Caracterização dos movimentos pendulares nas principais vias de acesso à cidade de Lisboa

Ricardo Ferreira
(2016020798)

Orientadores:

Mateus Mendes | ISEC

Nuno Lavado | ISEC

Laboratório de Dados Urbanos de Lisboa | LxDataLab

Licenciatura em Engenharia Informática

Ramo de desenvolvimento de aplicações

Instituto Politécnico de Coimbra

Instituto Superior de Engenharia de Coimbra

Julho de 2022

RESUMO

Este relatório apresenta um estudo que tem como objetivo analisar os fluxos diários de tráfego nas horas de ponta em Lisboa, considerando fatores como o calendário escolar, períodos de férias e a ocorrência de períodos de chuva nos principais pontos de entrada e saída da cidade. Utilizando dados de telemóveis, investigamos como as pessoas se movimentam durante as horas de ponta da manhã (7:00h-10:00h) e da tarde (17:00h-20:00h), explorando também outros períodos, como dias intensos de tráfego elevado, como a sexta-feira e o domingo. Além disso, aplicamos modelos preditivos para antecipar cenários futuros de mobilidade. Esses modelos foram baseados em redes neurais recorrentes (*RNNs*), em particular o *GRU*. Demostramos que esses modelos podem prever a intensidade do tráfego móvel com alguma precisão, com erros percentuais inferiores a 10%. Desta forma, permitindo antecipar cenários futuros de mobilidade e identificar pontos críticos de congestionamento nas principais vias de acesso à cidade em diferentes momentos da semana.

Palavras-chave: Fluxos diários de tráfego, Pontos de entrada e saída da cidade, Horas de ponta, Lisboa, Dados de telemóveis, Calendário escolar, Períodos de férias, Períodos de chuva, Mobilidade, Modelos preditivos, GRU (*Gated Recurrent Unit*), Pontos críticos de congestionamento.

ABSTRACT

This report presents a study aimed at analyzing the daily traffic flows during peak hours in Lisbon, considering factors such as the school calendar, vacation periods, and the occurrence of rainy periods at the city's main entry and exit points. Using mobile data, we investigated how people move during the morning (7:00 am - 10:00 am) and afternoon (5:00 pm - 8:00 pm) peak hours, as well as other periods with high traffic, such as Fridays and Sundays. Additionally, we applied predictive models to anticipate future mobility scenarios. These models were based on recurrent neural networks (RNNs), specifically the GRU (Gated Recurrent Unit). We demonstrated that these models could predict mobile traffic intensity with some accuracy, achieving error rates below 10%. Thus, enabling the anticipation of future mobility scenarios and the identification of critical congestion points on the main city access routes at different times of the week.

Keywords: *Daily traffic flows, City entry and exit points, Peak hours, Lisbon, Mobile data, School calendar, Vacation periods, Rainy periods, Mobility, Predictive models, GRU (Gated Recurrent Unit), Critical congestion points.*

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ACRÓNIMOS

CNNs – Convolutional Neural Networks

DeepTP – Deep Traffic Predictor

GNNs – Graph Neural Networks

GRU - Gated Recurrent Unit

IoT – Internet of things

IPMA – Instituto Português do Mar e da Atmosfera

ITS – Intelligent Transportation Systems

LSTM – Long short-term Memory

LxDataLab – Laboratório de Dados Urbanos de Lisboa

MLP – Multilayer perceptron

RNNs – Recurrent Neural Networks

SAEs - Stacked Autoencoders

STFGNN - Spatio-Temporal Fusion Graph Neural Networks

STGCN - Spatio-Temporal Graph Convolutional Networks

STGNN – Spatio-Temporal Graph Neural Networks

TAZs – Temporary Autonomous Zone

1 INTRODUÇÃO

A tecnologia de comunicações móveis revolucionou a forma como pessoas e empresas obtêm e partilham informações sobre mobilidade, proporcionando conjuntos massivos de trajetórias registadas, dados longitudinais e precisão em tempo real. Neste projeto, em colaboração com o Laboratório de Dados Urbanos de Lisboa e o Centro de Gestão e Inteligência Urbana de Lisboa, vamos analisar a gestão da mobilidade em Lisboa através de dados móveis, com foco na questão do planeamento quotidiano da cidade.

1.1 Enquadramento

O objetivo deste projeto é analisar os fluxos diários de tráfego nas horas de ponta em Lisboa, considerando fatores como o calendário escolar, períodos de férias e ocorrência de períodos de chuva nos principais pontos de entrada e saída da cidade. Utilizaremos dados de localização de telemóveis para entender como as pessoas se movimentam durante as horas de ponta da manhã (7:30h-10:00h) e da tarde (17:00h-19:30h). Além disso, vamos explorar outros períodos, incluindo dias intensos de tráfego elevado, sextas e domingos, e usar modelos preditivos para antecipar cenários futuros. Dessa forma, conseguimos obter uma compreensão mais completa dos padrões de mobilidade ao longo do dia, permitindo-nos identificar pontos críticos de congestionamento nas principais vias de acesso à cidade em diferentes momentos da semana.

De acordo com o desafio proposto pelo Laboratório de Dados Urbanos de Lisboa (LxDataLab) [12], esta iniciativa entre diversas instituições de investigação e de ensino superior, na qual o ISEC também está envolvido, visa utilizar ferramentas de análise de dados para extrair conhecimento especializado sobre o tráfego automóvel no município de Lisboa. Deste modo, o objetivo é desenvolver soluções analíticas que possam melhorar vários aspetos da cidade, tais como o planeamento, a segurança, a mobilidade, a gestão operacional e de emergência.

1.2 LxDataLab - Laboratório de Dados Urbanos de Lisboa

Nos últimos anos, vários desafios têm sido propostos no campo da análise de dados urbanos para melhorar a mobilidade e a qualidade de vida nas cidades. Alguns desses propostos pelo LxDataLab [10], incluem o desenvolvimento de indicadores de tráfego mais precisos e eficientes, a identificação de coberturas verdes para promover a sustentabilidade urbana, a utilização de dados de telemóveis para compreender e otimizar a mobilidade na cidade, e a análise dos movimentos pendulares e da movimentação de pessoas em zonas de diversão noturna. Esses desafios exigem a aplicação de técnicas avançadas de análise de dados para identificar os seus padrões específicos, e consequentemente melhorar a gestão urbana, o planeamento dos transportes e a qualidade de vida dos cidadãos.

O LxDataLab [10] é uma iniciativa da Câmara Municipal de Lisboa (CML), que visa aproveitar a vasta quantidade de dados disponíveis no município em diversas áreas de atuação, como transportes, urbanismo, obras, sinalização viária, edifícios, áreas verdes, tratamento de resíduos e painéis solares. De facto, desenvolve soluções analíticas avançadas, recorrendo a técnicas para inferência estatística, aprendizagem automática e outras formas de análise de dados, para melhorar o planeamento, a resiliência, a segurança, a mobilidade, a gestão operacional e a resposta a emergências na cidade de Lisboa.

Através da parceria com diversas instituições de pesquisa e ensino superior, incluindo o Instituto Superior de Engenharia de Coimbra (ISEC), permite que essas instituições tenham acesso às Bases de Dados existentes e proponham soluções analíticas, diagnósticos e previsões para diversos dados urbanos. Essa colaboração ativa contribui significativamente para a resolução dos problemas da cidade, tornando-a mais eficiente e sustentável.

1.3 Objetivos e plano de trabalhos

Os objetivos do projeto estão delineados na página oficial do LxDatalab [10]. Através da análise dos fluxos diários, principalmente entre a manhã (7:30h-10:00h) e tarde (17:00h-19:30h), nos principais pontos de entrada e saída da cidade de Lisboa, pretende-se atingir os seguintes objetivos:

- Caraterizar o volume total de entradas e saídas da cidade durante o período da hora de ponta;

- Caraterizar o volume de entradas e saídas da cidade durante o período da hora de ponta para cada um dos 11 pontos de entrada e saída;

- Comparar com outros períodos do dia;

- Relacionar o ponto anterior com variáveis como os períodos de aulas ou férias e a existência de pluviosidade;

- Análise das zonas de destino daqueles que saem da cidade;

- Análise das zonas de origem daqueles que entram na cidade.

O projeto consiste no seguinte plano de trabalhos:

- T1 – Estado da arte – recorrendo a artigos, livros, tutoriais, blogs, *github* realizar uma revisão do estado da arte dos desenvolvimentos nesta área;

- T2 – Estudo do *dataset* e ferramentas de análise de dados – caraterizar o *dataset*, aprender a utilizar *python*, bibliotecas associadas a este domínio e outras ferramentas necessárias;

- T3 – Análise exploratória dos dados – estudo dos dados usando ferramentas de análise adequadas;

- T4 – Relatório – Escrita de relatório detalhado e resumo do mesmo para publicação dos principais resultados.

Na tabela 1 é apresentado a seguinte calendarização para as tarefas:

Tabela 1 - Calendarização das tarefas

Meses	Tarefas	Metas
N	Estado da arte	INI - Familiarizar com a última pesquisa na área (INI + 4 Semanas)
N+1	Estudo do dataset e ferramentas	M1 - Selecionar o dataset e as ferramentas que serão usadas para a análise de dados (INI + 12 Semanas)
N+2	Análise de dados	M2 - Executar a análise de dados e gerar resultados (INI + 14 Semanas)
N+3	Relatório	M3 - Escrever o relatório com os resultados da análise de dados (INI + 18 Semanas)
N+4	Apresentação	M4 - Apresentar os resultados da análise de dados para o público (INI + 20 Semanas)

2 ESTADO DA ARTE

Neste capítulo é realizado um estudo do estado de arte, focando o estudo do fluxo de tráfego e as suas diferentes técnicas de recolha de dados com base em artigos, como está descrito na tabela 3. Efetivamente, também é analisado como estas técnicas podem ser relacionadas com o uso de smartphones como sensores para captar dados relativos à mobilidade.

2.1 Metodologia

A tabela 3 revela o estudo na base de dados do *Google Scholar*, onde foi realizada pesquisas com diversas palavra-chave. Inicialmente, com a palavra-chave *"traffic flow"* que apresentou 3 740 000 resultados. Depois começou-se a experimentar adicionar outras palavras-chave, que são mais específicas em termos de utilização, designadamente "*mobile data*" e *"machine learning"*. Estas apresentaram menos resultados, mas ainda devolveram um número substancial de artigos, variando entre 1 780 000 para *"traffic flow mobile data"* e 4 230 para *"machine learning traffic flow mobile data"*. Globalmente, estes números sugerem que existe uma quantidade significativa de investigação em curso na área do fluxo de tráfego e dos dados móveis.

Os artigos analisados foram publicados entre 2019 e 2023, e focam principalmente na gestão eficaz do tráfego em situações críticas, com o objetivo de ajudar a reduzir o congestionamento de tráfego, os custos de combustível e a poluição atmosférica, através do estudo do fluxo de veículos.

Efetivamente, os artigos escolhidos para o estado da arte são baseados na data de publicação mais recente (desde 2023) e nas palavras-chave "*traffic flow*", "*traffic flow mobile data*", "*machine learning traffic flow mobile data* e "*traffic prediction based on phone location*", tendo em conta os artigos mais citados e depois de ler 20/30 *abstracts*, escolheram-se os que mais se adequavam ao problema.

Tabela 2 - Número de artigos no google scholar com base numa palavra-chave

Google Scholar	
Palavras-Chave	**Número de artigos**
"traffic flow"	3 740 000
"traffic flow mobile data"	1 780 000

"traffic flow based on phone data"	8 690
"machine learning traffic flow mobile data"	4 230
"mobility data traffic flow"	6 320
"traffic prediction based on phone location"	6 080

2.2 Fluxo de Tráfego

Atualmente, o estudo do fluxo de tráfego é um campo de investigação com muitos desenvolvimentos em curso. Desta forma, os investigadores têm desenvolvido uma variedade de modelos e teorias para descrever e analisar o fluxo de tráfego. Estes modelos podem apresentar desde simples equações matemáticas a simulações mais complexas, como por exemplo o estudo de matrizes origem-destino que servem para determinar as rotas de desvio ideais nas redes urbanas [13], algoritmos de *"Machine Learning"* para classificar o congestionamento do tráfego [4] ou utilização de *frameworks* como a *"Promotion"* para prever o fluxo do tráfego [15], e assim conceber e gerir sistemas de transporte que sejam eficientes, seguros e sustentáveis.

Para além disto, a crescente utilização de telemóveis e o número de dispositivos conectados também contribuiu significativamente para o crescimento de tráfego móvel, disponibilizando uma riqueza de informações que podem ser utilizadas para prever padrões de tráfego, tais como a localização e a velocidade dos dispositivos na rede rodoviária [9][5].

2.2.1 Matriz Origem-Destino como sensor de tráfego

A matriz origem-destino (OD) na gestão do tráfego fornece uma imagem da distribuição espacial, com células individuais representadas através do número de viagens efetuadas entre um par de zonas de análise de tráfego (TAZs) [13]. Estes modelos estáticos e dinâmicos podem ser utilizados no planeamento de um sistema de transporte, de modo a otimizar o controlo e gestão do tráfego, dando particular atenção à necessidade e possibilidade de atualização frequente dos dados de entrada na sequência de alterações no volume do fluxo de tráfego real na rede rodoviária urbana.

2.2.2 Classificação do fluxo de tráfego através de algoritmos de *machine learning*

Efetivamente, para ajudar a melhorar as condições de tráfego, um grupo de investigadores decidiu avaliar o comportamento e eficácia de vários algoritmos de "machine learning" para classificar o fluxo de tráfego [4]. Concluíram que os Sistemas de Transporte Inteligentes (ITS) podiam ser uma alternativa associada à infraestrutura rodoviária com dispositivos IoT para captar informação.

Na verdade, estes resultados da classificação são influenciados por fatores através de afinação de parâmetros, o tamanho do conjunto de dados em termos de registos, atributos selecionados, e a percentagem de dados utilizados para formação e validação. Além disso, a análise de diferentes ambientes urbanos, incluindo intersecções, autoestradas e estradas suburbanas também é fundamental para identificar os classificadores mais eficazes.

2.2.3 *PROMOTION FrameWork* para prever o fluxo tráfego

A capacidade de prever o tráfego é uma das aplicações mais significativas em tecnologias de transporte inteligentes. A precisão das previsões é crucial para garantir a eficiência dos transportes públicos e para evitar congestionamentos. Tendo em conta algoritmos para prever o tráfico, a abordagem da *framework PROMOTION* é uma abordagem baseada em *deep learning* para a previsão do fluxo de tráfego, que utiliza redes neurais gráficas (GNNs) para modelar relações de espaço-tempo entre diferentes segmentos rodoviários numa rede de transportes [15].

De facto, esta *framework* explora técnicas inteligentes de pré-processamento baseado em gráficos, onde cada nó representa um segmento de estrada e as arestas representam a conectividade entre eles. O modelo GNN é treinado para aprender as características de cada nó e aresta, o que capta as relações espaço temporais entre os diferentes segmentos rodoviários, melhorando a precisão da previsão do fluxo de tráfego para otimizar sistemas inteligentes de transporte (ITS) e planeamento urbano.

2.2.4 *Smartphones* como sensores de fluxo de tráfego

O elevado uso de *smartphones* levou ao desenvolvimento de novas formas de recolha de dados sobre as condições de tráfego. Um desses métodos é a utilização de *smartphones* como sensores de tráfego, o que oferece várias vantagens, tais como ser rentável, proporcionando uma alternativa mais barata aos sensores de tráfego tradicionais e oferecendo uma cobertura mais ampla de dados em tempo real, como informações precisas e atualizadas sobre as condições de tráfego. Desta forma, para identificar com precisão a estrada que um condutor está a circular, o uso de algoritmos de *map-matching* [5] revelou que a trajetória dos veículos em movimento tem caraterística de memória

curta e assim é provável identificar o próximo segmento de estrada que o condutor vai entrar [9] para prever com precisão o fluxo do tráfego.

Tabela 3 - Análise comparativa de artigos baseado no fluxo do tráfego

Artigo	**Ano**	**Objetivo**	**Técnicas Aplicadas**	**Resultados**
[13]	2023	Gestão eficaz do tráfego em situações críticas.	Matriz Origem-Destino LSTM	Método proposto poderia ter uma implementação prática em sistemas de atribuição dinâmica de tráfego em tempo real para aplicações ITS, com média de MAPE em 7.18% (LSTM) e 6.80% (DLNa).
[4]	2023	Análise do fluxo de veículos para ajudar a reduzir o congestionamento de tráfego, os custos de combustível e a poluição atmosférica.	Decision Tree Extra-tree k-Near Neighbors Random Forest MLP	Resultados de classificação mais elevados foram obtidos no conjunto de dados da Primavera, o que coincide com as restrições de mobilidade estabelecidas nas cidades devido à pandemia de Covid-19, onde a Decision Tree obteve a maior precisão com 99.89% na experiência.
[15]	2022	Utilização de gráficos convulsionais para a previsão do fluxo de tráfego urbano num ambiente com Internet de alta velocidade.	STGCN STGNN STFGNN GraphWaveNet Promotion	A promotion quando comparada com as técnicas de previsão do fluxo de tráfego de base supera as outras soluções (STGCN, STGNN, STFGNN e GraphWaveNet).
[5]	2019	Demonstrar o desempenho do algoritmo Map-Matching e relacionar a precisão da estimativa e a frequência usada com os dados das torres de localização.	Map-Matching Virtual Inductive Loop	Os resultados desta experiência mostram que o map-matching obteve uma precisão média na ordem dos 60% a 85%. A precisão da rota estimada estende uma equação linear proporcional à frequência de amostragem do local.

3 CRISP-DM

O CRISP-DM é uma metodologia comum de análise de dados. É importante compreender as fases do processo CRISP-DM a fim de executar corretamente e com qualidade a análise de dados. Efetivamente, a análise de dados é um processo sistemático que envolve um conjunto de atividades para descobrir padrões, que é o objetivo final da análise de dados.

O modelo CRISP-DM consiste em seis fases, compreensão do negócio (*business understanding*), compreensão dos dados (*data understanding*), preparação ou pré-processamento dos dados (*data preparation*), construção de modelos (*modeling*), testes e avaliação (*evaluation*) e implementação (*deployment*).

