Título

Relatório de Estágio

Nome aluno

(Número)

Orientadores:

Nome | ISEC

Nome | Empresa

Licenciatura em Engenharia Informática

Ramo de …

Instituto Politécnico de Coimbra

Instituto Superior de Engenharia de Coimbra

Julho de 2021

# Agradecimentos

…

# Resumo

… este documento pretende apenas servir de linha orientadora, de uso opcional, para a elaboração dos relatórios de estágio… a estrutura e formatação finais dos relatórios deverão ser definidos caso a caso pelos alunos e respetivos orientadores…

***Palavras‑chave:*** *Big Data, Serialização, Spring Boot.*

# Abstract

…

***Keywords:*** *Big Data, Serialization, Spring Boot.*

**Índice**

[Agradecimentos i](#_Toc138323990)

[Resumo iii](#_Toc138323991)

[Abstract v](#_Toc138323992)

[Índice de figuras xi](#_Toc138323993)

[Índice de tabelas xiii](#_Toc138323994)

[Acrónimos xv](#_Toc138323995)

[1 Introdução 1](#_Toc138323996)

[1.1 Entidade de acolhimento 2](#_Toc138323997)

[1.2 Instituto Superior de Engenharia de Coimbra 2](#_Toc138323998)

[1.3 Objetivos e plano de trabalhos 2](#_Toc138323999)

[1.4 Estrutura do relatório 2](#_Toc138324000)

[2 Estado da Arte 5](#_Toc138324001)

[2.1 Metodologia 5](#_Toc138324002)

[2.2 Fluxo de Tráfego 6](#_Toc138324003)

[2.2.1 Matriz Origem-Destino como sensor de tráfego 6](#_Toc138324004)

[2.2.2 Classificação do fluxo de tráfego através de algoritmos de *machine learning* 6](#_Toc138324005)

[2.2.3. *Promotion FrameWork* para prever o fluxo tráfego 7](#_Toc138324006)

[2.2.3. Crescimento *IoT* e a sua utilidade para melhorar o fluxo do tráfego 7](#_Toc138324007)

[2.2.4. *Smartphones* como sensores de fluxo de tráfego 7](#_Toc138324008)

[3. CRISP-DM 9](#_Toc138324009)

[3.1. Compreensão do negócio 9](#_Toc138324010)

[4. *Dataset* 11](#_Toc138324011)

[4.1. Contexto 11](#_Toc138324012)

[4.2. Conteúdo 11](#_Toc138324013)

[4.2.1 Conjunto 1 – Entradas e Saídas Lisboa 12](#_Toc138324014)

[4.2.2 Conjunto 2 - Identificação dos 11 pontos de entrada e saida de Lisboa 13](#_Toc138324015)

[4.2.3 Conjunto 3 - Observações das estações meteorológicas do IPMA 13](#_Toc138324016)

[4.3. Qualidade do conteúdo 14](#_Toc138324017)

[4.3.1. Conjunto 1 14](#_Toc138324018)

[3.5.1.1. Análise Registos 17](#_Toc138324019)

[4.3.2. Conjunto 3 19](#_Toc138324020)

[5. Análise Exploratória 22](#_Toc138324021)

[5.1. Caraterizar o volume total de entradas e saídas da cidade durante o período das horas de ponta 22](#_Toc138324022)

[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 24](#_Toc138324023)

[5.3. Comparar com outros períodos do dia 25](#_Toc138324024)

[5.4. Relacionar variáveis como os períodos de aulas ou férias e a existência de pluviosidade 26](#_Toc138324025)

[5.5. Análise das zonas destino e origem da cidade 26](#_Toc138324026)

[5.5.1. Dias Uteis vs. Outros Dias 26](#_Toc138324027)

[5.5.2. Horas de Ponta vs. Outras Horas 29](#_Toc138324028)

[5.6. Série Temporal dos dados 30](#_Toc138324029)

[5.7. Autocorrelação 30](#_Toc138324030)

[6. Gated Recurrent Unit (GRU) 31](#_Toc138324031)

[6.1. GRU vs LSTM 31](#_Toc138324032)

[6.2. Mecanismo 32](#_Toc138324033)

[6.3. Implementação 32](#_Toc138324034)

[*7.* Conceitos de *Machine Learning* 35](#_Toc138324035)

[7.1. Pré-processamento dos dados 35](#_Toc138324036)

[7.2. Arquitetura do modelo 35](#_Toc138324037)

[7.3. Treinar o modelo 35](#_Toc138324038)

[8. Testes de Previsão e Resultados 37](#_Toc138324039)

[8.1. Testes para as entradas de cada eixo 37](#_Toc138324040)

[8.1.1. Lag 24 vs Lag 12 37](#_Toc138324041)

[8.1.2. Lag 24x7 (1 semana) para prever 24h à frente 39](#_Toc138324042)

[39](#_Toc138324043)

