Predição do risco de acidente rodoviário através de métodos de mineração de dados

David Almeida Dias

Dissertação para obtenção do Grau de Mestre em

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientadores/Supervisors:
Professor Doutor Alexandre Bernardino
Professor Doutor José Silvestre Serra da Silva

Júri:
Presidente/Chairperson: Prof. Pedro Filipe Zeferino Aidos Tomás
Orientador/Supervisor: Prof. Alexandre José Malheiro Bernardino
Vogais/Members of the Committee: Prof. Ricardo Adriano Ribeiro
Prof. Henrique Martins dos Santos Cunha

Novembro/2022

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Agradecimentos

A presente dissertação de mestrado simboliza uma das últimas etapas de um percurso de seis anos na Academia Militar. Não teria o mesmo sucesso sem o precioso apoio de várias pessoas. Correndo o risco de não mencionar alguns contributos, pretendo expressar os meus agradecimentos aos seguintes:

Aos meus orientadores, Professor Alexandre Bernardino e Professor José Silvestre Serra da Silva, a quem agradeço por todo o tempo despendido para aconselhamento e orientação, e pela compreensão e empenho demonstrados ao longo desta etapa.

À Academia Militar e à GNR, pelo contributo que tiveram na minha formação e no meu desenvolvimento enquanto aluno e militar e pela constante preocupação em garantir as melhores condições possíveis ao meu desenvolvimento enquanto pessoa e, também, as melhores condições possíveis para o desenvolvimento deste trabalho.

A todos os meus amigos e camaradas, em especial ao José Almeida, Pedro Martins, Ricardo Moura, Alexandre Yu Jin, Miguel Fernandes, Rafael Mota e Tiago Mota, pelos conselhos e motivação dada ao longo da realização deste trabalho.

A toda a minha família, os que cá estão e os que recentemente partiram. Em especial, aos meus pais, aos meus dois irmãos e à minha sobrinha, que me apoiam em todas as derrotas e a quem dedico todas as minhas vitórias.

A todos o meu profundo e eterno agradecimento!

David Almeida Dias

Resumo

Neste trabalho é proposta uma ferramenta de auxílio ao policiamento guiado por informações, através de um sistema de predição do risco de acidentes de viação. O sistema aplica as várias etapas do processo de descoberta de conhecimento em bases de dados na base de dados da Guarda Nacional Republicana (GNR), onde se encontram várias participações de acidentes.

Para além das participações de acidentes a GNR forneceu também dados relativos a contraordenações que contêm tanto a quantidade de fiscalizações realizadas, como o número de condutores com excesso de álcool, excesso de velocidade, entre outras contraordenações. Para complementar os dados fornecidos pela GNR, foram exploradas outras bases de dados disponíveis publicamente, como por exemplo dados meteorológicos e os calendários anuais com informação relativa a feriados e festividades.

Existem vários estudos na literatura com objetivos semelhantes. Todos eles abordam o problema como um problema de classificação supervisionada. Alguns dos estudos utilizam métodos clássicos, outros estudos utilizam métodos de aprendizagem profunda através de diferentes redes neurais profundas com arquiteturas distintas.

Neste trabalho, foram testados métodos de regressão supervisionada, entre eles o KNN, a árvore de decisão, a regressão linear, regressão de lasso, regressão de ridge e a rede neural tradicional. Após dividir-se os dados pelos diferentes tipos de localização, obteve-se um modelo com uma exatidão de 89% para as autoestradas, através do algoritmo de rede neural.

Palavras-Chave: predição de risco; acidentes de viação; classificação supervisionada; rede neural.

Abstract

This work proposes a tool to assist policing guided by information, through a system that predicts the risk of road accidents. The system applies the several stages of the knowledge discovery process in databases in the Nacional Guard (GNR) database, where several accident reports can be found.

In addition to accident reports, the GNR also provided data on administrative offenses that contained several inspections carried out, as well as the number of drivers with excess alcohol, speeding, among other types of administrative offenses. To complement this data, other available databases were provided by the GNR, such as data and calendars, data and data were explored with information related to holidays.

There are several studies in the literature with similar objectives. They all approach the problem as a supervised classification problem. Some studies use classical methods, other studies deep learning methods through different deep neural networks with different architectures.

In this work, supervised regression methods were tested, including KNN, decision tree, linear regression, lasso regression, ridge regression and the traditional neural network. After dividing the data by the different types of location, a model was obtained with an accuracy of 89% for the highways, through the neural network algorithm.

Keywords: risk prediction; road accidents; supervised classification; neural network.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Siglas e Acrónimos

CFS	*Correlation-based Feature Selection*
GNR	Guarda Nacional Republicana
KDD	*Knowledge Discovery in Databases*
KNN	*K nearest neighbors*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error*
PCA	*Principal Component Analysis*
RBA	*Relief-based feature selection*
SBS	*Sequential Backward Selection*
SBFS	*Sequential Backward Floating Selection*
SFS	*Sequencial Forward Selection*
SFFS	*Sequential Forward Floating Selection*
SVM	*Support Vector Machine*

1. Introdução

Os acidentes rodoviários causam várias mortes por ano e têm como consequência danos económicos e físicos para as vítimas e para o Estado. As ações de prevenção por parte das forças de segurança têm sido focadas naquilo a que se deu o nome de Policiamento Guiado por Informações. Atendendo a que sempre que existe um acidente rodoviário na zona de cobertura da GNR, os dados relativos ao mesmo são guardados na base de dados da GNR, tornando possível a existência de uma base de dados na qual é possível descobrir padrões e criar conhecimento. As técnicas de mineração de dados têm vindo a evoluir, sendo aplicadas em cada vez mais em problemas do mundo real. Existem sistemas de predição de risco de acidente que são usados por instituições ligadas ao trânsito em vários países.

