Dados e Aprendizagem Automática

Trabalho Prático

Ano Letivo 2022/2023

Grupo 13

Cristiano Pereira, PG50304
João Martins, PG50463
Jorge Lima, PG50506
Rúben Santos, PG50733

Mestrado em Engenharia Informática
Universidade do Minho, Braga, Portugal
11 de novembro de 2022

Resumo

Serve o presente documento de guia para a compreensão das decisões tomadas ao longo do desenvolvimento do trabalho prático da Unidade Curricular de Dados e Aprendizagem Automática. Nesta secção iremos descrever os traços gerais deste documento, nomeadamente o seu conteúdo.

Decidimos iniciar este relatório com a secção (1) Introdução. Nesta secção iremos contextualizar o problema apresentado no enunciado, descrever o objetivo do mesmo e ainda descrever a metodologia de trabalho utilizada no desenvolvimento do trabalho prático.

A secção (2) Dataset 1: Paris Airbnb Reviews acompanha o processo de análise e preparação dos dados do dataset escolhido pelo grupo. Nesta secção começamos por apresentar o resultado do nosso estudo do negócio, seguida de uma análise estatística dos dados. É dedicada uma secção para descrever todas as alterações que fizemos e terminamos com uma comparação de várias técnicas de aprendizagem.

À semelhança da secção anterior, a secção (3) Dataset 2: Incidents segue a mesma metodologia, sendo que, ao contrário do anterior, este dataset foi disponibilizado pela equipa docente.

Por fim, incluímos a secção (4) Conclusão, onde expomos as dificuldades encontradas ao longo do desenvolvimento do projeto assim como os resultados de aprendizagem obtidos.

Conteúdo

1 Introdução

Nesta secção iremos contextualizar e expor os objetivos propostos pelo enunciado. Irá ser apresentada a metodologia de desenvolvimento do projeto.

1.1 Contextualização

O projeto da Unidade Curricular de Dados e Aprendizagem Automática surge numa fase bastante relevante do nosso percurso académico. Fase esta ávida de mentes criativas e com habilidade de, metodicamente, delinear e cumprir etapas de um processo de estudo e análise de qualquer disciplina. Este projeto é, portanto, fundamental para consolidar as nossas habilidades de gestão de projetos, nomeadamente projetos de Machine Learning, seguindo a metodologia apresentada na secção 1.3

1.2 Objetivos

Este projeto tem como objetivo principal desenvolver as nossas capacidades de gestão de um projeto de Aprendizagem Automática. Para tal é necessário adotar uma metodologia, estudar o negócio alvo, estudar o dataset e manipular os dados de forma a gerar conhecimento. É também necessária a apresentação de um modelo ótimo que resolva o problema, acompanhado por uma breve comparação entre os restantes modelos subótimos.

1.3 Metodologia

Para organizar este projeto foi seguida uma metodologia CRISP-DM ou, melhor ainda, uma *simplified* CRISP-DM. Simplificada pois algumas das etapas desta metodologia não se aplicam ao nosso contexto académico enquanto que outras requerem uma abordagem mais focada na obtenção/criação do dataset, o que também não se aplica nesta situação, dado que este é descarregado da internet (ou fornecido pela equipa docente).

Desta forma, para ambos os datasets, iniciamos com uma secção de estudo do negócio, dividida nos objetivos e nos critérios de sucesso. De seguida, na secção estudo dos dados, começamos por referir a proveniência do dataset, descrevemos os seus atributos e, principalmente, exploramos as tendências dos dados oferecendo, sempre que possível, conclusões acerca do que observamos. Terminamos esta secção com uma análise da qualidade dos dados. Na próxima secção, preparação dos dados, selecionamos que atributos devem prosseguir para o modelo e explicamos todas as manipulações que fazemos ao dataset. Na secção de modelação apresentamos os modelos que melhor se adequam ao problema, assim como a técnica de particionamento utilizada e uma breve avaliação dos resultados obtidos. Por fim, na secção de avaliação de resultados, analisamos se os critérios de sucesso definidos inicialmente foram cumpridos.

2 Dataset 1: Paris Airbnb Reviews

2.1 Estudo do Domínio

Alugar uma casa é uma decisão difícil e requer uma quantidade considerável de reflexão e pesquisa. Os critérios a ter em conta no que diz respeito à avaliação do preço de uma casa são bastante variados, p.ex, pode depender da sua estrutura física como o número de quartos ou casas de banho, da sua idade ou até da região em que está inserida. Dito isto, é difícil avaliar o preço de uma casa, sendo que muitas vezes o *preço certo* é decidido pelas forças do mercado livre.

Airbnb é um serviço online comunitário para as pessoas anunciarem, descobrirem e reservarem acomodações e meios de hospedagem, com mais de seis milhões de listagens na plataforma, 150 milhões de usuários e com mais de 1 bilião de reservas até a data. É estimado que esta plataforma detém 20% da indústria de aluguer de casas de férias, logo podemos considerar este estudo, um estudo sobre a indústria em si.

2.1.1 Objetivos

O objetivo principal é desenvolver um modelo de aprendizagem automática capaz de oferecer uma estimativa razoável do preço do aluguer de propriedades imobiliárias, com base nos dados fornecidos pelo dataset escolhido, apresentados na secção (2.2) Estudo dos Dados.

2.1.2 Critérios de Sucesso

Para as questões apresentadas acima, obtemos os resultados na secção (2.2.3) Exploração dos Dados. Os resultados não dependem do modelo desenvolvido, e contribuem apenas para uma melhor compreensão do negócio.

Pelo contrário, o objetivo principal (prever o preço das casas), depende do modelo desenvolvido. Como tal, decidimos que, inicialmente, um bom critério de sucesso para validar o objetivo principal será obter previsões com menos de 15% de erro. Este critério pode ser alterado ao longo do desenvolvimento.

2.2 Estudo dos Dados

As fases seguintes de desenvolvimento, presentes nesta secção, servem para aprofundar ainda mais o entendimento do *dataset*. Partimos de uma breve descrição dos dados (2.2.2) para uma análise mais profunda (2.2.3). Analisamos as relações entre os diferentes atributos e ficamos aptos para tomar decisões acerca de como preparar os dados, na secção (2.3) Preparação dos Dados. Terminamos com uma análise da qualidade dos dados do *dataset* (2.2.4).

2.2.1 Coleção dos Dados

O *dataset* adquirido pelo grupo foi escolhido do repositório de datasets *inside airbnb*. Este pode ser consultado seguindo o seguinte http://insideairbnb.com/.

2.2.2 Descrição dos Dados

O *dataset* é constituído pelo seguinte conjunto de atributos:

1. *id*: Identificador do anúncio de aluguer.
2. *name*: Nome do anúncio.
3. *host_id*: Identificador do host da propriedade.
4. *host_name*: Nome do host da propriedade.
5. *neighbourhood*: Bairro no qual está localizada a propriedade.
6. *neighbourhood_group*: Grupo no qual o bairro se insere.
7. *latitude*: Latitude da propriedade.
8. *longitude*: Longitude da propriedade.
9. *room_type*: Tipo de quarto disponível.
10. *price*: Preço do aluguer, por noite.
11. *minimum_nights*: Número mínimo de noites a alugar.
12. *number_of_reviews*: Número de reviews da propriedade.
13. *last_review*: Data da última review.
14. *reviews_per_month*: Número de reviews por mês, em média.
15. *calculated_host_listings_count*: Número de propriedades que o host tem.
16. *availability_365*: Número de dias disponíveis por ano.
17. *number_of_reviews_ltm*: Número de reviews nos últimos doze meses.
18. *license*: Licença da propriedade.htbp

Este *dataset* é constituído por 61365 entradas (linhas), cada uma representando uma listagem de uma habitação.

2.2.3 Exploração dos Dados

Iremos, ao longo desta secção, explorar os dados, nomeadamente através da distribuição das variáveis e dos padrões e relacionamentos que identificamos como pertinentes. Podemos na figura 1 observar a distribuição enviesada dos dados. No dataset original a assimetria (*skewness*) é de 19,41, indicando uma "longa cauda no lado direito". Na figura estão presentes três gráficos, o da esquerda representa o dataset original, o do meio limita o preço a 2000€ por noite e o da direita limita o preço a 1000€.