[8.1.3. Testes para Feriados e Fins de Semana 40](#_Toc138324044)

[8.2. Testes para a soma das entradas de cada eixo 40](#_Toc138324045)

[8.2.1. Lag 24 vs Lag 12 40](#_Toc138324046)

[8.2.2. Lag 24x7 (1 semana) para prever 24h à frente 43](#_Toc138324047)

[8.2.2. Testes para Feriados e Fins de Semana 43](#_Toc138324048)

[8.3. Requisitos 44](#_Toc138324049)

[8.4. Arquitetura 44](#_Toc138324050)

[8.5. Tecnologias e ferramentas usadas 45](#_Toc138324051)

[8.6. Implementação 45](#_Toc138324052)

[8.7. Testes e validação 45](#_Toc138324053)

[8.7.1. Método A 46](#_Toc138324054)

[8.7.2. Método B 46](#_Toc138324055)

[9. Mais um capítulo 47](#_Toc138324056)

[10. Conclusões e trabalho futuro 49](#_Toc138324057)

[10.1. Conclusões 49](#_Toc138324058)

[10.2. Trabalho Futuro 49](#_Toc138324059)

[Referências [3] 51](#_Toc138324060)

[Anexos 53](#_Toc138324061)

[10.2.1.1.1.1.1.1.1. Proposta de Estágio 1](#_Toc138324062)

[10.2.1.1.1.1.1.1.2. Especificação de Requisitos 1](#_Toc138324063)

# Índice de figuras

Figura 1.1 Explicação da figura 1

Figura 2 - Modelo CRISP-DM. Fonte:[3] 11

# Índice de tabelas

[Tabela 4.1 – Tecnologias e ferramentas usadas40](#_Toc68777812)

# Acrónimos

…

ITS – Intelligent Transportation Systems

…

# Introdução

A tecnologia dos telemóveis revolucionou a forma como as pessoas e as empresas obtêm e partilham informação sobre mobilidade, fornecendo conjuntos massivos de trajetórias registadas, dados longitudinais e o seu potencial e precisão em tempo real. Este projeto é beneficiado pelo CGIUL e pretende analisar a gestão da mobilidade, através de dados de telemóveis, nomeadamente na questão do planeamento do quotidiano na cidade de Lisboa. Efetivamente, é fornecido conjuntos de dados de telemóveis referentes ao volume de pessoas que entram e saiam dos 11 principais pontos de entrada e saída da cidade nas horas de ponta da manhã (7:30h-10:00h) e da tarde (17:00h-19:30h), que geram congestionamentos nas principais vias de acesso. O objetivo do projeto é caraterizar esses fluxos diários durante os dois períodos referidos e a sua relação com fatores como calendário escolar, períodos de férias e ocorrência de temporal.

Contextualização… depois de ler os parágrafos que compõem este início de capítulo, deve ficar‑se a entender que existe uma determinada área/contexto, uma determinada empresa que pretende fazer ou já faz algo nessa área e que apresentou, no âmbito da unidade curricular de Projeto ou Estágio da licenciatura em Engenharia Informática (LEI) [1] do Instituto Superior de Engenharia de Coimbra (ISEC) [2], uma proposta de estágio com um determinado objetivo… que este documento pretende descrever o trabalho realizado…

…a referir no texto as figuras, quando existirem, conforme apresentado na Figura 1.1…

[](https://www.google.pt/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwiW-4ij3_zSAhXB0RoKHYa2AJkQjRwIBw&url=http://mariohenriquemartins.com.br/enquadramento-o-que-voce-destaca-em-seu-argumento/&psig=AFQjCNEBh3yreV0BxghV2mg3qhOIzvt7pA&ust=1490912295373271)

Figura 1.1 Explicação da figura

… as subsecções seguintes são apenas exemplificativas…

## Entidade de acolhimento

… a usar referências quando se mencionar algo dito por outra pessoa [2]…

## Instituto Superior de Engenharia de Coimbra

…

## Objetivos e plano de trabalhos

Através da análise dos fluxos diários 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;

… referir a proposta original, que se encontra em anexo, e apresentar o que realmente aconteceu…

## Estrutura do relatório

…

…

# 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. Efetivamente, também é analisado como estas técnicas podem ser relacionadas como o uso de smartphones como sensores para captar dados relativos à mobilidade.

## Metodologia

A pesquisa foi realizada principalmente usando a base de dados do *Google Scholar*, onde foi realizada pesquisa com a palavra-chave *"traffic flow"* e esta apresentou 3 740 000 resultados. Depois comecei a experimentar outras palavras-chave, que são mais específicas em termos de utilização, designadamente “mobile data” e *“machine learning”*. Estas apresentaram menos resultados, mas ainda devolvem 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.