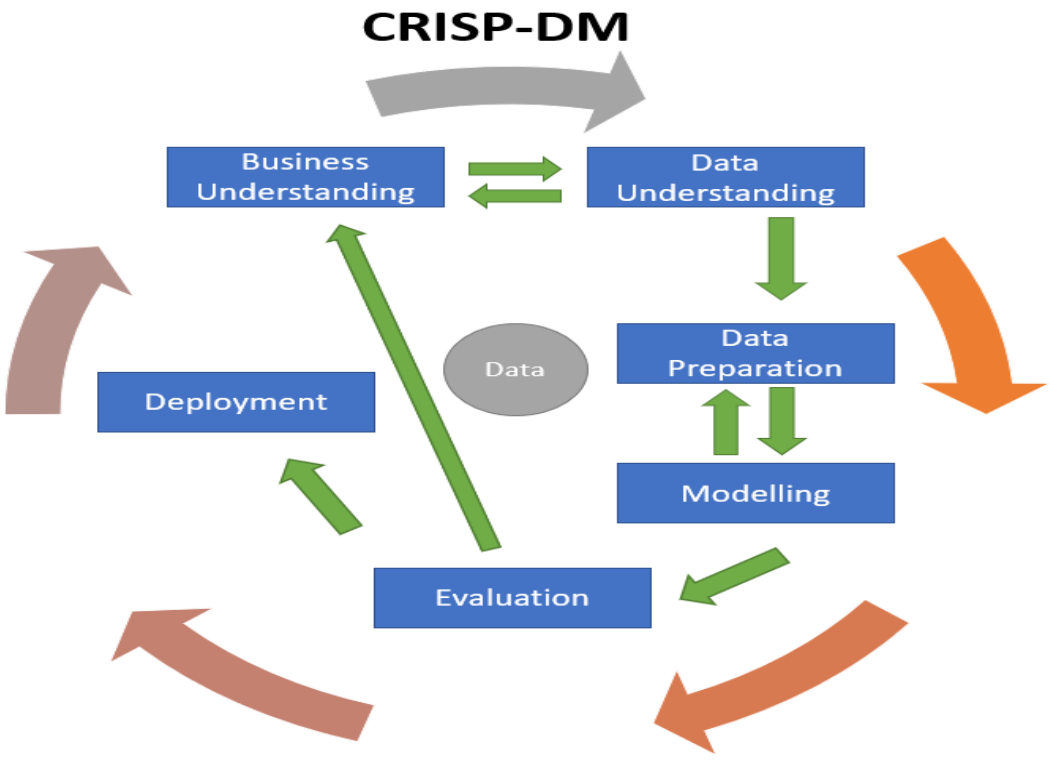


Figura 1 - Modelo CRISP-DM. Fonte:[3]

3.1. Compreensão do negócio

A primeira fase do modelo CRISP-DM é a compreensão do negócio, que é uma fase de alto nível para compreender as necessidades de gestão e os objetivos e requisitos do negócio. Pode ser desenvolvido um plano de projeto com base nestes objetivos e definido um orçamento de alto nível ou uma estimativa de custos para o realizar.

3.2. Compreensão dos dados

A segunda fase do modelo CRISP-DM, é a compreensão dos dados. Esta fase envolve a identificação de dados relevantes de diferentes fontes, a recolha dos dados iniciais, a familiarização com os dados, a identificação de problemas de qualidade, a descoberta de novos conhecimentos sobre os dados e a deteção de subconjuntos interessantes para formar hipóteses, através de tabelas ou gráficos.

3.3. Preparação ou Pré-processamento dos dados

A terceira fase do modelo CRISP-DM é a preparação dos dados. Na fase de preparação, os dados identificados na fase anterior são carregados num ambiente de análise e passam por um processo de tratamento, transformação, limpeza e seleção, para construir um conjunto final de dados a ser usado. Com efeito, esta etapa é muitas vezes a que consome mais tempo e esforço no processo CRISP-DM, cerca de 80% do tempo total, pois os dados do mundo real geralmente são incompletos, com falhas, contêm ruído e inconsistências, pelo que precisam de ser corrigidos e melhorados para a análise posterior.

3.4. Construção ou Modelação de modelos

Os dados tratados e selecionados são usados na próxima etapa, que é a construção do modelo matemático ou computacional. Esta etapa é também chamada modelação. Aqui são selecionadas as técnicas e tarefas a serem aplicadas para atender aos objetivos do negócio. Na fase de construção do modelo, pode-se escolher vários modelos possíveis e fazer avaliação comparativa entre eles.

3.5. Testes e Avaliação

Na fase de teste e avaliação, o modelo construído é testado e avaliado para ver se atende aos objetivos empresariais e para avaliar a precisão e eficiência do modelo em relação aos métodos alternativos. Essa é uma fase crítica, porque se não for descoberto nenhum padrão válido e útil, a mineração de dados pode ser considerada uma perda de tempo e esforço.

3.6. Implementação

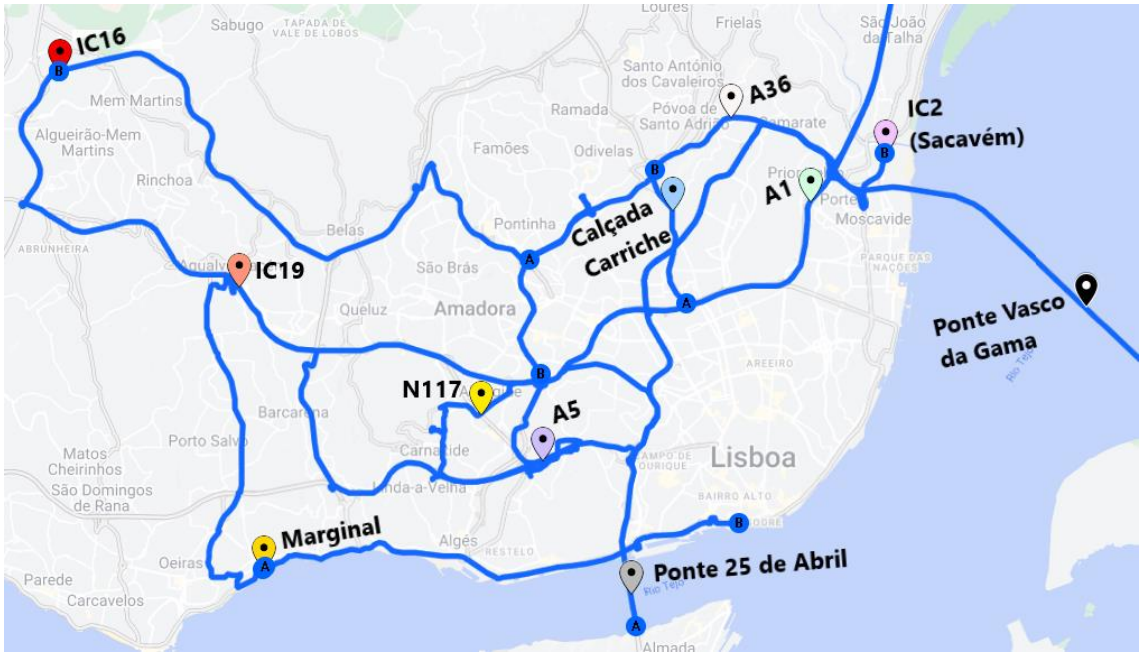
Na fase de implementação, o modelo de mineração de dados selecionado e avaliado na fase anterior é posto à prova. A implementação pode envolver o uso de software específico ou ferramentas de programação, além de testes e validações contínuas entre várias iterações de cada etapa ou na totalidade do processo, para garantir que o modelo esteja a funcionar corretamente, para que no final a documentação de todas as etapas do processo de implementação seja rigorosa e fiável para futuras referências.

4 DATASET

4.1 Contexto

O conjunto de dados extrapolados fornecidos pela Vodafone contém mais de 1 milhão de registos na pasta "*Número de telemóveis que entram e saem da cidade de Lisboa*" entre setembro de 2021 e dezembro de 2022, em particular com um período de 5 e 10 minutos entre setembro de 2021 e março de 2022 e 15 minutos entre abril e dezembro de 2022 para os 11 eixos principais da cidade de Lisboa, com o intuito de fazer uma estimativa global.

Há também registos na pasta "*Observações das estações meteorológicas do IPMA*" entre janeiro de 2020 e dezembro de 2022, com período de 1 hora.



4.2 Conteúdo

O conteúdo é dividido por 4 conjuntos, nomeadamente: i) tabela dos registos móveis representado pelo conjunto1; ii) localização dos pontos de entrada dos eixos, conjunto 2; iii) mapeamento dos troços da via representado por 11 registos, sem quaisquer metadados ou outros ficheiros que os represente; e por fim iv) os dados meteorológicos do IPMA, representados no conjunto 4.

Os dados do IPMA passam para o conjunto 3 para cumprir a ordenação, uma vez que não existem dados disponíveis para o mapeamento dos troços.

4.2.1 Conjunto 1 – Entradas e Saídas Lisboa

As variáveis do Conjunto1 são:

Eixo: nome eixo da cidade que efetuou o registo

Datetime: dia e hora do registo (dd/mm/aaaa hh:mm)

extract_year_2: ano do registo

extract_month_3: mês do registo

extract_day_4: dia do registo

C12: Número de entradas no eixo

C13: Número de saídas no eixo

Na tabela 4 apresenta-se estatísticas referente às variáveis de entrada e saída (C12 e C13) com 769669 registos no total, num período de amostragem de 14/09/2021 até 31/12/2022 e um tamanho de 55.7 Mb

Tabela 4- Descrição Geral do Conjunto 1

Variáveis	**C12**	**C13**
Média	144.51	141.12
Desvio Padrão	174.15	166.52
Mínimo	0.0	0.0
25%	28.72	29.43
Mediana	87.44	87.79
75%	195.53	193.05
Máximo	10270.72	10204.88

O valor máximo da coluna vem com a parte decimal devido à extrapolação dos dados, o que significa que o método de extrapolação que está a ser usado produz um valor ligeiramente maior do que o valor máximo real, seja por existir ruído nos dados, limites de extrapolação ou erro no método de extrapolação.

4.2.2 Conjunto 2 - Identificação dos 11 pontos de entrada e saída de Lisboa

As variáveis do Conjunto2 são:

id_eixo_viario: id do eixo

Eixo: nome do eixo

longitude: coordenada longitudinal do eixo

latitude: coordenada latitudinal do eixo

4.2.3 Conjunto 3 - Observações das estações meteorológicas do IPMA

As variáveis do Conjunto3 são:

fecha: Data da última leitura

estacion: N.º da estação - 01200535 - Lisboa Geofísico, 01200579- Lisboa Gago Coutinho, 01210762 - Lisboa Tapada da Ajuda

humidade: Humidade relativa média do ar

iddireccvento: Direção média do vento - 0 a 9 graus, 0: Norte, 1: Nordeste, 2: Leste, 3: Sudeste, 4: Sul, 5: Sudoeste, 6: Oeste, 7: Noroeste, 8: Norte, onde 0 e 9 representam a mesma direção

intensidadeventokm: Intensidade média do vento em quilômetros por hora

pressão: Pressão atmosférica

radiação: Radiação solar

temperatura: Temperatura

precacumulada: Precipitação acumulada

position: Coordenadas e tipo de entidade geográfica (ponto, linha ou polígono). Exemplo: {"coordinates": [-9.14965278,38.74],"type":"Point"} Lisboa Geofísico e {"coordinates": [-9.12830278,38.78],"type":"Point"} Lisboa Gago Coutinho.

Na tabela 5 apresenta-se estatísticas adicionais referentes às variáveis humidade, iddireccvento, intensidadeventokm, pressão, radiação, temperatura e precacumulada com 998 206 registos no total, num período de amostragem de 16/01/2021 até 8/09/2022 e um tamanho de 106 Mb.

Tabela 5 - Descrição Geral Conjunto 3

Variáveis	humidade	iddireccvento	intensidadeventokm	pressao	radiacao	temperatura	precacumulada
Média	67.63	4.46	-9.76	981.81	562.29	16.20	-4.60
Desvio Padrão	23.30	3.47	44.33	191.47	974.20	11.59	20.90
Mínimo	-99.0	0.0	-99.0	-99.0	-99.0	-99.0	-99.0
25%	55.0	1.0	1.8	1014.4	0.0	13.0	0.0
Mediana	71.0	5.0	9.7	1017.8	0.0	17.0	0.0
75%	83.0	8.0	15.1	1022.0	894.0	20.0	0.0
Máximo	100.0	9.0	40.0	1035.2	3883.0	40.0	16.0

É importante referir que apesar de existirem duas estações no ficheiro IPMA para prever o tempo, no projeto vamos trabalhar os dados como um todo, porque as estações são relativamente próximas uma da outra.

4.3 Qualidade do conteúdo

No caso de estudo deste *dataset*, o objetivo é analisar dados sobre o número de telemóveis que entram e saem de um determinado eixo da cidade de Lisboa e a Vodafone detém 60% do mercado. Os dados apresentam valores decimais, devido a um fator de correção dos outros 40% do mercado.

Para além disto, existe uma incongruência nos dados do IPMA, porque os meta dados no ficheiro vêm com a data de 2018 e no ficheiro *excel* vão desde 2020 a 2022, o que pode levantar dúvidas em relação à consistência destes dados.

4.3.1. Conjunto 1

Com base na tabela 5, que retrata a quantidade de registos para cada mês (setembro, outubro, novembro e dezembro) de 2021 e a diferença de registos com frequência de 5 em 5 min e 15 em 15 min para cada um dos 11 eixos, foi possível chegar a algumas conclusões sobre a consistência dos registos.

Em primeiro lugar, os registos mostram inconsistência a nível dos dias de amostra, particularmente em setembro de 2021, período no qual inicia dia 14 e acaba dia 30, fevereiro de 2022, de dia 1 a 8 e março de 2022, de dia 18 a 31.

Além disso existe também registos inconsistentes a nível do período da amostra, como é observado ao longo do ano de 2021. De facto, a média de registos com período de 5 min foi de 66.6%, enquanto os registos com período de 15 min corresponderam a 33.4%.

Tabela 6 - Análise registos conjunto1 2021

VODAFONE_EIXOS													
PGIL_VODAFONE_EIXOS_2021													
Mês	**# Registos**	**A1**	**A5**	**A36 Túnel do Grilo**	**Calçada de Carriche**	**IC2**	**IC16**	**IC19**	**Marginal**	**N117**	**Ponte 25 Abril**	**Ponte Vasco Gama**	**Datas Registadas**
Setembro	53844	4930	4947	4941	4882	4983	4841	4877	4862	4854	4868	4869	14/09/2021 até 30/09/2021 00:00 às 23:55
***(5min)**	35977	3288	3299	3312	3287	3319	3209	3258	3272	3237	3248	3248	
****(15min)**	17867	1642	1648	1629	1595	1664	1632	1619	1590	1617	1620	1611	
Outubro	95898	8690	8626	8680	8784	8710	8729	8712	8751	8817	8687	8712	1/10/2021 até 31/10/2021 00:00 às 23:55
***(5min)**	63733	5722	5711	5786	5783	5841	5843	5751	5834	5902	5748	5812	
****(15min)**	32165	2968	2915	2894	3001	2869	2886	2961	2917	2915	2939	2900	
Novembro	79288	7218	7152	7191	7245	7189	7204	7221	7239	7237	7166	7226	1/11/2021 até 31/11/2021 00:00 às 23:55
***(5min)**	52811	4821	4773	4824	4813	4786	4778	4812	4803	4795	4782	4824	
****(15min)**	26477	2397	2379	2367	2432	2403	2426	2409	2436	2442	2384	2402	
Dezembro	97240	8840	8830	8800	8876	8816	8824	8958	8884	8872	8741	8799	1/12/2021 até 31/12/2021 00:00 às 23:55
***(5min)**	64874	5880	5926	5847	5948	5861	5907	5960	5958	5921	5791	5875	
****(15min)**	32366	2960	2904	2953	2928	2955	2917	2998	2926	2951	2950	2924	

– *Registos com 5 min de diferença* ** – *Registos com 15 min de diferença

No que diz respeito a 2022, tabela 6, também falta consistência nos dias e na frequência. Através da tabela podemos verificar a variação no número de registos entre janeiro e março, bem como a discrepância de dias em falta em fevereiro e março, com apenas datas válidas entre dia 1 e 8, e entre dia 18 e 31, respetivamente. De maneira que, entre abril e dezembro, os registos são consistentes, visto que existe o mesmo número de registos para

cada eixo e as datas registadas são congruentes. Além disto, os registos até março apontam uma média de registos de 66.6% com período de 5 min e 33.4% com período de 15 min e 100% de registos com período de 15 min para os restantes meses do ano.

Tabela 7 - Análise registos conjunto1 2022

VODAFONE_EIXOS													
PGIL_VODAFONE_EIXOS_2022													
Mês	**Registos**	**A1**	**A5**	**A36 Túnel do Grilo**	**Calçada de Carriche**	**IC2**	**IC16**	**IC19**	**Marginal**	**N117**	**Ponte 25 Abril**	**Ponte Vasco Gama**	**Datas Registadas**
Janeiro	98005	8849	8941	8902	8865	8816	8917	8927	8959	8895	8921	8887	1/01/2022 até 31/01/2022 00:00 às 23:55
***(5min)**	65505	5920	5950	6005	5897	5953	5949	6043	5950	5959	5936	5943	
****(15min)**	32500	2929	2991	2897	2968	2964	2978	2899	3009	2936	2985	2944	
Fevereiro	19525	1777	1773	1762	1789	1761	1790	1761	1784	1750	1787	1791	01/02/2022 até 08/02/2022 00:00 às 23:55
***(5min)**	12937	1186	1176	1163	1190	1165	1169	1169	1177	1157	1195	1190	
****(15min)**	6588	591	597	599	599	596	621	592	607	593	592	601	
Março	42163	3826	3844	3802	3849	3845	3867	3888	3796	3812	3833	3801	18/03/2022 10:45 até 31/02/2022 00:00 às 23:55
***(5min)**	28166	2568	2567	2561	2548	2553	2567	2603	2495	2573	2572	2559	
****(15min)**	13997	1258	1277	1241	1301	1292	1300	1285	1301	1239	1261	1242	
Abril	31328	2848	2848	2848	2848	2848	2848	2848	2848	2848	2848	2848	01/04/2022 até 30/04/2022 00:00 às 23:55
***(5min)**	-	-	-	-	-	-	-	-	-	-	-	-	
****(15min)**	31328	2848	2848	2848	2848	2848	2848	2848	2848	2848	2848	2848	
Maio	31713	2883	2883	2883	2883	2883	2883	2883	2883	2883	2883	2883	01/05/2022 até 31/05/2022 00:00 às 23:55
***(5min)**	-	-	-	-	-	-	-	-	-	-	-	-	
****(15min)**	31713	2883	2883	2883	2883	2883	2883	2883	2883	2883	2883	2883	
Junho	31680	2880	2880	2880	2880	2880	2880	2880	2880	2880	2880	2880	01/06/2022 até 30/06/2022 00:00 às 23:55
***(5min)**	-	-	-	-	-	-	-	-	-	-	-	-	
****(15min)**	31680	2880	2880	2880	2880	2880	2880	2880	2880	2880	2880	2880	
Julho	32736	2976	2976	2976	2976	2976	2976	2976	2976	2976	2976	2976	01/07/2022 até 31/07/2022
***(5min)**	-	-	-	-	-	-	-	-	-	-	-	-	
****(15min)**	32736	2976	2976	2976	2976	2976	2976	2976	2976	2976	2976	2976	

													00:00 às 23:55
Agosto ***(5min)** ****(15min)**	32736 - 32736	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	2976 - 2976	01/08/2022 até 31/08/2022 00:00 às 23:55
Setembro ***(5min)** ****(15min)**	31152 - 31152	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	2832 - 2832	01/09/2022 até 30/09/2022 00:00 às 23:55
Outubro ***(5min)** ****(15min)**	29579 - 29579	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	2689 - 2689	01/10/2022 até 31/10/2022 00:00 às 23:55
Novembro ***(5min)** ****(15min)**	30057 - 30057	2732 - 2732	2733 - 2733	2733 - 2733	2732 - 2732	2733 - 2733	2732 - 2732	2733 - 2733	2733 - 2733	2733 - 2733	2733 - 2733	2733 - 2733	01/11/2022 até 30/11/2022 00:00 às 23:55
Dezembro ***(5min)** ****(15min)**	32725 - 32725	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	2975 - 2975	01/12/2022 até 31/12/2022 00:00 às 23:55

**– Registos com 5 min de diferença ** – Registos com 15 min de diferença*

4.3.1.1. Análise de registos do conjunto 1

As tabelas 8 e 9 mostram o número de registos esperados e validados a cada 15 min e 1 hora, de setembro de 2021 a dezembro de 2022. O número de registos esperados é a previsão de registos nesse período, o número de registos validados é a quantidade confirmada e a percentagem de registos omissos é a proporção dos registos esperados não gerados ou validados, dividida pelo total esperado.