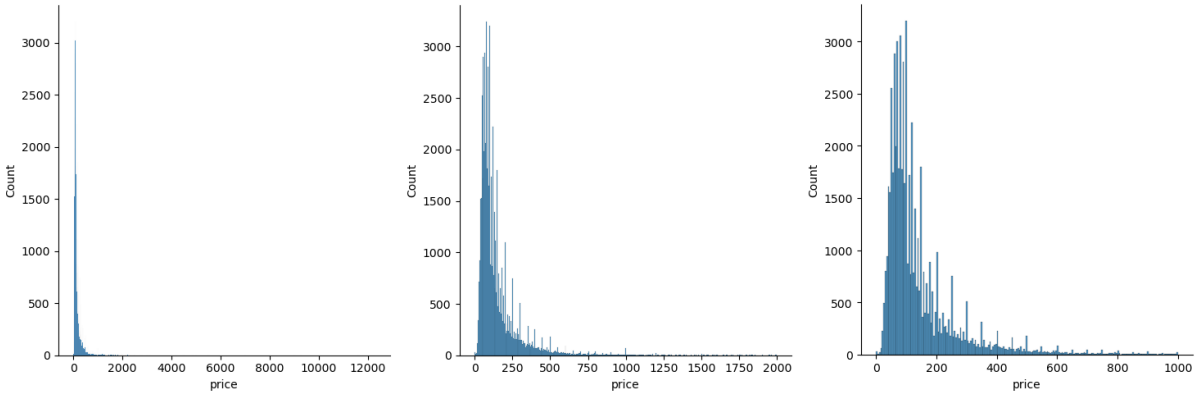

Figura 1: Distribuição do preço com diferentes limites de preço máximo (sem limite, 2000€, 1000€).

Podemos desde já perceber que a maior parte dos alugueres tem preços entre os 75 e 200 euros. Comprovamos esta observação com a figura abaixo, que contém o box plot do dataset original (esquerda) e o box plot para preços inferiores a 1000 euros.

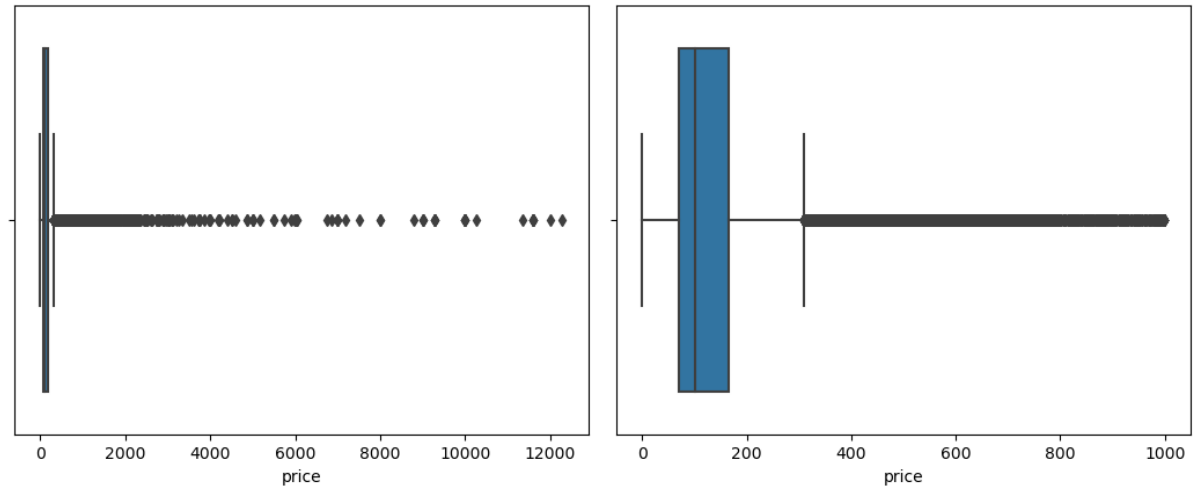

Figura 2: Box Plot do preço com diferentes limites de preço máximo (sem limite e 1000€).

Com ambas as figuras anteriores conseguimos também perceber que o dataset tem uma quantidade elevada de outliers, que terão de ser tratados na fase de preparação de dados (secção 2.3).

A coluna *neighbourhood*, que indica o bairro de Paris onde a propriedade está localizada, temos uma distribuição de valores mais aceitável. Temos zonas mais populadas do que outras mas, no geral, os diferentes valores do dataset estão bem representados e podemos assumir a sua validade.

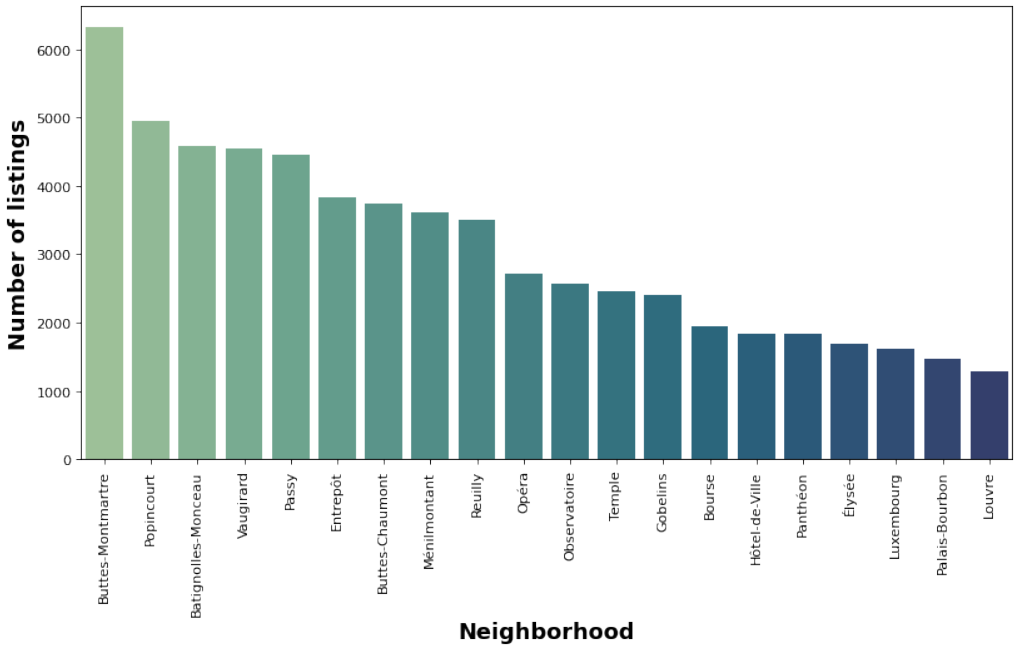

Figura 3: Distribuição da coluna *Neighbourhoods*.

Como podemos ver na figura 3, os três bairros mais representados no dataset são o *Buttes-Montmartre*, o *Popincourt* e o *Batignolles-Monceau*, com 6317, 4947 e 4590 propriedades, respetivamente.

Podemos observar na figura 4 a distribuição do preço de cada aluger, por bairro. Os bairros estão ordenados por ordem decrescente de número propriedades (da esquerda para a direita).