Tabela 1 - 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 |

## 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 [4], algoritmos de *“Machine Learning”* para classificar o congestionamento do tráfego [5] ou utilização de *frameworks* como a *“Promotion”* para prever o fluxo do tráfego [6], 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 [7][8].

### 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) [4]. 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, um estudo em 2017 constatou que o custo dos engarrafamentos de trânsito para os condutores americanos foi estimado em 179 milhões de dólares [4], como tal, para ajudar a melhorar as condições de tráfego os investigadores decidiram avaliar o comportamento e eficácia de vários algoritmos de *“machine learning”* para classificar esse fluxo e 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.

### *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 [6].

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.

### Crescimento *IoT* e a sua utilidade para melhorar o fluxo do tráfego

Foi estimado que cerca de 77,5 Exa bytes de dados móveis serão consumidos por mês até 2022, o que representa 71% de todo o tráfego IP [7]. Consequentemente, é essencial prever com precisão a carga de tráfego móvel para gerir eficientemente o grande volume de tráfego móvel. Por exemplo, foi desenvolvido um modelo baseado em "*Machine Learning*" chamado *Deep Traffic Predictor* (DeepTP), que utiliza RNNs, LSTM, CNNs e modelos LSTM para prever a carga de tráfego móvel, como também foi descoberto que o fluxo pode ser previsto com alta precisão, otimizando a alocação de recursos e a qualidade da circulação.

### *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* [8] 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 [7] para prever com precisão o fluxo do tráfego.

Tabela 2 - Análise comparativa de artigos baseado no fluxo do tráfego

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Artigo** | **Ano** | **Objetivo** | **Técnicas Aplicadas** | **Resultados** |
| [4] | 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*). |
| [5] | 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. |
| [6] | 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*). |
| [7] | 2022 | Previsão da quantidade de tráfego móvel | *LSTM*  *GRU*  *Random Forest*  *Linear Regression*  *Support Vector Machine* | Os esquemas baseados no *RNN* foram mais precisos que os esquemas convencionais baseados em *“Machine Learning”,* especialmente o método *GRU* que melhorou o desempenho 42.91% em termos de *RMSE* e 1.90% em resultados *R2.* |
| [8] | 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. |

# 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.

## 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. De facto, é desenvolvido um plano de projeto com base nestes objetivos e é 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 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, ruidosos e inconsistentes e precisam 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 ou modelação, onde 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 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 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.

Uma imagem com diagrama

Descrição gerada automaticamente

Figura 2 - Modelo CRISP-DM. Fonte:[3]

# *Dataset*

## Contexto

O conjunto de dados fornecidos pela Vodafone contém mais de 1 milhão de registos de “*Número de telemóveis que entram e saem da cidade de Lisboa*”. Há também “*Observações das estações meteorológicas do IPMA*” entre setembro de 2021 e dezembro de 2022. No que diz respeito à obtenção dos dados, existem 11 eixos principais usados para manter o registo do fluxo de entrada e saída de telemóveis.

## Conteúdo

//Esclarecer que falta conjunto 3

O conteúdo é dividido por 4 conjuntos, nomeadamente: i) grelha 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 outro ficheiro que os represente; e por fim iv) os dados meteorológicos do IPMA, representados no conjunto 3.

## 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

Abaixo apresenta-se estatísticas adicionais 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 - 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 |

## 4.2.2 Conjunto 2 - Identificação dos 11 pontos de entrada e saida 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

intensidadeventokm: Intensidade média do vento

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.

Abaixo 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 - 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 |

## 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

Em relação a 2021, os registos são bastante inconsistentes, porque são diferentes para cada eixo e particularmente em setembro, que os registos se iniciam no dia 14 e acabam em outubro no dia 1. Para além disto, os registos deste ano indicam uma média de registos de 66.6 % de 5 min e 33.4 % de 15 min.