Tabela 8 - Registos Esperados vs. Registos Validados (15 min)

Registos 15 min				
Ano	**Mês**	**Esperado**	**Validado**	**% Omissos**
2021	Setembro	31680	17669	44.23

2021	Outubro	32736	31031	5.20
2021	Novembro	31680	25819	18.50
2021	Dezembro	32736	31200	4.69
2022	Janeiro	32736	32015	2.20
2022	Fevereiro	29568	6503	78.00
2022	Março	32736	13731	58.05
2022	Abril	31680	31328	1.11
2022	Maio	32736	31713	3.12
2022	Junho	31680	31680	0.00
2022	Julho	32736	32736	0.00
2022	Agosto	32736	32736	0.00
2022	Setembro	31680	31152	1.66
2022	Outubro	32736	29579	9.64
2022	Novembro	31680	30057	5.12
2022	Dezembro	32736	32725	0.03

Para obter uma análise mais detalhada dos registos, procedemos com a consolidação dos arquivos CSV de todos os meses de 2021 e 2022 num único ficheiro. Esse processo foi realizado com o intuito de facilitar a análise e tornar os dados mais coerentes e consistentes, como é o caso da função *downsample*, que é um processo de redução do número de amostras de um conjunto de dados, sem perder as informações essenciais.

De maneira que, aplicamos o *downsample* dos dados para uma taxa de amostragem de 4 por hora (registos com um período de 5 min para registo de um período de 15 min) e, posteriormente, realizamos o *downsample* para um período de 1 h, através da função *resample*. Essa técnica permitiu representar os dados de maneira mais significativa e apropriada para as nossas análises, tornando-os mais fáceis de interpretar e comparar.

Na verdade, nos registos de 15 minutos, esperávamos encontrar 4 entradas a cada 15 min para cada uma das 24 h do dia, multiplicado pelo número de dias em cada mês e pelos 11

eixos em consideração. De maneira semelhante, para os registos com período de 1 hora, aguardávamos encontrar uma entrada para cada uma das 24 horas do dia, multiplicado pelo número de dias em cada mês e também pelos 11 eixos.

A análise revela um padrão semelhante de registos omissos consistente tanto para a tabela 8, quanto para a tabela 9.

Por outro lado, pode se não notar as diferenças entre as tabelas em setembro de 2021, com 44.23% de registos ausentes na tabela 8 e 43.48% na tabela 9. Isso acontece, porque só existe dados para metade desse mês. No entanto, em janeiro de 2022 já consegue-se ver a diferença do efeito do *downsampling* de 2.2% para 0%.

Tabela 9 - Registos Esperados vs. Registos Validados (1 h)

Registos 1 h				
Ano	**Mês**	**Esperado**	**Validado**	**% Omissos**
2021	Setembro	7920	4477	43.48
2021	Outubro	8184	7997	2.28
2021	Novembro	7920	6729	15.04
2021	Dezembro	8184	8117	0.82
2022	Janeiro	8184	8184	0.00
2022	Fevereiro	7392	1668	77.43
2022	Março	8184	3531	56.85
2022	Abril	7920	7832	1.11
2022	Maio	8184	7942	2.96
2022	Junho	7920	7920	0.00
2022	Julho	8184	8184	0.00
2022	Agosto	8184	8184	0.00
2022	Setembro	7920	7788	1.66
2022	Outubro	8184	7436	9.14

2022	Novembro	7920	7535	4.86
2022	Dezembro	8184	8184	0.00

Figura 2 - Registos Validados vs. Esperados por mês Conjunto 1 mais a diferença

Na figura 3 apresentam-se os registos validados em contraste com os esperados após o *downsampling* para a frequência de hora em hora. Observa-se efetivamente, a diferença nos meses de setembro e novembro de 2021 e fevereiro e março de 2022

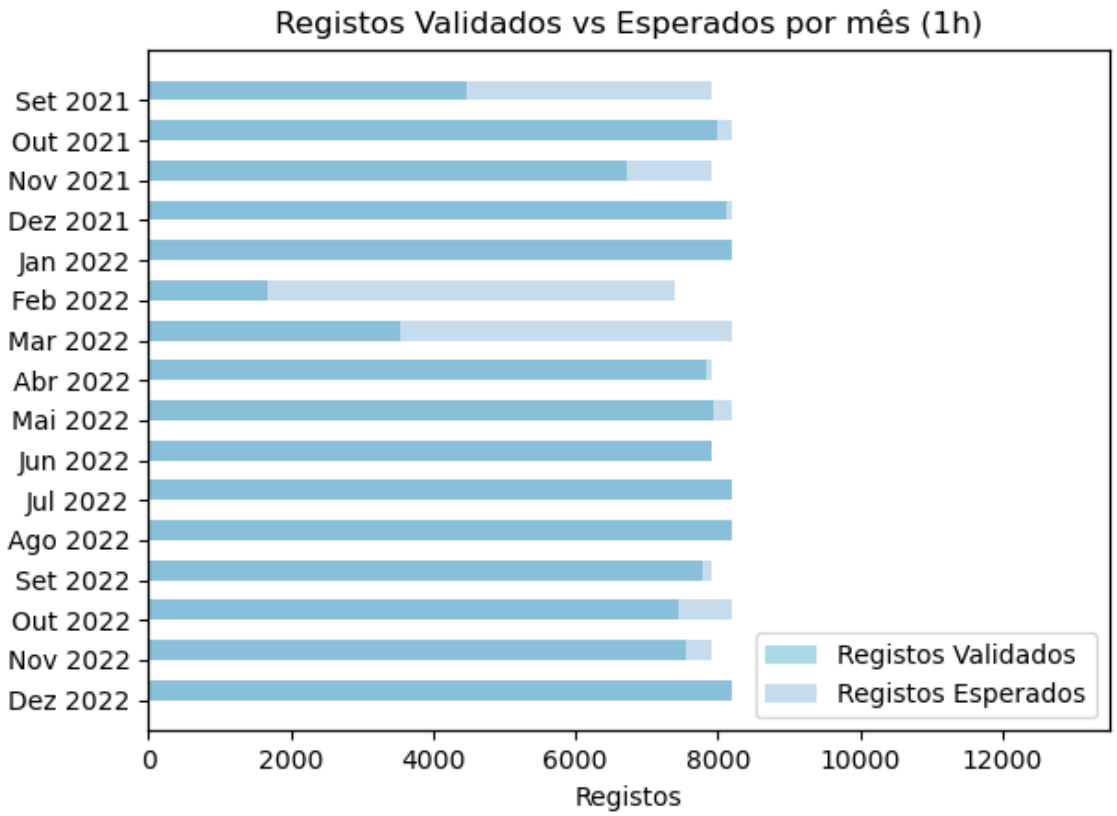


Figura 3 - Registos Validados vs. Esperados por mês Conjunto 1

4.3.2. Conjunto 3

O IPMA desempenha um papel fundamental na compreensão e monitoramento do clima em Portugal, desta forma disponibilizou um conjunto de dados (conjunto 3) para

relacionar o fluxo de tráfego com a ocorrência de pluviosidade, com uma frequência diária de hora em hora.

PGIL_IPMA_METEO_OBS_16.01.2020_a_08.09.2022.csv

Registos: 998 206

Datas Registadas: 16/01/2020 até 08/09/2022

Datas. Úteis: 14/09/2021 até 08/09/2022

Na verdade, através da aplicação da função *drop_duplicates* da biblioteca pandas no ficheiro foram apresentados 931 376 registos com exatamente os mesmos valores, cerca de 93.3%, e, portanto, é suposto o *dataset* apresentar apenas 66 830 registos que representem dados úteis para o problema, cerca de 6.7%. No entanto, ainda assim existem dados inconsistentes com valores nulos e incoerentes, com o valor -99.0 disposto por várias colunas. De modo que é necessário estudar os dados com valores inválidos para a mesma data na mesma estação. Com efeito, após este estudo foram validados 45 442 registos, o que representa 4.7% do *dataset*, conforme representado na figura 4.

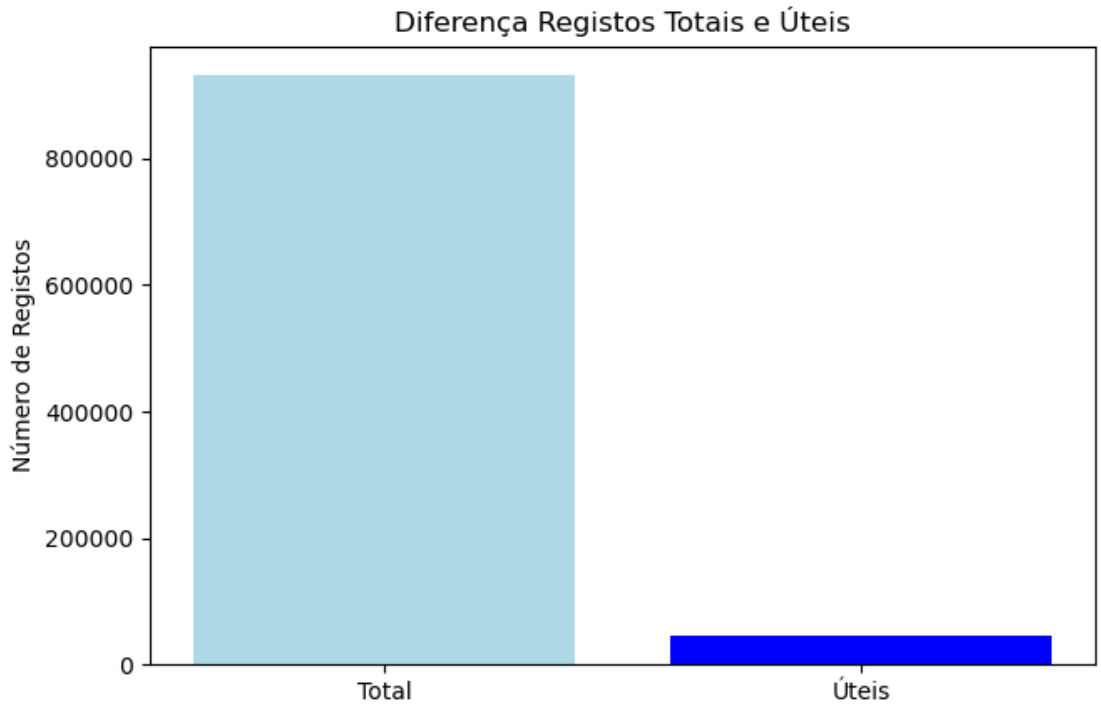


Figura 4 - Diferença Registos Totais e úteis Conjunto 3

Tabela 10 - Descrição Geral Conjunto 3 Dados Úteis

Variáveis	humidade	iddireccvento	intensidadeventokm	pressao	radiacao	temperatura	precacumulada
Média	68.3	5.5	12.27	1018.46	727.27	17.14	0.03
Desvio Padrão	18.1	2.99	6.19	5.34	1020.59	5.40	0.35
Mínimo	12.0	0.0	0.0	996.1	0.0	1.0	0.0
25%	55.0	2.0	7.9	1015.0	0.0	13.0	0.0
Mediana	70.0	6.0	11.9	1018.0	60.0	17.0	0.0
75%	83.0	8.0	16.6	1021.8	1270.5	20.0	0.0
Máximo	100.0	9.0	40.0	1035.2	3883.0	40.0	16.0

Depois de estudar os dados do IPMA voltou-se a realizar a descrição geral do conjunto 3, como apresenta a tabela 9.

4.3.2.2. Análise de registos do conjunto 3

A tabela 11 mostra o número de registos esperados e validados a cada hora, de janeiro de 2021 a dezembro de 2022.

Tabela 11 - Registos Esperados vs. Validados (IPMA)

Registos 1 H				
Anos	**Mês**	**Esperado**	**Validado**	**%Omissos**
2021 e 2022	Janeiro	2976	2933	1.44
2021 e 2022	Fevereiro	2688	2681	0.26
2021 e 2022	Março	2976	2960	0.53
2021 e 2022	Abril	2880	2842	1.32
2021 e 2022	Maio	2976	2974	0.07
2021 e 2022	Junho	2880	2698	6.32
2021 e 2022	Julho	2976	2972	0.07
2021 e 2022	Agosto	2976	2974	0.07
2021 e 2022	Setembro	2880	1779	35.07
2021 e 2022	Outubro	2976	1422	52.22
2021 e 2022	Novembro	2880	1440	50.00
2021 e 2022	Dezembro	2976	1470	50.06

Os dados do IPMA também foram *downsampled* para um período de 1 hora. O *downsampling* resultou na perda de alguns registos. A percentagem de registos perdidos foi maior nos meses de setembro, outubro, novembro e dezembro de 2022, com cerca de 50% de omissos. Isso ocorre porque esses meses tiveram mais inconsistências nos dados do que nos outros meses.

Na figura 5 observamos que a diferença de omissos entre setembro e dezembro de 2022 é próxima de 50%, o que vai dificultar bastante a análise do tráfego com ocorrência de pluviosidade.

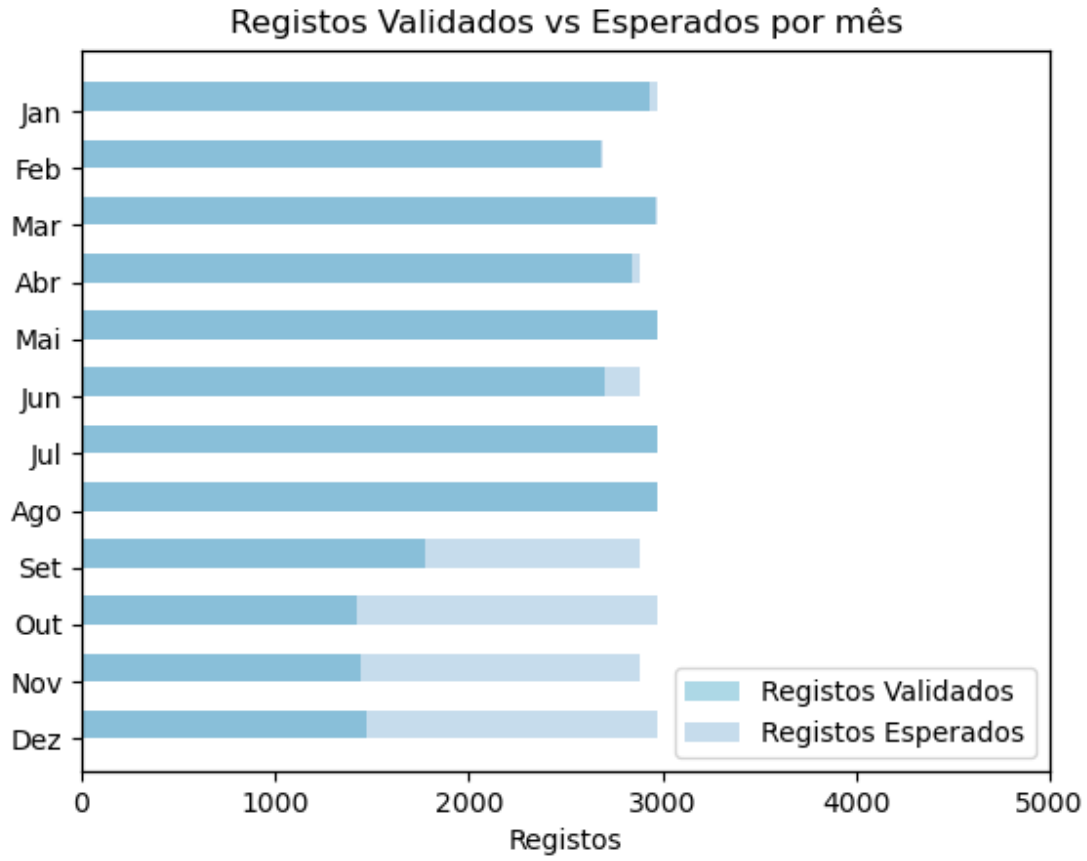


Figura 5 - Registos Validados vs. Esperados por mês Conjunto 3 por mês em 2021 e 2022

A figura 5 apresenta os registos validados em contraste com os esperados do conjunto 3 por mês. No gráfico, a diferença de registos omissos é notável para os meses setembro, outubro, novembro e dezembro.