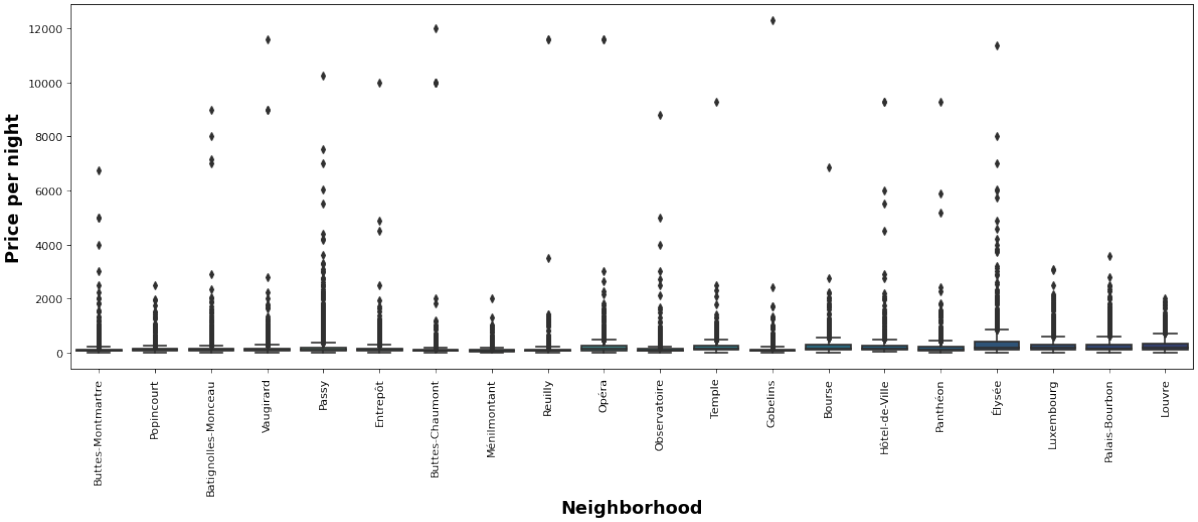

Figura 4: Preço do aluguer, por *Neighbourhoods*.

Os bairro mais caro, de acordo com os dados, é o *Élysée* onde, em média, o preço de um aluguer ronda os 360€ por noite. De seguida vem o *Louvre*, com uma média de 277€ por noite.

No que diz respeito à coluna *Room Type*, existem 4 tipos de quartos no dataset:

- *Entire home/apt*
- *Private room*
- *Shared room*
- *Hotel room*

A distribuição destes valores pode ser consultada no gráfico abaixo.

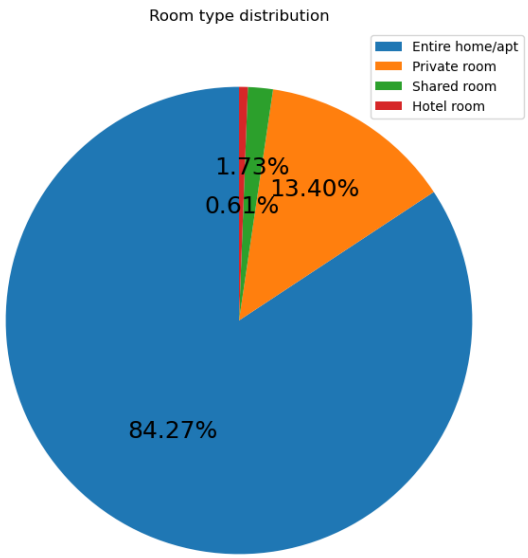

Figura 5: Distribuição da coluna *Room Type*.

A grande maioria dos anúncios são, de facto, de casas inteiras ou apartamentos (84%). Existem, no dataset, cerca de 0.6% (374) hotéis. Destes, cerca de 30% (125) estão no bairro *Élysée* e outros 30% (121) estão no bairro *Opéra*. Ou seja, os bairros mais caros tendem também a ter mais quartos de hotel. Fizemos o mesmo com os quartos partilhados e verificamos uma presença significativa em bairros mais baratos.

Dadas as colunas *availability_365* (número de dias que a propriedade está disponível por ano) e o preço, é possível estimar um valor esperado de receita, para cada *host*. Acontece que, nesta nossa análise, descobrimos que cerca de 40% das propriedades estão indisponíveis e como tal não estão a gerar rendimentos. Não é claro o porquê deste valor ser tão alto mas pode dever-se ao simples facto de os proprietários não estejam, de momento, interessados a alugar.

Os proprietários são também parte importante do conjunto de dados. Através da nossa análise do número de *reviews*, conseguimos detetar casos em que uma determinada propriedade tem 50 vezes mais *reviews* do que a média para aquele bairro. Cada *host* tem em média 31.14 *reviews* e 14.4 anúncios na plataforma. O segundo número é particularmente revelador e indica que a maior parte dos proprietários tem múltiplas propriedades, talvez até dependendo diretamente dos rendimentos gerados pela plataforma. De facto, os dez *hosts* com mais *reviews* têm, em média, 100x mais *reviews* do que a média. Em

relação à disponibilidade anual, apenas 6% das propriedades dos *hosts* mais populares estão indisponíveis durante o ano, face à média de 52% de disponibilidade. Isto pode dever-se ao facto de que os *hosts* mais populares tem mais tempo para gerir as suas propriedades, dado que se dedicam a tempo inteiro ao negócio em questão. Como seria de esperar, estes *hosts* aguardam receitas na casa dos 4.78M de euros, face à média de 31k euros.

Na figura 6 podemos visualizar os valores em falta no dataset. A coluna *neighbourhood_group* não tem valores, as colunas *last_review* e *reviews_per_month* tem 19,35% dos valores em falta e, por fim, a coluna *license* tem cerca de 40% de valores em falta.

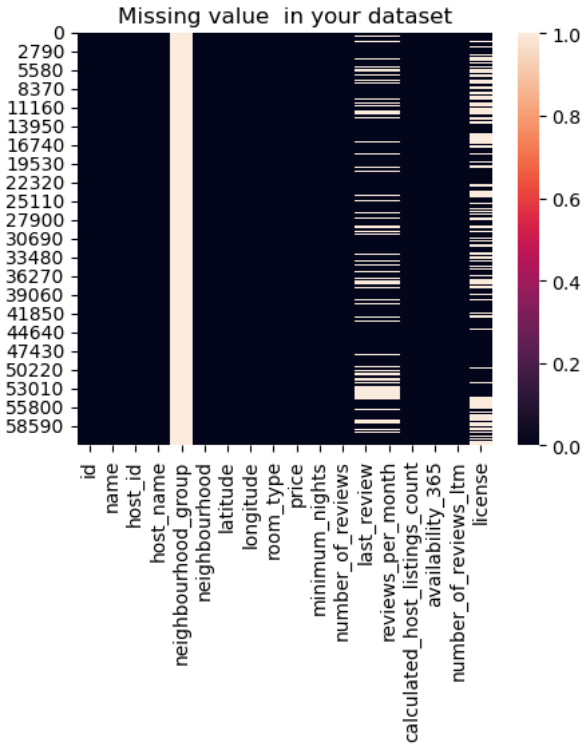

Figura 6: Valores em falta do dataset.

No *notebook* que acompanha o relatório, são ainda apresentados mais informações e observações sobre o conjunto de dados.

2.2.4 Qualidade dos Dados

O objetivo desta seção é avaliar a qualidade dos dados explorados previamente e identificar quaisquer problemas ou possíveis problemas que possam afetar a precisão ou a confiabilidade do modelo. De seguida, serão descritas as características dos dados, incluindo sua integridade, consistência e precisão.

A integridade dos dados refere-se à proporção de valores em falta no conjunto de dados. Neste projeto, descobrimos que o *dataset* contém valores em falta, tal como demonstrado na figura 6. A consistência dos dados refere-se ao grau em que os dados são uniformes

e estão em conformidade com um formato específico. Neste projeto, descobrimos que os dados são consistentes e estão de acordo com o formato esperado. Não encontramos nenhum problema com dados inconsistentes. O mesmo em relação à precisão, que refere o grau em que os dados estão corretos e livres de erros. No entanto, existe um elevado número de *outliers*, nomeadamente da classe preço. Esta classe também apresenta uma distribuição muito enviesada, como é possível ver na figura 1.

Concluindo, os dados do projeto são de qualidade mas requerem um tratamento cuidadoso de forma a resolver quaiquer problemas que afetem o desempenho dos modelos. Dito isto, segue-se a descrição da fase de processamento e modelação.