Tabela 4 - Análise registos conjunto1 2021

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **VODAFONE\_EIXOS** | | | | | | | | | | | | | |
| **PGIL\_VODAFONE\_EIXOS\_2021** | | | | | | | | | | | | | |
| **Mês** | **# Registos** | **A1** | **A5** | **A36 Túnel do Grilo** | **Calçada de Carriche** | **IC2 (Sacavém)** | **IC16** | **IC19** | **Marginal** | **N117 (Cabos Ávila)** | **Ponte 25 Abril** | **Ponte Vasco Gama** | **Datas Registadas** |
| Setembro  **\*(5min)**  **\*\*(15min)** | 53844  35977  17867 | 4930  3288  1642 | 4947  3299  1648 | 4941  3312  1629 | 4882  3287  1595 | 4983  3319  1664 | 4841  3209  1632 | 4877  3258  1619 | 4862  3272  1590 | 4854  3237  1617 | 4868  3248  1620 | 4869  3248  1611 | 14/09/2021 até 1/10/2021 00:00 às 23:55 |
| Outubro  **\*(5min)**  **\*\*(15min)** | 95898  63733  32165 | 8690  5722  2968 | 8626  5711  2915 | 8680  5786  2894 | 8784  5783  3001 | 8710  5841  2869 | 8729  5843  2886 | 8712  5751  2961 | 8751  5834  2917 | 8817  5902  2915 | 8687  5748  2939 | 8712  5812  2900 | 1/10/2021 até 31/10/2021 00:00 às 23:55 |
| Novembro  **\*(5min)**  **\*\*(15min)** | 79288  52811  26477 | 7218  4821  2397 | 7152  4773  2379 | 7191  4824  2367 | 7245  4813  2432 | 7189  4786  2403 | 7204  4778  2426 | 7221  4812  2409 | 7239  4803  2436 | 7237  4795  2442 | 7166  4782  2384 | 7226  4824  2402 | 1/11/2021 até 31/11/2021 00:00 às 23:55 |
| Dezembro  **\*(5min)**  **\*\*(15min)** | 97240  64874  32366 | 8840  5880  2960 | 8830  5926  2904 | 8800  5847  2953 | 8876  5948  2928 | 8816  5861  2955 | 8824  5907  2917 | 8958  5960  2998 | 8884  5958  2926 | 8872  5921  2951 | 8741  5791  2950 | 8799  5875  2924 | 1/12/2021 até 31/12/2021 00:00 às 23:55 |

No que diz respeito a 2022, falta consistência entre janeiro e março com o número de registos variável e as datas registadas em fevereiro entre dia 1 e 8 e março entre dia 18 e 31. 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% de 5 min e 33.4% de 15 min e 100% de registos de 15 min para os restantes meses do ano.

Tabela 5 - Análise registos conjunto1 2022 **\*–** Registos com 5min de diferença **\*\* –** Registos com 15min de diferença

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **VODAFONE\_EIXOS** | | | | | | | | | | | | | |
| **PGIL\_VODAFONE\_EIXOS\_2022** | | | | | | | | | | | | | |
| **Mês** | **Registos** | **A1** | **A5** | **A36 Túnel do Grilo** | **Calçada de Carriche** | **IC2 (Sacavém)** | **IC16** | **IC19** | **Marginal** | **N117 (Cabos Ávila)** | **Ponte 25 Abril** | **Ponte Vasco Gama** | **Datas Registadas** |
| Janeiro  **\*(5min)**  **\*\*(15min)** | 98005  65505  32500 | 8849  5920  2929 | 8941  5950  2991 | 8902  6005  2897 | 8865  5897  2968 | 8816  5953  2964 | 8917  5949  2978 | 8927  6043  2899 | 8959  5950  3009 | 8895  5959  2936 | 8921  5936  2985 | 8887  5943  2944 | 1/01/2022 até 31/01/2022 00:00 às 23:55 |
| Fevereiro  **\*(5min)**  **\*\*(15min)** | 19525  12937  6588 | 1777  1186  591 | 1773  1176  597 | 1762  1163  599 | 1789  1190  599 | 1761  1165  596 | 1790  1169  621 | 1761  1169  592 | 1784  1177  607 | 1750  1157  593 | 1787  1195  592 | 1791  1190  601 | 01/02/2022 até 08/02/2022 00:00 às 23:55 |
| Março  **\*(5min)**  **\*\*(15min)** | 42163  28166  13997 | 3826  2568  1258 | 3844  2567  1277 | 3802  2561  1241 | 3849  2548  1301 | 3845  2553  1292 | 3867  2567  1300 | 3888  2603  1285 | 3796  2495  1301 | 3812  2573  1239 | 3833  2572  1261 | 3801  2559  1242 | 18/03/2022 10:45 até 31/02/2022 00:00 às 23:55 |
| Abril  **\*(5min)**  **\*\*(15min)** | 31328  -  31328 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 2848  -  2848 | 01/04/2022 até 30/04/2022 00:00 às 23:55 |
| Maio  **\*(5min)**  **\*\*(15min)** | 31713  -  31713 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 2883  -  2883 | 01/05/2022 até 31/05/2022 00:00 às 23:55 |
| Junho  **\*(5min)**  **\*\*(15min)** | 31680  -  31680 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 2880  -  2880 | 01/06/2022 até 30/06/2022 00:00 às 23:55 |
| Julho  **\*(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 | 01/07/2022 até 31/07/2022 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 |

## 3.5.1.1. Análise Registos

Tabela 6 - Registos Esperados vs. Registos Validados (15min)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Registos 15min** | | | | |
| **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 |

Tabela 7 - Registos Esperados vs. Registos Validados (1h)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Registos 1h** | | | | |
| **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 |

Uma imagem com gráfico

Descrição gerada automaticamente

Figura - Registos Validados vs. Esperados por mês Conjunto 1

## 4.3.2. Conjunto 3

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. Uteis: 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 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 dissecar os dados com valores inválidos para a mesma data na mesma estação. Com efeito, após esta dissecação foram validados 45 442 registos, o que representam 4.7% do *dataset*.