5 ANÁLISE EXPLORATÓRIA

Nesta secção, a exploração de dados permite uma compreensão mais flexível do fluxo de tráfego diário entre as horas de ponta da manhã e da tarde na cidade de Lisboa. Ao analisar o conjunto de dados, é possível extrair informações valiosas sobre o volume total de entradas e saídas, padrões de tráfego de pontos de entrada e saída individuais dos 11 eixos e flutuações ao longo do dia e a influência de fatores externos, como ser fim de semana, feriado ou dia de chuva, afeta o volume nas horas de ponta. As informações sobre os destinos de quem entra ou sai da cidade têm implicações para o planeamento urbano e a gestão dos transportes.

5.1 Caraterizar o volume total de entradas e saídas da cidade durante o período das horas de ponta

A primeira pergunta do desafio é caraterizar o volume total de entradas e saídas da cidade durante o período das horas de ponta nos períodos da manhã entre as 7 as 10 horas, e da tarde entre as 17 e as 20 horas. Nas figuras 6 e 7 é possível observar a respetiva caraterização por mês.

A - Para períodos de ponta da manhã (7:00h-10:00h)

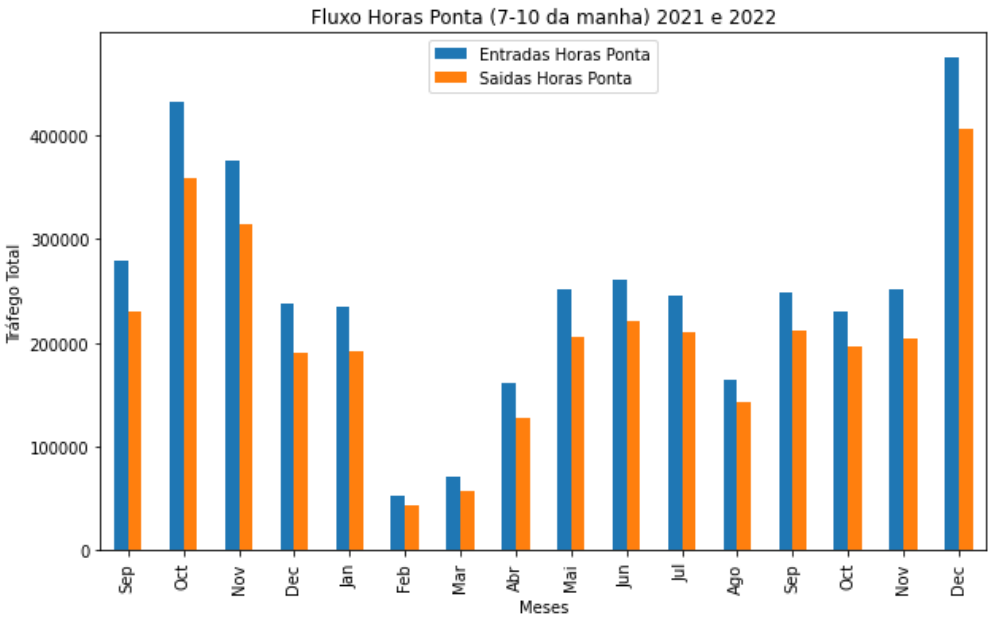


Figura 6 - Caraterização entradas e saídas nas horas de ponta da manhã (7:00h-10:00h)

B - Para períodos de ponta da tarde (17:00h-20:00h)

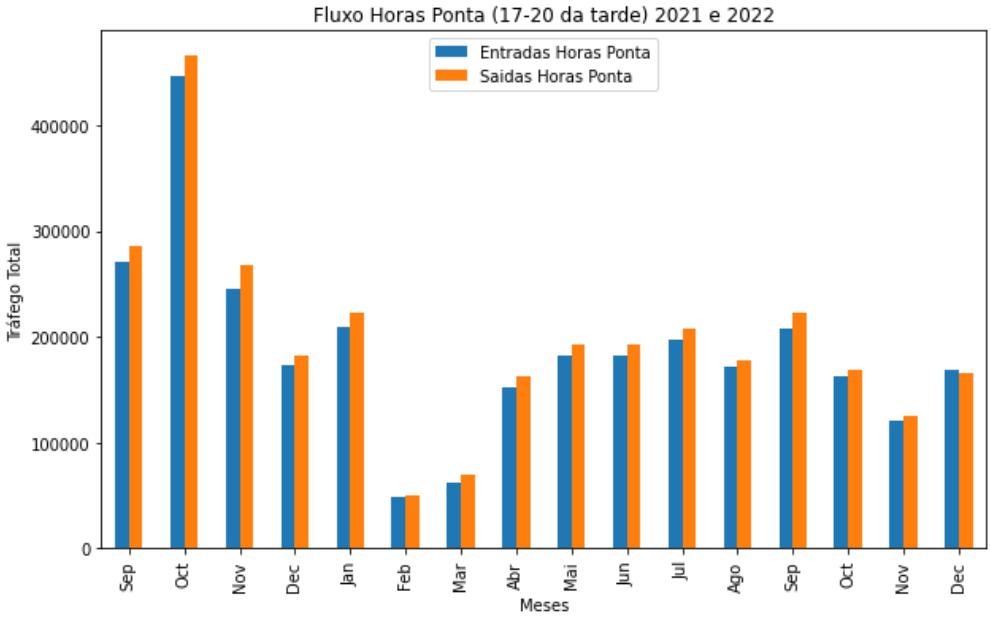


Figura 7 - Caraterização entradas e saídas nas horas de ponta da tarde (17:00h-20:00h)

De acordo com os dados mostrados nos gráficos das figuras 6 e 7, houve uma quantidade semelhante de tráfego durante o horário de pico da manhã e da tarde, entre setembro e novembro de 2021 e 2022. No entanto, em dezembro de 2022 na figura 6, ocorreu um aumento notável de tráfego no período, quase 3 vezes mais, no horário de pico da manhã, em relação a dezembro de 2021. Existem vários fatores que podem ter contribuído para esse aumento. Por exemplo, é possível que as condições climatéricas tenham sido mais favoráveis e apropriadas para as pessoas usarem transporte próprio para o trabalho. Por outro lado, foram levantadas todas as restrições de viagem da COVID-19 em 13 de dezembro de 2022 e mais pessoas regressaram às suas casas e retornaram ao escritório, levando a mais tráfego.

5.2 Caraterizar o volume total de entradas e saídas da cidade durante o período das horas de ponta para cada um dos 11 pontos de entrada e saída

Na segunda pergunta do desafio, vamos analisar o volume total de tráfego durante o horário de pico em cada um dos 11 pontos de entrada e saída da cidade. As figuras 8 e 9

apresentam detalhadamente hora a hora a frequência de tráfego para cada eixo em todos os meses do *dataset*.

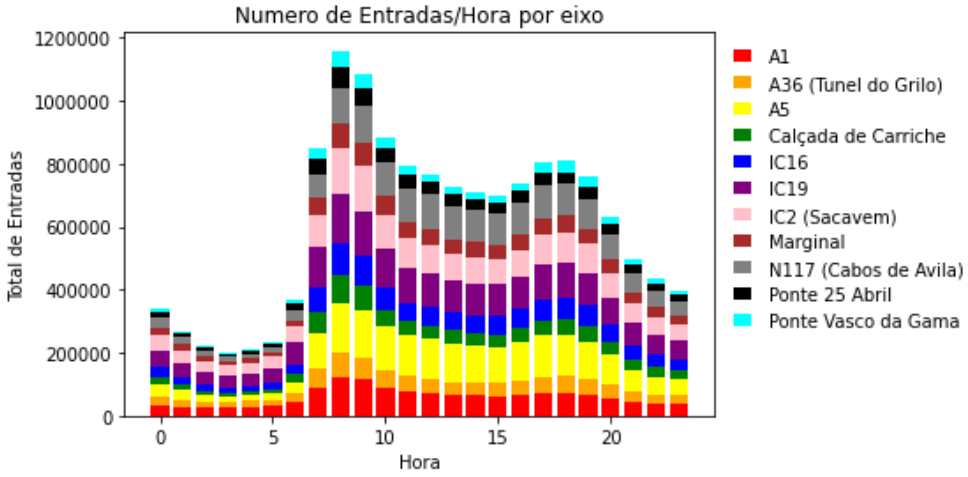


Figura 8 - Caraterização de entradas nas horas de ponta por eixo

Com base na análise do gráfico da figura 8, pode-se concluir que existe uma tendência significativa em relação às entradas na cidade, nomeadamente na A1, A5, IC19, IC2 e na N117 (Cabos de Ávila), sendo estes os pontos com mais volume de entradas, especialmente nas horas de pico da manhã 8h e 9h e nas de tarde 17h e 18h. É algo esperado, dado que estes eixos estão situados no núcleo central da cidade.

Além disso, a análise do gráfico revela que o pico da manhã apresenta o maior número, com 1 100 000 entradas, enquanto o pico da tarde regista uma queda significativa para 60 000 entradas às 20h, resultando numa redução de 45.5% em relação ao pico da manhã.

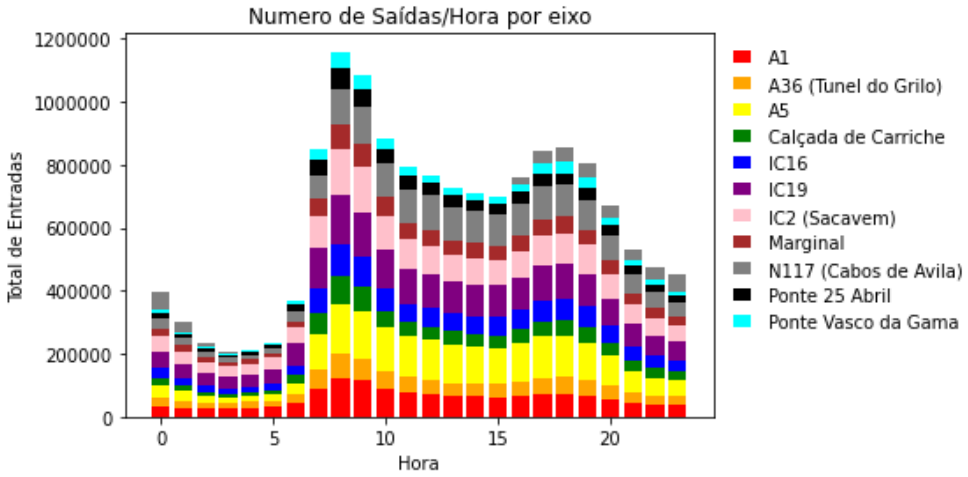


Figura 9 - Caraterização saídas nas horas de ponta por eixo

Efetivamente, relativamente à análise do gráfico das saídas na figura 9, continua-se a observar que os eixos A1, A5, IC19, IC2 e na N117 (Cabos de Ávila) são os pontos de saída mais utilizados durante as horas de pico das 8h às 9h e das 17h às 18h. Para além disto, o período de maior volume de saídas ocorre na parte da manhã, com cerca de 1 100 000 saídas e às 20h da tarde com cerca de 70 000 saídas, o que representa uma quebra de 36.7% em relação ao pico da manhã.

5.3 Comparar com outros períodos do dia

Nesta secção, para fazer a comparação com outros períodos do dia, foram selecionados três períodos para análise, como está representado na figura 10. De facto, a análise foi realizada com base na imagem das entradas, porque tanto as entradas como as saídas apresentam valores muito semelhantes.

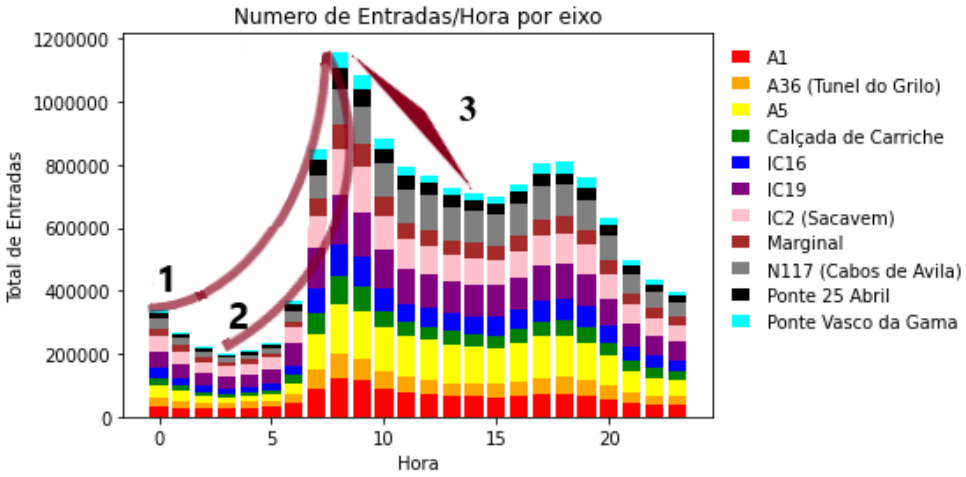


Figura 10 - Comparação com outros períodos do dia por eixo

No período da noite, representado pelos números 1 e 2, os resultados mostram que à meia-noite o tráfego é 30% em relação à hora de ponta com mais volume, uma diferença de 70% e 15% às 3 da manhã, uma diferença de 85%. No que diz respeito ao período escolhido da tarde, número 3 e hora 14, o tráfego representa 41% em relação à hora de ponta com mais volume, uma diferença de 59%.

Em relação aos eixos, tanto da parte da noite como da tarde, o eixo onde a diferença é mais notável é na Ponte Vasco da Gama a representar 18% à meia-noite e 7% às 3

da manhã, do volume da hora de ponta com mais tráfego e 41% da hora de ponta com mais volume às 14 horas.

5.4 Relacionar variáveis como os períodos de aulas ou férias e a existência de pluviosidade

De forma a tirar melhor partido dos dados foi decidido escolher o período escolar vs. existência de pluviosidade, porque tem mais registos escolares e mais informação. No entanto, sendo que existem 406 dias de escola validados para 35 dias de precipitação, a interpretação dos dados pode ser questionável.

Para além disto, é importante referir que a chuva pode ser caraterizada em dois tipos, chuva contínua e aguaceiros de chuva, variando na sua medição [8]. Visto que falta essa informação nos dados fornecidos e existem poucos dias de chuva para análise, o gráfico da figura 10 faz uma quantificação média geral.

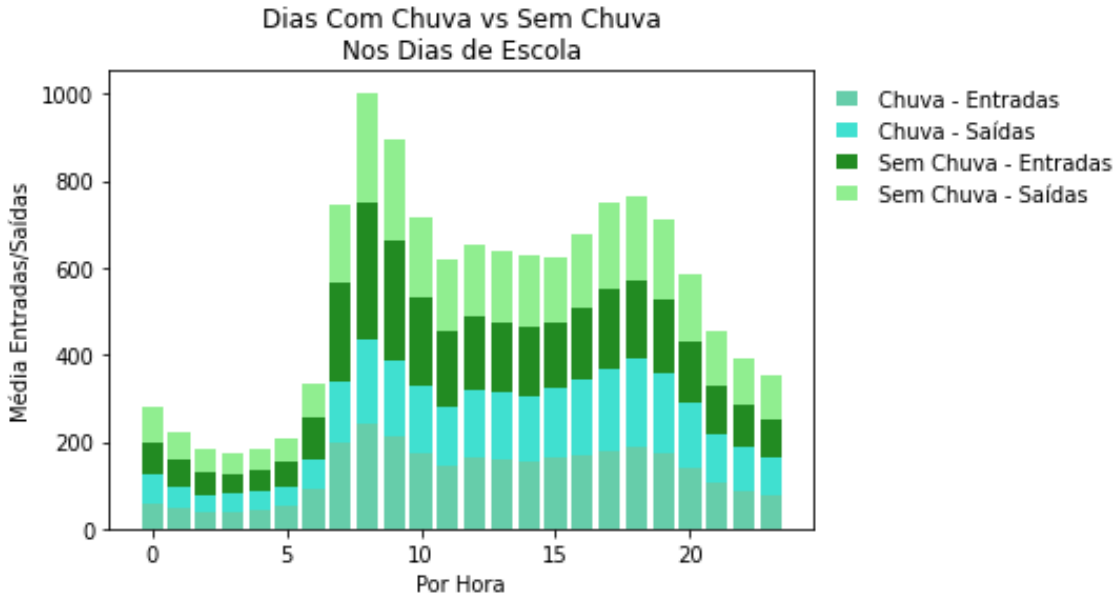


Figura 11 - Período de aulas vs. existência de pluviosidade

Quando observamos o gráfico da figura 11, verifica-se que a média de entradas/saídas em relação aos dias de chuva é igual, o que revela que mesmo com falta de dados existe congestionamento quando chove, no entanto, é difícil fazer uma qualificação objetiva, porque os dados são insuficientes para a questão do problema.

5.5 Análise das zonas destino e origem da cidade

Para esta análise consideramos os dias intensos aqueles que têm utilidade para o nosso problema, do ponto de vista da análise do tráfego, ou seja, o encerramento da semana de trabalho na sexta e a preparação da próxima no domingo e os outros dias, incluindo sábado, foram tratados como igual.

Foi estabelecido a comparação na quantidade de tráfego (entradas e saídas) nas horas de ponta das 7 às 10h da manhã e das 17 às 19h da tarde e períodos escolares de janeiro até junho e setembro até dezembro.

5.5.1. Dias Intensos vs. Outros Dias

Através da observação dos gráficos da figura 12 e 13, verifica-se que a quantidade de tráfego é semelhante com uma média de 30000 unidades paras as duas pontes, exceto no mês de dezembro onde foram levantadas as restrições de viagem do covid-19. Além disso, o tráfego é aproximadamente metade nas sextas e domingos, principalmente em julho e dezembro. Uma das razões para isto acontecer é o facto de representarem meses com tendência para tirar férias.