1.1. Motivação

Os métodos de mineração de dados podem ser utilizados numa base de dados de acidentes rodoviários para extrair conhecimentos que possam de alguma forma ajudar a guiar o policiamento e desta forma melhorar as técnicas de prevenção e campanhas de sensibilização por parte das forças de segurança. O autor desta dissertação, como engenheiro eletrotécnico militar da GNR, com interesse na área de aprendizagem automática, e preocupado com questões de Segurança Nacional, vê neste tema uma forma de juntar ambos os interesses, aumentando o seu conhecimento técnico. Após os estágios realizados ao longo da sua formação, o tema do policiamento guiado por informações revelou-se de extrema importância para a organização e planeamento de missões, devido à elevada falta de pessoal e de meios que existe atualmente na instituição.

1.2. Dificuldades

Neste trabalho foram processados os dados disponibilizados pela GNR correspondendo aos anos de 2019 a 2021 no distrito de Setúbal. Para além dos dados serem maioritariamente categóricos, que por norma são mais difíceis de analisar, existem vários dados incompletos e dados incorretos, pelo que foram necessárias técnicas para colmatar estes problemas. Não foi possível obter informação relativa à geo-localização dos acidentes já que estes dados não existiam para grande parte dos acidentes, que é uma variável bastante importante e usada em vários trabalhos encontrados na revisão de literatura que demonstraram ter sucesso.

1.3. Objetivos

Este trabalho tem como objetivo desenvolver uma ferramenta de auxílio ao policiamento guiado por informações, relativamente à área do trânsito. Para isso, foram testados vários algoritmos de mineração de dados que já demonstraram ter sucesso quando aplicados a conjuntos de dados semelhantes. Estas ferramentas foram aplicadas na base de dados da GNR, que contém várias participações de acidentes. Para complementar os dados fornecidos pela GNR, serão exploradas outras bases de dados disponíveis publicamente, como por exemplo dados meteorológicos e calendário anual.

1.4. Estrutura da Dissertação

A presente dissertação encontra-se dividida em seis capítulos, e está organizada da seguinte forma:

- **Capítulo 1 – Introdução:** neste capítulo é descrito a motivação do trabalho apresentado nesta dissertação de mestrado, os objetivos e a estrutura da dissertação;
- **Capítulo 2 – Enquadramento Teórico:** neste capítulo são explanados conceitos importantes relativos às diferentes etapas do processo de descoberta de conhecimento em bases de dados;
- **Capítulo 3 – Trabalhos Relacionados:** neste capítulo é feito um estudo do estado da arte sobre os diversos métodos de mineração de dados aplicados em bases de dados relativas a acidentes de viação;
- **Capítulo 4 – Metodologia:** neste capítulo é definida e proposta a metodologia com vista à consecução dos objetivos da dissertação;
- **Capítulo 5 – Resultados e Discussão:** neste capítulo são descritas as bases de dados utilizadas. São feitas as etapas de pré-processamento dos dados e são aplicados diversos algoritmos propostos na metodologia. Cada experiência é acompanhada pela sua respetiva análise e discussão;
- **Capítulo 6 – Conclusões:** neste capítulo são apresentadas as conclusões deste trabalho, consolidando assim os objetivos propostos. São também apresentados os possíveis trabalhos futuros.

2. Enquadramento Teórico

Neste capítulo são explicados os diferentes conceitos abordados ao longo do trabalho a ser desenvolvido.

2.1. Descoberta de Conhecimento em Bases de Dados (KDD)

A tecnologia atual permite o armazenamento de grandes e múltiplas bases de dados. A análise desses dados é muitas vezes útil, no entanto, é impraticável sem o auxílio de ferramentas computacionais. Daqui surgiu o processo de Descoberta de Conhecimento em Bases de Dados (KDD, do inglês: *Knowledge Discovery in Databases*), representado na Fig. 1, que tem como objetivo identificar padrões válidos e potencialmente úteis em dados e informações, de forma a gerar conhecimento, utilizando ferramentas computacionais [1], [2].

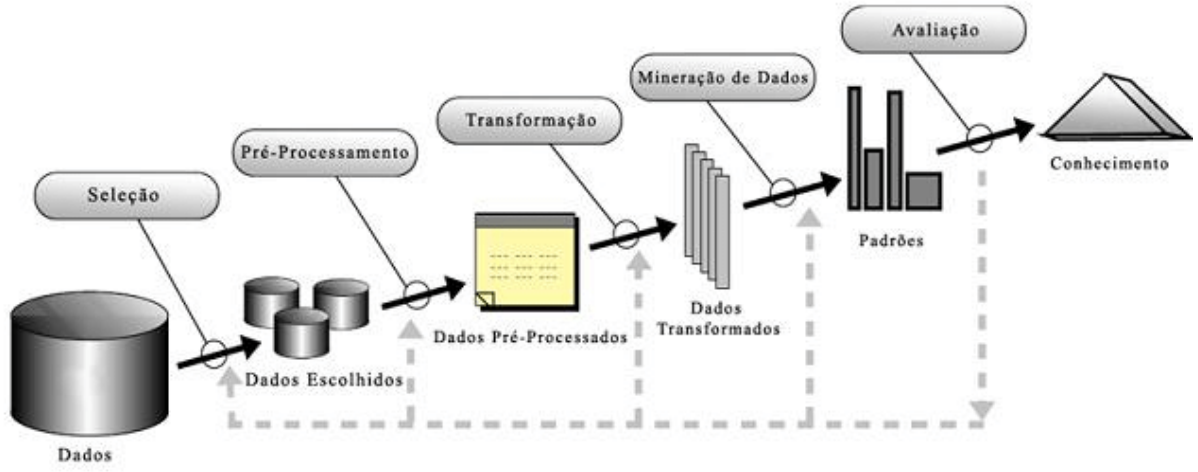

Figura 1 - Etapas que compõe o processo KDD, adaptado de [3].

