de infância influência de maneira distinta a mobilidade individual. Uma pessoa estaria mais disposta a efetuar uma viagem de longa duração por um amigo de infância do que por um colega de trabalho, sendo notório o peso que a duração e a força da relação estabelecida entre as pessoas tem na escolha das viagens efetuadas [11].

Para além da rede social estabelecida existem fatores que também condicionam a mobilidade individual, pessoas com um rendimento mensal baixo usam com mais frequência os transportes públicos e a bicicleta [12] em contraste com pessoas com um rendimento superior utilizam com mais frequência táxis e veículos privados. O rendimento mensal também influência o motivo da viagem [13], quando o rendimento mensal é baixo as viagens efetuadas são maioritariamente para o local de trabalho ou para se encontrar com pessoas da sua rede social. Por outro lado um rendimento mensal alto permite realizar viagens com propósitos menos essenciais como ir a um centro comercial às compras, interações sociais ou até mesmo viagens de negócios.

Outra componente que se deve ter em conta ao analisar a mobilidade individual é a energia consumida por esta atividade. Estima-se que cerca de dois terços do consumo total de energia de transporte[3]. Realizando uma comparação entre consumos de energia por passageiro-quilómetro para diversos modos de transporte, o consumo energético automóvel é o dobro do consumo de um autocarro e dez vezes superior ao consumo de um transporte ferroviário (de superfície ou subterrâneo) [14]. Conclui-se assim, que a energia consumida por habitante devido ao modo de transporte está relacionada, em parte, com os atributos da forma urbana, sendo este consumo superior numa cidade com maior poderio financeiro, motorizada, expandida e policêntrica.

[3] https://www.iea.org/impagr/cip/archived_bulletins/issue_no23.htm

2.2.2 Modelo e Análise Multinível das redes pessoais

Após estudar o efeito que as relações estabelecidas com as pessoas que nos rodeiam nas nossas deslocações diárias, surge a necessidade de avaliar a relação entre os transeuntes e o contexto urbano que o rodeia. Para tal, recorre-se ao modelo multinível, [15] onde o que melhor se adequa ao presente projeto é o modelo multinível com dois níveis, um nível para a análise do individuo e outro para relacionar este com o contexto em que o indivíduo se insere.

O primeiro nível é o designado *ego-network* em que *ego* é uma pessoa específica. Este nível é constituído pelas características do *ego* e da estrutura da sua rede pessoal constituída pelas pessoas que fazem parte do seu quotidiano, familiares entre outros. O segundo nível denominado *ego-alter,* consiste nos indivíduos (*alte)* que possuem algum tipo de relação com o *ego,* nas características de cada *alter* e da relação que existe entre o *ego* e o *alter* (Marijtje et al., 1999). De forma a criar esta rede de relações é avaliada a proximidade e a força da ligação que o *ego* possui com os seus *alter* (Carrasco et al., 2009). Os dados necessários são adquiridos através de uma série de perguntas efetuadas ao *ego* de forma a averiguar os *alter* deste e o tipo de relação que estes possuem [16]. Embora esta metodologia seja um pouco antiquada e desgastante para os intervenientes, apresenta-se como sendo eficaz na obtenção da informação necessária para a construção do modelo.

O principal objetivo deste modelo é estudar fenómenos onde os dados possuem uma estrutura hierárquica em que não se pode assumir que possuímos observações independentes [17]. O facto de permitir separar o indivíduo do contexto urbano em que este se insere, permite avaliar melhor os fatores que influenciam a sua mobilidade ao cruzar informação entre os dois níveis, sendo esta última uma informação para o projeto ao permitir uma melhor compreensão da influência do meio urbano na mobilidade individual.

2.3 Projetos de Mobilidade Urbana

2.3.1 S.M.A.R.T – Future Urban Mobility

Projeto do programa MIT *Singapore* tem como missão desenvolver novos paradigmas para o planeamento, *design* e operação de sistemas de mobilidade do futuro. Pretende focar-se não só nos passageiros como também nos transportes de mercadorias com o objetivo de melhorar a sustentabilidade e o bem-estar da sociedade. Serão alvo destas inovações: os transportes públicos, os veículos privados, bicicletas, peões, entre outros. Este projeto encontra-se em desenvolvimento em linha com o objetivo do projeto ECO-Circuitos e, em particular, deste trabalho.

O *LIVE Singapure*! é um subprojecto do *Future Urban Mobility* e tem como objetivo oferecer às pessoas informação em tempo real relativo à situação urbana da cidade, de forma a que possam tomar decisões informadas em sintonia com o ambiente em que se encontram. O facto de permitir ao utilizador tomar decisões com base em informação apresentada em tempo real relativamente ao estado do trânsito, horários dos transportes públicos, permite a escolha da melhor rota e modo de transporte a usar, com base na sua experiência e conhecimentos pessoais. Este é um subprojecto com funcionalidades interessantes para o projeto ECO-Circuitos, seguem-se alguns exemplos[4]:

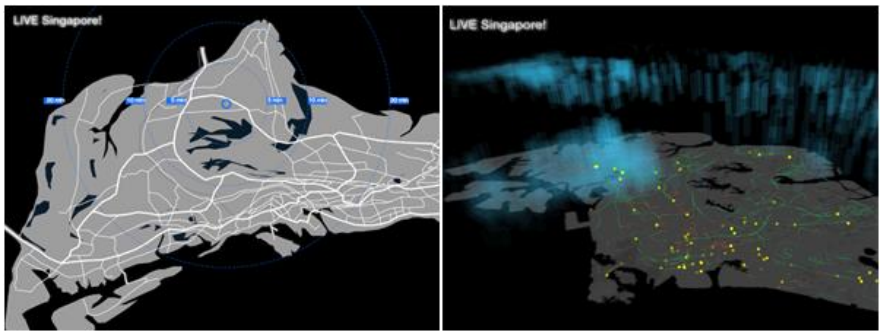


Figura 1 - Exemplos do projeto LIVE Singapore!

[4] Figuras retiradas de: http://senseable.mit.edu/livesingapore/visualizations.html

Open Cities

O *Open Cities* trata-se de um projeto de desenvolvimento cofundado pela União Europeia, que visa o desenvolvimento de aplicações e metodologias inovadoras a aplicar no sector público para que este se possa enquadrar melhor num cenário em que as cidades inteligentes fazem uso de serviços de internet. Tem como temas principais o uso de *open data*[5], *crowdsourcing*[6] e a mobilidade urbana.

Com a duração de trinta meses, este projeto é financiado pelo *ICT Policy Support Programme*[7] e visa o cumprimento de cinco objetivos principais:

- Encontrar as melhores práticas e ideias sobre como aplicar o *Open Innovation*[8] no sector público;
- Obter uma melhor compreensão da gestão de plataformas tecnológicas no âmbito do *Open Innovation*;
- Validar a utilização de plataformas em todos os países europeus para *crowdsourcing, open data* e redes de fibra ótica;
- Estimular o desenvolvimento de serviços de internet avançados;
- Entender de que forma o *Living Labs*[9] poderiam ser aplicados eficazmente para promover a adoção da inovação em cidades inteligentes.

Como resultado final, espera-se que o projeto ofereça três tipos diferentes de contribuições: i) Um novo entendimento de como abordar o *Open Innovation* por parte do setor público em direção da construção de uma cidade inteligente. b) Plataformas funcionais para *Open Data* e *Open Networks* abrangendo várias cidades europeias. c) Serviços de internet avançados fornecidos por *developers* que usam a plataforma.

Esta forma de contribuir vai de encontro com os objetivos do projeto agregador Tice.Mobilidade, onde também são fornecidas as ferramentas para quem quiser contribuir com ideias e aplicações inovadoras a fim de melhorar a mobilidade urbana dos restantes utilizadores.

---

[5] Dados que se encontram disponíveis para o público.
[6] Modelo de produção que utiliza a inteligência e os conhecimentos coletivos e voluntários para resolver problemas.
[7] http://ec.europa.eu/information_society/activities/ict_psp/index_en.htm
[8] Parte do projeto que se dedica à procura de metodologias inovadoras e orientadas ao utilizador para serem aplicadas no setor público.
[9] Encarregues de efetuar experiências em ambientes reais de forma a ver como os utilizadores reagem às novas metodologias.
[7] Imagens retiradas do sítio: http://senseable.mit.edu/livesingapore/visualizations.html

2.3.2 *DevChallenge*

O *DevChallenge* consiste num concurso lançado pelo Tice.Mobilidade com o objetivo de promover o desenvolvimento de aplicações para a plataforma *One.Stop.Transport*[10] *(OST)* no âmbito da mobilidade urbana. As aplicações serão posteriormente avaliadas e às três melhores será atribuído um prémio. Trata-se de uma iniciativa interessante que fomenta o espírito de equipa para desenvolver uma aplicação inovadora que faça uso das ferramentas pela plataforma *OST*.

O ponto forte deste desafio é facto de após ser submetida e tornada pública pelos criadores, esta pode ser experimentada pelos utilizadores registados na plataforma *OST*, podendo assim apresentar possíveis erros e até sugestões que visam o melhoramento da aplicação criada.

Entre as aplicações que concorreram destacam-se duas: CrowdMaps e SMTUC Ubique. A primeira diz respeito a uma construção colaborativa do mapa, em que cada um pode contribuir com informação para o seu enriquecimento. A segunda permite consultar os horários da companhia de transportes públicos de Coimbra, denominada SMTUC. Ambas com o objetivo de melhorar a dinâmica urbana, são aplicações de interesse para o projeto.

2.3.3 *Fireball*

Este projeto consiste na criação de um mecanismo de coordenação através do qual uma de rede de cidades inteligentes em toda a Europa se envolve numa colaboração a longo prazo. Foi concebido como uma resposta a uma situação em que diferentes participantes no domínio da investigação de serviços de internet, estão a funcionar isoladamente das restantes utilizando as suas próprias metodologias.

O desenvolvimento do projeto é realizado com três objetivos em mente:

- Conseguir uma coordenação a nível europeu de metodologias e abordagens no domínio do *Living Lab* e do *FIRE*;
- Aproveitar os recursos disponíveis a nível europeu para explorar as oportunidades que a internet do futuro tem para oferecer;

[10] https://www.ost.pt/

- Assegurar a coordenação do desenvolvimento e partilha das melhores práticas de inovação na internet do futuro em cidades e setores piloto.

Com a melhoria da comunicação entre cidades inteligentes e com a ambição de fomentar a criação de serviços de internet inovadores, *Fireball* apresenta-se como sendo um projeto interessante na área de mobilidade urbana e onde as cidades apresentam um papel fundamental para que este seja concretizado com sucesso. As abordagens tomadas para quebrar o isolamento que existe na troca de informação e serviço entre cidades, revela-se interessante para um trabalho futuro a realizar pelo projeto ECO-Circuitos.

2.4 Sumário sobre deteção do modo de transporte

Ao aprofundar o conhecimento relativamente às dinâmicas urbanas surge como evidente a necessidade de, no projeto, criar um algoritmo capaz de identificar com precisão qual o meio de transporte a ser usado pelos utilizadores através dos dados recolhidos nos sensores do *smartphone*. Esta é uma temática importante, que tem vindo a ser intensivamente explorada nos anos recentes, mas que toma um significado especial com o aparecimento dos *smartphones* e dos sensores neles integrados. Estes dispositivos permitem a recolha de dados do GPS e do acelerómetro (entre outros), e é através dos dados fornecidos por estes sensores que se torna possível inferir qual o meio de transporte que está a ser utilizado pelo utilizador, como se pode verificar nos estudos realizados por [18,19,20].

São, no entanto, reconhecidas algumas dificuldades no que toca à distinção entre alguns meios de transporte, como é o caso do autocarro e do automóvel [19], tanto nas abordagens que utilizam os dados GPS para a inferência do modo, como nas que fazem uso dos dados recolhidos pelo acelerómetro.

Nas secções que se seguem serão apresentados estudos de caso diferentes. Num deles, apenas é utlizado o GPS e os dados extraídos deste para a deteção do meio de transporte. Num outro apenas foi utilizado o acelerómetro e os dados recolhidos deste sensor. É apresentada uma experiência em que a informação recolhida pelo GPS é combinada com dados relativos à rede de transportes, nomeadamente informação relativa a paragens de autocarros e a localização destes em tempo real. De seguida é descrita uma abordagem interessante em que para a inferência do meio de transporte são utilizados os

dados recolhidos pelo GPS e os que são gerados pelo acelerómetro. Por fim são apresentadas algumas aplicações cuja temática diz respeito à mobilidade dos utilizadores.

2.4.1 Estudo de caso: GPS

A abordagem seguida em [18] passa numa primeira instância pela divisão da viagem em dois segmentos distintos: pedestre e não pedestre. Para conseguir esta divisão, os autores baseiam-se na velocidade do utilizador ao separar os segmentos que ultrapassem uma determinada velocidade dos restantes. Os passos seguintes têm em conta a duração dos segmentos, isto é, se a duração de um determinado segmento for inferior a um limite, então este será agregado ao segmento anterior. Por outro lado, se a duração de um segmento for inferior ao limite, então este será considerado um segmento incerto e caso se verifique um número consecutivo destes segmentos, estes serão agregados num segmento classificado como não pedestre.

Após os segmentos serem divididos, é recolhida informação do GPS para proceder a classificação dos mesmos. Os dados recolhidos pelo GPS são:

- Velocidade Média
- As três velocidades mais elevadas
- Variância da velocidade
- Taxa de mudança de orientação (do original, *Heading Change Rate)*
- Taxa de mudança de velocidade *(*do original, *Velocity Change Rate)*
- Taxa de paragem (do original, *Stop Rate)*

Os três últimos parâmetros são os mais significativos para a distinção das diferentes classes, sendo a taxa de mudança de orientação a frequência com que o utilizador muda a sua direção tendo em conta a distância percorrida. A taxa de paragem pode ser visto como a quantidade de vezes que o utilizador atinge uma velocidade abaixo de um limite sobre a distância percorrida. Por fim, temos a taxa de mudança de velocidade, que se refere ao número de vezes que a velocidade entre dois pontos ultrapassa um determinado limite sobre a distância percorrida.

Foram utilizados vários algoritmos de classificação, sendo *Decision Tree* [21] o algoritmo que obteve os melhores resultados na distinção das seguintes classes: pedestre, carro, autocarro, bicicleta. Os resultados obtidos pelo estudo apresentam uma acuidade de 76.2% com uma precisão de 51.6% e *recall* de 81,8%.

2.4.2 Estudo de caso: Acelerómetro

Os mecanismos que envolvem a análise dos dados recolhidos deste sensor são um pouco mais complexos que o anterior. Inicialmente, o sinal é dividido em janelas de cinco segundos cada uma com 250 amostras [22] onde cada uma se sobrepõem 50%, sendo de seguida aplicado um Fast Fourier Transform[11] de 250 pontos de forma a extrair os coeficientes de Fourier[12]. O passo seguinte consiste em retirar a seguinte informação dos componentes obtidos pela aplicação da FFT:

- Magnitude dos 250 componentes;
- Energia do sinal;
- Média do sinal;
- Variância do sinal.

O modelo escolhido para efetuar a classificação foi o SVM[13] (*Support Vector Machine*), uma vez que é de utilização comum na análise e reconhecimento de padrões. Os resultados obtidos foram considerados satisfatórios apresentando uma precisão média de 93.88%. Numa outra abordagem realizada por [23], utilizou uma janela maior (512 amostras) com o mesmo valor de sobreposição. Os autores utilizaram apenas 32 coeficientes da FFT e a variância do sinal como *features* da classificação dos seguintes modos: autocarro, metro, pedestre, bicicleta, comboio, carro, estacionário e mota. Também o classificador é distinto da abordagem apresentada anteriormente, sendo a *Decision Tree C4.5* o escolhido para a classificação dos modos. Dada a quantidade elevada dos modos considerados, a acuidade obtida foi significativamente mais baixa do que a anterior, 82.15%. Não deixa de ser no entanto uma abordagem interessante a considerar para o algoritmo a desenvolver neste trabalho.