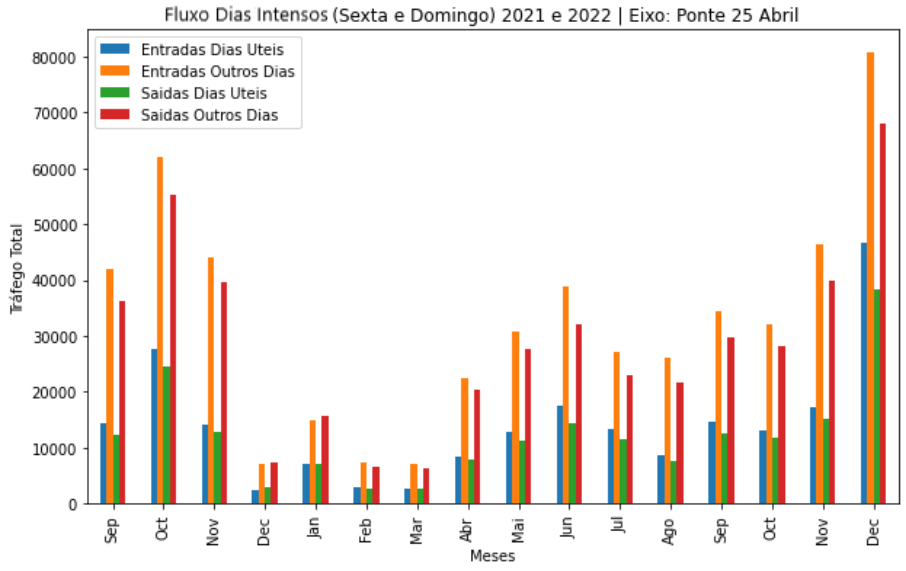


Figura 12 – Tráfego dias intensos vs. outros dias nas horas de ponta por mês para a Ponte 25 de Abril

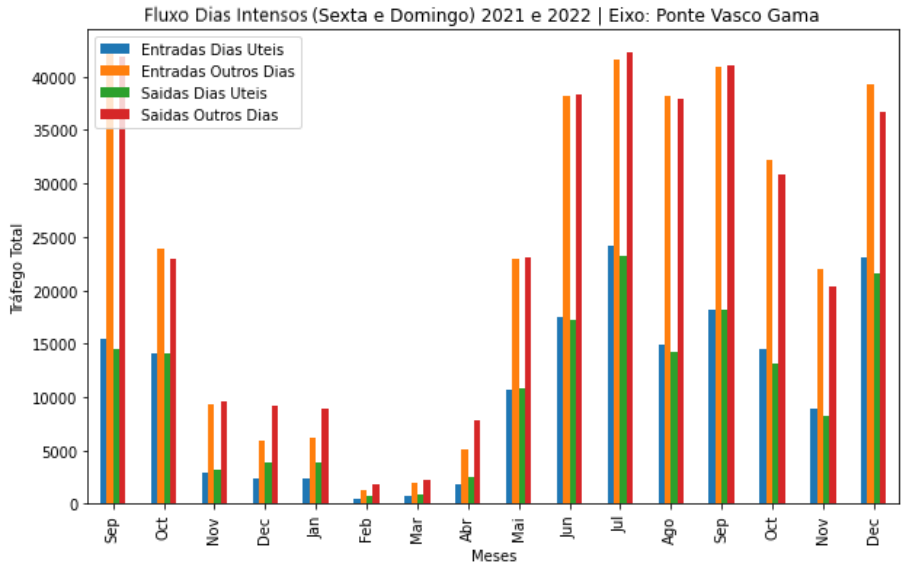


Figura 13 - Tráfego dias intensos vs. outros dias nas horas de ponta por mês para a Ponte Vasco da Gama

5.5.2. Horas de Ponta vs. Outras Horas

Entender o fluxo nas horas de ponta pode ser útil para encontrar soluções para o problema de mobilidade nas grandes cidades do mundo. Em Lisboa, o facto de muitas pessoas quererem deslocar-se todos os dias à mesma hora, pode vir de duas razões. Uma é a conveniência de escolher livremente o percurso pretendido. A segunda é a flexibilidade para garantir a deslocação devido a horários de trabalho variáveis.

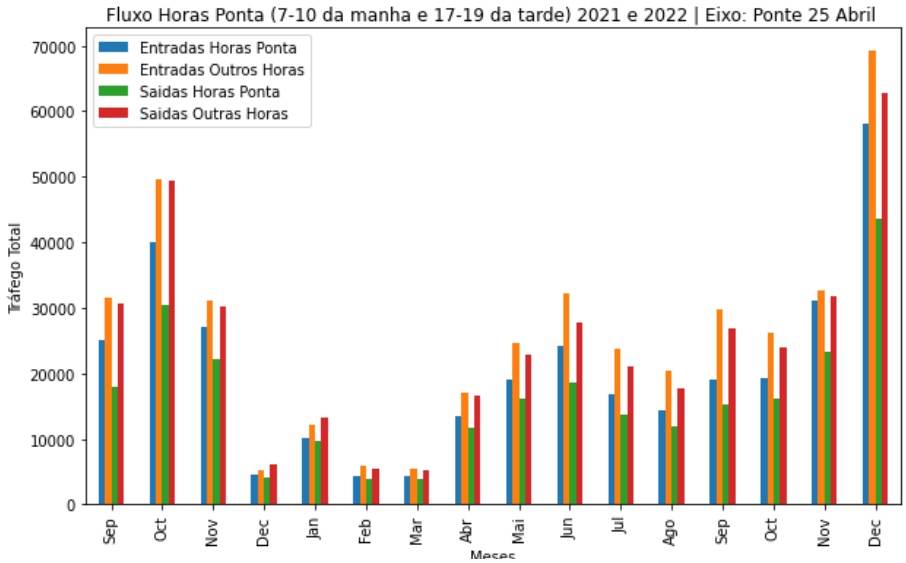


Figura 14 - Tráfego nas horas de ponta por mês para a Ponte 25 de Abril

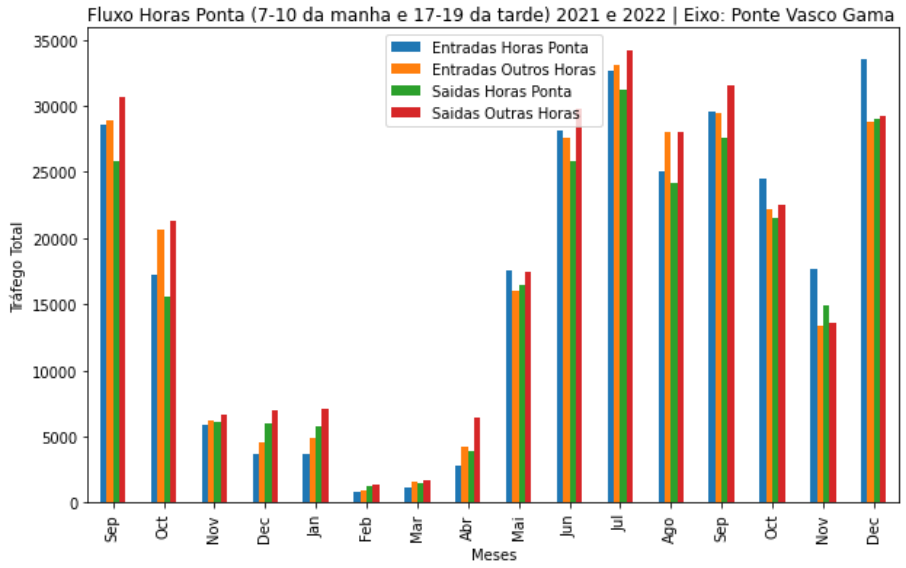


Figura 15 - Tráfego nas horas de ponta por mês para a Ponte Vasco da Gama

Através da análise dos gráficos das figuras 14 e 15, verifica-se uma quantidade de entradas superior nas horas de ponta para os meses de verão na ponte Vasco da Gama., uma diferença com cerca de 10000 unidades. No entanto, para os meses de novembro e dezembro de 2022 acontece o inverso para a ponte 25 de Abril, tendo mais 10000 unidades para novembro e mais 30000 para dezembro.

Para além disso, existem mais saídas noutras horas do que entradas para a ponte Vasco da Gama. Uma possibilidade para isto acontecer é que as pessoas preferem entrar pela ponte 25 de Abril e sair na Vasco da Gama, onde existe uma grande área residencial e industrial para as pessoas chegarem ao trabalho ou para casa.

5.6 Série Temporal dos dados

Visualizar a série temporal dos dados dá uma noção mais clara dos dados que vão ser trabalhados, neste caso das variáveis C12 e C13, entradas e saídas respetivamente. Desta forma, observar os vários padrões e tendências que os dados apresentam é fundamental para desenvolver com mais detalhe e precisão modelos preditivos, como é o caso do modelo usado no projeto, GRU, que foi desenhado para lidar com dados sequenciais. A Figura 16 mostra o comportamento de cada um dos 11 eixos de forma abrangente e revela com clareza a descontinuidade encontrada no capítulo 4, referente à compreensão dos dados.

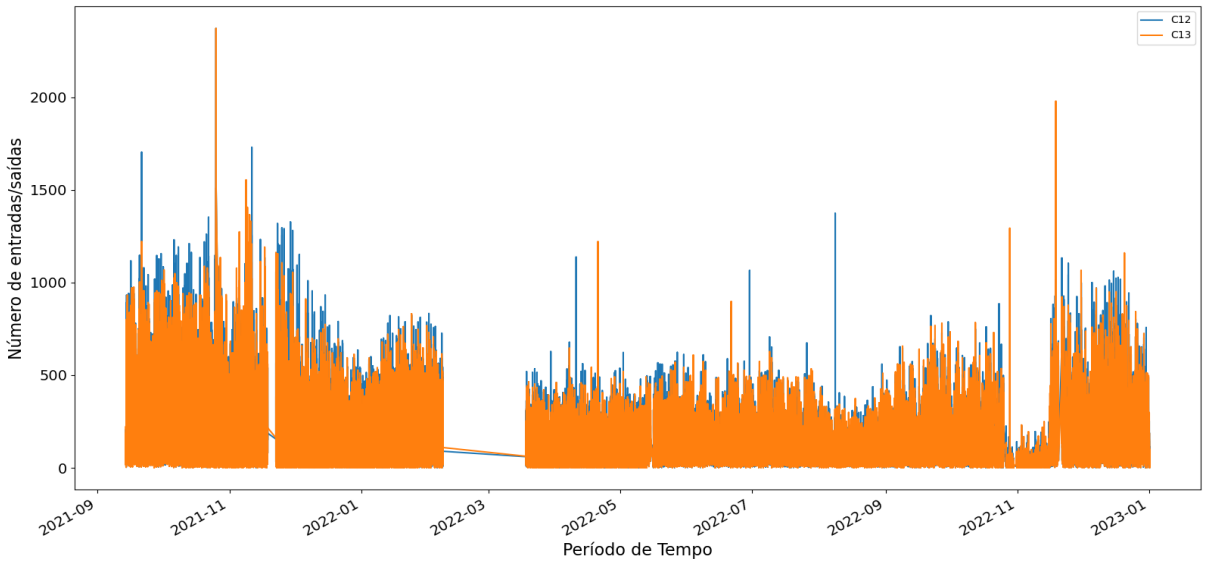


Figura 16 - Série temporal do conjunto de dados

5.7 Autocorrelação

Depois de observar a série temporal, é importante fazer a autocorrelação para medir a relação da variável que se pretende analisar e os seus valores anteriores em diferentes momentos. Por outras palavras, quantificar a similaridade entre observações nos diversos períodos.

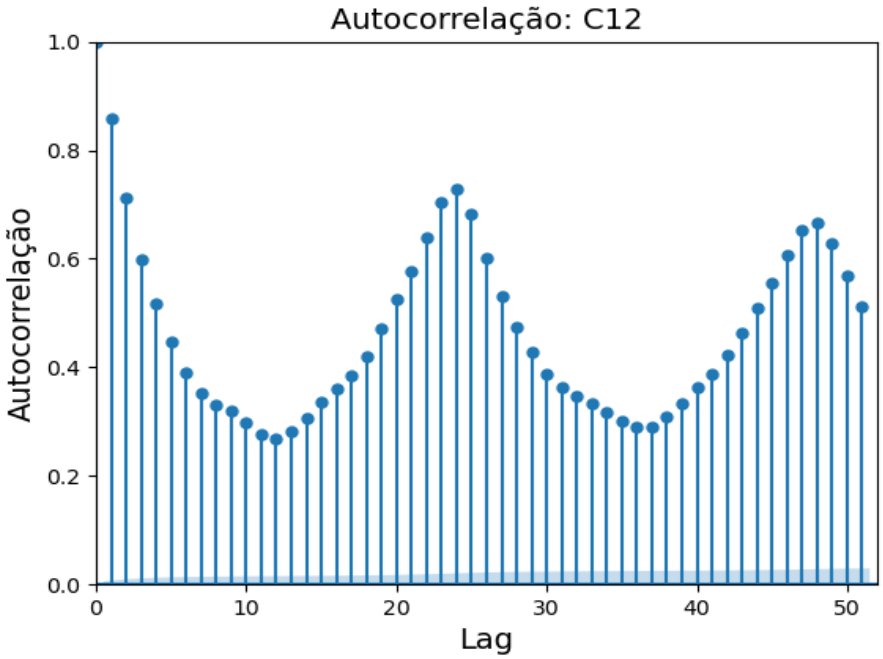


Figura 17 - Autocorrelação para a variável C12 (entradas), com um período de amostragem de 1 hora

A Figura 17, apresenta o gráfico da autocorrelação da variável C12, calculado com um período de 1 hora, para uma janela de até 51 horas. O resultado, como seria de esperar, verifica uma elevada correlação com o período de 24 horas antes – isto é, há um padrão diário bastante forte, mostrando que o tráfego de um dia está fortemente correlacionado com o tráfego do dia anterior à mesma hora.

6 GATED RECURRENT UNIT (GRU)

Gated Recurrent Unit (GRU) é uma versão recente das Redes Neurais Recorrentes (RNN), que foi criada pelo professor Kyunghyun Cho e os seus colegas na universidade de Nova York em 2014 e que tem vindo a ganhar muita popularidade [3] .

6.1 Arquitetura

A arquitetura GRU, utilizada neste projeto, é uma variante das redes neuronais recorrentes que se destaca na captura de dependências de longo prazo em dados sequenciais. Ao contrário das RNNs tradicionais, o GRU introduz mecanismos de portas que controlam o fluxo de informações dentro da rede. Estas portas, que consistem em portas de reposição e atualização, permitem à GRU reter e atualizar seletivamente a informação, permitindo-lhe captar eficazmente padrões relevantes em sequências mais longas. Isso torna a GRU adequada para tarefas de previsão de séries temporais em que as dependências de longo prazo desempenham um papel crucial.

6.2 GRU vs LSTM

As GRUs são semelhantes às redes Long Short-Term Memory (LSTM) no sentido de que são capazes de lidar com dependências de longo prazo em dados sequenciais, mas usam menos parâmetros e são computacionalmente mais flexíveis. Por exemplo, se o input na rede for a frase "A minha comida favorita é", o GRU não será capaz de entender o significado total da frase, se fizer parte de um contexto mais extenso, porque a LSTM usa uma porta adicional que armazena mais informação. Por outras palavras, o GRU é uma versão leve da LSTM, que combina memória de longo e curto prazo no seu estado oculto [7].

6.3 Mecanismo

A partir da figura 18, observa-se que o GRU compreende duas portas, a porta de atualização (*update*) e a porta de reposição (*reset*). A porta de atualização lembra o quanto da memória passada deve ser "retido", enquanto a porta de reposição controla o quanto da memória passada deve ser "esquecido".

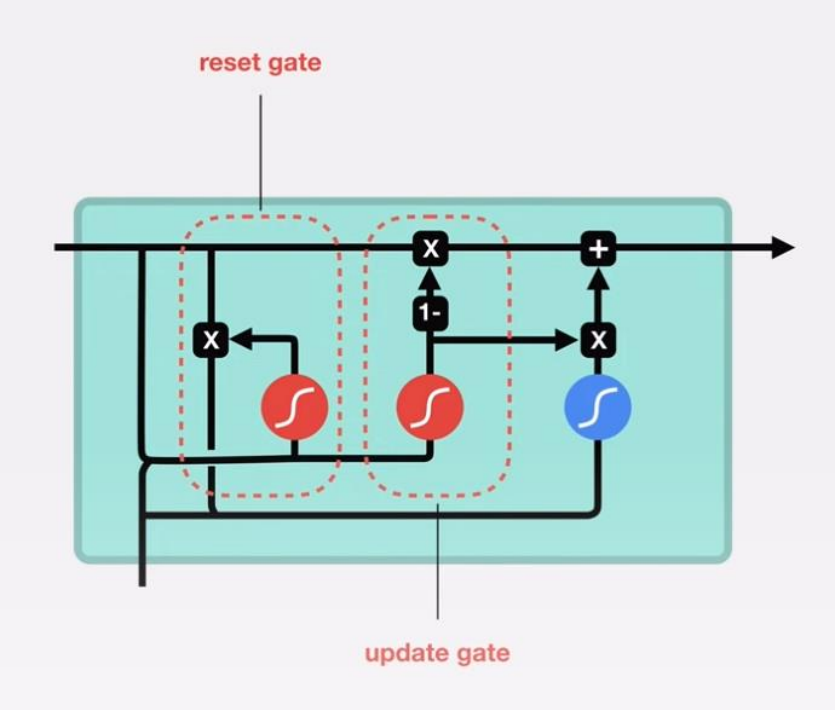


Figura 18 - Arquitetura interna de uma unidade GRU

Para cada elemento da sequência de input, cada camada calcula as seguintes funções:

$$rt = \sigma(Wirxt + bir + Whrh(t-1) + bhr) \quad (1)$$
$$zt = \sigma(Wizxt + biz + Whzh(t-1) + bhz) \quad (2)$$
$$nt = tanh\big(Winxt + bin + rt * (Whnh(t-1) + bhn)\big) \quad (3)$$
$$ht = (1 - zt) * nt + zt * h(t-1) \quad (4)$$

Onde ht é o estado oculto no tempo t, xt é o input no tempo t, Whzh(t−1) é o estado oculto da camada no tempo t-1 ou no tempo *0* e rt, zt são as portas de reinicialização e atualização respetivamente. σ é a função sigmoid e * é o símbolo da multiplicação [12].