Pode-se separar este processo em diferentes etapas. Apesar de algumas discrepâncias entre as diferentes fontes de pesquisa relativamente ao nome dado a cada uma das etapas ou ao número de etapas, consideraram-se as seguintes: 1 - Seleção de Dados / Definição do problema; 2 - Pré-Processamento de dados; 3 - Mineração dos dados; 4 - Avaliação e Representação dos Resultados [1]–[5].

Resumidamente, o significado de cada etapa pode ser dado por:

1- **Seleção de dados / Definição do problema**: Nesta etapa define-se o domínio dos dados disponíveis, identifica-se que informações e dados são relevantes e quais os objetivos da descoberta de conhecimento [1]–[5].
2- **Pré-Processamento**: Esta segunda etapa tem como objetivo preparar os dados para os algoritmos da etapa seguinte, nomeadamente efetuar a limpeza dos dados, integração dos dados, redução dos dados e transformação/normalização dos dados [1], [4] e [5].
3- **Mineração dos dados**: Alguns autores referem-se à Mineração de Dados e ao processo de Descoberta de Conhecimento em Bases de Dados como sinónimos. No entanto, considerou-se a Mineração de Dados como uma etapa desse processo de KDD, tal como em [1]–[3],

[5]. É nesta etapa que se aplicam algoritmos sobre os dados em busca de conhecimento, ou seja, de extrair padrões nos dados. A escolha do algoritmo a ser aplicado depende do tipo de tarefa a ser realizada.

Os diferentes algoritmos podem ser inseridos em diferentes técnicas de mineração. Autores como [2], [3], [5], definem as seguintes técnicas: Classificação, Regressão, Agrupamento, Sumarização, Modelo de Dependência e Deteção de sequências, mudanças e desvios. No entanto, [1] e [4] definem as mesmas técnicas que os anteriores mas acrescentam também as técnicas de Associação.

4- **Avaliação e representação dos resultados**: Nesta etapa são interpretados os modelos obtidos e são utilizadas métricas de avaliação de modo a estimar a qualidade dos resultados. Por fim devem-se utilizar ferramentas de visualização dos dados obtidos como saída.

2.1.1. Seleção de dados / Definição do Problema

A primeira etapa do processo de Descoberta de Conhecimento em Bases de Dados consiste em definir o objetivo de acordo com o ponto de vista do cliente e entender os conhecimentos prévios necessários sobre o objeto de estudo. Com isto é possível definir o conjunto de dados alvo [3]. O analista de dados, ou seja, a pessoa que irá criar os modelos através de algoritmos, deve realizar entrevistas com especialistas da área especifica do problema. É comum agrupar os dados numa única tabela bidimensional, porque a maioria dos algoritmos de Mineração de Dados pressupõe que os dados estejam organizados dessa forma [1].

Segundo [1] a junção dos dados numa única tabela pode ocorrer de duas formas:

a) Junção Direta – Todas as características e registos da base de dados são incluídos sem que qualquer análise seja feita.

b) Junção Orientada – O especialista no domínio da aplicação, em parceria com o analista de dados, escolhe as características e os registos que tenham potencial ou mais potencial para influir na seleção de características [4] [5].

2.1.2. Pré-Processamento

O trabalho realizado nesta etapa pode ser reduzido se forem criadas várias regras para a entrada dos dados na base de dados [1]. Normalmente isso não acontece e, por isso, as técnicas de pré-processamento, quando aplicadas antes da mineração, podem melhorar substancialmente a qualidade dos padrões encontrados e/ou o tempo necessário para a execução do algoritmo de mineração [4]. Isto porque os dados podem muitas vezes ser incompletos e inconsistentes ou conter ruído [5].

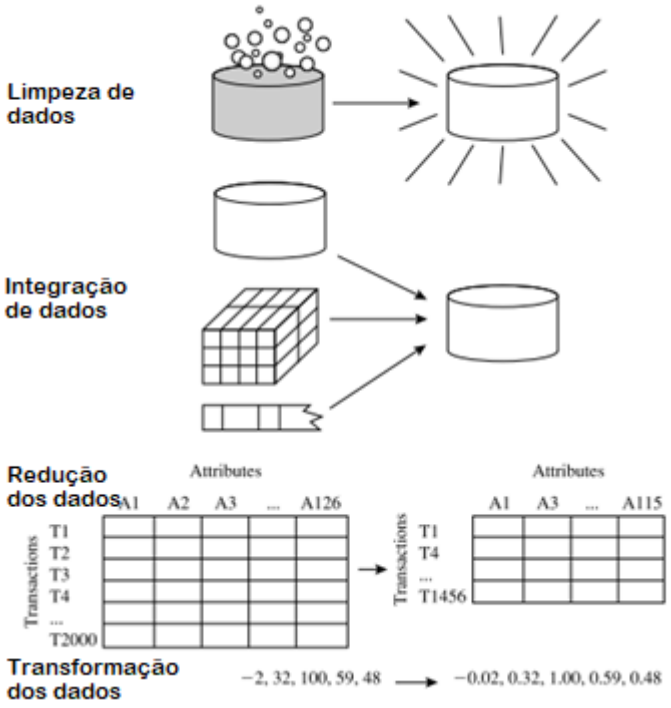

Figura 2 - Tipos de pré-processamento, adaptado de [4]

a) Limpeza dos dados

A fase de limpeza dos dados envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores desconhecidos e redundantes. Alguns métodos de limpeza de dados, para dados incompletos, que foram considerados de [1] e [4] são:

- **Preenchimento com Medidas Estatísticas** – substituir um valor em falta pela média ou moda. Para dados com distribuições normais a média deve ser usada. Para distribuições não normais é comum usar a mediana;
- **Preenchimento com métodos de mineração de dados** – são usados algoritmos de mineração de forma a sugerir os valores mais prováveis a serem utilizados para substituir valores nulos. Isto pode ser feito a partir de métodos de regressão, ferramentas *Bayesianas*, ou árvores de decisão.

Existem também outros algoritmos que são utilizados para completar dados em falta, como por exemplo, extensões do algoritmo de Análise dos Componentes Principais (PCA, do inglês: *Principal Component Analysis*) [6]. O preenchimento com métodos de mineração de dados é o método mais frequente, pois é aquele que usa mais informação dos dados já presente na base de dados.

Relativamente à limpeza de dados para dados com ruído ou inconsistentes, são apresentados de seguida alguns algoritmos que se focam neste problema [1], [4]:

- **Métodos de Visualização –** Através de histogramas e outros tipos de representações visuais de dados, é possível detetar *outliers* e eliminá-los manualmente, ou substituir os seus valores por métodos descritos anteriormente;
- **Partição dos dados em células (*Bins*) –** consiste em repartir os dados e agrupá-los em conjuntos pequenos. Desta forma é possível suavizar os mesmos, alterando os seus valores, utilizando para isso os valores da sua vizinhança, por exemplo, através da média de cada conjunto. Para além da média, existem outras formas de suavizar os dados de cada conjunto que podem ser vistas em [1], pág. 34-36 e [4], pág.89-90.

b) Integração dos dados

A integração de dados envolve agregar mais informações aos registos existentes através da consulta de bases de dados externas e ver como estes influenciam os resultados. Para o caso de acidentes de viação, pode-se, por exemplo, integrar dados meteorológicos na tabela bidimensional e entender como estes influenciam os resultados. Existem vários desafios ao integrar os dados tais como: entender como os diferentes dados devem ser unidos; normalização dos diferentes tipos de dados; calcular a correlação entre os mesmos; e eliminar dados duplicados [1] [4].

c) Redução dos dados

Relembrando que os dados vêm em tabelas bidimensionais, estes podem ser reduzidos horizontalmente ou verticalmente [1]. A redução horizontal consiste na eliminação de amostras, isto é, de linhas da tabela de dados, como demonstrado na Tab. 1. As operações consideradas mais relevantes de **redução horizontal** mencionadas em [1] são:

- **Segmentação do banco dados** – por exemplo, analisar apenas os clientes de uma certa cidade;
- **Agregação de Informações** – por exemplo, somar todas as compras de um cliente, obtendo as despesas totais.

Tabela 1- Exemplo de eliminação horizontal de dados, em que apenas se considera os clientes da cidade de Lisboa.

Nome	Despesas	Cidade
Luís	200.0	Lisboa
Miguel	300.0	Lisboa
Pedro	100.0	Porto

Já a redução vertical consiste na eliminação de colunas da tabela bidimensional, como por exemplo eliminar a coluna “Despesas” na Tab. 1. A seleção de características (do inglês: *feature selection*), é essencial para reduzir a complexidade do algoritmo e melhorar a performance do mesmo. Algumas operações de **redução vertical** mencionadas por E. Goldshmidt et al. [1] , S. Agarwal [4], U. M. Khaire et al. [7] e Canedo et al.[8] são:

- **Eliminação direta de características** – Devem ser eliminadas características que sirvam de identificação das linhas, pois são diferentes para todas as linhas. Podem ser eliminadas características com valor constante em todas as linhas caso estes não adicionem nenhuma informação útil à descoberta de conhecimento. Para as restantes características que não têm valores constantes, nem são características de identificação, a eliminação direta deve ser utilizada apenas quando o conhecimento prévio permite ter uma certeza quase absoluta que uma certa característica não é relevante para o problema;
- **Abordagem dependente do modelo –** consiste em executar o algoritmo de mineração iterativamente ao mesmo tempo que se adiciona ou retira características, de forma a verificar quais as combinações de características que resultam em melhores predições. Durante a iteração, existem várias formas de adicionar ou remover características (conforme indicado em [1], pág. 29-31);
- **Algoritmos de seleção de características –** consiste em algoritmos criados para selecionar as características mais relevantes e que mais influenciam uma variável alvo.

d) Transformação dos dados

Consiste em trabalhar os dados de forma a obter melhores resultados na mineração através de diferentes técnicas como a **criação de características** a partir das existentes, a **codificação dos dados**, **normalização dos dados** e a **agregação dos dados.**

A codificação dos dados pode ser Numérica-categórica, que divide valores de características contínuos em intervalos codificados ou Categórica-Numérica, que representa valores de características categóricas por códigos numéricos; Sendo que a maioria das variáveis da tabela fornecida são categóricas com múltiplas variáveis, existem três possibilidades de codificação Categórica-Numérica a serem utilizadas, que são: codificação determinada; técnicas algorítmicas; técnicas automáticas [9]. Para o âmbito deste trabalho, apenas se utilizará a codificação determinada em que os dados podem ser utilizados por outro tipo de algoritmos de mineração. Uma maneira de identificar uma codificação determinada é que os valores codificados serão os mesmos todas as vezes que aplicarmos a técnica [9].