2.4.3 Estudo de caso: GPS + Dados externos

Este estudo inclui dados relativamente à rede de transportes em tempo real, como por exemplo, a localização dos autocarros em tempo real e a localização das paragens de autocarro [20].

[11] Consiste na decomposição do sinal em várias frequências de forma rápida usando o método das Transformadas de Fourier
[12] Que não são mais do que os sinais das funções sin(x) e cos(x) obtidas pela decomposição.
[13] http://www.support-vector-machines.org/

A inclusão destes dados juntamente com os dados recolhidos do GPS leva a um aumento considerável na precisão de deteção de transportes, como podemos observar comparando os valores da Tabela 1 com os da Tabela 2.

Tabela 1 - Precisão da inferência do modo de transporte sem dados externos[14]

	Naive Bayes	*Decision Tree*	*Random Forest*
Autocarro	47.0	40.6	56.5
Estacionário	100	100	100
Pedestre	94.7	93.8	92.7
Carro	42.3	43.5	58.1
Bicicleta	70.2	68.8	71.4
Precisão Média (%)	70.84	69.34	75.74

Tabela 2 - Precisão da inferência do modo de transporte com dados externos[15]

	Naive Bayes	*Decision Tree*	Random Forest
Autocarro	85.0	88.3	88.3
Estacionário	100	100	100
Caminhar	96.7	94.7	96.8
Carro	78.2	85.1	88.9
Bicicleta	88.9	85.5	88.9
Precisão Média (%)	89.76	90.72	92.58

[14] Retirado da literatura [20]
[15] Retirado da literatura [20]

As *features* dos dados GPS utilizadas são as seguintes:

- Velocidade média;
- Aceleração média;
- Orientação;
- Precisão média das coordenadas GPS;

Relativamente aos dados em tempo real é recolhida a seguinte informação:

- *Candidate bus closeness*, ou seja, o autocarro que se encontra mais próximo do utilizador;
- Distância média em relação ao autocarro;
- Distância média em relação à paragem de autocarro.

O primeiro ponto refere-se ao método que identifica qual o autocarro que se encontra mais perto do utilizador num dado momento. Para isso, é calculada a distância Euclidiana entre um conjunto de pontos referentes à localização do utilizador e os pontos dos autocarros que se encontrem dentro da mesma janela temporal dos pontos do utilizador. Aquele que apresentar a menor distância será considerado o *candidate bus*.

O segundo ponto é semelhante ao método descrito anteriormente. No entanto, em vez de ser calculada distância euclidiana ao longo de uma janela temporal, a distância é apenas calculada para um determinado instante.

O terceiro diz respeito à média da distância do utilizador em relação às paragens de autocarro que se encontram no sistema de dados.

Depois de estarem reunidos todos os dados necessários, estes serão enviados para o algoritmo de classificação. Os algoritmos escolhidos foram o *Random Forest* [24] devido à elevada precisão com que este efetua as classificações e pela eficiência que apresenta mediante um grande conjunto de dados. Foi também escolhido o modelo de Redes Bayesianas [25] pela sua capacidade de gerar automaticamente predições ou decisões, mesmo na situação de inexistência de algumas peças de informação.

Ao comparar os valores das Tabelas 1 e 2, torna-se claro o ganho na precisão de inferência do modo de transporte quando se junta aos dados do GPS, os dados externos fornecidos em tempo real. Embora no nosso projeto não seja possível obter dados como a localização do autocarro em tempo real, é possível adquirir a posição das paragens de

autocarro. Essa informação aliada aos dados recolhidos pelo GPS já oferece uma boa precisão a nível de deteção do modo de transporte.

2.4.4 Estudo de caso: GPS + Acelerómetro

No estudo efetuado por [26], é efetuada a fusão dos dados recolhidos pelos dois sensores (GPS e acelerómetro) para efetuar a distinção entre as seguintes classes: autocarro, metro, pedestre, bicicleta, comboio, carro, estacionário e mota. O algoritmo desenvolvido baseia-se na informação recolhida pelas torres GSM para detetar se o utilizador se encontra fora de um local fechado ou não.

Os dados que foram considerados significativos para a classificação das classes são:

Tabela 3 - Dados extraídos pelos sensores

GPS	Acelerómetro
Velocidade	Variância do sinal
	Energia do sinal
	Coeficientes da FFT

Recolhidos os dados necessários, estes são adicionados ao algoritmo de classificação, que neste estudo é o DT-DHMM. Trata-se de um algoritmo que combina o *Decision Tree* com o *Discrete Hidden Markov Model* para o cálculo das transições entre estados. Este algoritmo de classificação obteve, segundo os seus autores, uma precisão de 98,8% na classificação das diferentes classes, sendo portanto um método interessante e, se possível, a considerar para o nosso projeto.

2.5 Aplicações para *smartphone*

As aplicações desenvolvidas para *smartphones* são fundamentais para a recolha dos dados necessários para a identificação dos modos de transporte utilizados nas deslocações dos utilizadores. Nas secções que se seguem, apresentamos um conjunto de aplicações, que embora tenham objetivos diferentes, são utilizadas para fornecer ao utilizador informação útil relativamente ao seu tipo de mobilidade.

2.5.1 *Move*

Esta aplicação Android trata-se de um trabalho realizado por um aluno de Mestrado da Faculdade de Ciências e Tecnologias da Universidade de Coimbra, como parte do seu estágio num subprojecto do *iTeam* denominado *Greenhomes*[16]. Sempre que seja possível, a aplicação recolhe os dados necessários (GPS, wi-fi, cell-ID, acelerómetro, *bluetooth*), tentando não se exceder o gasto de bateria do *smartphone*. Permite ainda o envio dos dados recolhidos para uma base de dados do projeto de modo automático.

O facto de os utilizadores poderem visualizar e corrigir os dados diariamente pela aplicação numa plataforma do projeto disponível *online*, apresenta-se como sendo uma funcionalidade bastante interessante. O principal problema desta aplicação é mesmo o consumo de bateria, uma vez que o utilizador não conseguirá utilizar o *smartphone* mais de quinze horas com a aplicação em funcionamento (este valor varia consoante a utilização do *smartphone* por parte do utilizador e consoante o modelo do dispositivo), sendo que este terá de ser um ponto de foco muito importante na aplicação a ser desenvolvida para o nosso projeto.

[16] (Projeto "*Greehomes* – iTEAM", disponível em https://www.cisuc.uc.pt/publication/show/2763)

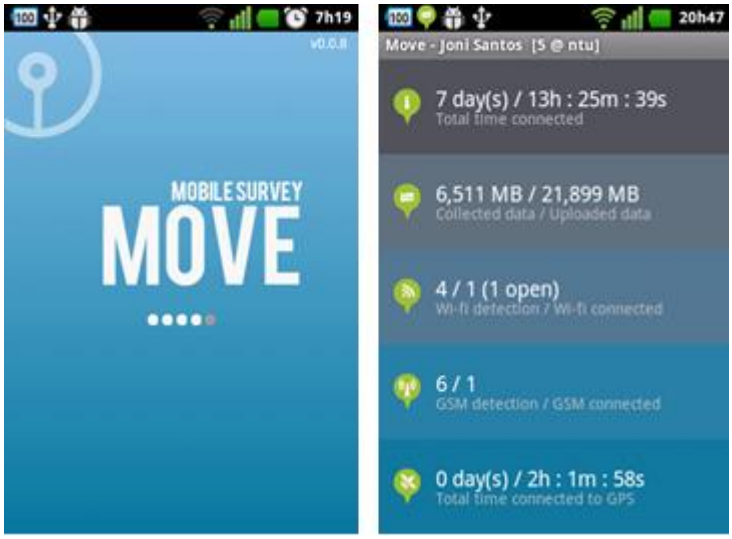


Figura 2 - Capturas de ecrãs da aplicação Move[17]

2.5.2 *Lifemap*

Desenvolvida em 2011 pela Universidade de Yonsei, Coreia do Sul, apresenta-se como sendo uma aplicação Android que permite visualizar e gerir a rotina do quotidiano do utilizador. Tem a limitação de não apresentar os percursos efetuados diariamente, no entanto indica os pontos de paragens, tempos de estadia e a ligação entre eles. Tal como a aplicação anterior, faz uso de vários sensores do *smartphone:* acelerómetro, *Bluetooth*, Termómetro, *Wi-Fi* e *GSM*. De referir que estes sensores nem sempre se encontram simultaneamente ativos a recolher dados.

Possuí alguns mecanismos que têm como objetivo diminuir a carga da aplicação sob a bateria, para isso a aplicação deteta se o utilizador se encontra em modo estacionário. Caso isso se verifique, todos os sensores mencionados, à exceção do acelerómetro, são desativados.

[17] Imagens cedidas por Jóni Santos (joni@student.dei.uc.pt)

Figura 3 - Capturas de ecrãs da aplicação *Lifemap*[18]

2.5.3 *Automatic*

Esta aplicação foi lançada em 2013 nos Estados Unidos da América, funciona em *Android* e *iPhone*. Esta tem o objetivo de indicar o padrão de condução de um utilizador, sugerindo modificações instantâneas durante uma viagem, com vista a uma diminuição no consumo de combustível. Tem como limitação o facto de só funcionar para carros a gasolina. Oferece também informação relativamente aos locais visitados pelo utilizador, informando-o se executou uma condução económica (travagens e acelerações repentinas, excesso de velocidade).

A aplicação funciona apenas se for realizada uma ligação entre o *smartphone* e o computador de bordo do veículo via *Bluetooth*. Outra funcionalidade interessante é o facto de a aplicação realizar automaticamente chamadas de emergência, se o utilizador tiver um acidente. O acelerómetro tem um papel importante nesta funcionalidade ao medir a intensidade do acidente permitindo assim saber qual o grau de gravidade.

[18] http://appaggie.com/2011/08/28/lifemap/

Figura 4 - Captura de ecrã da aplicação *Automatic*[19]

2.5.4 *Moves*

Desenvolvida pela empresa finlandesa *ProtoGeo Oy*, esta aplicação encontra-se disponível para *Android* e *Iphone*. Esta aplicação tem como objetivo construir um diário de mobilidade do utilizador, de forma a que este ao final do dia possa verificar no mapa o percurso das deslocações efetuadas, o tempo de estadia nos locais visitados e o modo de transporte utilizado nas deslocações realizadas. Todos estes registos são realizados de forma automática, isto é, não é necessário qualquer tipo de interação por parte do utilizador para registar a sua atividade.

Esta aplicação faz uso do *GPS* do *smartphone* para registar os percursos realizados pelo utilizador, assim como o modo de transporte e o tempo de estadia. No entanto, a quantidade de modos de transporte detetados pela aplicação são bastante reduzidos, permitindo apenas identificar o modo pedestre, bicicleta e correr, sendo os restantes modos de transportes classificados na mesma categoria, neste caso, Transporte. O acelerómetro é utlizado para calcular o número de passos do utilizador nas suas deslocações. Através do *Wi-Fi* ou da ligação de dados, a aplicação recolhe informação para poder identificar os locais visitados pelo utilizador.

Sendo esta uma aplicação que faz uso do *GPS*, verifica-se um uso mais acentuado da bateria do *smartphone*. Apesar desta limitação, a aplicação possuí uma opção que permite otimizar o uso da bateria, prejudicando no entanto a precisão no registo dos percursos.

[19] http://www.automatic.com/

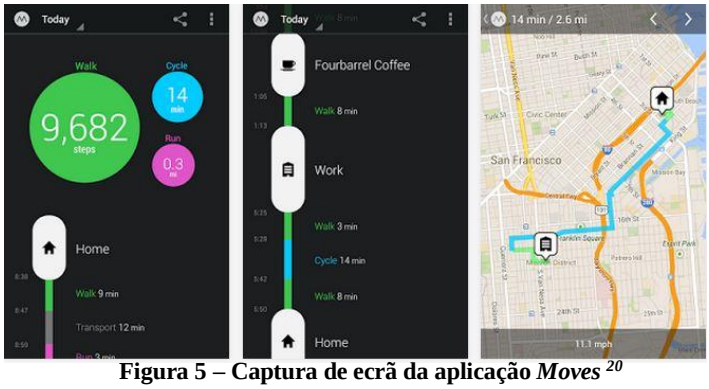


Figura 5 – Captura de ecrã da aplicação *Moves* [20]

2.5.5 *CO2GO*

Desenvolvida pelo *SENSEable City Laboratory* do *MIT* [21], é uma aplicação que tem como objetivo a deteção do modo de transporte, para poder apresentar as emissões de dióxido de carbono emitidas ao longo da deslocação efetuada. Com esta abordagem pretende-se que o utilizador comece a tomar decisões conscientes do modo de transporte a utilizar, reduzindo assim a emissão de gases poluentes para a atmosfera.

A aplicação permite distinguir uma grande variedade de modos de transporte: autocarro, carro, metro, comboio, mota, bicicleta e pedestre. Os dados recolhidos pelo acelerómetro são interpretados em tempo real pelo algoritmo. O *GPS* é utilizado para o cálculo da distância viajada, de forma a contribuir para a precisão do algoritmo e para indicar no mapa o percurso realizado na deslocação. Os utilizadores posteriormente podem partilhar as suas rotas de baixa emissão de dióxido de carbono com os restantes utilizadores.

Figura 6 - Captura de ecrã da aplicação CO2GO [22]

[20] http://www.moves-app.com/
[21] http://senseable.mit.edu/
[22] http://senseable.mit.edu/co2go/

2.5.6 Comparação das aplicações

As aplicações aqui apresentadas possuem objetivos diferentes. O principal objetivo da aplicação *Move*, consiste na recolha de dados de mobilidade através dos diferentes sensores do *smartphone*. A aplicação *Automatic* visa otimizar a condução do utilizador, de forma a ajudá-lo a economizar no gasto do combustível, por exemplo. Por outro lado, as aplicações *Lifemap, Moves e CO2GO* possuem funcionalidades semelhantes, nomeadamente a apresentação dos pontos de paragem do utilizador assim como o tempo de estadia e de deslocação entre eles. No entanto, as aplicações *Moves e CO2GO* apresentam funcionalidades mais completas que o *Lifemap*, ao permitir ao utilizador visualizar o seu diário de mobilidade, isto é, para além dos pontos de paragem também pode consultar os percursos diários realizados. Apesar de todas estas diferenças, as aplicações têm um ponto importante em comum, que é a otimização do uso da bateria, visto que o facto de fazerem uso de sensores como o GPS, leva a um desgaste mais rápido da vida útil da bateria. Segue-se a Tabela 4 onde são apresentadas comparações relativamente às funcionalidades das aplicações.