Uma imagem com gráfico

Descrição gerada automaticamente

Figura - Diferença Registos Totais e Uteis Conjunto 3

Tabela - Descrição Geral Conjunto 3 Dados Uteis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variáveis** | **humidade** | **iddireccvento** | **intensidadeventokm** | **pressao** | **radiacao** | **temperatura** | **precacumulada** |
| Média | 69.38 | 4.25 | -13.19 | 1018.15 | 536.38 | 17.01 | -3.99 |
| Desvio Padrão | 18.01 | 3.50 | 47.05 | 18.22 | 960.26 | 5.39 | 19.56 |
| Mínimo | 12.0 | 0.0 | -99.0 | -99.0 | -99.0 | 1.0 | -99.0 |
| 25% | 56.0 | 0.0 | 0.7 | 1014.8 | 0.0 | 13.0 | 0.0 |
| Mediana | 71.0 | 4.0 | 9.4 | 1018.0 | 0.0 | 16.0 | 0.0 |
| 75% | 84.0 | 8.0 | 15.1 | 1022.0 | 821.0 | 20.0 | 0.0 |
| Máximo | 100.0 | 9.0 | 40.0 | 1035.2 | 3883.0 | 40.0 | 16.0 |

Tabela 6 - Registos Esperados vs. Validados (IPMA)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Registos 1H** | | | | |
| **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 |

Uma imagem com gráfico

Descrição gerada automaticamente

Figura - Registos Validados vs. Esperados por mês Conjunto 3

# Análise Exploratória

De acordo com o desafio proposto pelo Lisbon Urban Data Lab (LxDataLab), esta iniciativa entre o ISEC e a Câmara Municipal de Lisboa (CML) visa utilizar ferramentas de análise de dados para extrair conhecimento especializado sobre o tráfego movel ao 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.

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

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

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

Figura - Caraterização entradas e saídas nas horas de ponta da manhã (7:30h-10:00h)

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

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

Figura - Caraterização entradas e saídas nas horas de ponta da tarde (17:00h-19:30h)

De acordo com os dados mostrados nos gráficos, 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, houve um aumento notável no horário de pico da manhã tráfego em relação a dezembro de 2021. Existem vários fatores que podem ter contribuído para esse aumento. Por exemplo, é possível que as mudanças climáticas tenham sido mais favoráveis e apropriadas para as pessoas dirigirem para o trabalho, bem como as restrições do COVID-19 podem ter sido menos restritas em dezembro de 2022 ou mais pessoas podem ter retornado ao escritório, levando a mais tráfego.

## 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

Uma imagem com captura de ecrã, texto, diagrama, Saturação de cores

Descrição gerada automaticamente

Figura - Caraterização entradas nas horas de ponta por eixo

Com base na análise do gráfico, pode-se concluir que existe uma tendência significante em relação às entradas na cidade, nomeadamente na Ponte Vasco da Gama e na Ponte 25 de Abril, sendo estes os pontos de entrada mais utilizados, especialmente durante as horas de pico das 8h às 9h e das 19h às 20h. 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 80 000 entradas, resultando numa redução de 33% em relação ao pico da manhã.

Uma imagem com captura de ecrã, texto, diagrama, Saturação de cores

Descrição gerada automaticamente

Figura - Caraterização saídas nas horas de ponta por eixo

Efetivamente, relativamente a análise do gráfico das saídas, pode-se concluir que o Eixo N117 (Cabos de Avila) é o ponto de saída mais utilizado durante as horas de pico das 8h às 9h e as Pontes Vasco da Gama e 25 de Abril durante as horas de pico das 19h às 20h. Para além disto, o período de maior volume de saídas ocorre na parte da manhã, com cerca de 1 100 000 saídas e a tarde 80 000 saídas, o que representa uma quebra de 33% em relação ao pico da manhã, como na análise do gráfico de entradas.

## Comparar com outros períodos do dia

Em comparação com os outros períodos do dia, os resultados mostram que, durante as horas de ponta, existe uma diferença cerca de 30% no número de entradas e saídas pelos 11 eixos com as horas da tarde. Em contrapartida, durante as horas noturnas, a diferença foi muito mais significativa, com uma variação de 75% a 65% no número de entradas e saídas, respetivamente.

## Uma imagem com texto, captura de ecrã, diagrama, Gráfico Descrição gerada automaticamenteRelacionar variáveis como os períodos de aulas ou férias e a existência de pluviosidade

Figura - Períodos de aulas ou férias vs. existência de pluviosidade

Na verdade, existem 2100 registos nos dias de escola divididos pelos 11 eixos, em relação aos 160 registos de chuva, sendo esta caraterizada entre chuva fraca entre 0 a 5 mm e chuva moderada entre 5 a 25mm.