2.3 Preparação dos Dados

A fase de preparação dos dados tem como objetivo principal produzir um *dataset* que será depois utilizado na fase da modelação, presente na secção (2.4) Modelação. Resumidamente, nesta secção iremos justificar o porquê de tomarmos determinadas decisões relativas ao formato do *dataset*.

2.3.1 Selecionar Dados

Como consequência direta da exploração de dados, foram retiradas de imediato as seguintes colunas:

- *neighbourhood_group*: Todos os valores estão em falta.
- *id*: Todos os valores são diferentes.
- *name*: O nome do anúncio não foi processado.
- *host_name*: Diretamente relacionado com a coluna *host_id*.
- *latitude* & *longitude*: Dado que propriedades são todas em paris, estes valores são todos bastante semelhantes.

Como consequência do processo iterativo de preparação e modelação, as seguintes colunas não contribuíram para o conhecimento do modelo de forma significativa e a sua remoção, para além de melhorar a generalização do modelo, melhoraram o seu desempenho.

- *host_id*: Identificador do proprietário do anúncio.
- *last_review*: Data da última *review*.

Desta forma, seguimos o processamento dos dados com 10 das colunas originais enunciadas na secção 2.2.2.

2.3.2 Tratamento dos Dados

Esta secção visa fornecer uma visão geral detalhada das etapas realizadas para preparar os dados para a modelação. Iremos abordar o tratamento de valores em falta e outliers, assim como a transformação de algumas colunas.

Neste *dataset* foram detetados valores em falta, nomeadamente nas colunas *reviews_per_month* e *license*, como indica na figura 6. No caso do número de *reviews* por mês, temos cerca de 20% de valores em falta. Dado o nosso entendimento do domínio, fruto do estudo dos dados, iremos substituir estes valores por zero, dado que, muito provavelmente, dizem respeito a propriedades que não tem avaliações. Esta decisão é fortificada pela não existência de zeros nesta coluna. O caso da coluna *license*, com quase 40% dos valores em falta, pode ser também entendido como propriedades que não necessitam de nenhuma licença para estarem a ser alugadas. Desta forma, a decisão mais correta neste contexto foi transformar a coluna *license* em valores binários, sendo que o valor 0 (zero) indica a inexistência de licensa e o valor 1 (um) indica a existência de uma licença.

Existem também colunas com valores categóricos nominais que necessitam de ser transformadas para valores numéricos nomeadamente, *neighbourhood* e *room_type*. Ambas as colunas têm um reduzido número de valores diferentes, sendo que foram mapeadas para valores inteiros e.g. "*Élysée*": 8, "*Hotel*": 4. Esta decisão é motivada pelo facto de que, muitos modelos, necessitarem de valores numéricos para funcionar (incluindo alguns que utilizamos na fase de modelação).

Em relação à forte presença de *outliers*, evidenciada pelos box plot apresentados na figura 2, foi tomada a decisão de excluir todas as linhas cujo módulo do Z-Score fosse superior a 5, ou seja, cujo preço esteja a uma distância superior ao preço médio mais (ou menos) cinco vezes o desvio padrão. Isto resultou numa perde de cerca de 4% do conjunto de dados. O valor limite (5) foi escolhido, face ao valor comummente escolhido (3), dado que os dados não estão normalmente distribuídos, como já previamente discutido.

Na figura abaixo, são apresentadas as correlações entre variáveis do *dataset*.

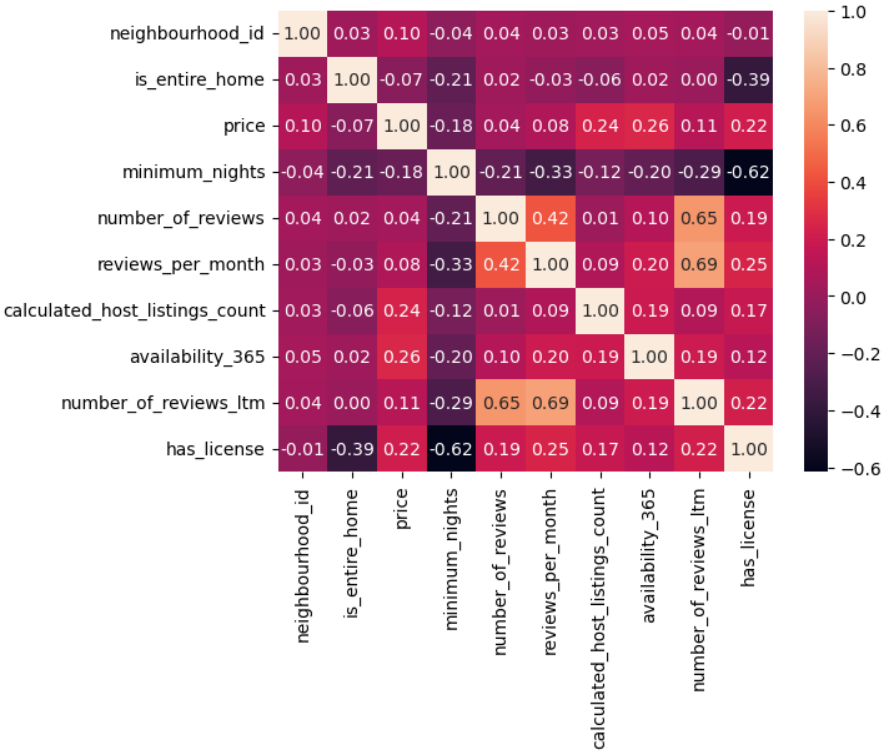

Figura 7: Correlação entre as colunas do dataset.

Reparamos que a coluna *number_of_reviews_ltm* tem uma forte correlação com ambas as colunas *number_of_reviews* e *reviews_per_month*. De forma a melhorar a generalização do modelo, decidimos remover a coluna *number_of_reviews_ltm*, sendo que o desempenho do mesmo não foi afetado.

2.4 Modelação

Nesta secção descrevemos o processo de seleção do modelo final, começando por selecionar uma técnica de modelação, seguido de descrever o método de particionamento de dados, os diferentes modelos a serem testados e, por fim, uma comparação dos resultados de cada modelo.

2.4.1 Técnica de Modelação e Particionamento dos Dados

Irão ser utilizadas vários modelos para esta fase de modelação, nomeadamente regressão linear múltipla e um conjunto de *ensembles*. Quanto aos *ensembles*, usamos dois modelos de *Boosting* (*AdaBoostRegressor* e *GradientBoostingRegressor*) e um modelo de *Bagging* (*RandomForestRegressor*). Os modelos de *Boosting* envolvem treinar múltiplas instâncias de um modelo base sobre os mesmos dados, mas alterando, em cada iteração, o peso de cada exemplo. Desta forma os erros vão sendo iterativamente corrigidos. Por outro lado, os modelos de *Bagging* requerem treinar múltiplas instâncias do modelo base sobre diferentes subconjuntos do conjunto de dados, amortizando desta forma os erros cometidos.

Foi também utilizada um modelo de *GridSearch* para testar diferentes configurações de hiperparâmetros, nomeadamente para o *GradientBoostingRegressor*, dado que foi o modelo que obteve melhores resultados no geral.

O *dataset* foi particionado com valores representativos de cada classe e com aproximadamente 10% dos exemplos dedicados para o teste do modelo.

2.4.2 Resultados e Modelo Final

Os resultados dos modelos podem ser consultados ao longo desta secção, graficamente. São apresentadas as seguintes métricas:

- R2 Score
- MAE
- RMSE

O R2-Score mede a proporção da variância entre a variável dependente e as independentes. Esta métrica oferece uma ideia do quão bem o modelo está a prever os testes, sendo que quanto mais próximo de 1, melhores são as previsões. Na figura conseguimos confirmar um R2-Score de 0.83 para o melhor modelo, o *GradientBoostingRegressor*.