Tabela 4 – Comparação das aplicações para *smartphone*

	MOVE	***LIFEMAP***	***AUTOMATIC***	***MOVES***	***CO2GO***
SENSORES	GPS, wi-fi, cell-ID, acelerómetro, *bluetooth*	Acelerómetro, *bluetooth*, termómetro, *wi-fi, GSM*	Acelerómetro, *Bluetooth*,	Acelerómetro, GPS, *Wi-fi, GSM*	Acelerómetro, GPS
MODOS DE TRANSPORTE	X	X	X	Pedestre, correr, bicicleta	Autocarro, carro, metro, comboio, mota, bicicleta e pedestre
OBJETIVO	Recolher dados de mobilidade	Apresentar os pontos de paragem do utilizador	Otimizar a condução do utilizador	Apresentar o diário de mobilidade do utilizador	Apresentar o diário de mobilidade do utilizador juntamente com a emissão de

					carbono da deslocação
SISTEMA OPERATIVO	Android	Android, Iphone	Android, Iphone	Android, Iphone	Android

2.6 Classificadores

Em geral, os métodos de classificação dividem-se entre supervisionados e não supervisionados. No primeiro caso o *dataset* de treino é constituído por amostras que se encontram etiquetadas já com o resultado correto. Desta forma os classificadores supervisionados avaliam o *dataset* a fim de criar uma função de inferência que permita lidar com novos dados, classificando corretamente os dados que não possuam etiqueta a identifica-lo. No segundo caso o *dataset* não se encontra identificado com qualquer etiqueta, sendo o objetivo dos classificadores deste tipo encontrar caraterísticas escondidas que permitam distinguir e dividir os dados em grupos distintos. Algoritmos de *clustering* são um exemplo deste tipo de classificação

No trabalho a desenvolver serão utilizados algoritmos que pertencem ao grupo da classificação supervisionada. Em primeiro lugar é apresentado o *Naïve Bayes*, seguido do *Random Forest* e do *J48*. Estes classificadores foram escolhidos em detrimento de outros, visto que os resultados obtidos na literatura [20, 27] foram bastante bons, com uma precisão média de 92.58%. Para além disso, dado que o principal objetivo é a integração deste algoritmo com a aplicação móvel, algoritmos de classificação mais complexos como é o caso de redes neuronais ou *support vector machines* não foram considerados para este trabalho. Estes exigem um poder computacional maior o que levaria a um gasto mais acentuado da bateria do *smartphone*.

2.6.1 *Naïve Bayes*

Trata-se de um classificador estatístico que faz uso de todos os atributos, partindo de dois pressupostos [28]:

- Todos os atributos são igualmente importantes;
- Os atributos são estatisticamente independentes, isto é, saber o valor de um atributo não diz nada relativamente ao valor de outro.

É devido a esta última premissa que os cálculos das probabilidades dos acontecimentos apresentam um custo computacional reduzido, visto que faz uso do teorema de *Bayes* para o cálculo das probabilidades a atribuir a cada classe.

Este teorema trata-se da manipulação matemática de probabilidades condicionais e faz uso da seguinte fórmula:

$$P(C|F_1 \cdots F_n) = \frac{P(C) \cdot P(F_1 \cdots F_n|C)}{P(F_1 \cdots F_n)} \quad (1)$$

Onde C representa a variável dependente que diz respeito à classe em questão e $F_1 \cdots F_n$ diz respeito aos atributos, o elemento condicionado. Dado que o denominador acaba por tomar por ser constante ao longo dos cálculos, esta propriedade reduz a complexidade do classificador permitindo que a classificação das classes seja feita num tempo reduzido.

Relativamente à forma como o classificador lida com atributos que não possuam valores, o valor atribuído à probabilidade desses atributos é de zero. De forma a evitar que ocorra o erro de divisão por zero no cálculo do teorema de teorema de *Bayes*, adiciona-se o valor um à combinação de cada classe-atributo, evitando assim que o valor desta seja zero.

Este classificador destaca-se na área de classificação de documentos [29]. Esta área consiste em identificar corretamente em que categoria se encontra um determinado documento, por exemplo: desporto, negócios. *spam*, entre outros. O maior desafio desta é o facto de ter de lidar com uma grande quantidade de *features* no *dataset*, onde a simplicidade deste algoritmo se apresenta como sendo uma vantagem relativamente a outros como é o caso das *decision trees*, redes neuronais e *support vector machines* [30].

2.6.2 *Random Forest*

Trata-se de um conjunto de árvores de decisão que ao receber um conjunto de dados atuam como um sistema de votos, ou seja, no final o resultado que obtiver mais votos por parte das árvores será considerado o resultado final.

A geração destas árvores ocorre em três passos:

- O *dataset* de treino é usado na totalidade para a construção das árvores;

- São escolhidos aleatoriamente M *features* para obter a melhor combinação para efetuar a divisão dos nós da árvore;
- Cada árvore gerada cresce até atingir o valor máximo, visto que neste classificador as árvores não são podadas.

Ao contrário dos restantes classificadores aqui apresentados, o *Random Forest* não necessita que se separe um *dataset* de treino e outro de teste. Internamente o algoritmo efetua uma divisão no *dataset* [24].

Outra característica interessante que este classificador apresenta é a remoção de valores considerados atípicos. A abordagem tomada para conseguir isto consiste na avaliação do valor de proximidade do caso que está a ser avaliado relativamente aos restantes dados, se o valor de proximidade for pequeno então será removido do *dataset.*

Este classificador apresenta como principais vantagens os seguintes pontos:

- É eficiente em lidar com *datasets* de dimensões elevadas;
- Oferece uma estimativa de que variáveis são importantes na classificação;
- Calcula a proximidade entre pares de casos que podem ser usados posteriormente para *clustering*, identificar valores atípicos ou simplesmente para ter visão mais interessante dos dados em causa;
- Apresenta uma acuidade elevada comparativamente aos restantes algoritmos de classificação.

Estamos perante um classificador robusto que permite obter bons resultados mesmo quando o *dataset* possuí classes cujos valores possuem atributos sem valor atribuído. É também uma boa opção caso o *dataset* em questão possua valores atípicos que não foram filtrados anteriormente. Estas características fazem com que este seja um classificador bastante usado especialmente na área de bioinformática, pelo facto de ser um classificador que pode ser aplicado a uma vasta gama de tipos de dados [31].

2.6.3 *J48*

Esta é uma implementação em *Java* do algoritmo de classificação C4.5 [32]. Ao contrário do classificador anterior, apenas é gerada uma árvore de decisão. O funcionamento deste algoritmo consiste na análise em cada nó, escolher o atributo dos dados que permite uma melhor divisão em subconjuntos mais pequenos. A medida utilizada nesta é o ganho a nível de informação que se consegue através da divisão efetuada que consiste na subtração da entropia que existe antes e depois da divisão:

$$Ent(S_1 \cdots S_n) = -\sum_{c=1}^{N} p_c \cdot log_2\, p_c \qquad (2)$$

Onde p_c representa a frequência relativa da classe C no conjunto total S_n , a divisão que se apresentar com um maior ganho a nível de informação será utilizado na criação de um novo nó na árvore.

Outra diferença relativamente ao algoritmo anterior, diz respeito à poda das árvores. Neste classificador são realizados dois tipos de podas: pré-poda e pós-poda. No caso da primeira, a construção da árvore é interrompida quando os atributes são irrelevantes. No caso da segunda, a árvore é construída até ao fim e são removidos apenas alguns ramos da árvore.

As vantagens apresentadas por este classificador são:

- Fácil de implementar;
- Consegue lidar com informação que contenha "ruído";
- Os modelos gerados são de fácil interpretação.

Apesar das vantagens utilizadas, é um classificador que quando perante um *dataset* que contém pouca variação nos seus valores obtém um fraco desempenho a classificar os dados. No entanto é um classificador bastante utilizado em classificações baseadas em regras, isto é, fazem uso de um conjunto de regras do tipo *If – Then*, podendo assim ser aplicado em diversas áreas.

Capítulo 3

Objetivos e Abordagem

Os objetivos do presente estágio vão de encontro aos objetivos do ECO-Circuitos, um subprojecto do Projeto QREN TICE.Mobilidade – Sistema de Mobilidade centrado no Utilizador, referidos anteriormente. O principal objetivo do ECO-Circuitos visa a criação de um sistema capaz de traçar padrões individuais de mobilidade urbana e descrever esse perfil ao utilizador com informação financeira, temporal e ambiental associadas, fruto das suas opções de mobilidade. Se possível, pretende-se ainda identificar alternativas mais eficientes para o utilizador, tanto em termos económicos como ambientais. Neste contexto, o presente estágio propõe o desenvolvimento de um algoritmo inteligente capaz de, com base em dados dos sensores do *smartphone*, inferir qual o modo de transporte utilizado em cada uma das deslocações do utilizador. Assim, o objetivo principal deste trabalho é o de poder contribuir com uma ferramenta essencial para o reconhecimento de padrões individuais de mobilidade urbana. Posteriormente numa fase mais avançada do projeto, pretende-se integrar o algoritmo desenvolvido com a aplicação Android caso o impacto na bateria dos dispositivos não seja muito grande. Em termos de possíveis aplicações adicionais, estes padrões, quando agregados por zonamento e anonimizados, constituem um contributo importante para empresas e instituições responsáveis pelo planeamento urbano, para que este seja feito de uma forma mais eficiente.

A fim de cumprir os objetivos propostos, a abordagem seguida foi dividida em quatros etapas identificadas na Figura 7:

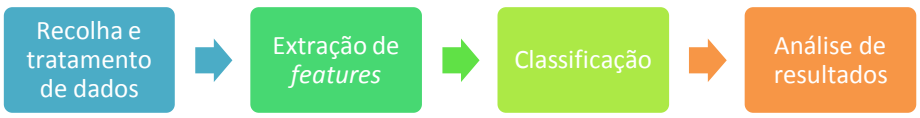


Figura 7 - Etapas da abordagem

Numa primeira etapa efetuou-se a recolha de dados de referência considerados necessários para a definição e validação do algoritmo de inferência. Os dados foram obtidos a partir do sensor de GPS e do acelerómetro de *smartphone*s e foram adquiridos com a ajuda da aplicação Android desenvolvida pela copromotora do projeto. Estes dados de referência foram posteriormente tratados de forma a remover possíveis erros de classificação por parte dos utilizadores. Numa etapa seguinte foram analisadas as *features* necessárias extrair dos dados tratados para que seja possível distinguir os diferentes modos de transporte. Na etapa da classificação procede-se à análise dos classificadores que melhor se adequam para a inferência do modo de transporte. Por fim, são analisados os resultados obtidos com o objetivo de validar e posteriormente melhorar o algoritmo de inferência dos modos de transporte.

O trabalho realizado ao longo do primeiro semestre incidiu principalmente no aprofundamento do estado de arte. Em primeiro lugar, foi necessário pesquisar literatura genérica sobre mobilidade urbana e perfis de mobilidade, de forma a entender o funcionamento da dinâmica urbana. De seguida foi dado início à pesquisa de artigos relacionados com a temática de inferência de modos de transporte. Foi também efetuado um estudo relativamente aos classificadores que melhor se adequavam às necessidades do projeto.

No segundo semestre, verificou-se um atraso no acesso a dados com a qualidade necessária devido a problemas na frequência de recolha de dados da aplicação Android, impedindo assim o início da análise e tratamento dos dados recolhidos pelos respetivos sensores (GPS e acelerómetro). Apesar da busca de *datasets* de estudos anteriores para mitigar este problema, também não foram encontrados dados que se enquadrassem com as necessidades do projeto. Quando os problemas da aplicação foram resolvidos, teve inicio um período de recolha intensiva de dados por parte dos voluntários. Em primeiro lugar foram analisados os dados recolhidos pelo GPS e encontradas as características necessárias para serem usadas pelo algoritmo de inferência do modo de transporte. De seguida procedeu-se à análise dos dados recolhidos pelo acelerómetro. Verificada a sua inadequação computacional devida aos tempos de desempenho apresentados, decidimos deixar para trabalho futuro o aprofundar este estudo de modo a tentar: i) perceber a real importância destes dados para melhorar o classificador baseado em GPS; ii) caso aferida essa importância desenhar um algoritmo mais apropriado para o fim em causa. Por fim, foi analisada a melhor maneira de segmentar a viagem do utilizador de forma a inferir

com precisão os meios de transportes utilizados nas suas deslocações e desenvolvido o algoritmo responsável por lidar com os erros de classificação. Após concluído o desenvolvimento do algoritmo, procedeu-se à integração deste com a aplicação web desenvolvida e com a aplicação do *smartphone*.

A Figura 8 ilustra de que forma as tarefas foram divididas ao longo dos semestres.

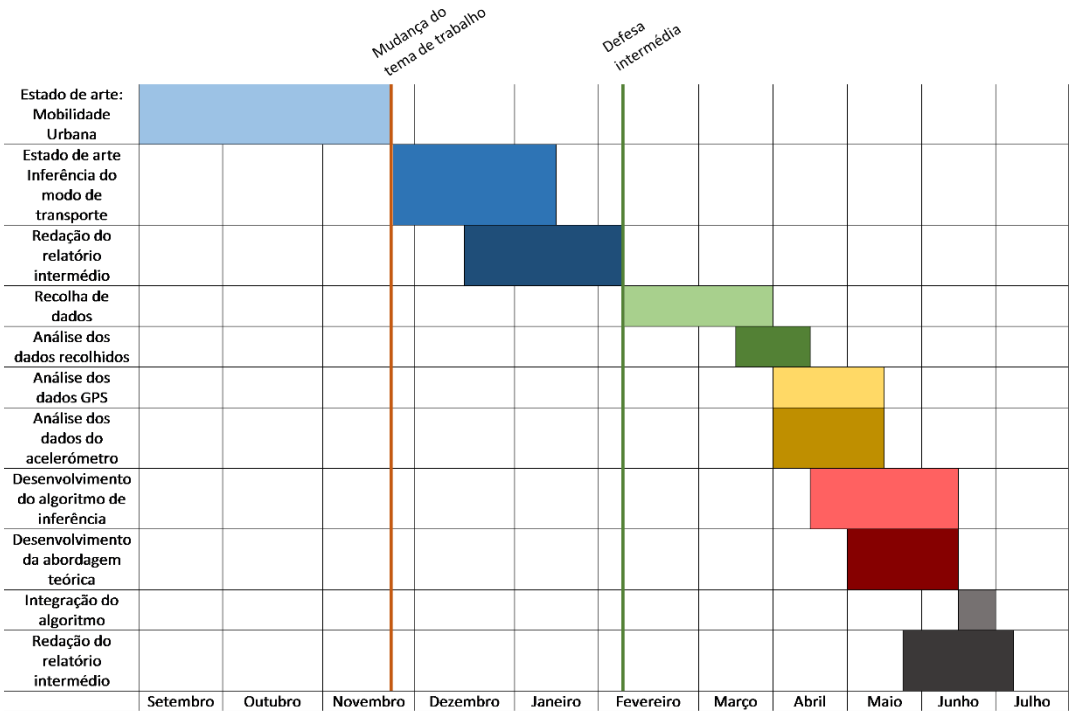


Figura 8 - Planificação do trabalho realizado

Capítulo 4

Implementação

Neste capítulo será apresentado o trabalho de desenvolvimento produzido ao longo do ano letivo, assim como as considerações e decisões tomadas em cada fase. Começamos por descrever o processo de recolha de dados e as *features* que posteriormente serão extraídas, primeiro dos dados recolhidos pelo GPS e, de seguida, os do acelerómetro. Será, então, explicada em detalhe a abordagem escolhida para a inferência dos modos de transporte utilizando os dados GPS. Segue-se uma abordagem teórica para a inferência dos modos de transporte através da junção dos dados recolhidos pelo acelerómetro e pelo GPS. Por fim será apresentada a aplicação *web* desenvolvida que permite aos utilizadores visualizar os seus dados de mobilidade e informação relativamente às viagens efetuadas, na versão presente, utilizando apenas os dados obtidos pelo GPS.

4.1. Recolha, tratamento de dados e problemas encontrados

A recolha de dados é um processo importante no projeto, visto que as fases seguintes dependem da quantidade e qualidade dos dados obtidos. A aplicação Android utilizada na recolha destes dados foi distribuída por um grupo de vinte voluntários em quatro cidades: Coimbra, Guimarães, Porto e Lisboa durante aproximadamente dois meses. Antes de dar início às suas viagens, o utilizador indica o modo de transporte e no final indica o motivo da viagem realizada (Figura 9).