Observando o gráfico, verifica-se que média de entradas/saídas em relação aos dias de chuva é cerca de metade, no entanto é difícil fazer uma qualificação objetiva, porque os dados são insuficientes para a questão do problema.

## Análise das zonas destino e origem da cidade

Relativamente à análise das zonas destino e origem da cidade, foi estabelecido a comparação na quantidade de tráfego (entradas e saídas) entre dias uteis como sexta e domingo, horas de ponta das 7 às 10 da manhã e das 17 às19 da tarde e períodos escolares de janeiro até junho e setembro até dezembro.

## 5.5.1. Dias Uteis vs. Outros Dias

Uma imagem com gráfico

Descrição gerada automaticamenteUma imagem com gráfico

Descrição gerada automaticamenteAtravés da observação dos gráficos, verifica-se que a quantidade de tráfego é aproximadamente metade nas sextas e domingos, de modo que é importante apontar algumas razões para o elevado fluxo de tráfego nestes eixos, nomeadamente através de padrões de deslocação, turísticos e de eventos. Os padrões de deslocação podem surgir do facto de ser a capital, pois engloba uma grande rede comercial e residencial, os padrões turísticos indicam ao facto de ser a área mais popular de Portugal e os padrões de eventos revelam a realização de concertos, festivais, atividades desportivas, entre Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamenteoutros.

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

Figura - Tráfego nos dias uteis para as pontes 25 de Abril e Vasco da Gama

## 5.5.2. Horas de Ponta vs. Outras Horas

Estudos mostram que entender 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 conveniência, de modo a escolher livremente o percurso pretendido. A segunda é a flexibilidade para garantir a deslocação devido a horários de trabalho variáveis.

Uma imagem com gráfico

Descrição gerada automaticamente

Uma imagem com gráfico

Descrição gerada automaticamente

Figura - Tráfego nas horas de ponta para as pontes 25 de Abril e Vasco da Gama

Através da análise dos gráficos, verifica-se que a quantidade de tráfego nas horas de ponta ronda 75% do tráfego na maioria dos eixos, exceto na Ponte 25 de Abril e na Ponte Vasco da Gama. De facto, estas pontes fazem a ligação para a área da grande Lisboa e desempenham um papel vital no suporte das atividades diárias da cidade. No entanto, é possível observar que existe uma distribuição desigual de tráfego, onde existem mais entradas na Ponte 25 de Abril e mais saídas na Ponte Vasco da Gama. Uma possibilidade para isto acontecer é que do outro lado da ponte existe uma grande área residencial e industrial para as pessoas chegarem ao trabalho ou para casa, entre outros fatores

## Série Temporal dos dados

Uma imagem com diagrama, captura de ecrã, Gráfico, design

Descrição gerada automaticamenteFazer a série temporal dos dados dá uma noção mais clara dos dados que vão ser trabalhados, neste caso das variáveis C12 e C13. 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 sequencias.

Figura - Série temporal do conjunto de dados

## 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.

Uma imagem com texto, file, captura de ecrã, Gráfico

Descrição gerada automaticamente

Figura - Autocorrelação para a variável C12 (entradas)

Acima, apresenta-se o gráfico da autocorrelação da variável C12 e o resultado, como seria de esperar, verifica uma elevada correlação com o período de 24 horas antes.

# Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU), uma versão mais recente das Redes Neurais Recorrentes (RNN), que foi criada pelo professor [Kyunghyun Cho](https://arxiv.org/search/cs?searchtype=author&query=Cho%2C+K) e os seus colegas na universidade de Nova York em 2014 e que tem vindo a ganhar muita popularidade.

## 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 LSTM usa porta adicional que armazena mais informação. Por outras palavras, o GRU é uma versão leve do LSTM que combina memória de longo e curto prazo no seu estado oculto.

## Mecanismo

Efetivamente, o Gru compõe duas portas, a porta de atualização e a porta de reposição. A porta de atualização lembra o quanto da memória passada deve ser “retida”, enquanto a porta de reposição sabe o quanto da memória passada deve ser “esquecida”.

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

**//Inserir nova equação (word) para poder fazer referência**

Onde ***ht***​ é o estado oculto no tempo t, ***xt***​ é o input no tempo t, **Whz​h(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.