Uma técnica de codificação determinada é a codificação de rótulo (do inglês: Label) é simplesmente atribuir um valor inteiro a cada valor possível de uma variável categórica. Por exemplo, se um conjunto de dados contém uma variável categórica com valores extraídos do conjunto {“baunilha”, “chocolate”, “morango”}, a codificação de rótulo poderá atribuir os valores mapeados do conjunto {0, 1, 2}, respetivamente. O problema desta técnica é que, caso a variável categórica não apresente uma hierarquia/ordem, muitos algoritmos de mineração vão atribuir uma distância/importância diferente entre os diferentes valores. Por exemplo, o algoritmo poderá considerar que a distância entre 0 e 2 é uma distância maior do que entre 0 e 1. Caso exista uma hierarquia entre as variáveis categóricas, como o conjunto {“fraco”, “médio”, “forte”}, pode-se utilizar a codificação de rótulo, captando assim a relação entre as diferentes categorias. No entanto, é necessário encontrar a

distância certa entre as diferentes categorias de modo que uma unidade não tenha uma relevância demasiado exagerada ou reduzida em relação à distância numérica de valores de outras variáveis [9].

Outra técnica determinada é a codificação *One-hot*. Expressamos a codificação *One-hot* formalmente da seguinte forma: Seja x uma variável aleatória categórica discreta com n valores distintos x1, x2, …, xn. Então, a codificação *One-hot* de um determinado valor xi é um vetor v onde cada componente de v é zero, exceto para o componente número i, que tem o valor 1. Por exemplo, suponha que temos alguma variável aleatória x que recebe valores do conjunto S = {a, b, c}. Seja x1 = a , x2 = b e x3 = c . Uma codificação *One-hot* para x é: a = (1, 0, 0) , b = (0, 1, 0) e c = (0, 0, 1) [9].

No entanto este tipo de codificação pode trazer problemas, principalmente para problemas de regressão como é o caso deste trabalho. Isto porque através desta codificação podem surgir problemas de Multicolinearidade, em que as variáveis independentes possuem relações lineares aproximadamente exatas. Para resolver este problema a variável x do exemplo do parágrafo anterior pode ser apenas caracterizada por duas das categorias diferentes de 0, como por exemplo: a = (1, 0), b = (0, 1) e c = (0, 0). Sendo assim, considerando k como o número de categorias existentes para uma certa variável categórica, em vez de k vetores para cada variável categórica, passam agora a ser criados k-1 vetores [9].

Um passo importante, também realizado na transformação de dados é a **normalização dos dados**, em que se ajustam os valores de cada característica de forma que fiquem na mesma escala entre si, pois existem algoritmos de mineração que podem considerar valores mais elevados como mais importantes de forma tendenciosa; Por fim, a **Agregação de dados**, para análise de dados em vários níveis de abstração, por exemplo, os dados de vendas diárias podem ser agregados de modo a obter os valores totais mensais e anuais [1] [4].

2.1.3. Mineração de dados

A mineração de dados é frequentemente dividida em três grupos principais. A **aprendizagem supervisionada** que ocorre quando um algoritmo aprende através dos dados disponíveis, em que esses dados já têm uma saída associada. Por exemplo, se o objetivo de um problema de mineração de dados for prever o género masculino ou feminino através da imagem de um rosto, para este tipo de aprendizagem seria necessário ter um conjunto de rostos já com o género devidamente identificado, de forma que o algoritmo, através desse conjunto de imagens, conseguisse criar um modelo para prever novas imagens. É importante distinguir problemas de regressão em que os dados para os quais queremos prever o valor são valores numéricos, de problemas de classificação em que os dados são valores categóricos [4][10][11][12][13].

Em segundo, a **aprendizagem não supervisionada** tem uma definição semelhante à supervisionada. No entanto, apesar de também aprender através dos dados disponíveis, estes dados não têm uma variável de saída associada. Estes métodos são uteis para fornecer novos padrões, variáveis de saída ou relações entre variáveis que não seriam facilmente detetados de outra forma. Um exemplo são os algoritmos de agrupamento (do inglês: *Clustering*) que agrupam as entradas em classes [4][10][11][12].

Por último, a **aprendizagem por reforço**, que está ligada a problemas em que o algoritmo deve tomar decisões. Tal como na aprendizagem não supervisionada, os algoritmos recebem dados que não têm uma variável de saída associada, no entanto o algoritmo vai aprendendo através de tentativa e erro, recebendo uma resposta para cada decisão que toma. Um exemplo deste tipo de aprendizagem é quando um computador aprende a jogar um jogo sozinho [10].