Figura 9 - Escolha do modo de transporte e motivo da viagem

A partir do momento que o utilizador inicia a sua viagem, a aplicação começa a recolher automaticamente os dados os respetivos sensores, nomeadamente do GPS e do acelerómetro. A frequência com que os sensores registam os dados sofreu alterações ao longo do projeto até ser encontrada a frequência ideal. No caso do GPS, a informação é recolhida a cada segundo e no caso do acelerómetro verifica-se uma frequência de recolha de 50 Hz, isto é, num segundo são recolhidas cinquenta amostras por parte deste sensor. Toda esta informação é armazenada na memória interna do *smartphone* e posteriormente enviada para uma base de dados remota que se encontra a cargo da *SmartMove*.

O processo encontra-se descrito na Figura 10, onde as etapas a azul se referem a ações realizadas automaticamente pela aplicação e as etapas a verde dizem respeito às interações do utilizador com a aplicação.

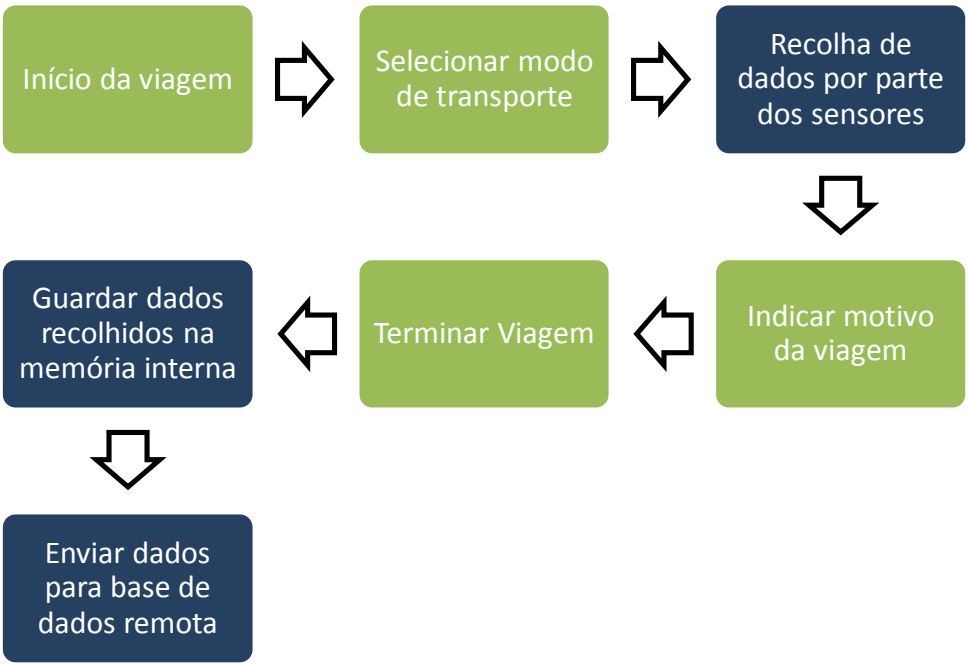


Figura 10 - Descrição do processo de recolha de dados

Este processo de desenvolvimento da aplicação realizou-se ao longo de aproximadamente dois meses, durante os quais trabalhámos no sentido de perceber aperfeiçoar a recolha de dados necessária, quer ao projeto como um todo, quer ao nosso método de inferência particular. Assim, foram recolhidos os dados de mobilidade que foram (e serão) utilizados para treinar o classificador.

A Tabela 5 apresenta a duração e a distância total dos trajetos percorridos relativamente a cada um dos diferentes modos de transporte relativamente ao conjunto de dados utilizados para o estudo. Os dados utilizados dizem respeito à recolha efetuada pelos vinte voluntários ao longo de dois meses. Embora o número de voluntários seja razoável, a sua mobilidade realiza-se à base do modo Pedestre e do modo Carro, não sendo possível de momento incluir outros meios de transporte como mota ou até mesmo bicicleta elétrica. Através desta tabela é possível observar o peso relativo que cada meio de transporte tem nas distâncias e nos tempos absolutos quando em relação ao *dataset* como um todo.

Tabela 5- Estatísticas dos dados recolhidos

	Pedestre	Carro	Autocarro	Bicicleta	Total
Duração Total	26h 16min 18seg (18.95%)	93h 38min 10seg (67.56%)	15h 48min 14seg (11.40%)	2h 53min 24seg (2.09%)	138h 36min 4seg (100%)
Distância Total	115,58 Km (2.08%)	5100,51 Km (92%)	301,95 Km (5.45%)	26,33 Km (0.47%)	5544,37 Km (100%)

Com o objetivo de lidar com possíveis classificações erradas por parte dos voluntários, ou seja, situações em que efetuou uma viagem na qual utilizou o modo Pedestre e no entanto classificou como modo Carro, por exemplo, foi criado um conjunto de filtros obtidos através de uma análise empírica aos dados de mobilidade que fazem parte do *dataset* apresentado anteriormente de voluntários pré-determinados. Os segmentos cuja velocidade máxima não se encontre dentro dos valores adequados a cada modo de transporte (definidos pelos filtros) são removidos do conjunto de segmentos a ser usados para treinar o classificador, por exemplo um segmento classificado com o modo Carro que registe uma velocidade máxima inferior a 3m/s, é retirado *dataset*. Na Tabela 6 encontram-se os valores utilizados pelos filtros aplicados aos diversos modos de transporte.

Tabela 6 - Valores usados para filtrar a amostra

Modo de Transporte	Velocidade (m/s)
Pedestre	[0,3]
Carro	[3,60]
Autocarro	[3,30]
Bicicleta	[3,10]

Após a aplicação destes filtros verificou-se uma redução considerável na quantidade de dados relativos ao modo Pedestre. Esta diminuição pode ser explicada pelo facto de os utilizadores se esquecerem de alterar o modo de transporte quando trocam do modo Pedestre para o modo Carro, por exemplo. Os restantes modos de transporte obtiveram uma redução abaixo de 1%, visto que no caso do modo Carro e Autocarro torna-se mais difícil verificar quando ocorre um erro de classificação devido às suas semelhanças relativamente às velocidades registadas em meios urbanos. Os valores obtidos após a aplicação dos filtros acima mencionados podem ser consultados na Tabela 7, onde também é apresentada a percentagem de dados que foram retirados da amostra relativamente a cada modo de transporte.

Tabela 7 - Estatísticas dos dados recolhidos após aplicação dos filtros

	Pedestre	Carro	Autocarro	Bicicleta	Total
Duração Total	18h 58min 41seg (27.76%)	93h 19min 17seg (0.34%)	15h 41min 30seg (0.71%)	2h 53min 24seg (0.0%)	125h 19min 32seg (9.58%)
Distância Total	63,39 Km (45.15%)	5099,11 Km (0.03%)	301,42 Km (0.18%)	26,33 Km (0.0%)	5544,37 Km (0.98%)

O processo de recolha de dados em bruto é também a fase mais suscetível a atrasos, visto encontrar-se fortemente dependente do desenvolvimento da aplicação Android por parte da empresa *SmartMove* e da disponibilidade dos voluntários para efetuar a recolha de dados.

Ao longo desta etapa verificaram-se alguns problemas com a aplicação Android, nomeadamente, verificámos a frequência com que os dados eram recolhidos era demasiado baixa para os fins em vista. Embora fosse possível fazer a distinção entre o modo Pedestre e o modo Carro, devido ao facto de ambos serem caracterizados por um padrão de mobilidade muito distinto, o mesmo não se verificava ao tentar fazer a distinção entre o modo Carro e modo Autocarro.

Outra situação que levou ao atraso do desenvolvimento do algoritmo de inferência, foi o facto de os dados não estarem a ser enviados corretamente para o servidor. Em alguns casos a aplicação deixava de funcionar fazendo com que se perdessem os dados da viagem em causa do utilizador, logo baixando a quantidade de amostragem de dados ou interrompendo viagens, tornando-as em 'ruído'.

Todos estes problemas levaram a que os dados da base de dados remota fossem retirados até que a aplicação se encontrasse a recolher e enviar os dados corretamente, resultando num atraso em aceder a dados com a qualidade e quantidade necessária para passar à etapa seguinte de desenvolvimento.

Foram realizados esforços de forma a poder prosseguir com o desenvolvimento do algoritmo. No caso do GPS, foi encontrado um *dataset* disponibilizado pelo projeto *Geolife* onde à semelhança do que acontece neste projeto, os voluntários indicavam qual o meio de transporte utilizado nas suas deslocações diárias. Após analisar este *dataset*, concluímos que o mesmo não podia ser utilizado no nosso projeto devido a dois problemas fundamentais:

- O *dataset* não inclui a velocidade instantânea nos pontos de GPS recolhidos. Sendo esta uma característica essencial para o cálculo das nossas *features* e por conseguinte para a construção do nosso classificador, só por si inviabiliza a utilidade deste *dataset* no nosso trabalho;
- Foram encontrados casos em que a informação das secções etiquetadas pelos utilizadores com o modo de transporte não se encontravam *dataset* fornecido.

No caso dos dados do acelerómetro apenas foi encontrado um *dataset*, mas este não se encontrava enquadrado com os objetivos do projeto uma vez que tinha como principal objetivo o reconhecimento das atividades do utilizador e não do modo transporte que utilizava para se deslocar.

Visto que nenhuma das soluções encontradas se enquadrava com este projeto, registou-se um período de espera durante o qual foram oferecidas sugestões para o melhoramento da aplicação Android.

4.2. Extração de *features*

Nesta secção serão apresentadas as *features* extraídas dos dados filtrados. Em primeiro lugar são apresentados os cálculos necessários para a obtenção das *features* a partir do GPS. Segue-se a apresentação do método adotado para a extração das *features* do acelerómetro.

4.2.1. Extração de *features* de dados de GPS

A abordagem para extrair as características necessárias para a distinção dos modos de transporte realizou-se com base no trabalho descrito na Secção 2.4.1. As *features* calculadas a partir das viagens dos voluntários foram:

- Taxa de mudança de orientação
- Taxa de mudança de velocidade
- Taxa de paragem (do original, *Stop Rate)*
- Velocidade Média
- Velocidade Máxima
- Taxa de proximidade da paragem de autocarros

No caso das três primeiras *features*, seguiu-se uma abordagem diferente dos autores [18] ao considerar o número total de pontos GPS recolhidos na viagem em vez de considerar a distância total percorrida. Esta escolha prende-se com o facto de, assim, ser exigido um menor número de cálculos do que aqueles efetuados pela abordagem em [18], obtendo, no final, resultados coerentes comparativamente aos obtidos pela abordagem referida. Ainda relativamente às três primeiras características é necessário definir valores limite para efetuar os cálculos necessários para a obtenção do valor das *features.* Como tal, foram testados vários valores para averiguar qual o melhor valor a atribuir a esses valores limite. Os valores encontrados podem ser consultados na Tabela 8.

Tabela 8 - Valores atribuídos aos limites

Feature	**Limite**	**Valor**
Heading Change Rate	Hc	20
Velocity Change Rate	Vr	3.4
Stop Rate	Vs	1
Bus Stop Closeness Rate	Bs	20

Segue-se uma explicação mais detalhada relativamente a cada uma destas características e a sua importância no algoritmo de inferência.

4.2.1.1. Taxa de mudança de orientação

Dado que a mobilidade dos carros se encontra restringida à estrada, estes apresentam menos flexibilidade em relação à mudança repentina da sua orientação. Pelo contrário, os peões apresentam uma maior variação na orientação do seu trajeto à medida que se deslocam. Desta forma a taxa de mudança de orientação tem como objetivo medir a frequência com que o sentido de orientação muda ao longo da viagem, ao considerar o número de vezes que a orientação do utilizador é superior ao limite Hc descrito na Tabela 8. Esta *feature* é calculada da seguinte forma:

$$\mathrm{HCR}=\frac{\mathrm{Pc}}{\#Amostras} \qquad (1)$$

Onde Pc representa o número de pontos GPS cuja orientação é superior ao limite Hc. Trata-se de uma *feature* que permite facilitar a distinção entre o modo Pedestre e os modos motorizados como é o caso do modo Carro e Autocarro, visto que para o modo Pedestre é esperado um valor mais elevado do que nos modos referidos.

4.2.1.2. Taxa de paragem

Ao analisar os padrões de paragens dos diferentes meios de transporte, torna-se claro que é o modo Autocarro que apresenta um maior número de paragens ao longo de uma viagem. Embora tanto o modo Carro como Pedestre possam também apresentar um número significativo de paragens, normalmente o padrão de paragens apresentado será distinto [18].

O *stop rate* tem em conta o número de vezes que a velocidade do utilizador é inferior ao limite Vs, definido na Tabela 8, ao longo de um segmento:

$$SR=\frac{Ps}{\#Amostras} \tag{2}$$

Onde Ps representa a quantidade de pontos GPS cuja velocidade é inferior ao limite Vs. Esta *feature* é utilizada para distinguir o modo Autocarro do modo Carro, visto que no caso do modo Autocarro o facto de ter de efetuar mais paragens para deixar entrar ou sair passageiros, o valor esperado para este modo seja superior ao valor obtido pelo modo Carro.

4.2.1.3. Taxa de mudança de velocidade

Tal como acontece com *stop rate*, diferentes modos de transporte apresentam padrões na mudança de velocidade diferentes. Esta mudança é avaliada ao contar o número de vezes que a velocidade é superior ao limite Vr, definido na Tabela 8, ao longo de um segmento:

$$VCR=\frac{Pv}{\#Amostras} \tag{3}$$

Onde Pv representa o total de pontos GPS cuja velocidade é superior ao limite Vr. Trata-se de uma taxa que tem como objetivo facilitar a distinção entre os modos motorizados como o modo Carro e Autocarro do modo Pedestre. Visto que quando a deslocação é realizada pelo modo Pedestre apresenta uma velocidade mais ou menos constante, o valor desta taxa para este modo será mais baixa do que o valor registado para os restantes modos referidos.

4.2.1.4. Taxa da proximidade da paragem de autocarros

Esta *feature* relaciona a distância do utilizador relativamente à paragem de autocarro que está mais próxima do seu trajeto. No caso dos nossos testes, foram utilizadas as coordenadas geográficas das paragens de Coimbra obtidas através da plataforma *One.Stop.Transport*, enquanto que as coordenadas das paragens necessárias para usar dados obtidos noutros pontos do país, foram obtidas através do *OpenStreetMap*[23]. Ao viajar de autocarro, e de acordo com [20], esta taxa tende a ser maior do que nos restantes modos de transporte sendo calculada contando o número de vezes que a distância entre o utilizador e a paragem de autocarro é inferior ao limite Bs, definido na Tabela 8, ao longo de uma viagem:

$$\mathrm{BSCR}=\frac{\mathrm{PS}}{\text{Duração Total}} \tag{4}$$

Onde PS representa a quantidade de pontos GPS cuja distância Euclidiana é inferior ao limite Bs. Com esta *feature*, pretende-se facilitar a distinção entre o modo Autocarro e o modo Carro, visto que são esperados valores mais elevados para o modo Autocarro.

As zonas de concentração de paragens de autocarro utilizadas nos nossos testes podem ser visualizas na Figura 11.