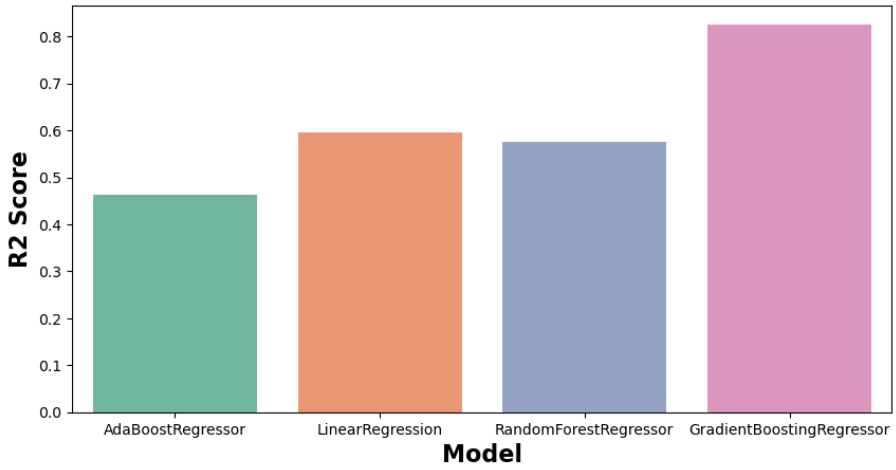

Figura 8: *R2 Score*.

De seguida podemos ver os erros MAE e RMSE, sendo que os modelos mantêm o mesmo esquema de cores e os gráficos estão na mesma escala.

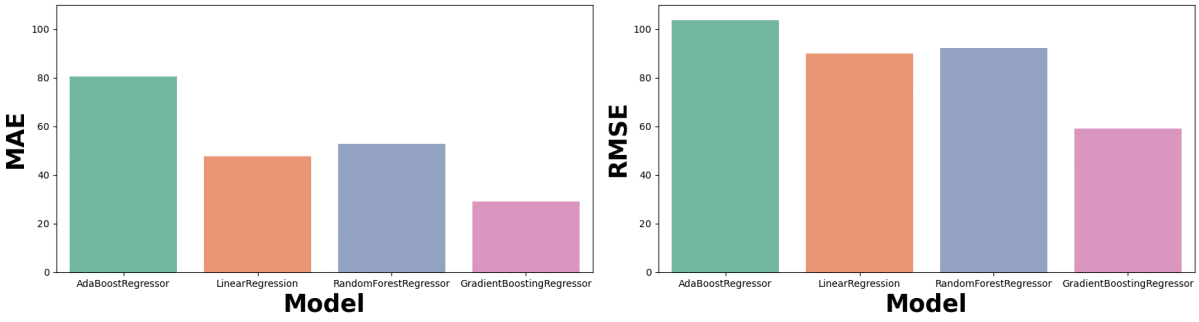

Figura 9: *Mean Absolute Error* (MAE) e *Root Mean Squared Error* (RMSE).

O modelo com melhor R2-Score mantêm também o menor erro, sendo que tem um MAE de 29,19€ e um RMSE de 59,12€. No que diz respeito ao valor médio do aluguer de uma propriedade, o MAE conta com 20.5% de erro e o RMSE com 41,6% de erro.

2.5 Avaliação de Resultados

Em suma, o objetivo inicialmente definido de obter um erro inferior a 15% não foi cumprido. Neste momento, o erro em relação ao preço médio, tendo em conta o MAE, ronda à volta dos 20%. Geralmente o RMSE é superior ao MAE dado que tem em conta a direção do erro (positivo ou negativo) e penaliza erros maiores. No entanto, isto quer dizer que ainda existem outliers nos dados e que estão estão a provocar este erro bastante superior ao MAE, que por sua vez é mais robusto a valores extremos. O modelo final é aquele que fornece melhores resultados: *GradientBoostingRegressor*. O grupo fica satisfeito com o resultado e o processo de tratamento e modelação do conjunto de dados.

3 Dataset 2: Incidents

3.1 Estudo do Domínio

Acidentes rodoviários são a principal causa de morte de crianças e jovens em todo o mundo, com 26% das mortes que ocorrem globalmente, têm origem em estradas. Sendo assim é importante estudar que os fatores mais influenciam estes acidentes e tentar diminuir a sua frequência. Analisar e estudar um dataset com informações sobre estas circunstâncias é uma boa maneira de prever situações de acidentes e que fatores levam a essas situações. Neste trabalho vamos estudar um dataset com acidentes rodoviários em Guimarães.

3.1.1 Objetivos

O objetivo principal é desenvolver um modelo de aprendizagem automática capaz de prever casos de acidente com a maior eficácia possível, com base nos dados fornecidos pelo dataset escolhido, apresentados na secção (3.2) Estudo dos Dados. No contexto da unidade curricular de Dados e Aprendizagem Automática vamos utilizar a plataforma kaggle para competir e comparar os modelos construidos, com o objetivo de obter o melhor modelo possível.

3.1.2 Critérios de sucesso

Para as questões de negócio apresentadas acima, um bom critério de sucesso para validar o objetivo principal será obter previsões com menos de 10% de erro.

No entanto este critério pode ser alterado ao longo do desenvolvimento, devido a comparação com os outros modelos feita na plataforma kaggle.

3.2 Estudo dos Dados

Esta secção tem como objetivo aprofundar o entendimento do conjunto de dados. Começamos por descrever o processo de obtenção do *dataset* (3.2.1), passando para uma breve descrição dos dados (3.2.2). Na secção (3.2.3) Exploração dos Dados, apresentamos uma análise estatística do conjunto de dados, apresentando a sua correlação e construindo bases para tomar decisões na secção (3.3) Preparação dos Dados. Terminamos com uma análise da qualidade dos dados do *dataset* (3.2.4).

3.2.1 Coleção dos Dados

Este *dataset* foi fornecido pela equipa docente e, como tal, não foi necessária nenhuma pesquisa para o obter.

3.2.2 Descrição dos Dados

O *dataset* é constituído pelo seguinte conjunto de atributos:

1. *city_name*: Nome da cidade em causa.
2. *record_date*: Timestamp associado ao registo.
3. *magnitude_of_delay*: Magnitude do atraso provocado pelos incidentes que se verificam no record_date correspondente.
4. *delay_in_seconds*: Atraso, em segundos, provocado pelos incidentes que se verificam no record_date correspondente.
5. *affected_roads*: Estradas afetadas pelos incidentes que se verificam no record_date correspondente.
6. *luminosity*: Nível de luminosidade.
7. *avg_temperature*: Valor médio da temperatura para o record_date.
8. *avg_atm_pressure*: Valor médio da pressão atmosférica para o record_date.
9. *avg_humidity*: Valor médio de humidade para o record_date.
10. *avg_wind_speed*: Valor médio da velocidade do vento para o record_date.
11. *avg_precipitation*: Valor médio de precipitação para o record_date.
12. *avg_rain*: Avaliação qualitativa do nível de precipitação para o record_date.
13. *incidents*: Indicação acerca do nível de incidentes rodoviários que se verificam no record_date.

Este *dataset* é constituído por 5000 entradas (linhas) para treino e 1206 entradas (linhas) para teste na plataforma kaggle, umas representando um acidente e outras não, sendo estas últimas o grupo de controlo do dataset.

3.2.3 Exploração dos Dados

Iremos, ao longa desta secção, explorar os dados, nomeadamente através da distribuição das variáveis e dos padrões e relacionamentos que identificamos como pertinentes.

Podemos na figura 10 observar a distribuição dos dados categóricos. Na figura estão presentes quatro gráficos. No primeiro conseguimos constatar que o atraso provocado pelos incidentes geralmente não é definido e em certos casos causa grandes atrasos, quase nunca acontecendo a situação de causar um atraso moderado. A distribuição da luminosidade contém as proporções esperadas para um dia normal.Também podemos verificar que grande parte das entradas ocorrem sem a influência de chuva. A coluna incidents, target do projeto, apresenta uma distribuição equilibrada, em que temos um bom conjunto de "entradas de controlo", em que não houveram acidentes. É de notar que a coluna *test-only* são os dados para teste que não contém a informação dos acidentes, como seria de esperar.

Em relação a estes dados no playbook fazemos mais análises, principalmente, comparando o valores destas colunas em relação à coluna target (incidentes).