6.4 Implementação

O código abaixo implementa uma classe Gated Recurrent Unit (GRU) [2]. A classe GRU contém métodos para inicializar os pesos e parâmetros da através da função "__init__" e executa os GRU através da função "forward" como passagem direta. A função sigmoide é a função de ativação utilizada pelas portas GRU.

```
import numpy as np

def sigmoid(x):
    """Sigmoid função de ativação."""
    return 1 / (1 + np.exp(-x))
```

```
class GRU:
    """
    Inicializa pesos e parâmetros de aprendizagem

    Argumentos:
    - input_dim: Dimensão dos vetores de entrada
    - hidden_dim: Dimensão dos vetores do estado oculto.

    """
    __init__ (self, input_dim, hidden_dim):
    self.input_dim = input_dim
    self.hidden_dim = hidden_dim

    # Calcula matrizes e pesos dos vetores para a porta de atualização (z)
    self.W_z = np.random.randn(input_dim, hidden_dim)
    self.U_z = np.random.randn(hidden_dim, hidden_dim)
    self.b_z = np.zeros((1, hidden_dim))

    # Calcula matrizes e pesos dos vetores para a porta de reposição (r)
    self.W_r = np.random.randn(input_dim, hidden_dim)
    self.U_r = np.random.randn(hidden_dim, hidden_dim)
    self.b_r = np.zeros((1, hidden_dim))

    # Calcula matrizes e pesos dos vetores para a nova porta gerada
    self.W = np.random.randn(input_dim, hidden_dim)
    self.U = np.random.randn(hidden_dim, hidden_dim)
    self.b = np.zeros((1, hidden_dim))

def forward (self, X):
        """
        Executa a função forward através de passage direta

        Argumentos:
        - X: Sequência dos vetores de entrada.

        Retorna:
        - H: Sequência dos estados ocultos.
        - Z: Sequência das saídas da porta de atualização.
        - R: Sequência das saídas da porta de reposição.
        """

      T = X.shape[0]  # Número de iterações
      H = np.zeros((T + 1, self.hidden_dim))  # Estados ocultos
      Z = np.zeros((T, self.hidden_dim))  # Saídas da porta de atualização
      R = np.zeros((T, self.hidden_dim))  # Saídas da porta de reposição
```

```
    H [0] = np.zeros((1, self.hidden_dim))  # Estado oculto inicial

for t in range(T):
    x_t = X[t] # Entrada na iteração t

    # Calcula porta de atualização (z), reposição (r) e a nova porta gerada (h_tilde)
    z_t = sigmoid(x_t @ self.W_z + H[t] @ self.U_z + self.b_z)
    r_t = sigmoid(x_t @ self.W_r + H[t] @ self.U_r + self.b_r)
    h_tilde_t = np.tanh(x_t @ self.W + (r_t * H[t]) @ self.U + self.b)

    # Atualiza o estado oculto através da equação 4
    H[t+1] = (1 - z_t) * H[t] + z_t * h_tilde_t

    Z[t] = z_t # Armazena a porta de atualização para esta iteração
    R[t] = r_t  # Armazena a porta de reposição para esta iteração

return H[1:], Z, R
```

7 CONCEITOS DE *MACHINE LEARNING*

7.1 Pré-processamento dos dados

Antes de treinar o modelo, os dados são submetidos a etapas essenciais de pré-processamento para garantir melhor desempenho. Estes passos incluem a normalização ou o escalonamento das características de entrada, o que coloca os dados num intervalo normalizado e impede que determinadas características dominem o processo de aprendizagem. Além disso, os dados são divididos em conjuntos de treino e teste separados, o que permite avaliar o desempenho do modelo em dados não vistos. Essa separação ajuda a avaliar a capacidade do modelo de generalizar novas observações.

De facto, foi usado no projeto o *MinMaxScaler*, onde o pré-processamento de dados é utilizado para dimensionar (escalonar) os valores de um conjunto de dados para um intervalo específico, geralmente entre 0 e 1. De modo que, realiza a subtração do valor mínimo do conjunto de dados a cada valor original e, em seguida, divide a diferença pelo intervalo entre o valor mínimo e o máximo.

Conceitos Principais:

Dados de treino e teste: As variáveis *train* e *test* caraterizam o *path* que aponta para os conjuntos de dados de treino e teste, respetivamente. Estes conjuntos de dados contêm as características de entrada (X) e os valores-alvo (Y) para treinar e avaliar o modelo.

Lag e ***Gap***: Neste contexto, o *lag* representa o número de passos de tempo anteriores que são utilizados como entrada para prever o valor futuro, enquanto o *gap* refere-se à janela de previsão. Por exemplo, se o *lag* for definido como 24 e o *gap* como 1, o modelo utiliza os dados dos 24 passos de tempo anteriores como entrada para prever 1 hora à frente no tempo.

Scaler: A principal função de um *scaler* é garantir que todas as características numéricas têm magnitudes comparáveis, o que ajuda a evitar que as características com escalas maiores dominem o processo de aprendizagem em detrimento das características com escalas mais pequenas. Isto é, efetuar a normalização dos dados.

7.2 Treinar o modelo

Para treinar um modelo preditivo em *Python*, devem ser realizados certos passos para avaliar os dados de um determinado conjunto de dados. Em primeiro lugar, começa-se por pré-processar os dados, o que implica separá-los em dois ficheiros: conjunto de treino e de teste. Normalmente, o conjunto de treino compreende 70 a 80% dos dados, enquanto o conjunto de teste 20 a 30%. Os dados de teste podem ainda dividir-se em teste e

validação. Em seguida, cria-se uma instância de modelo sequencial e incorpora-se a camada GRU com os parâmetros desejados, incluindo o número de entradas e a forma de entrada. Em seguida, especifica-se a *loss function*, o *optimizer* e a avaliação métrica para o processo de formação. Depois, treina-se o modelo através da função *fit*, especificando o número de *epochs* (iterações) e o *batch size*. Por fim, avalia-se o desempenho do modelo nos dados de teste com a função *evaluate*, que fornece métricas de perda e precisão. Desta forma, é possível utilizar o modelo treinado para fazer previsões em dados novos e que ainda não foram visualizados.

Conceitos principais:

Epochs: Uma *epoch* é uma iteração completa através de todo o conjunto de dados de treino. Durante cada *epoch*, os pesos do modelo são ajustados com base no algoritmo *optimizer* e na *loss function*. O número de *epoch*s determina quantas vezes o modelo será treinado em todo o conjunto de dados.

Batch Size: O *batch size* refere-se ao número de exemplos de treino usados na passagem para frente/para trás da rede neural durante cada *epoch*, o que pode afetar a velocidade e a estabilidade do processo de treino. *Batch size* maiores podem resultar num treino mais rápido, mas exige mais memória. *Batch size* menores fornecem atualizações de peso mais frequentes, mas podem tornar o processo de treino mais lento.

Model: O parâmetro *model* representa o objeto do modelo de rede neuronal que será treinado e validado.

Optimizer: É um algoritmo ou técnica que ajusta os parâmetros do modelo de forma a minimizar o erro ou a *loss function*. O objetivo é encontrar os valores ótimos para os parâmetros do modelo que melhor se ajustem aos dados de treino e generalizem bem para dados não vistos.

Loss Function: Através de uma função como o erro quadrático médio (*MSE*) é medido a diferença média entre os valores previstos e reais, servindo como uma medida do desempenho do modelo. Podem ser usadas outras funções, como o erro absoluto médio (*MAE*), por exemplo.

Early Stopping: *Early Stopping* é usado no processo de treino, porque ajuda a evitar o ajuste excessivo, interrompendo o treino se a função de erro de validação parar de melhorar durante um determinado número de *epochs*.

Learning Rate: Determina a taxa de atualização dos pesos de um modelo durante o treino. Por outras palavras, controla a velocidade com que um modelo aprende com os dados.

Métricas de avaliação: As métricas de avaliação, como o erro percentual absoluto médio (*MAPE*), são calculadas durante o treino ou teste, e podem ser usadas para avaliar o desempenho do modelo.

Save Model: O modelo treinado é guardado no disco como um formato de arquivo h5 para uso futuro. Isto permite carregar o modelo e fazer previsões sem ter de o treinar novamente.

7.3 Avaliar o Modelo

Para avaliar o modelo, usa-se um modelo pré-treinado 'gru', já com conjunto de dados concebido para executar as previsões. Desta forma, para compreender melhor a sua estrutura, existe uma imagem que representa visualmente a arquitetura do modelo para proceder às previsões através dos dados de teste e por fim avaliar o seu desempenho e precisão.

Load_model: A função load_model é responsável por carregar um modelo pré-treinado a partir de um ficheiro guardado ou os pesos e a arquitetura de um modelo treinado. Esta função permite trazer um modelo treinado de volta à memória para utilização posterior.

Plot_model: A função plot_model é utilizada para visualizar a arquitetura ou estrutura de um modelo. Ela gera uma representação gráfica das camadas, conexões e fluxo de dados do modelo e ajuda a compreender a conceção do modelo, a identificar os seus componentes e a analisar o fluxo de informação.

Predict: A função de *predict* é utilizada para efetuar previsões utilizando o modelo carregado. Recebe dados de entrada e gera previsões ou probabilidades de saída com base nos padrões e relações aprendidos no modelo.

8 TESTES DE PREVISÃO E RESULTADOS

Neste capítulo são apresentados os testes e resultados obtidos, bem como o modelo que foi aplicado durante a validação dos mesmos.

8.1 Arquitetura do modelo

A arquitetura do modelo é representada na figura 19 por um *InputLayer*, com uma sequência de comprimento 12 e 1 característica de entrada, que a processa através de duas camadas *GRU*, GRU_1 e GRU_2, reduzindo o comprimento da sequência a uma única representação vetorial. A seguir vem a camada *dropout* para ajudar na regularização do modelo e evitar *overfitting* e uma camada *dense* para dar a previsão final de saída.

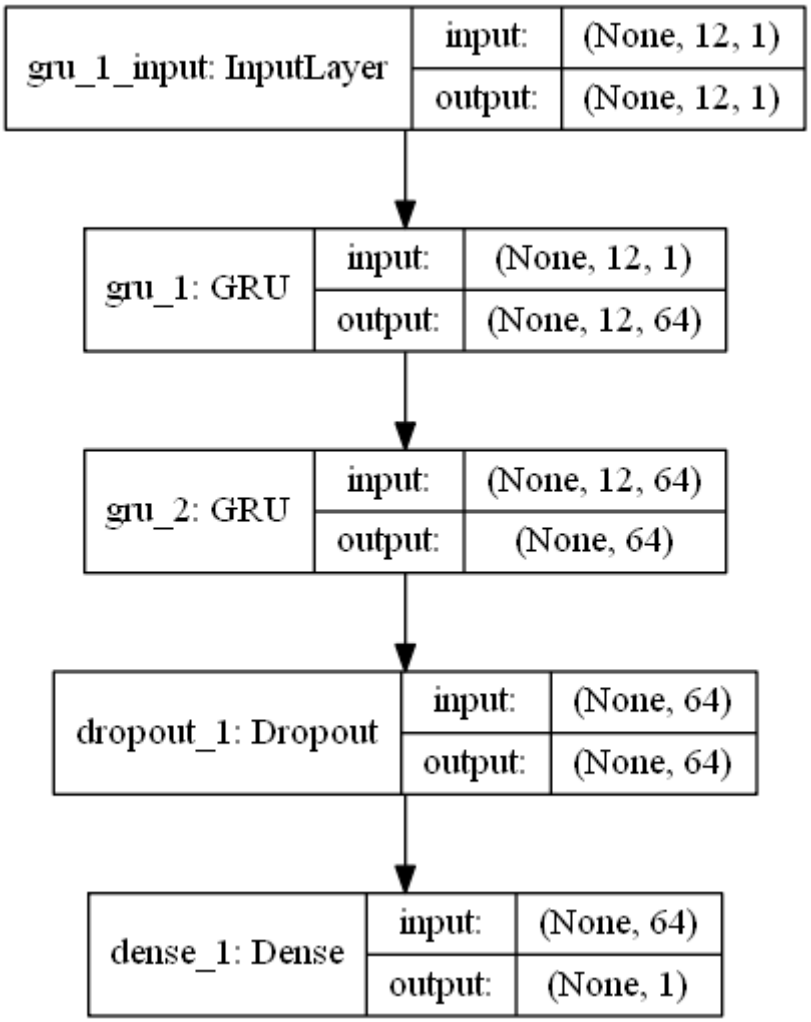


Figura 19 - Arquitetura Modelo GRU

8.2 Primeiros testes

Para iniciar os testes, começou-se por determinar o *lag* referência de acordo com a figura 17 e foi fornecido código base, obtido no repositório GitHub [1], em python. Este código serviu como ponto de partida para começar a fazer os testes. Tendo em conta que se tratava da fase inicial, manteve-se os parâmetros iniciais como o *batch size* a 256, número de *epochs* 600, a *loss function MSE*, o *optimizer RMSprop* e avaliação métrica *MAPE*.

Numa primeira abordagem, o conjunto de dados previsto surgiu misturado para todos os eixos, tanto para o conjunto de treino, como para o de teste, que apresentava a mesma informação. Esta abordagem resultou num gráfico com elevado nível de inconsistência e falta de coerência para responder eficazmente às questões do problema.

8.3 Testes para as entradas de cada eixo

Após os testes iniciais, a estratégia foi mudada para tratar cada eixo individualmente, reservando 76% dos registos para o treino, entre 14 de setembro de 2021 e 31 de agosto de 2022 e 24% para o teste, entre 1 de setembro e 31 de dezembro de 2022.

Além disso, para a gerar melhor performance no modelo foram usados entre 100 e 600 *epochs*, 64 e 256 para o *batch size*, o *optimizer Adam* com *learning rate* de 0.001, *loss function MSE* e avaliação métrica *MAPE*.

8.3.1. *Lag* 24 vs *Lag* 12

É importante considerar tanto o *lag* 24 quanto o *lag* 12, visto que se trata da previsão de tráfego, onde existem padrões diários recorrentes, como se observa na figura 17. Com efeito, realizar testes preditivos para identificar o melhor *lag* que maximize a precisão do modelo e a captura destes padrões é relevante para fazer uma análise fundamentada. Além disso, é importante considerar o *gap* (prever à frente) para 1, 2, 4 e 8 horas, porque ao analisar o *gap*, é possível identificar áreas onde o modelo está a ser subestimado ou superestimado.

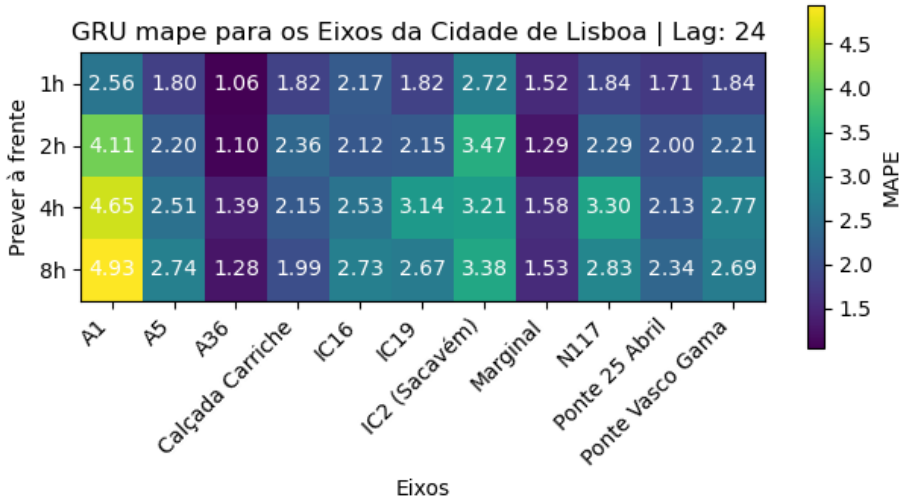


Figura 20 - MAPE da previsão do tráfego para os Eixos da Cidade de Lisboa | Lag: 24

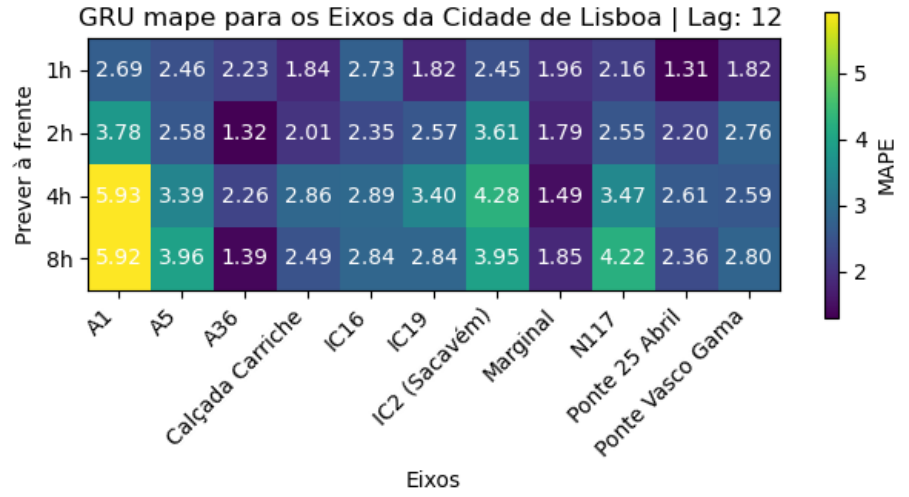


Figura 21 - MAPE da previsão do tráfego para os Eixos da Cidade de Lisboa | Lag: 12

De facto, ao analisar as Figuras 20 e 21, foi observada uma diferença média de 0.86pp no MAPE entre as previsões com um *lag* de 24 e um *lag* de 12, para todos os eixos. Uma diferença de 0.86pp no *MAPE* pode ser considerada dentro da faixa de variação esperada e não ter um impacto prático relevante nas decisões tomadas com base nas previsões. No entanto, noutras situações, uma diferença de 0.86pp pode ser considerada significativa, especialmente se estiver acima de uma margem de erro estabelecida.

Também observamos que existe uma diferença média de 2.07pp na A1, 3.14pp na A5, 2.37pp na A36, 0.88pp em Carriche, 1.26pp na IC16, 0.85pp na IC19, 1.51pp na IC2, 1.17pp na Marginal, 2.14pp na N117, 0.3pp na Ponte 25 de Abril e 0.46pp na Ponte Vasco da Gama. De modo que, verifica-se uma correlação do *MAPE* com os eixos com mais tráfego, nomeadamente na A1, A5, IC2 e N117. Podemos então concluir que quanto maior o tráfego maior é o erro percentual médio absoluto (*MAPE*) e mais difícil obter uma previsão de qualidade. Esta conclusão sugere que a precisão das previsões pode ser afetada pelo volume de tráfego. Em vias com maior fluxo de veículos, é provável que haja maior complexidade e imprevisibilidade no comportamento do tráfego, o que pode levar a um aumento do erro nas previsões.

De seguida são apresentadas as figuras de 22 e 23, onde foram selecionados 4 gráficos dos eixos com maior *MAPE* para prever 8 horas à frente na última semana de dezembro de 2022.