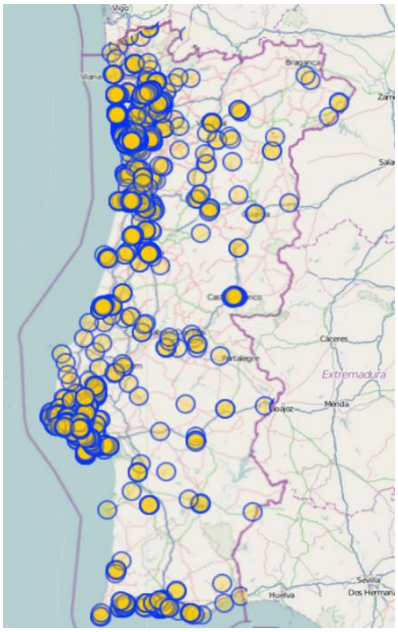


Figura 11 - Paragens de autocarro utilizadas no cálculo da feature

[23] http://www.openstreetmap.org/

4.2.2. Extração de *features* dos dados do Acelerómetro

Dado que o GPS é um sensor que nem sempre se encontra ativo devido à elevada carga que este tem na bateria do *smartphone*, bem como devido a problemas técnicos de georreferenciação dos próprios equipamentos, considerou-se que talvez estes cortes pudessem ser compensados através da fusão com os dados obtidos pelo acelerómetro para identificar o modo de transporte do utilizador.

Baseando esta exploração inicial na referência [19], e tal como foi feito pelos seus autores, ao longo do processo de recolha de dados não foi aplicada qualquer restrição relativamente à posição do *smartphone*. Devido aos bons resultados obtidos em [33], optou-se por calcular igualmente a aceleração total nos três eixos:

$$\hat{a}_{tot}(t) = \sqrt{\hat{a}_x(t)^2 + \hat{a}_y(t)^2 + \hat{a}_z(t)^2} \quad (5)$$

Onde $\hat{a}_{tot}$ representa a aceleração total calculada, e os componentes $\hat{a}_x$. $\hat{a}_y$ *e* $\hat{a}_z$ dizem respeito à aceleração obtida para os eixos x, y e z respetivamente.

Desta forma a orientação do *smartphone* deixa de ser problemática, visto que passa a ser considerada a soma da aceleração nos três eixos e não num eixo em particular. Desta forma, torna-se possível avançar para a fase seguinte.

Foi escolhido o tamanho da janela seguido da sobreposição das janelas, com o objetivo de analisar o valor da aceleração registada pelo acelerómetro nessas janelas temporais. Por um lado, janelas pequenas e uma sobreposição grande permitem mais instâncias para treinar o classificador. No entanto, as janelas pequenas não são suficientes para detetar as particularidades dos diferentes modos de transporte porque não é recolhida informação suficiente nessas janelas. Com isto em mente, e com base nos resultados obtidos em [22], optámos por aplicar uma janela de cinco segundos com uma sobreposição de cinquenta por cento à aceleração total obtida anteriormente.

De seguida é aplicada a *FTT* de 256 componentes em cada janela, de forma a obter os coeficientes necessários para efetuar os cálculos das seguintes *features*:

- Energia do sinal
- Média do sinal
- Entropia do sinal

As duas primeiras *features* foram escolhidas dados os bons resultados obtidos em [23] e a última pelos resultados obtidos em [34]. Estas são explicadas, de forma mais detalhada, nas secções seguintes.

4.2.2.1. Energia do sinal

Atividades como caminhar, correr e andar de bicicleta apresentam um sinal com maior energia do que modos de transporte como o Carro e Autocarro visto que se tratam de modos em que o *smartphone* regista mais movimentos, sendo portanto uma *feature* muito útil na distinção entre estes meios de transporte.

O valor da energia do sinal é calculado através da fórmula:

$$\text{Energia}=\sum_{i=1}^{n}\frac{{a_i}^2+{b_i}^2}{n} \tag{6}$$

Onde a *e* b dizem respeito aos coeficientes da parte real e imaginária respetivamente da FFT. Por fim, o elemento n representa o tamanho da janela.

4.2.2.2. Média do sinal

Esta é uma *feature* com um custo computacional baixo [35] e que permite a remoção de picos aleatórios e ruído do sinal. Pode ser obtido através da fórmula:

$$\text{Média}=\sum_{i=1}^{n}\frac{{a_i}^2}{n} \tag{7}$$

Onde a diz respeito ao coeficiente da parte real da FFT e o elemento n representa o tamanho da janela.

4.2.2.3. Entropia do sinal

A entropia é útil na distinção de sinais que possuem energias semelhantes mas que correspondem a padrões de mobilidade diferentes, isto é, visto que cada meio de transporte apresenta um valor característico para a entropia registada, esta torna-se uma medida útil na distinção entre o modo Carro e Autocarro cujos sinais possuem uma energia semelhante. A fórmula que permite o cálculo desta componente é dada por:

$$\text{Entropia} = \sum_{i=1}^{n} \frac{\sqrt{{a_i}^2 + {b_i}^2}}{\sum_{k=1}^{n} \sqrt{{a_k}^2 + {b_k}^2}} \quad (8)$$

Onde a *e* b dizem respeito aos coeficientes da parte real e imaginária respetivamente da FFT. Por fim, o elemento n representa o tamanho da janela.

A Figura 12 pretende esclarecer todo o processo descrito nesta secção, de forma a torna mais clara a maneira como as *features* necessárias são obtidas.

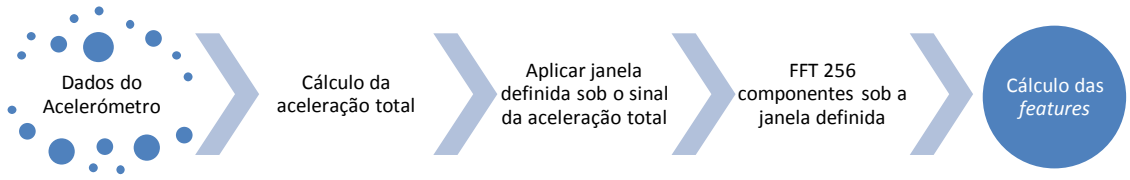


Figura 12 - Descrição do tratamento de dados do acelerómetro

4.3. Inferência do modo de transporte a partir dos sensores do *smartphone*

Nesta secção irá ser descrito o algoritmo desenvolvido para a inferência do modo de transporte utilizando os dados do GPS dos *smartphones* dos utilizadores da aplicação. De seguida, é apresentada uma abordagem teórica para o desenho de um possível algoritmo de inferência de modos de transporte onde são utilizados os dados GPS juntamente com os do acelerómetro.

4.3.1. Algoritmo de inferência: GPS

Depois de realizado o tratamento dos dados e extraídas as *features* necessárias, procede-se para a comparação da acuidade dos diferentes classificadores de forma a averiguar qual o que se adequa melhor à inferência dos modos de transporte, cujos resultados podem ser consultados na Secção 5.3.1. Como se pode aí observar, o

classificador que obteve melhores resultados foi o *W-RandomForest*, que corresponde à implementação do classificador Random Forest em *Java* pela biblioteca do *software Weka*[24], pelo que foi este o escolhido para a nossa implementação.

Segue-se a fase de segmentação das viagens dos utilizadores, onde optámos por realizar uma divisão fixa em segmentos de acordo com uma janela temporal. Foram testadas janelas de trinta segundos, um minuto, um minuto e trinta segundos e dois minutos. As duas últimas revelaram ser demasiado grandes e não permitiam detetar pontos em que se verificava uma mudança de modo de transporte. A primeira mostrou ser um período demasiado pequeno para recolher *features* significativas, fazendo com que os segmentos fossem classificados como provindo de meios de transporte errados. Posto isto, a janela que se mostrou utilizável para segmentar as viagens foi a janela correspondente a um minuto. A Figura 13 descreve o processo de classificação do modo de transporte, onde, a verde, se encontram as fases descritas acima e, a azul, a próxima etapa, que passamos a descrever

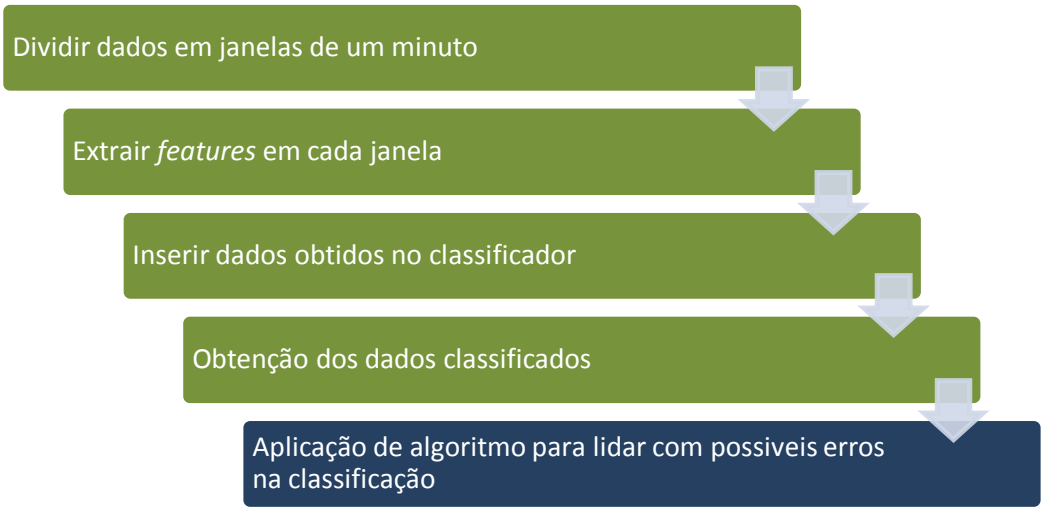


Figura 13 - Processo de classificação dos dados

Apesar do classificador treinado apresentar um bom desempenho (ver Secção 5.3.1), está naturalmente sujeito a erros nas classificações efetuadas. De forma a minimizar estes erros, foi desenvolvido um algoritmo que identifica possíveis erros e aplica as correções necessárias. Na Secção 5.4 encontram-se os testes realizados para averiguar o impacto que este algoritmo de correção tem na classificação final. Segue-se a Figura 14 que ilustra alguns casos detetados pelo algoritmo e as correção efetuadas.

[24] http://www.cs.waikato.ac.nz/ml/weka/

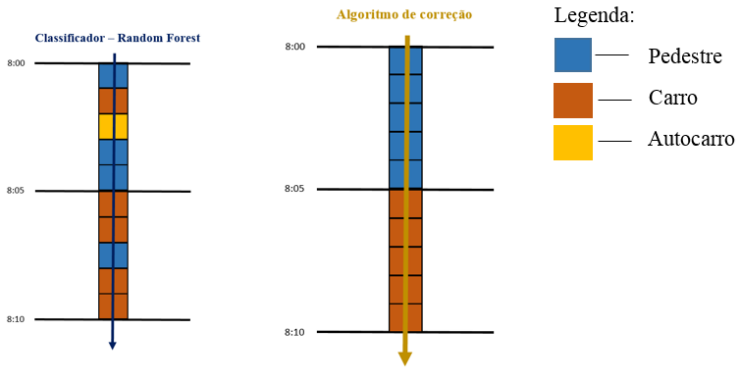


Figura 14 -Alguns exemplos de classificações erradas e respetivas correções

Segue-se a apresentação do pseudocódigo na Figura 15, que diz respeito à implementação dos algoritmos referidos anteriormente.

```
//ALGORITMO DE CLASSIFICALÃO
Verificar se o modelo treinado existe
    Se existir
        Carregar o modelo para uma estrutura
    Caso contrário
        Criar e treinar o modelo
        Guardar o modelo criado
Segmentar a viagem em janelas de 1 minuto
    Extrair features em cada janela
Classificar as instâncias obtidas
//ALGORITMO DE CORRECÇÃO
Verificar se é a primeira vez que o algoritmo analisa as instâncias obtidas
    Caso seja a primeira vez
        Se encontrar casos que se enquadrem com este cenário
            Carro
            Autocarro
            Carro
        (por exemplo)
            Alterar classificação para o modo de transporte Autocarro (neste caso)
        Se encontrar casos que se enquadrem com este cenário
            Pedestre
            Carro
            Autocarro
            Pedestre
        (por exemplo)
            Alterar a classificação do modo Carro e Autocarro para o modo Pedestre (neste caso)
    Caso seja a segunda vez
        Se o primeiro elemento da classificação for diferente do segundo
            Substituir o valor do primeiro elemento pelo valor que se encontra no segundo
        Se encontrar casos que se enquadrem com este cenário
            ...
            Pedestre
            Pedestre
            Pedestre
            Autocarro
            Carro
            Carro
            Carro
            ...
        (por exemplo)
            Substituir o modo Autocarro pelo modo Carro*
    Preparar o output do algoritmo
    Output[Modo=Pedestre, Tempo inicial=2013-06-06 19:45:36 , Tempo final=2013-06-06 20:16:44]

    *Visto que não há maneira de saber ao certo a que modo corresponde, optou-se por se substituir pelo
valor que se encontra a seguir
```

Figura 15 - Pseudocódigo dos algoritmos desenvolvidos

4.3.2. Abordagem teórica utilizando dados GPS e acelerómetro

Este é um trabalho que se ainda se encontra em desenvolvimento, visto que a componente relativa ao tratamento dos dados do acelerómetro se revelou computacionalmente inadequada devido à dimensão elevada da quantidade de dados que são necessários processar. No entanto, os resultados preliminares obtidos e apresentados

na Secção 5.3.2 referentes às classificações utilizando dados do acelerómetro, permitiram desenvolver uma abordagem teórica de um possível algoritmo juntando estes dois tipos de dados.

A abordagem proposta é semelhante à descrita na secção anterior, mas neste caso após efetuada a classificação da janela de um minuto iremos dividi-la em subjanelas de cinco segundos com uma sobreposição de cinquenta por cento, sob as quais iremos extrair as *features* do acelerómetro e classificar cada uma das subjanelas. O passo seguinte consiste em averiguar qual o modo de transporte obtido mais vezes na classificação dos segmentos do acelerómetro com a classificação do GPS, se estas forem diferentes o segmento será classificado com o meio de transporte obtido pelo acelerómetro, visto que o GPS se baseia maioritariamente na velocidade para classificar o modo de transporte. Já o acelerómetro é mais sensível a variações no movimento do utilizador permitindo assim distinguir mais facilmente uma pessoa parada a conversar com um amigo de um carro parado no trânsito, por exemplo. Caso sejam iguais mantém-se a classificação obtida pelo GPS. A Figura 16 visa clarificar a abordagem apresentada.

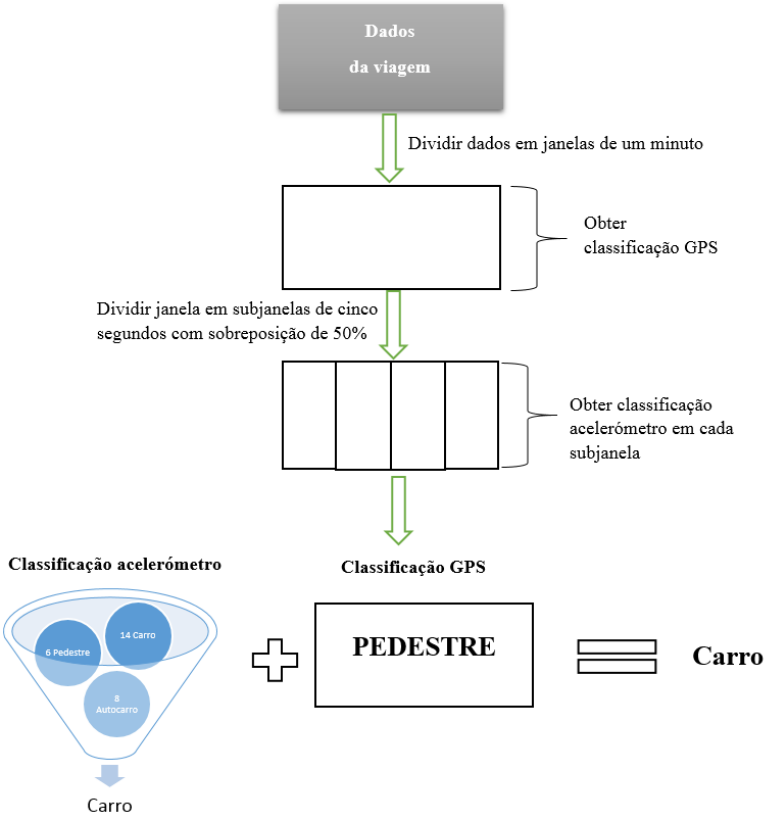


Figura 16 - Abordagem teórica usando dados GPS e do acelerómetro

4.4. Aplicação *Web* para visualização de mobilidade individual

Até ao momento os voluntários limitavam-se apenas a recolher dados sem poder visualizar qualquer tipo de informação relativamente à sua mobilidade. A aplicação *web* foi desenvolvida com o objetivo de oferecer alguma informação ao utilizador derivada da mobilidade identificada.