Para os objetivos deste trabalho os dados disponíveis têm uma saída associada e, por isso, considera-se o problema como uma classificação supervisionada. Por este motivo serão mencionadas apenas técnicas de aprendizagem supervisionada. Em [11] foram analisados 84 artigos que discutem diferentes técnicas de aprendizagem supervisionada e não supervisionada, em que o objetivo passou por encontrar uma definição para os diferentes termos e técnicas existentes. Concluiu-se que Árvores de Decisão, *Naive Bayes* e Máquina de vetores de suporte são as técnicas mais usadas nesses 84 artigos. Alguns dos algoritmos de classificação supervisionada são os seguintes:

A **árvore de decisão** é maioritariamente usada em problemas de classificação, apesar de também apresentar bons resultados para problemas de regressão. É uma ferramenta que é representada como uma árvore lógica tendo um conjunto de condições e conclusões representados sob a forma de nós e ramos. A árvore liga as condições e as conclusões, ou seja, os ramos e os nós. Usa estimativas e probabilidades para calcular resultados prováveis e ajuda a decidir qual a resposta que representa menor incerteza. A árvore é construída atribuindo cada valor possível no domínio de dados a uma classe com base no valor da variável de destino que é a mais comum na instância da iteração, ou seja, cada nó interno testa uma característica e a cada um dos nós é atribuído a uma classificação [4][10][11][12][13][14];

O classificador ***Naive Bayes***, de uma forma muito resumida, consiste em encontrar a probabilidade a posteriori a partir das probabilidades a priori e das probabilidades condicionadas para cada uma das classificações possíveis. Por exemplo, se o objetivo do classificador for detetar se uma mensagem de texto é ou não spam, através do conjunto de dados já fornecido (visto que é um tipo de aprendizagem supervisionada) é possível obter a probabilidade de uma certa palavra existir, dado que é spam e a probabilidade de uma certa palavra existir dado que não é spam, obtendo as probabilidades condicionadas. Também é possível obter a probabilidade de uma mensagem ser ou não spam, dividindo aquelas que são spam pelo conjunto total de mensagens e aquelas que não são spam pelo conjunto total de mensagens, obtendo a probabilidade a priori. Para classificar uma nova mensagem, através da fórmula de Bayes, basta multiplicar a probabilidade a priori pela probabilidade condicionada de cada palavra existente na nova mensagem, obtendo assim a probabilidade a posteriori para cada uma das classes. Por fim, escolhe-se a classe que obtém um valor maior de probabilidade a posteriori. Uma restrição deste algoritmo é que assume que as características são independentes entre si, mas mesmo com esta assunção apresenta bons resultados em vários problemas de classificação [1][4][10][11][12][13][14];

Máquina de vetores de suporte (SVM, do inglês: *Support Vector Machine*) é um algoritmo avançado que pode lidar com problemas de regressão e classificação, embora, no geral, apresente

melhores resultados para a classificação. Resumidamente, o algoritmo tenta encontra o plano ou hiperplano que divide o conjunto de dados e as respetivas classes. O objetivo passa por encontrar as margens de modo que a distância entre cada classe e a margem mais próxima seja maximizada levando ao menor erro de classificação possível [8][9][10] [14];

O **K-vizinhos mais próximos** (KNN, do inglês: *k-neirest neighbors)* é um dos algoritmos mais simples de classificação e em muitas situações apresenta erros maiores que os três algoritmos descritos anteriormente. O algoritmo armazena todos os registos disponíveis e para classificar uma nova instância, esta é comparada com a classe das K instâncias mais próximas. A classe que for representada mais vezes nesse conjunto de K instâncias já classificadas, é a classe atribuída a essa nova instância [8][9][10][11];

A **rede neural artificial** (ANN, do inglês: *Artificial Neural Network*), são algoritmos que se baseiam na estrutura dos neurónios do cérebro humano, mas numa escala mais pequena e simplificada. Estas redes são compostas por um conjunto de nós divididos em camadas. Por norma as redes neurais têm entre duas a três camadas. No caso de mais de 3 camadas e de ser criada uma rede neural mais complexa com ligações diferentes de uma rede normal comum, é considerado uma aprendizagem profunda (do inglês: *deep learning*). Considera-se que N é o número total de camadas e k o índice de uma certa camada. A primeira camada, k = 0, é a camada de entrada e a última camada, k = N, é a camada de saída, todas as restantes camadas intermediárias, k > 0 e k < N, são chamadas de camadas ocultas. Nas redes neurais tradicionais o trajeto da informação começa na primeira camada e, de seguida, passa pelas camadas ocultas, de forma sequencial. Por fim, os dados chegam à camada de saída. Não existem ligações no sentido inverso do trajeto da informação, ou seja, a informação vai da esquerda para a direita [11][15][16]. Existem variantes das redes neurais, algumas das quais são: Regressão Logística, Rede Neural de Propagação de Retorno, Rede Neural Convolucional e Rede Neural Recorrente [11][15].

a) Comparação entre os diferentes métodos

Tabela 2 - Comparação entre algoritmos de classificação supervisionada, adaptado de [10] e [11].