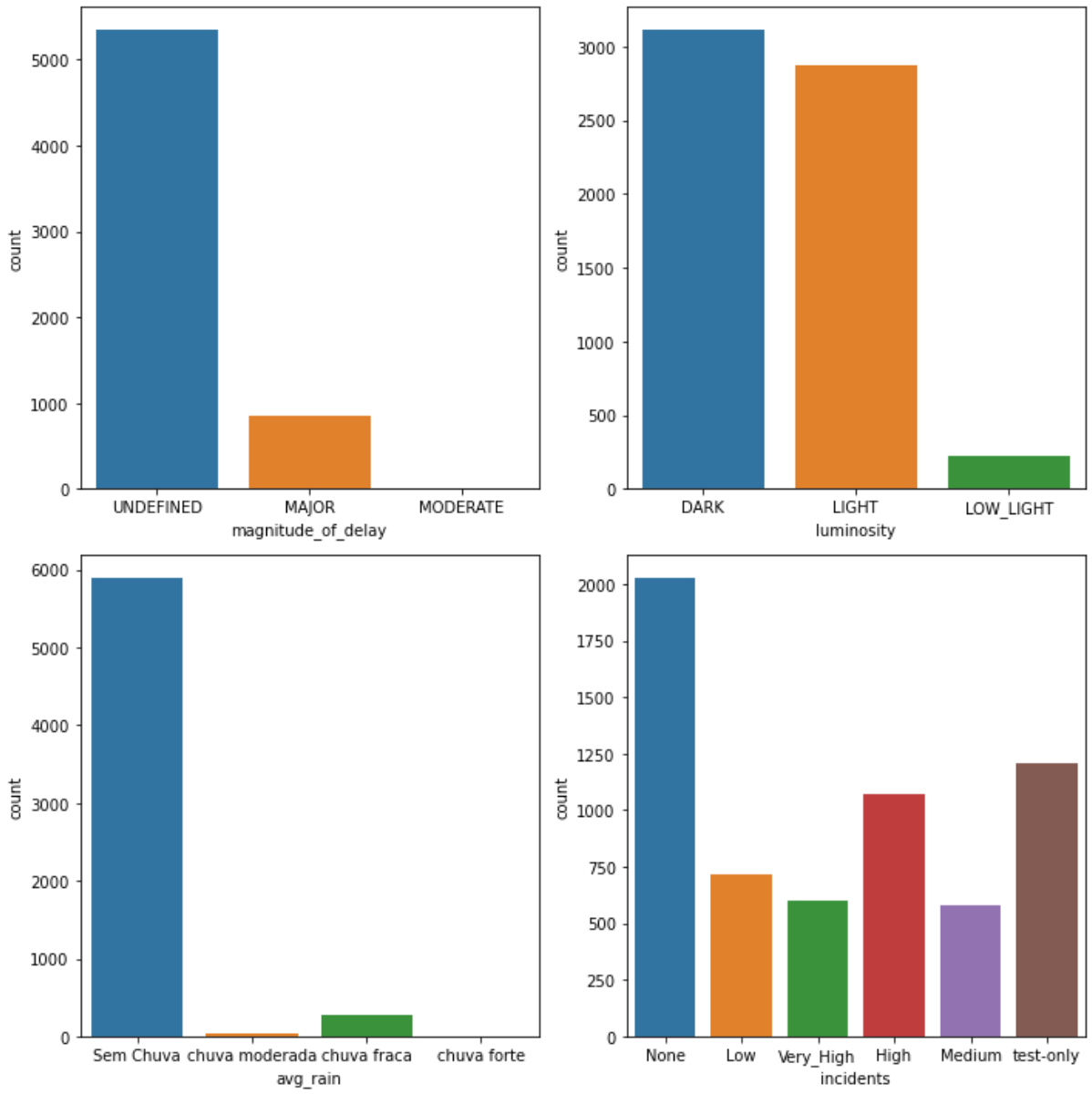

Figura 10: Distribuição dos dados categóricos

Com ambas as figuras seguintes conseguimos perceber que a coluna a representar a temperatura média é bem distribuida, no entanto, contém uma quantidade relevante de outliers, que poderemos ter de processar na fase de preparação de dados.

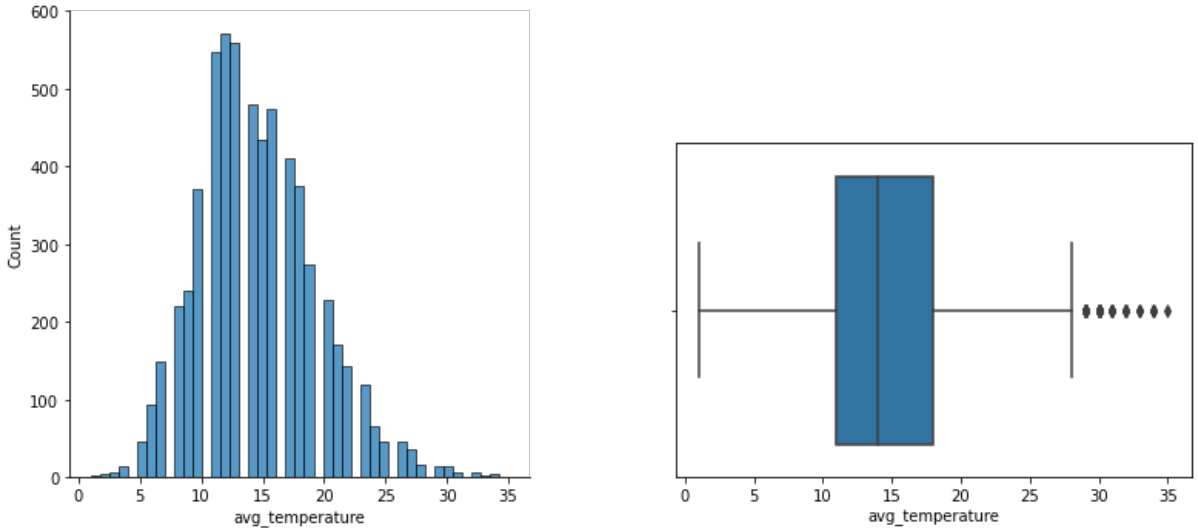

Figura 11: Distribuição e boxplot da temperatura média

Como podemos ver nas figuras a humidade média é geralmente bastante alta em grande parte do dataset. Após alguma pesquisa concluímos que é natural esta distribuição, pois todos estes casos se localizam em Guimarães, que se situa no norte de Portugal, cuja humidade relativa é geralmente alta.

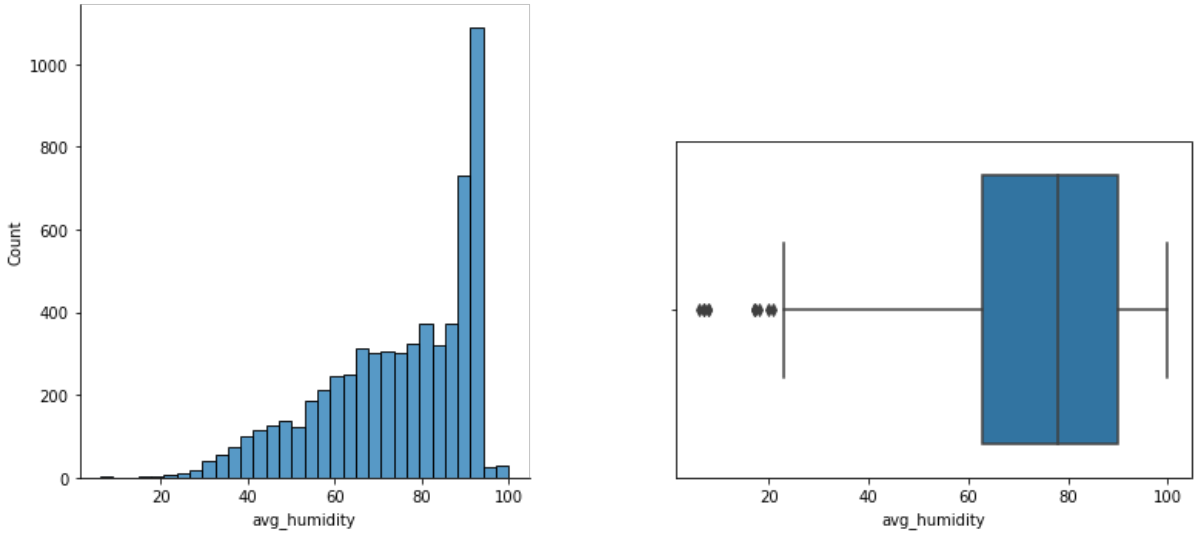

Figura 12: Distribuição e boxplot da humidade média

A coluna que representa a pressão atmosférica média tem uma distribuição de valores aceitável. A maioria das entradas está contida na região entre os 1000 e 1030 hPa. que são valores normais de pressão atmosférica para a região Norte de Portugal, estes valores são considerado valores de alta pressão, geralmente associado a tempo estável e bom o que corresponde ao observado na figura 10 na distribuição da chuva.