Por outro lado, esta aplicação serviu para verificar se os dados estão a ser recolhidos corretamente, permitindo quer validar o nosso algoritmo, quer oferecer *feedback* à empresa *SmartMove* e mitigar possíveis problemas encontrados na aplicação Android.

A aplicação foi desenvolvida com a Play Framework[25] e o facto de utilizar *Java* como linguagem de programação facilitou o seu desenvolvimento. É constituída por duas *tabs*, na primeira encontramos informação relativamente aos meios de transporte utilizados até ao momento (Figura 17). Desta forma o utilizador pode verificar qual o modo de transporte que usa mais nas suas deslocações.

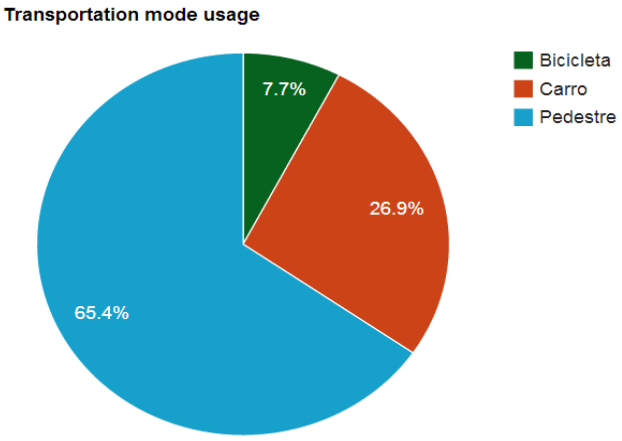


Figura 17 - Gráfico de utilização dos modos de transporte

[25] http://www.playframework.com/

De seguida é apresentado um gráfico (Figura 18) que mostra ao utilizador a utilização dos meios de transporte relativamente ao longo do período de um mês, tendo a opção de selecionar o mês para o qual deseja ver a sua informação de mobilidade.

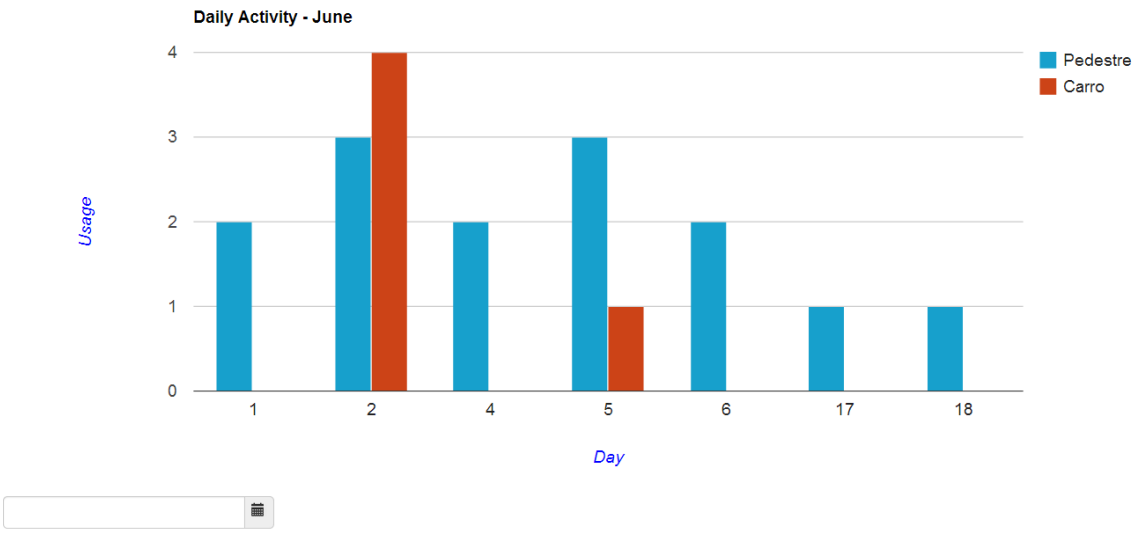
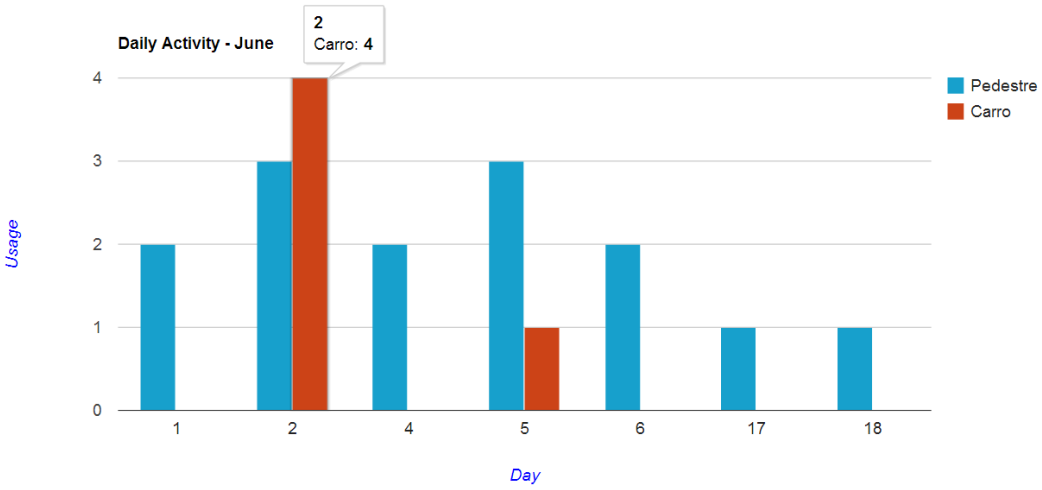


Figura 18 - Gráfico com informação de mobilidade relativo ao mês escolhido

Caso o utilizador queira obter informação mais detalhada relativamente a um dia específico do mês, basta selecionar o dia do modo de transporte sobre o qual quer obter mais informação (Figura 19).



ID	Mode	Date	Duration	Distance(km)	Average Speed (Km/h)
84	Carro	2013-06-02 20:05:25.0	0h4min41s	4.01	51.33
1677411746	Carro	2013-06-02 20:07:29.0	0h1min1s	0.48	28.2
85	Carro	2013-06-02 20:10:25.0	0h6min34s	3.54	32.32
-1866105055	Carro	2013-06-02 20:12:25.0	0h2min48s	1.58	33.86

Figura 19 - Gráfico e tabela com informação de mobilidade

Neste exemplo o utilizador selecionou o dia 2 do mês de Junho e deslocações efetuadas com o carro a servir de meio de transporte. Visto que nesse dia andou quatro vezes de carro, é lhe apresentada uma tabela com dados relativos a essas quatro viagens.

A segunda *tab* contém do lado esquerdo uma tabela onde podemos encontrar informação relativa a todas as viagens efetuadas pelo utilizador até ao momento, e do lado direito o mapa onde aparece o trajeto percorrido na viagem que o utilizador selecionar (Figura 20).

ID	Mode	Date	Duration	Distance(km)	Average Speed (Km/h)
9	Pedestre	2013-05-27 17:33:30.0	0h24min57s	2.44	5.87
10	Carro	2013-05-27 18:18:34.0	0h6min43s	1.53	13.62
59	Carro	2013-05-31 10:06:48.0	1h24min52s	110.78	78.32
-641404480	Pedestre	2013-05-31 19:06:38.0	0h4min1s	0.34	5.05
67	Pedestre	2013-05-31 19:10:40.0	0h6min42s	0.5	4.46
81	Pedestre	2013-06-01 14:31:45.0	0h7min3s	0.43	3.68
82	Pedestre	2013-06-01 14:52:14.0	0h18min59s	1.61	5.08
83	Pedestre	2013-06-02 18:30:16.0	0h3min49s	0.36	5.64
84	Carro	2013-06-02 20:05:25.0	0h4min41s	4.01	51.33
1677411746	Carro	2013-06-02 20:07:29.0	0h1min1s	0.48	28.2
85	Carro	2013-06-02 20:10:25.0	0h6min34s	3.54	32.32
-1866105055	Carro	2013-06-02 20:12:25.0	0h2min48s	1.58	33.86
86	Pedestre	2013-06-02 22:23:50.0	0h8min42s	0.82	5.62
1602844272	Pedestre	2013-06-02 22:25:51.0	0h4min13s	0.39	5.53
-2048700598	Pedestre	2013-06-04 14:00:56.0	0h27min29s	2.38	5.2
-1513383592	Pedestre	2013-06-04 19:38:17.0	0h33min38s	2.42	4.31
1802581297	Pedestre	2013-06-05 13:19:36.0	0h35min14s	3.48	5.93
630921200	Pedestre	2013-06-05 13:57:35.0	0h10min48s	0.94	5.2
-1442626339	Pedestre	2013-06-05 14:55:10.0	0h43min0s	4.63	6.46
136304363	Carro	2013-06-05 18:44:15.0	0h10min56s	4.55	24.97
1566214664	Pedestre	2013-06-06 19:44:43.0	0h7min28s	0.74	5.95
-1696643056	Pedestre	2013-06-08 19:58:20.0	0h16min16s	0.99	3.66
1807688550	Pedestre	2013-06-17 12:30:58.0	0h21min22s	1.94	5.46
1460970135	Pedestre	2013-06-18 17:24:31.0	0h9min48s	1.05	6.43
-804146394	Bicicleta	2013-08-06 18:27:23.0	0h4min25s	1	13.58
-54320987	Bicicleta	2013-08-06 18:32:31.0	0h13min28s	2.94	13.08

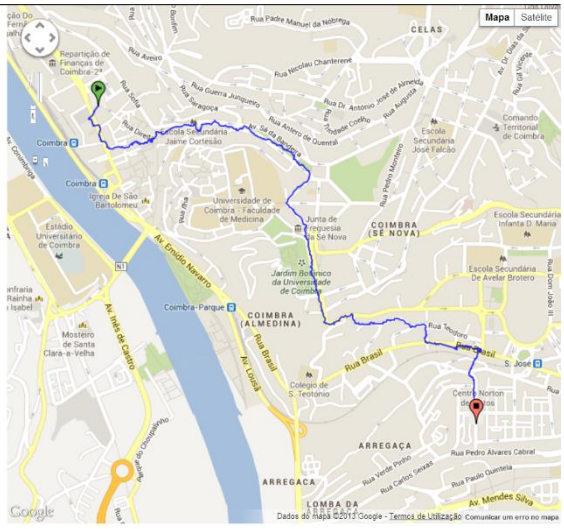


Figura 20 - Tabela e mapa com informação de mobilidade do utilizador

4.5. Integração do algoritmo

Concluída a fase de implementação do algoritmo de inferência, procedeu-se à etapa seguinte do projeto que consiste na integração deste algoritmo com a aplicação Android *mySteps* e com a aplicação web apresentada na Secção 4.4.

4.5.1. Integração do algoritmo na aplicação Android *mySteps*

O facto de o nosso algoritmo ter sido desenvolvido em *Java*, facilitou bastante esta integração. Nesta fase, procedeu-se á alteração do *output* gerado de forma a cumprir os requisitos colocados pelo *developer* responsável pelo desenvolvimento da aplicação. Inicialmente o algoritmo retornava o modo de transporte inferido juntamente com a hora de início e fim em que esse modo foi usado. No entanto com esta abordagem, o servidor

responsável por atribuir o modo de transporte às deslocações efetuadas não conseguia atribuir o modo de forma correta à respetiva secção. Foi então necessário adaptar o *output* para que no final o algoritmo devolver as coordenadas do trajeto associadas ao modo de transporte inferido, como por exemplo: latitude, longitude – modo de transporte.

Segue-se a Figura 21 que pretende clarificar de que forma o algoritmo se relaciona com a aplicação Android e com o servidor.

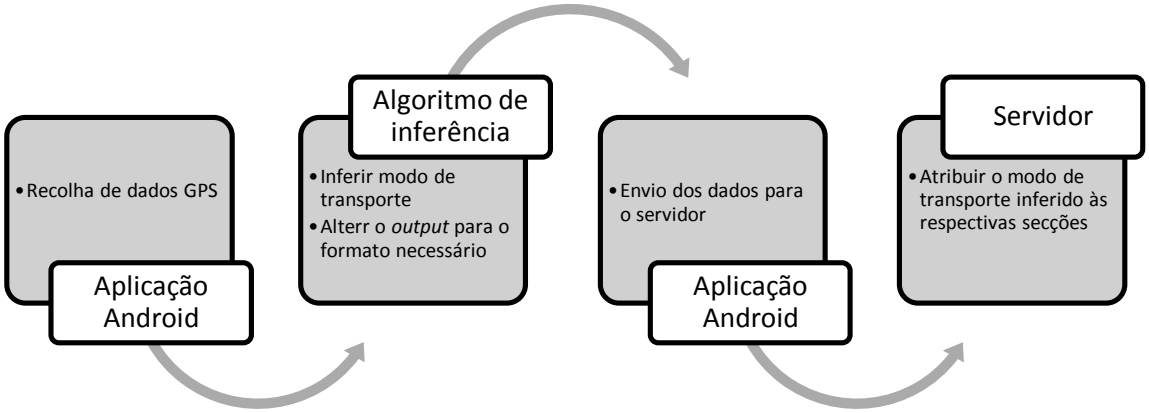


Figura 21 - interação do algoritmo com a aplicação Android e com o servidor

Após se completar este processo, a informação das deslocações é apresentada na aplicação. Desta forma o utilizador pode consultar as viagens realizadas, o modo de transporte usado, a duração das viagens como se pode verificar pela Figura 22.

Figura 22 - Captura de ecrã da aplicação *mySteps*

4.5.2. Integração do algoritmo na aplicação Web

Esta integração foi realizada com o objetivo de testar o algoritmo de inferência de modos de transporte. Desta forma a comparação dos resultados obtidos com o *ground truth* torna-se mais fácil de visualizar. Torna-se também mais fácil para os utilizadores dar *feedback* em relação ao algoritmo, podendo assim efetuar melhorias caso seja necessário. Para começar o utilizador deve colocar a data que quer analisar para que de seguida a informação relativamente às viagens realizadas nesse dia seja disponibilizada (Figura 23).

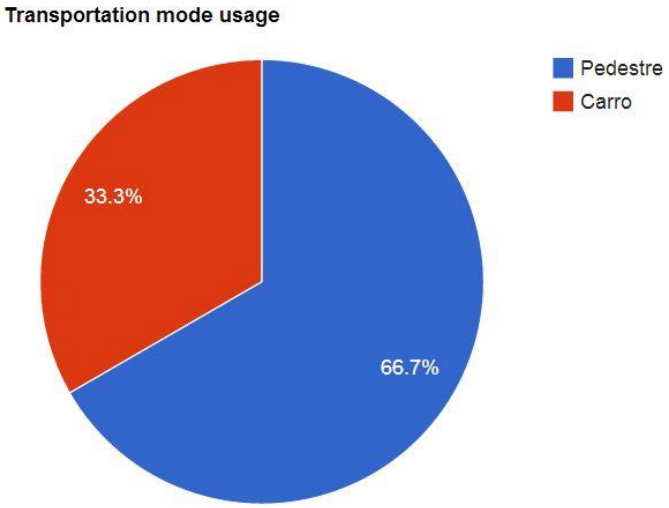


Figura 23 - Gráfico de utilização dos modos de transporte após a aplicação do algoritmo de inferência

À semelhança do gráfico apresentado na Figura 17, podemos observar os transportes utilizados no dia selecionado. No entanto, os modos apresentados neste gráfico são aqueles que foram identificados pelo algoritmo de inferência desenvolvido.