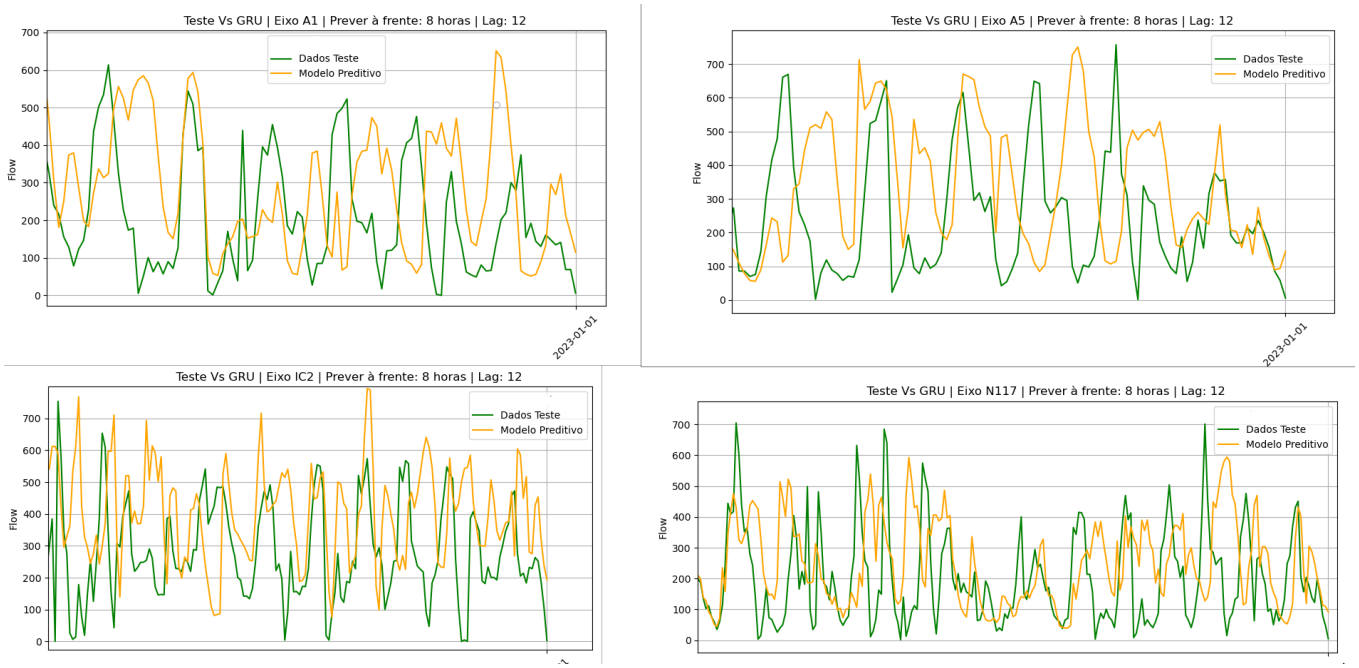


Figura 22 - Gráfico da previsão do tráfego para os eixos A1, A5, IC2 e N117 | Lag: 24

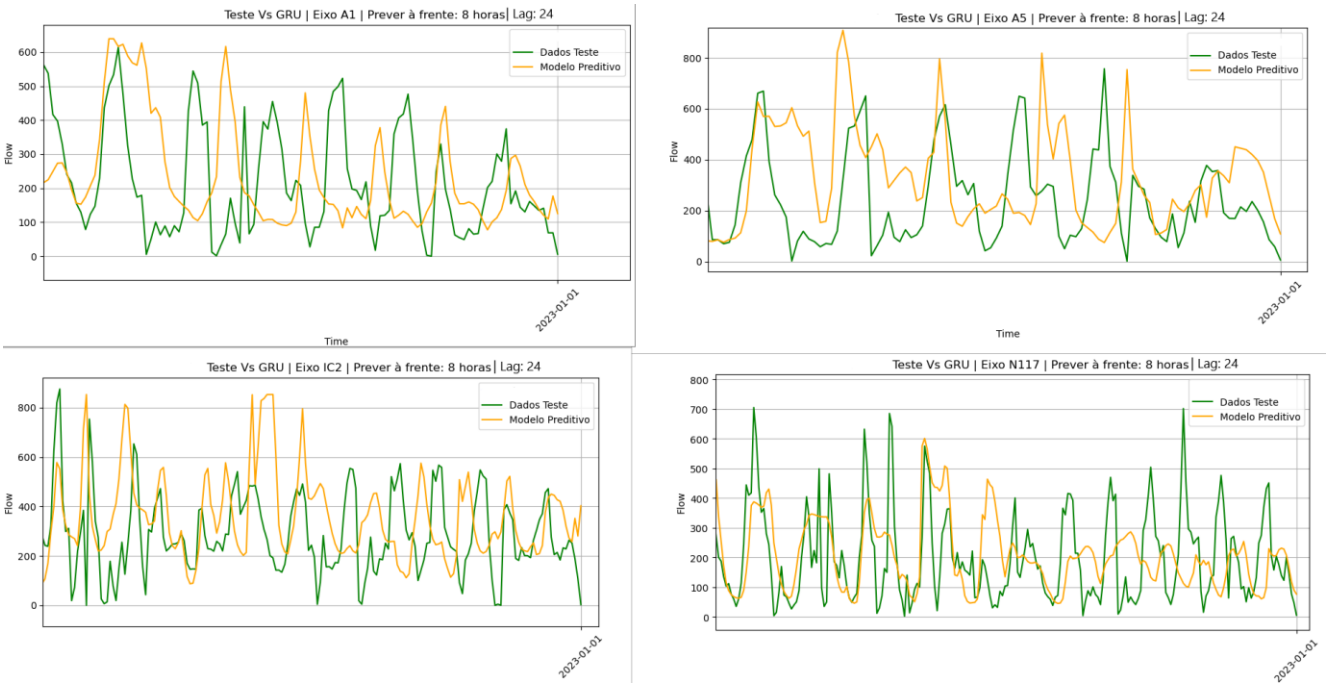


Figura 23 - Gráfico da previsão do tráfego para os eixos A1, A5, IC2 e N117 | Lag: 12

Na figura 23, o gráfico com *lag* 12 mostra flutuações mais notáveis nos dados de teste em relação ao modelo preditivo para prever 8 horas à frente. Essa diferença é resultado da consideração das observações do mesmo horário no dia anterior, capturando as correlações e padrões diários do tráfego. Para além disto, o intervalo de tempo no final de dezembro de 2022 revela que essa previsão é complexa, porque é a altura mais esperada do ano, tanto o Natal como a Passagem de Ano, e foi o ano onde levantaram todas as restrições de viagem da COVID-19.

8.3.2. *Lag* 1 semana (24x7) para prever 24h à frente

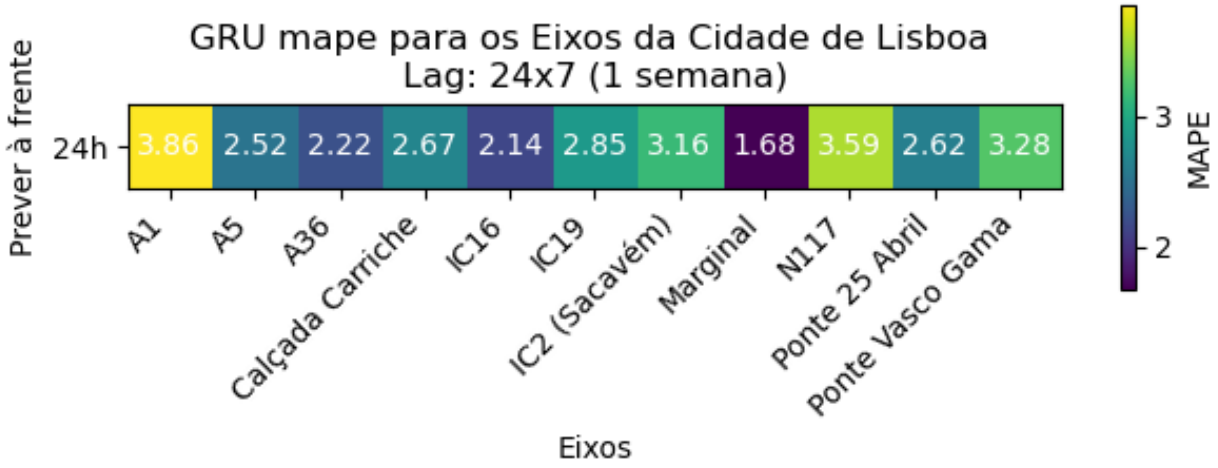


Figura 24 - MAPE da previsão do tráfego para os Eixos da Cidade de Lisboa | Lag: 1 semana

A figura 24 retrata a experimentação de uma semana de *inputs* para trás para prever 1 dia à frente e revela mais erro nas áreas de entrada da zona de Lisboa, como a A1, N117 e a ponte Vasco da Gama. Na verdade, como existe mais congestionamento nessas áreas, porque são pontos chave para movimentar a economia na cidade de Lisboa, o modelo apresenta dificuldades em apresentar resultados mais precisos.

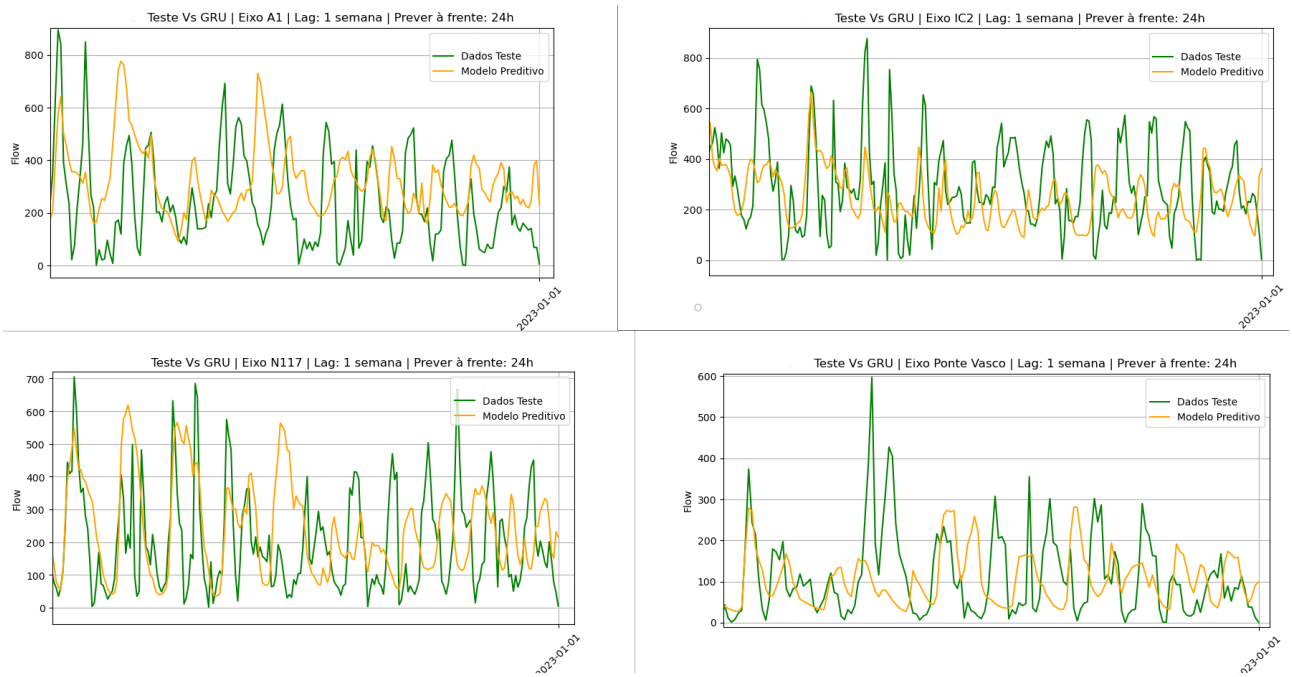


Figura 25 - Gráfico da previsão do tráfego para os eixos A1, IC2, N117 e Ponte Vasco Gama | Lag: 1 semana

Efetivamente, fez-se a seleção dos seguintes eixos: A1, IC2, N117 e Ponte Vasco Da Gama, com taxas de erro MAPE de 3.86%, 3.16%, 3.59% e 3.28%, respetivamente, conforme ilustrado na figura 24. Embora esses eixos apresentem erros relativamente baixos, os gráficos da figura 25 revelam grandes flutuações para um intervalo de tempo das duas últimas semanas de dezembro.

8.3.3. Testes para Feriados e Fins de Semana

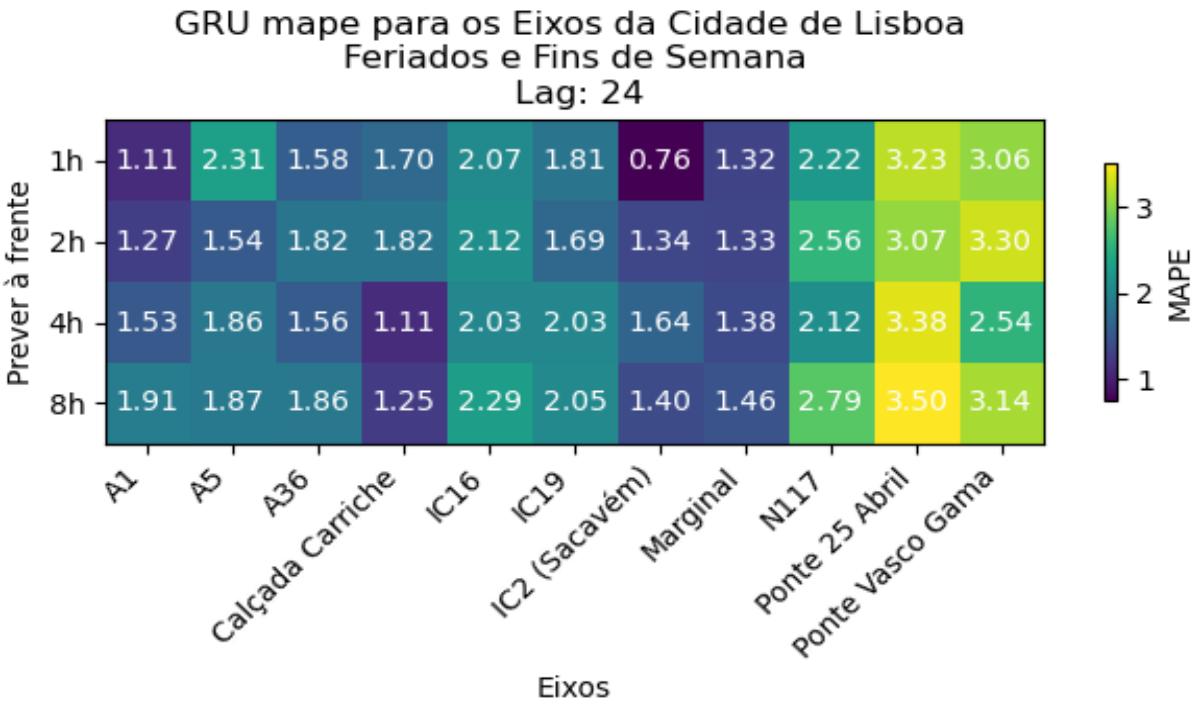


Figura 26 - MAPE para os Eixos da Cidade de Lisboa nos feriados e fins de semana | Lag: 24

É interessante observar que o *MAPE* é menor na maioria dos eixos menos nas pontes durante os fins-de-semana e feriados. Essa diferença de erro médio, é evidente em relação à figura 26 (lag 24), em particular 1.3pp para a Ponte 25 de Abril e 0.63pp para a Ponte Vasco da Gama. De modo que, fazendo melhor uso de *feature engineering*, sinalizando os registos associados a fins de semana e feriados, seria possível obter melhores resultados.

8.4 Testes para a soma das entradas de cada eixo

Após realizar os testes em cada eixo, foi determinado a soma dos valores de todos os eixos por registo, a fim de obter uma visão mais global.

8.4.1 *Lag* 24 vs *Lag* 12

No que diz respeito à soma das entradas, foi selecionado um período específico entre outubro e novembro para análise. Ao analisar o gráfico, torna-se evidente que, à medida

que a janela de previsão aumenta, o modelo encontra mais dificuldade em captar e prever com precisão nos picos dos dados, como é observado nas figuras 27 e 28.

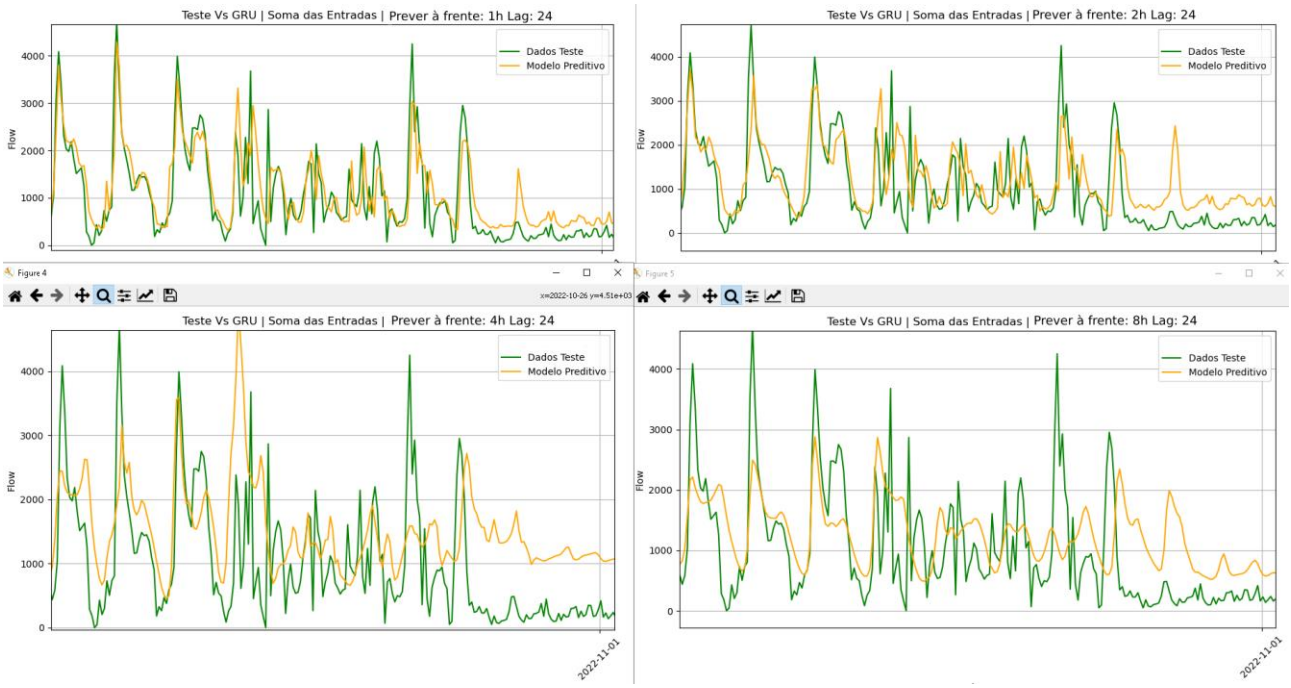


Figura 27 - Gráficos soma das entradas teste vs. modelo para as janelas de previsão de 1, 2, 4, 8 horas com lag 24

Tabela 12 - MAPE obtido na previsão da Soma das Entradas com Lag: 24

Prever à frente: 1h	**Prever à frente: 2h**	**Prever à frente: 4h**	**Prever à frente: 8h**
8.42	10.13	10.34	11.04

Quando o *gap* aumenta, o *MAPE* também aumenta, neste caso com uma diferença média, entre cada intervalo de tempo, de aproximadamente 0.71pp para *lag* 24 e 0.39pp para *lag* 12, como se observa nas tabelas 12 e 13.