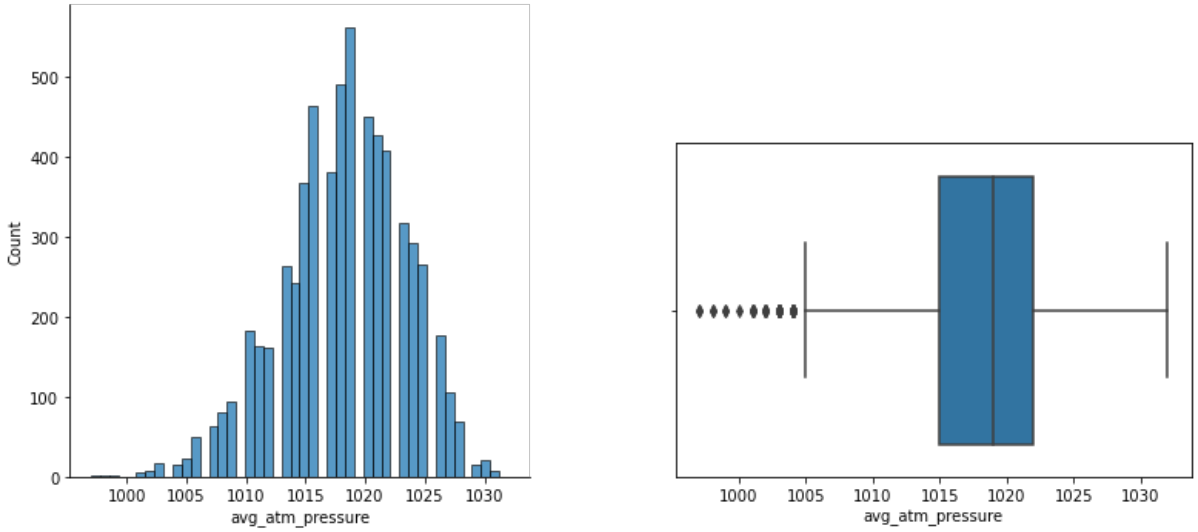

Figura 13: Distribuição e boxplot da pressão atmosférica média

Tal como nos gráficos anteriores os valores da velocidade média do vento correspondem as nossas expectativas, pois a velocidade média do vento na região Norte de Portugal é geralmente moderada, devido ao seu clima temperado.

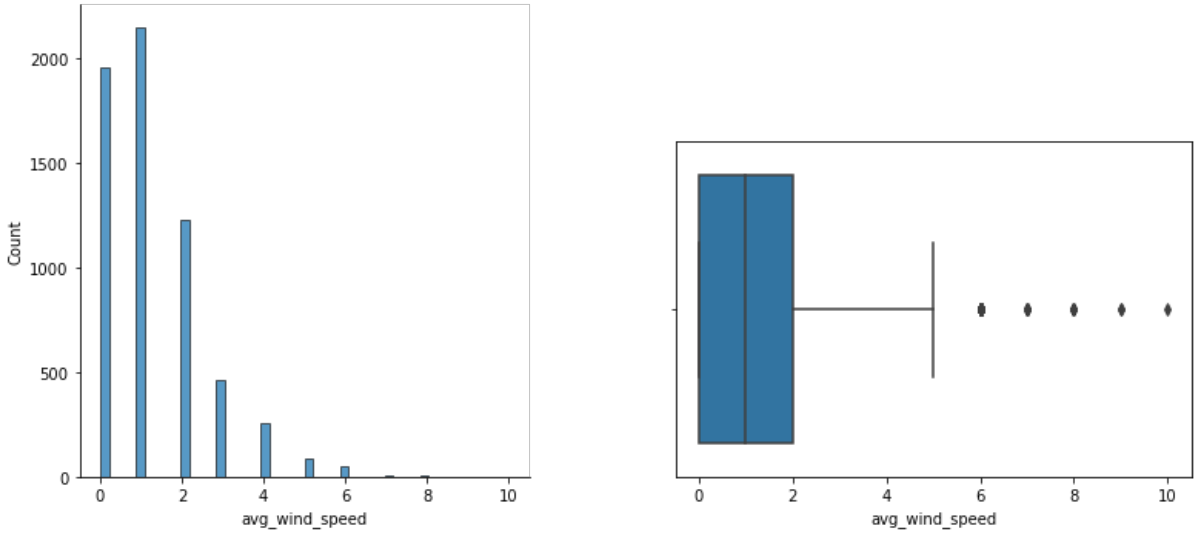

Figura 14: Distribuição e boxplot da velocidade do vento média

3.2.4 Qualidade dos Dados

O objetivo desta seção é avaliar a qualidade dos dados explorados previamente e identificar quaisquer problemas ou possíveis problemas que possam afetar a precisão ou a confiabilidade do modelo. De seguida, serão descritas as características dos dados, incluindo sua integridade, consistência e precisão.

A integridade dos dados refere-se à proporção de valores em falta no conjunto de dados, que, através de analises disponíveis no playbook, este dataset não apresenta.

A consistência dos dados refere-se ao grau em que os dados são uniformes e estão em conformidade com um formato específico, que em grande maioria é o que encontramos, exceto em alguns casos que vão ser tratados na secção 3.3.2. Quanto ao conjunto de outliers é preciso ter cuidado no contexto deste dataset, sabendo que um outlier pode corresponder a razão de causa do(s) acidente(s), ou seja, é importante ter atenção ao desempenho dos modelos com algum tratamento aplicado a estes.

Concluindo, os dados do projeto são de qualidade mas requerem um tratamento cuidadoso de forma a resolver quaisquer problemas que afetem o desempenho dos modelos. Dito isto, segue-se a descrição da fase de processamento e modelação.

3.3 Preparação dos Dados

A fase de preparação dos dados tem como objetivo principal produzir um *dataset* que será depois utilizado na fase da modelação, presente na secção (3.4) Modelação. Resumidamente, nesta secção iremos justificar o porquê de tomarmos determinadas decisões em relação ao formato do *dataset*.

3.3.1 Selecionar Dados

Maior parte das colunas no dataset são importantes e relevantes, exeto duas. Tanto as colunas *city_name* e *avg_precipitation* foram removidas divido a apeas conterem um único valor, tornando-as irrelevantes para o modelo. Também é de notar que não existem linhas duplicadas neste dataset e portanto nehuma entrada é retirada.

3.3.2 Feature Engineering

Em relação à criação de novas colunas, foi necessário fazer alterações a quatro colunas do dataset:

A coluna *affected_roads* contém uma lista de estradas afetadas numa dada entrada e como tal necessita de algum tratamento, dado que não é um valor númerico nem nominal. Entre algumas opções a que mostrou melhores resultados foi a substituição pela coluna *number_affected_roads* que contabiliza o número de diferentes estradas afetadas.

Da coluna *record_date* era possível retirar diferentes informações como hora do dia, altura do dia, ano, ... No entanto os melores resultados resultaram da criação de duas colunas: a coluna *month* e a coluna *day*.