Para além deste gráfico, o utilizador pode também consultar duas tabelas (Figura 24). Onde na tabela da esquerda se encontram as secções divididas pelo algoritmo de inferência, enquanto que à direita pode consultar o *ground truth* que possuí as secções divididas pelo utilizador através da aplicação Android *mysteps.*

Mode	Beginning	End
Pedestre	2013-06-02 18:31:16	2013-06-02 18:33:17
Carro	2013-06-02 20:05:43	2013-06-02 22:23:54
Pedestre	2013-06-02 22:24:54	2013-06-02 22:32:32

Mode (Real)	Beginning (Real)	End (Real)
Pedestre	2013-06-02 18:30:16.0	2013-06-02 18:34:05.0
Carro	2013-06-02 20:05:25.0	2013-06-02 20:16:59.0
Pedestre	2013-06-02 22:23:50.0	2013-06-02 22:32:32.0

Figura 24 - Tabelas com informação das secções obtidas pelo algoritmo e pelo *ground truth* respetivamente

Capítulo 5

Resultados e Análise

Neste capítulo são apresentadas os programas e o equipamento utilizado para realizar os testes efetuados, os resultados obtidos e a respetiva análise. São apresentadas as métricas utilizadas e o desempenho dos classificadores escolhidos utilizando dois *dataset* distintos: um com dados relativos ao GPS e outro com dados gerados pelo acelerómetro. De seguida é avaliado de que forma a segmentação e o algoritmo de correção influenciam o desempenho do algoritmo.

5.1. Programas e equipamento utilizado

Na fase de desenvolvimento dado que a linguagem escolhida foi *Java,* foi utilizado como ambiente de desenvolvimento o programa *Eclipse*[26]. Este foi escolhido em detrimento de outros, por oferecer todas as funcionalidades necessárias para o desenvolvimento e por ser uma ferramenta que já foi usada no desenvolvimento de projetos anteriores. Devido à linguagem escolhida foi usada a biblioteca do *software Weka versão 3.6.9,* uma ferramenta bastante usada em projetos de *data mining* e fácil de integrar com novos projetos.

Na fase de testes, o programa selecionado para efetuar a avaliação dos classificadores foi o *Rapid Miner*[27] *versão 5.3.013.* Este possuí uma interface gráfica fácil de utilizar e permite a integração de vários módulos, nomeadamente do *Weka.* Com este programa torna-se mais fácil de criar os cenários de teste e de interpretar os resultados obtidos. Os testes foram realizados num computador portátil cujas especificações se encontram na Tabela 9.

[26] http://www.eclipse.org/
[27] http://rapid-i.com/

Tabela 9 - Especificações técnicas da máquina de testes

PC 64BITS	
SISTEMA OPERATIVO	Windows 7 Premium
RAM	8GB
CPU	Intel i7 Quad-Core @ 2.2 Ghz
DISCO RÍGIDO	5400Rpm

5.2. Métricas de desempenho

Para que seja possível averiguar o desempenho dos classificadores escolhidos, definimos em primeiro lugar a matriz de confusão que permite representar os resultados possíveis da classificação como mostra a Tabela 10:

Tabela 10 - Matriz de confusão para classificação binária

		Valor Verdadeiro	
		Positiva	**Negativa**
Valor Previsto	**Positiva**	Verdadeiro Positivo (Vp)	Falso Positivo (Fp)
	Negativa	Falso Negativo (Fn)	Verdadeiro Negativo (Vn)

No entanto, a matriz apresentada é para problemas binários. Este é um problema em que a classificação atribui uma de quatro classes possíveis: pedestre, carro, autocarro, bicicleta. Podemos afirmar portanto que se trata de um problema multi-classe. Desta forma, quando a matriz de confusão é construída e para calcular os valores acima descritos, consideramos a classe que queremos analisar como a classe positiva e as restantes como classes negativas.

Através da matriz obtida torna-se possível calcular as métricas a utilizar na avaliação dos classificadores. As fórmulas utilizadas no seu cálculo foram retiradas de [36] e encontram-se descritas na Tabela 11.

Tabela 11 - Métricas de desempenho

Métrica	***Fórmula***
Recall	$\frac{Vp}{Vp + Fn}$
Precisão	$\frac{Vp}{Vp + Fp}$
$\boldsymbol{F_1}$	$\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
Acuidade	$\frac{Vp + Vn}{Vp + Fp + Fn + Vn}$

A métrica *recall* indica o número de segmentos inferidos pelo classificador que pertencem ao modo de transporte em questão, atingindo o seu valor máximo (100%) na ausência de Falsos Negativos. A precisão diz respeito ao número de segmentos inferidos pelo classificador que realmente pertencem ao meio de transporte em questão, atingindo o seu valor máximo (100%) na ausência de Falsos Positivos. No entanto isoladamente estas métricas não oferecem informação suficiente para avaliar o desempenho do classificador, como tal é utilizada uma métrica que relaciona estas duas métricas, denominada F_1. Para avaliar o desempenho global do classificador é calculado o macro-F_1, que consiste na média dos valores F_1 calculados para cada modo de transporte, tem como valor máximo 100%. A acuidade mede a proporção de resultados corretos inferidos pelo classificador, tem como valor máximo 100%. Visto que estamos perante um *dataset* que não se encontra balanceado (ver Tabela 5), optou-se por usar como medida para

avaliar o desempenho geral do classificador a macro-F_1, dado que a acuidade nestes casos pode ser enganadora, pois o classificador pode não conseguir prever corretamente um dado meio de transporte e ainda assim obter uma acuidade elevada.

5.3. Seleção de *features*

Para esta etapa foi utilizado o teste de Kruskal-wallis para avaliar as features mais discriminantes. Trata-se de um método não paramétrico que através do coeficiente do Chi-Quadrado, nos permite avaliar o poder discriminativo de cada *feature*. É analisado de seguida o valor do p-value, que deve ser baixo (normalmente abaixo de 0.005 [37]) de forma a averiguar se é ou não possível rejeitar a hipótese nula. Os resultados obtidos pelo teste encontram-se na Tabela 12.

Tabela 12-Resultados obtidos no teste de Kruskal-wallis

FEATURES	CHI-QUADRADO	P-VALUE
VELOCIDADE MÁXIMA	563,32	9.0061e-122
TAXA DE MUDANÇA DE VELOCIDADE	485,13	7.9432e-105
VELOCIDADE MÉDIA	477,89	2.9478e-103
TAXA DE PARAGEM	338,92	3.7364e-073
TAXA DA PROXIMIDADE DA PARAGEM DE AUTOCARROS	209,47	3.7820e-045
TAXA DE MUDANÇA DE ORIENTAÇÃO	37,33	3.9219e-008

As *features* selecionadas apresentam um poder discriminativo bastante bom, sendo a taxa de mudança de orientação a que apresenta o valor mais baixo. No entanto, o seu p-value permite-nos rejeitar a hipótese nula, isto é, esta *feature* permite diferenciar as classes apesar de ter um poder discriminativo muito baixo. Na Secção que se segue, serão apresentados os resultados obtidos avaliando os classificadores. Para tal, após encontrado o melhor classificador são considerados dois cenários: um onde são utilizados todas as *features* e outro onde apenas são consideradas as melhores, ou seja, é excluída a *feature* que diz respeito à taxa de mudança de orientação.

5.4. Desempenho dos classificadores

O método escolhido para avaliar o desempenho dos classificadores foi o *k-fold cross-validation* para garantir imparcialidade na divisão dos subconjuntos. Este método consiste na divisão do *dataset* em *k* subconjuntos, onde um dos subconjuntos será utilizado para testar os resultados do classificador. Para os testes realizados optou-se por um *k = 10*, visto tratar-se do valor utilizado mais comum [38]. Os dados contidos em cada subconjunto encontram-se estratificados para garantir que cada um contém o mesmo número de amostras por modo de transporte.

5.4.1. Desempenho dos classificadores sobre dados GPS

O *dataset* utilizado nesta fase de testes e avaliação é constituído apenas por dados recolhidos pelo GPS. Numa primeira fase exploratória, começamos por avaliar o desempenho do algoritmo implementado apenas os dados em bruto, isto é, sem aplicar os filtros definidos na Secção 4.1. As Tabelas 13, 14 e 15 apresentam os resultados obtidos pelos diversos classificadores nesta análise preliminar à adequação e robustez dos algoritmos de classificação.

Tabela 13 – Desempenho do classificador W-RandomForest com os dados do GPS

		W-RandomForest			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	93.90%	94.25%	82.32%	25%
	Precisão	90.50%	92.61%	89.40%	100%
	F_1	92.17%	93.42%	85.71%	40%
	Macro-F_1	77.83%			
	Acuidade	95.74%			

Tabela 14 - Desempenho do classificador W-J48 com os dados do GPS

		W-J48			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	95.31%	89.60%	77.44%	50%
	Precisão	88.26%	91.63%	82.47%	29%
	F_1	91.65%	90.60%	79.88%	36.36%
	Macro-F_1	74.62%			
	Acuidade	94.24%			

Tabela 15 - Desempenho do classificador W-NaiveBayes com os dados do GPS

		W-NaiveBayes			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	93.90%	88.27%	59.15%	0%
	Precisão	72.99%	89.46%	85.84%	0%
	F_1	82.14%	88.87%	70.04%	0%
	Macro-F_1	60.26%			
	Acuidade	91.78%			

Antes de passar para uma análise mais detalhada dos resultados obtidos foi, naturalmente, avaliado o desempenho dos classificadores utilizando desta vez os dados após serem filtrados. Os resultados obtidos podem ser consultados nas Tabelas 16, 17 e 18.

Tabela 16 - Desempenho do classificador W-RandomForest *com os dados do GPS tratados*

		W-RandomForest			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	100%	96.58%	89.17%	25%
	Precisão	99.46%	95.93%	89.74%	100%
	F_1	99.73%	96.25%	89.45%	40%
	Macro-F_1	81.36%			
	Acuidade	97.71%			

Tabela 17 - Desempenho do classificador W-J48 com os dados do GPS tratados

		W-J48			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	99.46%	92.71%	83.44%	50%
	Precisão	100%	93.78%	81.88%	28.57%
	F_1	99.73%	93.24%	82.65%	36.36%
	Macro-F_1	78%			
	Acuidade	96.11%			

Tabela 18 - Desempenho do classificador W-NaiveBayes com os dados do GPS tratados

		W-NaiveBayes			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	95.65%	90.89%	61.78%	0%
	Precisão	75.21%	90.68%	88.18%	0%
	F_1	84.21%	90.79%	72.66%	0%
	Macro-F_1	61.92%			
	Acuidade	92.86%			

Mesmo sem limpeza de dados, registaram-se valores bastante satisfatórios para os classificadores. No entanto, isto pode ser explicado pelo facto de estar a ser considerado um cenário ideal, onde os segmentos que dizem respeito aos diferentes modos de transporte se encontram bem delimitados.

Relativamente aos classificadores, o *W-NaiveBayes* embora se tenha revelado o classificador mais rápido, foi também o que apresentou piores resultados. O facto de a amostra ser tão desequilibrada leva a que modos de transporte como o Autocarro e a Bicicleta, que são os que se encontram em menor número no conjunto de dados, obtenham um valor menos elevado no cálculo das probabilidades de pertença a modo. Tendo como consequência uma taxa de erro elevada na classificação destes modos de transporte traduzindo-se numa redução significante no desempenho registado para este classificador.

Tanto o *W-RandomForest* como o *W-J48* obtiveram resultados bastante bons com o *W-RandomForest* a obter um desempenho melhor nos dois cenários. Este facto pode ser explicado pela uma maior robustez face a erros e valores discrepantes que este classificador apresenta, consequência da diversidade garantida ao selecionar aleatoriamente um dado número de *features* na fase de criar as árvores e pelo facto de na fase da classificação o resultado final ser obtido através da votação entre todas as árvores criadas. Ambos revelam no entanto uma dificuldade na identificação do modo Bicicleta. Enquanto que o *W-48* identifica mais facilmente este modo, a precisão com que o faz é extremamente baixo. Já o *W-RandomForest* embora apresente um *recall* mais baixo, isto é, mais dificuldades em identificar o modo Bicicleta, a precisão com que o faz é máxima. Analisando o valor F_1 obtido por ambos os classificadores podemos concluir que apesar de baixo, o *W-RandomForest* é o que obtém o melhor desempenho na deteção deste modo em relação aos restantes classificadores.

Assim, e após selecionar o classificador *W-RandomForest* para o restante desenvolvimento, podemos passar à avaliação do nosso algoritmo propriamente dita. Nesse sentido foi analisada a matriz de confusão apresentada na Tabela 19 com o objetivo de observar quais os modos de transporte em que a classificação obtém mais erros.

Tabela 19 - Matriz de confusão do classificador W-RandomForest sobre dados GPS utilizando todas as *features*

	Pedestre	Carro	Autocarro	Bicicleta
Pedestre	184	0	1	0
Carro	0	424	16	2
Autocarro	0	15	140	1
Bicicleta	0	0	0	1

Na matriz acima apresentada, podemos observar que o classificador apresenta uma maior dificuldade na distinção entre o modo Carro e o modo Autocarro. A semelhança no comportamento destes modos no meio urbano é a principal causa para explicar o resultado obtido. No que diz respeito ao modo Bicicleta a pouca quantidade de dados recolhidos não permitiu obter resultados conclusivos. Podemos no entanto afirmar que em geral os resultados são bastante bons.

Para finalizar, o classificador é avaliado utilizando apenas as melhores *features* selecionadas pelo teste de Kruskal-wallis e o impacto que estas têm na classificação. Ao comparar os resultados obtidos pela matriz de confusão apresentada na Tabela 20, permite-nos tirar conclusões relativamente ao impacto da escolha das *features* no classificador selecionado.

Tabela 20 - Matriz de confusão do classificador W-RandomForest sobre dados GPS utilizando as melhores *features* obtidas pelo teste de Kruskal-wallis

	Pedestre	Carro	Autocarro	Bicicleta
Pedestre	184	0	1	0
Carro	0	421	24	2
Autocarro	0	18	132	0
Bicicleta	0	0	0	2

Ao analisar as duas Tabelas, podemos observar que ao utilizar as melhores *features* o classificador obtém melhores resultados para a classe bicicleta. Relativamente

ao modo pedestre, este não sofreu qualquer alteração. No entanto, verifica-se uma ligeira dificuldade na distinção entre a classe carro e autocarro, Tendo em conta que estes são os modos de transporte mais utilizados pelos utilizadores, para o nosso classificador foi considerado o uso de todas as *features* para a classificação das diferentes classes. Relativamente ao modo pedestre, este não sofreu qualquer alteração.

5.4.2. Desempenho dos classificadores: Acelerómetro

O volume de dados gerado pelo acelerómetro é demasiado grande para ser utilizado na totalidade. Como tal, e para esta fase de testes preliminar, apenas foram considerados os dados recolhidos durante um dia e pertencentes apenas a um utilizador. Para verificar mais um modo, foram também acrescentados dados relativos a duas viagens utilizando o modo de transporte Bicicleta.