## Implementação

*import numpy as np*

***//Função de Ativação do GRU***

*def sigmoid(x):*

*"""Sigmoid activation function."""*

*return 1 / (1 + np.exp(-x))*

***//Inicializa pesos e parâmetros de aprendizagem***

*class GRU:*

*\_\_init\_\_(self, input\_dim, hidden\_dim):*

*self.input\_dim = input\_dim*

*self.hidden\_dim = hidden\_dim*

*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))*

*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))*

*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))*

***//Recebe sequência de entrada de vetores e retorna uma sequência de estados ocultos***

*def forward(self, X):*

*T = X.shape[0]*

*H = np.zeros((T + 1, self.hidden\_dim))*

*Z = np.zeros((T, self.hidden\_dim))*

*R = np.zeros((T, self.hidden\_dim))*

*H[0] = np.zeros((1, self.hidden\_dim))*

*for t in range(T):*

*x\_t = X[t]*

*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)*

*H[t+1] = (1 - z\_t) \* H[t] + z\_t \* h\_tilde\_t*

*Z[t] = z\_t*

*R[t] = r\_t*

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

Uma imagem com captura de ecrã, diagrama, file, Retângulo

Descrição gerada automaticamente

Figura - Flow Gates Modelo Gru

# Conceitos de *Machine Learning*

## 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 entradas, 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.

## Arquitetura do modelo

A arquitetura do modelo GRU (Gated Recurrent Unit), utilizada neste projeto, é uma variante das redes neuronais recorrentes (RNNs) 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.

## Treinar o modelo

**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 de otimização e na função de perda definida. 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.

***Lag***: Neste contexto, o *lag* refere-se ao número de passos de tempo anteriores utilizados como entrada para prever o valor futuro. Por exemplo, se *lag* = 24, o modelo usa os dados dos 24 passos de tempo anteriores como entrada para prever o valor do passo de tempo seguinte.

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

**Dados de treino e teste**: As variáveis *train* e *test* são caminhos que apontam 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.

***Loss Function***: Através da função de perda de erro quadrático médio (MSE) é medido a diferença média ao quadrado entre os valores previstos e reais, servindo como uma medida do desempenho do modelo.

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

***Save Model***: O modelo treinado é salvo 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.

***Learning Rate***: Determina o tamanho da etapa em que os parâmetros de um modelo são ajustados 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 e podem ser usadas para avaliar o desempenho do modelo.

# Testes de Previsão e Resultados

Nesta secção, através da análise do gráfico de autocorrelação da variável C12, representante dos inputs do conjunto de dados, foi determinado o *lag* referência (24) para começar a trabalhar o modelo.

Uma vez identificado o *lag*, foi executado o modelo para cada eixo, a fim de determinar qual dos eixos apresentava o menor *mape* para prever a próxima hora.

Para iniciar os testes, foi usado o código base obtido no repositório GitHub [9]. Este código serviu como ponto de partida para começar a fazer os testes. Os parâmetros iniciais que utilizados para *batch* e o número de *epochs* foram definidos como 256 e 600, respetivamente.

Além disso, é importante notar que, durante a fase de treino e teste, misturámos os conjuntos de dados para todos os eixos. No entanto, reconhecemos que esta abordagem resultou no facto de o gráfico apresentado não fornecer dados suficientes e coerentes para responder eficazmente às questões do problema.

## Testes para as entradas de cada eixo

## 8.1.1. Lag 24 vs Lag 12

Efetivamente, a estratégia foi mudada para tratar cada eixo individualmente no conjunto de dados de treino e de teste. Esta alteração serviu para melhorar a precisão e a fiabilidade do modelo. Além disso, para aumentar a robustez do modelo, foram incorporadas uma camada adicional *GRU* e uma camada de *dropout*. O tamanho do *batch* foi ajustado para 64 e foi introduzida uma loss function para mitigar o problema de *overfitting*, particularmente em relação às 600 épocas de treino projetadas. Estas modificações resultaram em melhorias notáveis no desempenho do modelo.

Uma imagem com texto, captura de ecrã, Tipo de letra, file

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, Tipo de letra, file

Descrição gerada automaticamente

Figura 17 - Mape para os Eixos da Cidade de Lisboa | Lag: 12

Figura 16 - Mape para os Eixos da Cidade de Lisboa | Lag: 24

De facto, foi observada uma diferença de mape na média de 0,86% entre as previsões à frente de *mape* com um *lag* igual a 24 e um *lag* igual a 12. Esta constatação sugere que a escolha do *lag* tem um impacto significativo na precisão das previsões do modelo. Além disso, também se observou que *mape* aumenta gradualmente à medida que o tamanho da previsão também aumenta. Esta observação está de acordo com o comportamento esperado do modelo e enfatiza ainda mais a importância de selecionar um tamanho de previsão adequado para ter mais precisão.

## 8.1.2. Lag 24x7 (1 semana) para prever 24h à frente

## Uma imagem com texto, captura de ecrã, diagrama, file Descrição gerada automaticamente

Figura 18 - Mape para os Eixos da Cidade de Lisboa | Lag: 1 semana

A experimentação de 1 semana de *inputs* para trás para prever 1 dia à frente 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.