Algoritmo	Vantagens	Desvantagens
Árvore de Decisão	Consegue lidar com dados incompletos, com ruído, com atributos redundantes e não linearmente separáveis; requer pouca complexidade computacional na classificação de novas instâncias; fácil de visualizar e compreender.	Tem dificuldade em lidar com dados que contenham muitas dimensões; pode facilmente fazer um sobreajuste (do inglês: *overfitting*) do modelo; A construção da árvore, ou seja, do modelo, tem alguma complexidade computacional.
Naive Bayes	Processa dados incompletos, com ruído, com atributos redundantes e não linearmente separáveis; requer pouca complexidade computacional na classificação de novas instâncias e na construção do modelo; requer pouca quantidade de dados para alcançar modelos com erros baixos de classificação.	Tem dificuldade em lidar com dados que contenham muitas dimensões ou uma grande quantidade de dados; tem mau desempenho quando existem várias dependências entre os diferentes atributos, já que o algoritmo assume independência entre os mesmos.
Máquina de Vetores de Suporte	Analisa com dados incompletos, com ruído, com atributos redundantes e com dados que contêm muitas dimensões ou grande quantidade de dados; requer pouca complexidade computacional na classificação de novas instâncias;	A construção do modelo tem bastante complexidade computacional; exceto algumas alternativas do algoritmo, este não tem solução em problemas não linearmente separáveis;
k-vizinhos mais próximos	Consegue lidar com dados com ruído, com atributos redundantes, dados não linearmente separáveis e dados de grandes dimensões ou grande quantidade de dados; requer pouca complexidade computacional na construção do modelo; é um algoritmo de fácil implementação.	Tem dificuldade em lidar com dados que contenham muitas entradas incompletas; requer bastante complexidade computacional na classificação de novas instâncias; é difícil estimar o valor correto de K de forma a obter o melhor modelo possível.
Redes Neurais	Processa dados incompletos, com ruído, com atributos redundantes, não linearmente separáveis, não-lineares ou dinâmicos e dados com fortes dependências entre variáveis; requer pouca complexidade computacional na classificação de novas instâncias.	A construção do modelo tem bastante complexidade computacional; A escolha da quantidade de camadas ocultas e do tamanho de cada camada oculta para o qual se obtém o melhor modelo possível é difícil de estimar e de interpretar.

2.1.4. Avaliação e representação dos resultados

A avaliação de um classificador é feita através de métricas com diferentes significados e interpretações. Diferentes métricas avaliam diferentes características do classificador induzidas pelo algoritmo de classificação. A exatidão (do inglês: *accuracy*) ou taxa de erro são as métricas mais comummente utilizadas [4][5][17].

Para o este trabalho é relevante estudar as métricas para classificadores mais conhecidas. A partir da Tab. 2 podem ser geradas várias métricas de avaliação, demonstradas na Tab. 3.

Tabela 3 – Matriz para classificação binária, adaptada de [17].

	Atual classe positiva	**Atual classe negativa**
Previsão Classe Positiva	Verdadeiro positivo (*vp*)	Falso negativo (*fn*)
Previsão Classe Negativa	Falso positivo (*fp*)	Verdadeiro negativo (*vn*)

Tabela 4 – Métrica de avaliação de classificadores, adaptada de [17].

Medida	**Fórmula**	**Foco de avaliação**
Exatidão (ex)	$\frac{vp + vn}{vp + fp + vn + fn}$	Em geral, a precisão métrica mede o rácio de predições corretas no número total de instâncias avaliadas.
Taxa de erro (err)	$\frac{fp + fn}{vp + fp + vn + fn}$	A taxa de erro mede a proporção de previsões incorretas sobre o número total de instâncias avaliadas.
Sensibilidade (ss)	$\frac{vp}{vp + fn}$	Esta medida é usada para medir a fração de padrões positivos que são classificados corretamente.
Especificidade (ep)	$\frac{vn}{vn + fp}$	Esta medida é usada para medir a fração de padrões negativos que são classificados corretamente.
Precisão (p)	$\frac{vp}{vp + fp}$	A medida de precisão é usada para medir os padrões positivos que são previstos corretamente a partir do total de padrões previstos em uma classe positiva
Retorno (r)	$\frac{vp}{vp + vn}$	O retorno é usado para medir a fração de padrões positivos que são classificados corretamente
F-score (mf)	$\frac{2 * p * r}{p + r}$	Esta medida representa a média entre os valores de retorno e precisão.
Media Geométrica (mg)	$\sqrt{vp * vn}$	Esta medida é usada para maximizar a taxa *tp* e a taxa *tn*, e simultaneamente manter ambas as taxas relativamente equilibradas.

Para além da exatidão e da taxa de erro, o F-score e a média geométrica também são frequentemente usadas como métricas de avaliação pois têm um desempenho melhor do que a

exatidão e do que a taxa de erro na otimização do classificador para problemas de classificação binária. As restantes métricas na Tab. 3 podem dar mais informação sobre o modelo obtido, no entanto, são inadequadas para selecionar a solução ótima, pois baseiam-se numa avaliação parcial (verdadeiros positivos ou negativos) [17]. Por fim, após a avaliação de performance, devem-se utilizar ferramentas de visualização dos padrões e modelos extraídos ou visualização dos dados [1]–[5].

3. Trabalhos relacionados

Os trabalhos relativos à aplicação de algoritmos de mineração de dados em bases de dados de acidentes de viação podem ser divididos em 3 áreas principais, as quais são: predição da severidade da lesão resultante do acidente de viação, estudos de deteção automática de acidentes em tempo real, por exemplo, através da análise do congestionamento do transito em tempo real e, por último, estudos de predição do risco ou localização provável de acidente por estrada ou região. Este trabalho tem um foco no último tema, predição do risco de acidente por estrada ou região.

Os acidentes de viação podem ser agrupados em função da sua causa: comportamentos do condutor, fatores climatéricos e características das estradas e das viaturas. O problema é abordado como uma classificação supervisionada.

Antes de abordar as diferentes formas de predizer o risco de acidente por estrada ou região, considerou-se importante analisar alguns das principais características encontradas na literatura que influenciam o acidente de viação.