Nas Tabelas 21, 22 e 23 podemos consultar os valores obtidos que permitem comparar o desempenho dos classificadores.

Tabela 21 - Desempenho do classificador W-RandomForest com os dados do acelerómetro

		W-RandomForest			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	86.9%	94.9%	85.5%	81.5%
	Precisão	87.2%	93.4%	85.6%	85.3%
	F_1	87.05%	94.14%	85.5%	83.36%
	Macro-F_1	87.51%			
	Acuidade	94.93%			

Tabela 22 - Desempenho do classificador W-J48 com os dados do acelerómetro

		W-J48			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	85.80%	94.40%	82.00%	77.10%
	Precisão	84.60%	91.90%	84.50%	83.30%
	F_1	85.20%	93.13%	83.23%	80.08%
	Macro-F_1	85.41%			
	Acuidade	94.08%			

Tabela 23 - Desempenho do classificador W-NaiveBayes com os dados do acelerómetro

		W-NaiveBayes			
	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	22.10%	81.50%	86.90%	1.10%
	Precisão	38.00%	70.90%	39.50%	13.90%
	F_1	27.95%	75.83%	54.31%	2.04%
	Macro-F_1	40.03%			
	Acuidade	78.56%			

Uma vez mais, o *W-NaiveBayes* foi o classificador a obter o pior desempenho, com o *W-RandomForest* e o *W-J48* a obterem valores muito próximos. Os classificadores obtiveram um desempenho satisfatório, sendo agora possível detetar mais facilmente e com alguma precisão o modo Bicicleta. Uma vez mais foi o *W-RandomForest* a obter o melhor desempenho, apresentando-se superior em todas as métricas medidas.

Selecionado o melhor classificador, foi observada a matriz de confusão (Tabela 24) para analisar, com mais algum detalhe, o seu comportamento na classificação dos modos de transporte.

Tabela 24 - Matriz de confusão do classificador W-RandomForest com dados do acelerómetro

	Pedestre	Carro	Autocarro	Bicicleta
Pedestre	20406	963	1426	685
Carro	708	56399	463	1829
Autocarro	1601	459	13526	238
Bicicleta	681	2580	380	16002

O número elevado de elementos relativos a cada modo de transporte, deve-se à divisão da amostra de dados de acordo com a abordagem apresentada na Secção 4.2.2. O resultado obtido para o modo Carro e o modo Bicicleta, vai de encontro ao que era esperado, visto que ambas são atividades que registam uma baixa energia na aceleração urbana [19]. Relativamente ao resultado obtido na classificação do modo Autocarro e a confusão com o modo Pedestre, pode ser ainda explicado pelo facto da linha do autocarro em questão passar por locais cujo pavimento não se encontra nas melhores condições, fazendo com que o número de oscilações registado pelo acelerómetro seja elevado. Deve notar-se que os resultados obtidos relativamente à classificação do modo Bicicleta são bastante mais precisos quando comparados com os que foram obtidos usando apenas os dados do sensor de GPS.

Em conclusão, estes resultados embora computacionalmente dispendiosos em tempo, são bastante promissores. Carecem, no entanto, de uma análise e trabalho mais

aprofundado, que não se coaduna com o tempo oficial de apresentação da dissertação, sendo, assim, deixados para um a fase futura.

5.5. Avaliação do algoritmo de inferência

Para avaliar o algoritmo desenvolvido, foi aplicado o método de segmentação descrito na Secção 4.3.1 às viagens realizadas durante um mês em que cada semana corresponde a dados de utilizadores distintos, para garantir diversidade na amostra e nos meios de transporte utilizados. Os resultados obtidos são comparados com o *ground truth* de forma a validar a classificação e avaliar a precisão com que estas são realizadas (Tabela 25). De seguida é aplicado o algoritmo de correção e o resultado é, de novo, comparado com o *ground truth* e ainda com a classificação inicial, com o objetivo de averiguar o impacto que o algoritmo tem na classificação final (Tabelas 26).

Tabela 25 - Desempenho do algoritmo de inferência sem aplicar o algoritmo de correção

	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	98.66%	76.98%	89.36%	25.32%
	Precisão	78.97%	93.49%	45.41%	90.90%
	F_1	87.72%	84.44%	60.22%	39.61%
	Macro-F_1	68%			
	Acuidade	89.63%			

Tabela 26 - Desempenho do algoritmo de inferência com o algoritmo de correção

	Classe	Pedestre	Carro	Autocarro	Bicicleta
Métricas	*Recall*	98.44%	87.48%	100.00%	22.15%
	Precisão	76.91%	96.07%	70.68%	92.11%
	F_1	86.35%	91.57%	82.82%	35.71%
	Macro-F_1	74.11%			
	Acuidade	92%			

Analisando os resultados obtidos, pode concluir-se que o algoritmo de correção produz um impacto positivo no desempenho final da classificação com uma melhoria de 6.11%. O ponto fraco do algoritmo encontra-se na classificação do modo Bicicleta, embora o faça com relativa precisão (92.11%) apresenta dificuldades em distinguir este modo dos restantes tendo por isso um *recall* extremamente baixo. Estes valores podem ser explicados pela pouca quantidade de dados que foram recolhidos relativamente a esse modo.

Com o objetivo de analisar melhor o comportamento do algoritmo de inferência juntamente com o algoritmo de correção, foi construída a matriz de confusão apresentada na Tabela 27.

Tabela 27 - Matriz de confusão do algoritmo de inferência juntamente com o algoritmo de correção

	Pedestre	Carro	Autocarro	Bicicleta
Pedestre	443	27	0	106
Carro	6	440	0	12
Autocarro	1	33	94	5
Bicicleta	0	3	0	35

A observação dos resultados obtidos permite analisar o tipo de mobilidade realizada pelos utilizadores. Onde o modo Pedestre e o modo Carro aparecem como os meios de transporte mais utilizados nas deslocações, já o modo Autocarro e o modo Bicicleta ocupam uma minoria na escolha dos utilizadores para efetuarem as suas viagens. Esta tendência por parte dos utilizadores apresenta-se como um obstáculo na obtenção de dados relativamente a estes meios de transporte minoritários. Embora o classificador não apresente dificuldades significativas na deteção do modo Autocarro, existe uma clara dificuldade na distinção do modo Pedestre com o modo Bicicleta. No entanto este resultado pode também ser explicado pelo facto do trajeto percorrido pelo utilizador ser semelhante a um percurso efetuado no modo Pedestre e que provavelmente foi realizado a uma velocidade reduzida. Apesar dos problemas apresentados, o algoritmo desenvolvido juntamente com o algoritmo de correção apresentam resultados bastante satisfatórios da deteção dos principais modos de transporte.

Capítulo 6

Conclusões e Trabalho futuro

Para que a aplicação *mobile* a desenvolver pelo projeto ECO-Circuitos funcione corretamente e seja bem aceite no mercado, tem de ser o menos intrusiva possível com os utilizadores. Para cumprir este objetivo, até ao momento no presente estágio foi desenvolvido um algoritmo capaz de inferir automaticamente o modo de transporte utilizado no momento das deslocações, através dos dados GPS recolhidos pelos *smartphones dos* utilizadores. Atualmente com o algoritmo desenvolvido é possível detetar quatro modos de transporte: modo Pedestre, modo Carro, modo Autocarro e o modo Bicicleta. Dos modos aqui apresentados, o algoritmo consegue identificar com bastante precisão os três primeiros modos, tendo no entanto como ponto fraco a deteção do modo Bicicleta apesar da aplicação do algoritmo de correção desenvolvido. Tal como foi referido anteriormente, este problema deve-se à quantidade de dados recolhidos deste modo de transporte ser extremamente baixa. Apesar desta debilidade, a identificação dos principais meios de transporte utilizados nas deslocações do dia-a-dia é feita com uma precisão elevada (91.34% sem contar com o modo Bicicleta).

Como trabalho futuro seria interessante considerar mais meios de transporte tais como a bicicleta elétrica, o comboio entre outros. Seria também interessante explorar melhor a abordagem teórica proposta nesta dissertação, onde é efetuada a fusão dos dados GPS com os dados recolhidos pelo acelerómetro com o objetivo de melhorar a precisão com que a classificação é realizada. Torna-se necessário realizar uma análise e um trabalho aprofundado para averiguar até que ponto os dados deste último sensor podem ser utilizados nesta nova abordagem. Seria também interessante desenvolver um módulo que efetue os cálculos das *features* acelerómetro e do GPS em tempo real, reduzindo consideravelmente o tamanho dos ficheiros que são enviados para o servidor da empresa *SmartMove*. No entanto antes de avançar com esta abordagem é necessário avaliar o impacto que estes cálculos têm na bateria e no desempenho geral do *smartphone.*

A integração do algoritmo na aplicação do *smartphone* foi um passo importante para o projeto, tornando-se possível averiguar que as bibliotecas utilizadas apenas são compatíveis com *smartphones* que possuam a versão 4.0 ou superior do sistema

operativo *Android*. No entanto, não foi possível obter *feedback* por parte de todos os voluntários relativamente ao impacto do algoritmo na bateria, conseguindo apenas o testemunho de dois voluntários que afirmam não ter notado diferença entre a versão sem o algoritmo e a versão com o algoritmo. Apresentando-se assim como uma abordagem viável, inovadora e interessante a considerar para o futuro.

Em suma, apesar das limitações do algoritmo apresentado nesta dissertação os resultados obtidos indicam que tanto o algoritmo de inferência como o algoritmo de correção desenvolvidos permitem obter bons resultados na identificação dos principais modos de transporte referidos anteriormente, sem que a vida útil da bateria seja significativamente afetada. Relativamente ao acelerómetro, embora não tenha sido possível efetuar uma análise aprofundada aos dados de mais utilizadores, os resultados aqui obtidos são bastantes promissores para o desenvolvimento de uma abordagem onde apenas só são considerados os dados deste sensor e um uma outra abordagem onde é efetuada a fusão dos dados deste sensor com os dados recolhidos pelo GPS.

Referências

[1] E. Murakami and D. P. Wagner, "Can using global positioning system (GPS) improve trip reporting?," *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 2–3, pp. 149–165, 1999.

[2] R. P. Paola A. Gonzalez, Jeremy S. Weinstein, Sean J. Barbeau, Miguel A. Labrador, Philip L. Winters, Nevine Labib Georggi, "Automating mode detection using neuronal networks and assisted GPS data collected using GPS-enabled mobile phones," in *15th World Congress on Intelligent Transportation Systems*, 2008, pp. 16–20.

[3] I. Cameron, T. . Lyons, and J. . Kenworthy, "Trends in vehicle kilometres of travel in world cities, 1960–1990: underlying drivers and policy responses," *Transport Policy*, vol. 11, no. 3, pp. 287–298, Jul. 2004.

[4] T. J. Lyons, J. R. Kenworthy, C. Moy, and F. dos Santos, "An international urban air pollution model for the transportation sector," *Transportation Research Part D: Transport and Environment*, vol. 8, no. 3, pp. 159–167, 2003.

[5] W. Anderson, "Urban Form, Energy and the Environment: A Review of Issues, Evidence and Policy," *Urban Studies*, vol. 33, no. 1, pp. 7–36, Feb. 1996.

[6] J. Kenworthy and C. Townsend, "The Millennium Cities Database for Sustainable Transport. (CDROM Database) International Union (Association) of Public Transport," *Brussels and Institute for Sustainability and Technology Policy (ISTP)*, 2001.

[7] T. Hau, "Transport for urban development in Hong Kong," *Communications for Urban Development*, 2002.

[8] W. H. K. Lam and M.-L. Tam, "Reliability of territory-wide car ownership estimates in Hong Kong," *Journal of Transport Geography*, vol. 10, no. 1, pp. 51–60, 2002.

[9] C. Kang, X. Ma, D. Tong, and Y. Liu, "Intra-urban human mobility patterns: An urban morphology perspective," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1702–1717, Feb. 2012.

[10] C. Morency, A. Paez, M. J. Roorda, R. Mercado, and S. Farber, "Distance traveled in three Canadian cities: Spatial analysis from the perspective of vulnerable population segments," *Journal of Transport Geography*, vol. 19, no. 1, pp. 39–50, Jan. 2011.

[11] C. Mccarty, "Structure in Personal Networks," *JoSS aritcle*, vol. 3.

[12] T. Litman, "*Evaluating Accessibility for Transportation Planning Measuring People ' s Ability To Reach Desired Goods and Activities*," no. January 2008. 2012.

[13] J. Rodrigue, "*The geography of transport systems*," Third Edit. New York: Routledge, 2013, p. 416.

[14] F. Le Néchet, "Urban spatial structure, daily mobility and energy consumption: a study of 34 European cities," *Cybergeo : European Journal of Geography*, no. 580, 2012.

[15] J.-A. Carrasco and E. J. Miller, "The social dimension in action: A multilevel, personal networks model of social activity frequency between individuals," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 1, pp. 90–104, Jan. 2009.

[16] S. Wasserman and K. Faust, "*Social network analysis: Methods and applications*," vol. 8. Cambridge university press, 1994.

[17] M. a. . van Duijn, J. T. van Busschbach, and T. a. . Snijders, "Multilevel analysis of personal networks as dependent variables," *Social Networks*, vol. 21, no. 2, pp. 187–210, Apr. 1999.

[18] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Ma, "Understanding mobility based on GPS data," *Web. In Proc. WWW 2008, ACM Press (2008), 247-256*, no. 49, 2008.

[19] B. Nham, K. Siangliulue, and S. Yeung, "Predicting mode of transport from iphone accelerometer data," 2008.

[20] L. Stenneth, P. Yu, O. Wolfson, B. Xu, and S. Morgan, "Transportation Mode Detection using Mobile Devices and GIS Information," 2011.

[21] B. Chandra and P. Paul V, "A Robust Algorithm for Classification Using Decision Trees," in *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, 2006, pp. 1–5.

[22] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data," *AAAI*, pp. 1541–1546, 2005.

[23] V. Manzoni, D. Maniloff, K. Kloeckl, and C. Ratti, "Transportation mode identification and real-time CO2 emission estimation using smartphones," pp. 1–12, 2010.

[24] L. B. Statistics and L. Breiman, "Random Forests," in *Machine Learning*, 2001, pp. 131–163.

[25] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," in *Machine Learning*, 1997.

[26] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks*, vol. 6, no. 2, pp. 1–27, Feb. 2010.

[27] D. Piggott, "Inferring Transportation Mode using Smartphone Sensor Data," 2011.

[28] L. Devroye, "*A probabilistic theory of pattern recognition*," vol. 31. springer, 1996.

[29] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification ?," vol. 5, no. 3, pp. 37–46, 2011.

[30] S. Chakrabarti, S. Roy, and M. V Soundalgekar, "Fast and Accurate Text Classification via Multiple Linear Discriminant Projections," *The VLDB Journal*, vol. 12, no. 2, pp. 170–185, 2003.

[31] Y. Qi, "*Random Forest for Bioinformatics*." Springer US, 2012, pp. 307–323.

[32] J. R. Quinlan, "*C4.5: programs for machine learning*." San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[33] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," *Pervasive Computing*, pp. 1–17, 2004.

[34] J. Ho, "Interruptions: using activity transitions to trigger proactive messages," Massachusetts Institute of Technology, 2004.

[35] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. P. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645–662, Mar. 2010.

[36] C. Silva, "Inductive Inference for large scale text classification," Universidade de Coimbra, 2008.

[37] J. D. Spurrier, "On the null distribution of the Kruskal–Wallis statistic," *Journal of Nonparametric Statistics*, vol. 15, no. 6, pp. 685–691, 2003.

[38] J. C. Mar and G. J. McLachlan, "Model-based clustering in gene expression microarrays: an application to breast cancer data," in *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003 - Volume 19*, 2003, pp. 139–144.