## Uma imagem com texto, captura de ecrã, Tipo de letra, número Descrição gerada automaticamente8.1.3. Testes para Feriados e Fins de Semana

Figura 19 - Mape para os Eixos da Cidade de Lisboa nos feriados e fins de semana | Lag: 24

É interessente observar a imprevisibilidade do tráfego nas pontes durante os fins-de-semana e feriados. De facto, a análise dos padrões de tráfego das pontes revela uma correlação entre fins-de-semana e feriados e a existência de congestionamento. Por outro lado, a disponibilidade de tempo para lazer, ocorrência de vários eventos e outras atividades surgem como fatores-chave que contribuem para este fenómeno. De modo que, fazer melhor uso de *feature engineering* seria possível obter melhores resultados.

## Testes para a soma das entradas de cada eixo

## 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 o tamanho da previsão aumenta, o modelo encontra maiores dificuldades em captar e prever com precisão os picos dos dados. De acordo com a tabela de *mape*, faz sentido que à medida que a previsão aumente, pois o *mape* também aumenta. Esse resultado é coerente com as expectativas, pois quanto maior for o tamanho de previsão à frente, maior a incerteza e a dificuldade de obter estimativas precisas.

Uma imagem com texto, diagrama, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 20 - Gráficos soma das entradas teste vs. modelo para as janelas de previsão de 1, 2, 4, 8 horas com lag 24

Tabela - Mape 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 |

Uma imagem com texto, diagrama, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura 21 - Gráficos soma das entradas teste vs. modelo para as janelas de previsão de 1, 2, 4, 8 horas com lag 12

Tabela - Mape 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 |

## Uma imagem com texto, file, Gráfico, diagrama Descrição gerada automaticamente8.2.2. Lag 24x7 (1 semana) para prever 24h à frente

Figura 22 - Gráficos soma das entradas teste vs. modelo para prever 24 à frente com 1 semana de lag

Apesar do gráfico apresentar um *mape* 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.

## Testes para Feriados e Fins de Semana

Uma imagem com texto, diagrama, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura 23 - 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 - 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 |

Para demonstrar as diferenças de cada janela de previsão, foi usado o intervalo de tempo que inclui os 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 tracejado verde na vertical.

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 a janela de previsão aumenta, como demonstra os gráficos.

## Requisitos

…

## Arquitetura

…

## Tecnologias e ferramentas usadas

… que se encontram descritas na Tabela 8.7.

Tabela 8.10 – Tecnologias e ferramentas usadas

|  |  |  |
| --- | --- | --- |
| **Nome** | **Descrição** | **Utilização no projeto** |
| Picasso [2] | Biblioteca *open source* que permite fazer *download* e *caching* de imagens de forma simples. | Utilizada para fazer o *download* das imagens de perfil do utilizador. |
|  |  |  |

## Implementação

…

## Testes e validação

…

### Método A

…

### Método B

…

# Mais um capítulo

…

# Conclusões e trabalho futuro

...

## Conclusões

...

## Trabalho Futuro

…

# Referências [3]

|  |  |
| --- | --- |
| [1] | “ISEC - Licenciaturas - Engenharia Informática,” [Online]. Available: https://www.isec.pt/PT/estudar/licenciaturas/EngInfor/. [Acedido em 8 4 2021]. |
| [2] | “ISEC - Instituto Superior de Engenharia de Coimbra | Instituto Politécnico de Coimbra,” [Online]. Available: https://www.isec.pt/PT/Default.aspx. [Acedido em 8 4 2021]. |
| [3] | C. Y. Wijaya, “CRISP-DM Methodology For Your First Data Science Project,” 26 Abril 2021. |
| [4] | R. Ż. Teresa Pamuła, “Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning,” 4 Novembro 2022. |
| [5] | B. M. Eddy Sanchezdela Cruz, “Urban Traffic Flow Identification by Comparing Machine Learning Algorithms,” p. 7, 7 Fevereiro 2023. |
| [6] | A. B. G. S. J. C.-W. L. Youcef Djenouri, “Hybrid graph convolution neural network and branch-and-bound optimization for traffic flow forecasting,” 22 Setembro 2022. |
| [7] | W. L. Fazel Haq Ahmadzai, “A mobile traffic load prediction based on recurrent neural network: A case of telecommunication in Afghanistan,” 28 Maio 2022. |
| [8] | E. J. K. P. L. R. B. Juan Martín, “Traffic Monitoring via Mobile Device Location,” 17 Outubro 2019. |
| [9] | L. "xiaochus", *https://github.com/xiaochus/TrafficFlowPrediction.* |
| [10] | M. G. Pytorch. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.GRU.html. |
| [11] | I. d. m. GRU. [Online]. Available: https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-4/. |
| [12] | I. G. f. gates. [Online]. Available: https://www.youtube.com/watch?v=8HyCNIVRbSU&ab\_channel=TheA.I.Hacker-MichaelPhi. |

# Anexos

Proposta de Estágio

… conteúdo Anexo A

… conteúdo Anexo A

… conteúdo Anexo A

Especificação de Requisitos

… conteúdo Anexo B

… conteúdo Anexo B

… conteúdo Anexo B