Também encontramos uma correlação basante alta entre as colunas *magnitude_of_delay* e *delay_in_seconds*, constatando que sempre que havia uma magnitude *"Undefined"* havia um delay de 0 segundos, e isto ocorria em 75% dos casos. Além disso 94% dos valores de *delay_in_seconds* são inferiores a 10 minutos. Sendo assim mapeamos para valores inteiros a coluna *delay_in_seconds* e.g. "*Undefined*": 1, "*Moderate*": 2 e "*Major*": 3. Após isso, para tirar proveito da informação de ambas as colunas decidimos multiplicar os valores de ambas criando uma nova e.g. *delay_coefficient* = *delay_in_seconds* * *delay_in_seconds*.

3.3.3 Tratamento dos Dados

Nesta secção visa fornecer uma visão geral detalhada das etapas realizadas para preparar os dados para a modelação. O tratamento de valores em falta não existe dado que não existem valores em falta.

O tratameto de outliers foi experimentado mas não resultou em melhores resultados nos modelos, isto porque provavelmente representam um grande fator na circunstância que levou ao acidente.

No entanto ocorreu a transformação de algumas colunas:

Existem colunas com valores categóricos nominais que necessitam de ser transformadas para valores numéricos nomeadamente, *luminosity* e *avg_rain*. Estas colunas têm um reduzido número de valores diferentes, sendo que foram mapeadas para valores inteiros e.g. "*Dark*": 0, "*Light*": 1 e "*Low_Light*": 2. Esta decisão é motivada pelo facto de que, muitos modelos, necessitarem de valores numéricos para funcionar (incluindo alguns que utilizamos na fase de modelação).

Por outro lado as colunas numéricas contém valores em unidades com escalas muito dispersas. Com o objetivo de não influenciar os modelas a interpretarem valores altos como valores de alta importância decidimos aplicar uma padronização, alterando a escala dos valores para estarem no intervalo [-1,1], as seguintes colunas: *avg_atm_pressure*, *avg_wind_speed*, *avg_humidity*, *avg_temperature* e *delay_coefficient*.

Na figura abaixo, são apresentadas as correlações entre variáveis do *dataset*.

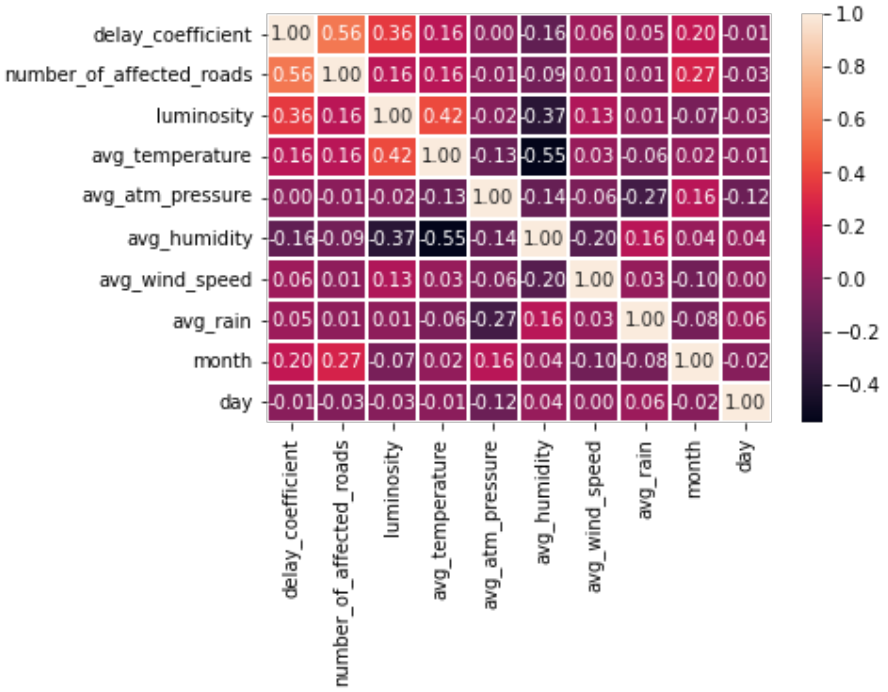

Figura 15: Correlação entre as colunas do dataset.

Como não encontramos grandes correlações entre as colunas decidimos aceitar estas alterações para a aplicação nos modelos.

3.4 Modelação

Neste subcapítulo vamos explicar o processo modelação, revelar a técnica de modelação, a maneira como particionamos os dados, seguido dos resultados e do modelo final e por fim a avaliação destes.

3.4.1 Método de Particionamento de Dados

Este dataset já está dividido em duas partes, treino e testes. No entanto o dataset de testes têm como objetivo ser usado na plataforma kaggle, e sendo esta limitada a 3 submissões diárias, decidimos dividir o dataset de treino para obter uma ideia dos melhores modelos e parâmetros, tirando máximo proveito das submissões na plataforma.

O *dataset* de treino foi particionado com 30% dos exemplos dedicados para o teste do modelo e os restantes 70% para treino.

3.4.2 Selecionar Técnica de Modelação

Inicialmente implementamos vários modelos de decisão e selecionamos aqueles que obtiveram resultados mais promissores para aperfeiçoamento dos parâmetros.

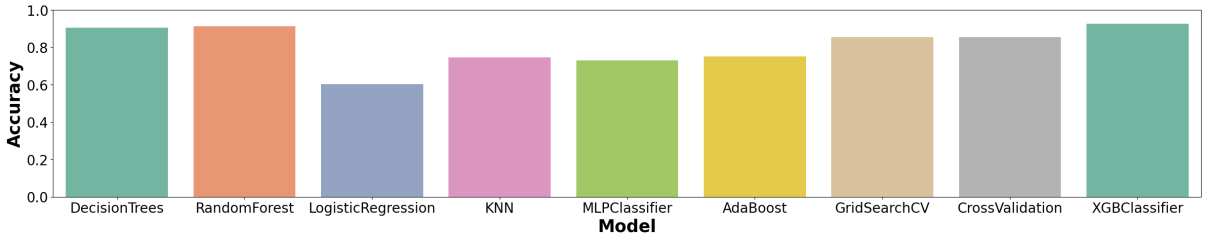

Figura 16: *Accuracy de vários modelos*

Além da analise da accuracy, no *notebook* encontramos gráficos com analise de métricas como o Recall, o F1-Score e a Precision. No entanto são sempre os mesmos 3 modelos que se destacam nestas avaliações, o *DecisionTreeClassifier*, *RandomForestClassifier* e o *XGBClassifier*. É de notar que no *notebook* também se encontra uma implementação de uma *Neural Network* que obteve resultados promissores, mas necessitava de bastante atenção e estudo para obter os parâmetros corretos.

3.4.3 Resultados e Modelo Final

Foi utilizado um modelo de *GridSearch* para testar diferentes configurações de hiperparâmetros, para os três modelos referidos anteriormente. No fim o modelo de seleção foi o

XGBClassifier obtendo melhores resultados no playbook e nas submissões da plataforma do kaggle, obtendo uma accuracy de 92.662%

3.5 Avaliação de Resultados

Em suma, o objetivo inicialmente definido de obter um erro inferior a 10% foi cumprido. No entanto os resultados de outros modelos de outros grupos na plataforma kaggle mostram que existia uma margem para melhorar, apesar da posição na leaderboard da plataforma ser aceitavél.

4 Conclusão

Concluímos este documento conscientes do trabalho desenvolvido e dos objetivos propostos pela equipa docente. Em relação ao que foi feito e é aqui materializado estamos relativamente satisfeitos, mas, na verdade, sentimos que poderíamos ter ido mais além na complexidade dos modelos. Em ambas as fases de estudo de dados fizemos uma boa análise e tiramos conclusões razoáveis. Para terminar, do ponto de vista da análise e interpretação dos dados, assim como da criação de modelos de aprendizagem automática, o grupo é grato à UC de DAA (e anteriores UCs relacionadas) pela forma que nos treinou e consciencializou para os problemas da área.