Figura 28 - Gráficos soma das entradas teste vs. modelo para as janelas de previsão de 1, 2, 4, 8 horas com lag 12

Tabela 13 - MAPE obtido na previsão da Soma das Entradas com Lag: 12

Prever à frente: 1h	**Prever à frente: 2h**	**Prever à frente: 4h**	**Prever à frente: 8h**
8.27	12.5	12.54	13.1

8.4.2. *Lag* 1 semana (24x7) para prever 24h à frente

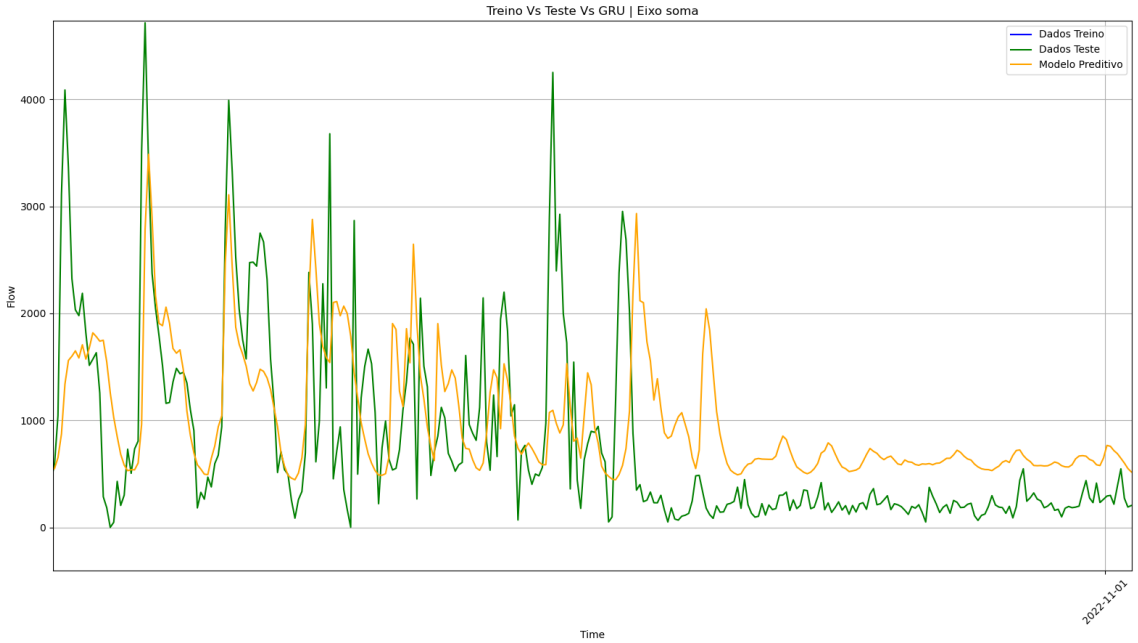


Figura 29 - Gráficos da previsão da soma das entradas teste vs. modelo para prever 24 à frente com 1 semana de lag

Apesar do gráfico da figura 29 apresentar um *MAPE de* 12.40%, observa-se que o modelo acompanha relativamente bem o padrão da amostra de teste para a soma das entradas de todos os eixos.

8.4.3 Testes para Feriados e Fins de Semana

Para demonstrar as diferenças de cada janela de previsão, foi usado o intervalo de tempo que inclui 6 feriados entre 21 de novembro (1º jogo de Portugal no mundial de 2022) e 31 de dezembro (final do ano), que são representados pelo traço interrompido verde na vertical, como apresenta a figura 30.

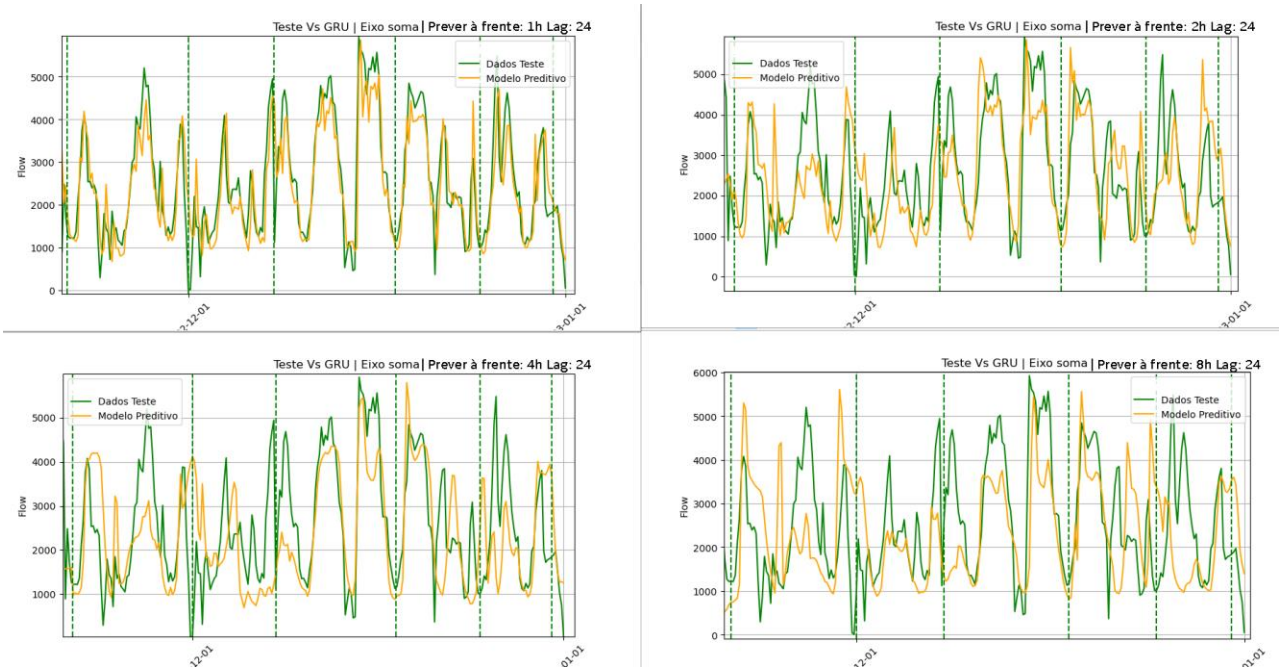


Figura 30 -Gráficos soma das entradas teste vs modelo para as janelas de previsão de 1, 2, 4, 8 horas com lag 24 para os feriados e fins de semana

Tabela 14 - MAPE Soma das Entradas com Lag: 24 nos feriados e fins de semana

Prever à frente: 1h	Prever à frente: 2h	Prever à frente: 4h	Prever à frente: 8h
1.18	1.74	1.91	2.26

É possível observar que o *MAPE* mostra resultados baixos na tabela para os testes realizados durante os feriados e fins-de-semana, no entanto o modelo encontra desafios em acompanhar a amostra teste à medida que o *gap* aumenta, como demonstra os gráficos.

8.5 Limitações do Modelo

O modelo atual não leva em conta informações contextuais como feriados e fins de semana, o que afeta a precisão das previsões de tráfego. Uma abordagem seria adicionar variáveis binárias que indicam se um dia é feriado ou fim de semana, através de *feature engineering*. Além disso, a representação temporal gerada pelas camadas GRU pode não ser suficientemente para capturar padrões complexos e de longo prazo. Modelos mais avançados, como redes neurais convolucionais de 1D, podem ser considerados para melhorar a captura de informações temporais [16]. A combinação dessas abordagens pode levar a previsões de tráfego mais precisas, considerando os fatores contextuais que sejam relevantes. Em alternativa, podem usar-se dois modelos, um para prever o tráfego durante a semana e outro para o fim de semana, como apresentado neste relatório.

9 CONCLUSÕES E TRABALHO FUTURO

Neste capítulo são apontadas as conclusões acerca do trabalho e sugestões futuras para continuar o trabalho.

9.1 Conclusões

Examinámos a previsão de tráfego através dos dados de localização de telemóveis, baseada em RNN, em particular o *GRU*, na qual foi estimada a quantidade de tráfego móvel com período de 1 h. O desempenho dos modelos de previsão de tráfego foi verificado utilizando os dados realistas recolhidos pelo LxDataLab, durante setembro de 2021 e dezembro de 2022. Em relação à avaliação do desempenho, confirmamos que a intensidade de tráfego móvel pode ser prevista com alguma precisão utilizando o modelo GRU. Esta conclusão é sustentada pelos erros percentuais que são relativamente baixos, em geral inferiores a 10% tanto para cada um dos 11 eixos como para a soma de todos, na previsão com 8 h de antecedência.

9.2 Trabalho Futuro

Como trabalho futuro, são apresentadas aqui, as seguintes sugestões:

- Incorporar modelo para diferenciar dias de semana dos feriados e fins de semana
- Explorar outros modelos, como LSTM e SAEs
- Explorar modelos para dados com períodos de amostragem de 5, 15, 30 min

REFERÊNCIAS [3]

[1] "xiaochus", L. (s.d.). https://github.com/xiaochus/TrafficFlowPrediction. [Acedido: 19-07-2023].

[2] Astonzhang. (s.d.). 10.2. Gated Recurrent Units (GRU). Obtido de d2l.ai: https://d2l.ai/chapter_recurrent-modern/gru.html#gated-recurrent-units-gru [Acedido: 19-07-2023].

[3] Cho, K. (11 de Dezembro de 2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

[4] Eddy Sanchezdela Cruz, B. M. (7 de Fevereiro de 2023). Urban Traffic Flow Identification by Comparing Machine Learning Algorithms. p. 7.

[5] Fazel Haq Ahmadzai, W. L. (28 de Maio de 2022). A mobile traffic load prediction based on recurrent neural network: A case of telecommunication in Afghanistan.

[6] gates, I. G. (s.d.). Obtido de https://www.youtube.com/watch?v=8HyCNIVRbSU&ab_channel=TheA.I.Hacker-MichaelPhi [Acedido: 19-07-2023].

[7] GRU, I. d. (s.d.). Obtido de https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-4/ [Acedido: 19-07-2023].

[8] IPMA. (s.d.). meteorologia/previsao. Obtido de ipma: https://www.ipma.pt/pt/educativa/faq/meteorologia/previsao/faqdetail.html?f=/pt/educativa/faq/meteorologia/previsao/faq_0033.html [Acedido: 19-07-2023].

[9] Lisboa, C. M. (s.d.). lisboaaberta.cm-lisboa. Obtido de Desafios: https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/desafios-teste [Acedido: 19-07-2023].

[10] lxdatalab. (s.d.). desafio-73-lxdatalab. Obtido de dados.cm-lisboa.pt: https://dados.cm-lisboa.pt/dataset/desafio-73-lxdatalab [Acedido: 19-07-2023].

[11] LxDataLab. (s.d.). LxDataLab Apresentação. Obtido de lisboaaberta.cm-lisboa.p: https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/apresentacao [Acedido: 19-07-2023].

[12] Pytorch, M. G. (s.d.). Obtido de https://pytorch.org/docs/stable/generated/torch.nn.GRU.html [Acedido: 19-07-2023].

[13] Teresa Pamuła, R. Ż. (4 de Novembro de 2022). Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning.

[14] Wijaya, C. Y. (26 de Abril de 2021). CRISP-DM Methodology For Your First Data Science Project.

[15] Youcef Djenouri, A. B.-W. (22 de Setembro de 2022). Hybrid graph convolution neural network and branch-and-bound optimization for traffic flow forecasting.

[16] Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., . . . Li, H. (22 de Agosto de 2019). T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction.

ANEXO

ANEXO A:

Proposta de Projeto

Ano Letivo de 2022/2023

2º Semestre

LxDataLab #73 - Movimentos pendulares nas principais vias de acesso à cidade de Lisboa, com base em dados de telemóveis

SUMÁRIO

A caracterização do tráfego no município de Lisboa é algo fundamental para o planeamento da vida na cidade, nomeadamente no que se refere ao volume de pessoas que nela entram diariamente nas horas de ponta da manhã (7:30h-10:00h) e da tarde (17:00h-19:30h), o que gera congestionamentos nas principais vias de acesso. Para os 11 principais pontos de entrada e saída da cidade existem dados que permitem conhecer o número de dispositivos móveis que entram e saem por cada um desses pontos a cada período de 15 minutos. O desafio que se coloca é tentar caraterizar estes fluxos diários durante os dois períodos referidos e a sua relação com fatores como os calendários escolares e a pluviosidade.

ESTAGIÁRIO (indicar o destinatário da proposta se já estiver definido)

Número de aluno: 2016020798

Nome completo: Ricardo Joaquim Gonçalves Ferreira

RAMO (indicar o(s) ramo(s) em que se enquadra)

■ Desenvolvimento de Aplicações

□ Redes e Administração de Sistemas

□ Sistemas de Informação

ENTREVISTA/PROCESSO DE SELEÇÃO (informar se o candidato indicado pelo DEIS ISEC será submetido a uma entrevista ou outro tipo de processo de seleção antes da sua admissão efetiva)

■ Não

□ Talvez Especificar:

□ Sim Especificar:

1. ÂMBITO

Este projeto com o Laboratório de Dados Urbanos de Lisboa (LxDataLab) surge no âmbito da parceria entre o ISEC e a Câmara Municipal de Lisboa (CML). Este surge da necessidade de extrair valor da informação disponível no município, com recurso a ferramentas avançadas de análise de dados e a recursos humanos especializados, para assim criar soluções de analítica capazes de melhorar o planeamento, resiliência, segurança, mobilidade, a gestão operacional e de emergência na cidade de Lisboa. Este projeto envolve o município e a academia num ecossistema inovador. A CML lança desafios, sendo este projeto um desses desafios. Devido à natureza dos dados, o aluno terá de aceitar e assinar os termos e condições do LxDataLab, nomeadamente assinando um acordo de confidencialidade para acesso aos dados.

2. OBJETIVOS

De acordo com o descritivo disponível1pretende-se conhecer o seguinte:

A - Para a períodos de ponta da manhã (7:30h-10:00h)

- Caraterizar o volume total de entradas e saídas da cidade durante o período da hora de ponta,

- Caraterizar o volume de entradas e saídas da cidade durante o período da hora de ponta para

cada um dos 11 pontos de entrada e saída,

- Comparar com outros períodos do dia,

- Relacionar o ponto anterior com variáveis como calendários escolares e a ocorrência de

pluviosidade,

- Análise das zonas de origem daqueles que entram na cidade,

- Análise das zonas de destino daqueles que saem da cidade.

B - Para a períodos de ponta da tarde (17:00h-19:30h)

- Caraterizar o volume total de entradas e saídas da cidade durante o período da hora de ponta,

- Caraterizar o volume de entradas e saídas da cidade durante o período da hora de ponta para

cada um dos 11 pontos de entrada e saída,

- Comparar com outros períodos do dia,

- Relacionar o ponto anterior com variáveis como os períodos de aulas ou férias e a existência

de pluviosidade,

- Análise das zonas de destino daqueles que saem da cidade,

- Análise das zonas de origem daqueles que entram na cidade.

Para além das análises exploratórias anteriores, pretende-se ainda outros resultados que surjam do interesse/criatividade/capacidade do aluno, nomeadamente a proposta de modelos preditivos para o número de telemóveis que entram e saem.

Dados a disponibilizar

- Conjunto1 - Número de telemóveis que entram e saem da cidade a cada 15 minutos nos 11

principais eixos de entrada na cidade de Lisboa - Eixos da cidade de Lisboa

- Conjunto2 - Identificação dos 11 pontos de entrada e saída de Lisboa

Conjunto3 - Observações das estações meteorológicas do IPMA de Lisboa: Geofísico, Gago Coutinho e Tapada da Ajuda

3. PROGRAMA DE TRABALHOS

O projeto consistirá nas seguintes atividades e respetivas tarefas:

 T1 – Estado da arte – recorrendo a artigos, livros, tutoriais, blogs, github realizar uma revisão do estado da arte dos desenvolvimentos nesta área;

 T2 – Estudo do dataset e ferramentas de análise de dados – caraterizar o dataset, aprender a utilizar python, bibliotecas associadas a este domínio e outras ferramentas necessárias;

 T3 – Análise exploratória dos dados – estudo dos dados usando ferramentas de análise adequadas;

 T4 – Relatório – Escrita de relatório detalhado e resumo do mesmo para publicação dos principais resultados.

4. CALENDARIZAÇÃO DAS TAREFAS

O plano de escalonamento das tarefas é apresentado em seguida (a adaptar em função do projeto/estágio)

	Tarefas		N		N+1		N+2		N+3		N+4		N+5	
T1	Estado da arte													
T2	Estudo do dataset e ferramentas													
T3	Análise de dados													
T4	Relatório													
	Metas	INI			M1				M2			M3		M4

INI	Início dos trabalhos
M1 (INI + 4 Semanas)	Tarefa T1 terminada
M2 (INI + 12 Semanas)	Tarefa T2 terminada
M3 (INI + 14 Semanas)	Tarefa T3 terminada
M4 (INI + 18 Semanas)	Tarefa T4 terminada
M5 (INI + 20 Semanas)	Tarefa T5 terminada

5. LOCAL, HORÁRIO DE TRABALHO E COMDIÇÕES OFERECIDAS

O trabalho decorrerá no ISEC e em teletrabalho com uma média de 35h/semana.

6. TECNOLOGIAS ENVOLVIDAS

python, bibliotecas diversas, linux shell scripting.

7. METODOLOGIA

Será criado um dossier do projeto, procurando-se desenvolver material relevante para o relatório desde as fases iniciais. O acompanhamento será feito através de reuniões a calendarizar, de acordo com o procedimento existente no ISEC.

ORIENTAÇÃO

DEIS-ISEC: Mateus Mendes

Email do orientador <mmendes@isec.pt>

Prof. Adjunto

DFM-ISEC: Nuno Lavado

Email do orientador <nlavado@isec.pt>

Prof. Adjunto