3.1. Seleção de características

Canedo et.al [8] realizaram uma revisão a algoritmos de seleção de características. Segundo os mesmos, nos últimos anos, os algoritmos de seleção de características tornaram-se componentes indispensáveis para excluir o número de características irrelevantes e redundantes. Existem duas abordagens principais na seleção de características: avaliação individual e avaliação de subconjunto. A avaliação individual atribuindo-lhes pesos de acordo com seus graus de relevância. Por outro lado, a avaliação de subconjuntos produz subconjuntos de características candidatas com base num certo algoritmo de mineração. Quando uma avaliação individual de características é obtida, é necessário estabelecer um limite para descartar aquelas menos relevantes para o algoritmo. Belanche et al. [18] optaram por descartar as características com pesos associados ao ranking que estavam a mais de duas variâncias da média.

Foram comparados doze algoritmos de seleção de características no trabalho de Canedo et.al [8] em diferentes tipos de dados sintéticos. Os conjuntos de dados estudados foram dados com problemas de correlação, dados com problemas de redundância de características e dados que incorporavam problemas de ruído. O melhor algoritmo para um conjunto de dados semelhante aos dados do presente trabalho (com poucas características comparativamente ao número de amostras) é: Seleção de recursos com base em relevo (RBA, do inglês: *Relief-based feature selection*).

Já no trabalho de Belanche et al. [18] também foram comparados vários algoritmos. Um dos algoritmos que mais se destacou foi também o RBA. Ainda no mesmo trabalho, outro algoritmo com

bons resultados foi a seleção direta sequencial (SFS, do inglês: *Sequencial Forward Selection*) e o algoritmo variante deste (SBS, do inglês: *Sequencial Backward Selection*).

Em 2015, Lin et al. [19] propuseram o algoritmo de associação (FP-T, do inglês: *Frequent Pattern Tree*), um método de pré-processamento dos dados, de forma a selecionar características que têm mais probabilidade de contribuir para a predição em problemas especificamente relacionado a acidentes de viação. O seu método de seleção de características teve mais sucesso para os modelos de mineração *K-vizinhos mais próximos* e *Naive Bayes*. A melhor performance atingida foi cerca de 61% numa amostra acidentes num segmento da autoestrada I-64 em Norfolk, Virgínia.

Os algoritmos de seleção de características que se optaram por utilizar nesta dissertação foram encontrados na literatura analisada e estão representados na Tab. 5.

Tabela 5 - Algoritmos de seleção de características.

RBA - Relief-based Feature Selection
SBS - Sequential Backward Selection

A aplicação destes algoritmos foi feita em dados sintéticos em que já se sabia à priori quais as características importantes. A avaliação da performance destes algoritmos foi feita através de equações que relacionam o número de características consideradas relevantes e não relevantes pelos algoritmos, comparando com as características são de facto relevantes e não relevantes [8].

Um fator importante encontrado na literatura, é a relação entre as condições climatéricas e o risco de acidente, mostrando a importância de incluir este tipo de dados no trabalho proposto. Já em 2003, Eisenberg et al. [20] estudaram a relação entre a precipitação e os acidentes de viação, concluindo que existe uma correlação entre os mesmos e chegando à conclusão que os períodos mais perigosos são os períodos de precipitação após grandes intervalos de seca, que pode ser explicado pela acumulação de óleos nas estradas. Mais tarde, em 2013, Bergel-Hayat et al. [21] estudaram a correlação entre várias características das condições climatéricas (precipitação, temperatura, força do vento, etc) e a quantidade de acidentes de viação em bases de dados que continham acidentes de França, Países Baixos e uma região de Atenas, por períodos de mais de 20 anos, concluindo que existe uma forte correlação entre as condições atmosféricas e o número de acidentes. Em 2016, Tamerius et al. [22] fizeram uma análise semelhante, mas mais complexa, incluindo cerca de 600000 acidentes do estado de Iowa, nos Estados Unidos da América, incluindo nos dados de precipitação as alterações que ocorrem nas diferentes horas do dia e não apenas uma média da precipitação diária. Neste estudo concluíram que o aumento de precipitação aumenta em cerca de 70% o risco de acidentes.

Relativamente ao comportamento humano, Febres et al. [23] concluiu que a não utilização do cinto de segurança resulta em acidentes com maior severidade para as vitimas e que as distrações e o tipo de estrada também influenciam a ocorrência de acidentes, através de um modelo de redes Bayesianas num conjunto de cerca de 241000 acidentes ocorridos em Espanha entre 2016 e 2017. Musile et al. [24] demonstrou que o consumo excessivo de álcool cria um aumento do risco de acidente.

Matín-Reyes et al. [25] encontraram uma forte relação entre o aumento do risco de acidente e a condução sem carta de condução ou carta de condução suspensa. Song et al. [26] provaram que o histórico de acidentes e de contraordenações de um condutor, como a passagem por um sinal vermelho, influenciam bastante o seu risco de participar num novo acidente.

Relativamente às condições de estrada e tipo de viaturas, Kumeda et al. [27] na tentativa de encontrar os fatores que mais influenciam um acidente, concluíram no seu estudo que as condições de luminosidade, o número de identificação da estrada e o número de veículos envolvidos são os principais fatores que influenciam os acidentes. O tamanho a amostra foi de 555 acidentes registados no Reino Unido, em 2016. As características analisadas foram os seguintes: condições meteorológicas, da estrada e de luminosidade, a hora do acidente, o número de veículos, a identificação da estrada e a severidade da lesão causada pelo acidente.

Na Tab. 6 podemos ver de forma mais organizada alguns das características mais comuns encontrados na literatura.

Tabela 6 - Características consideradas relevantes em estudos de acidentes de viação encontrados na literatura

Condições climatéricas	Precipitação; temperatura; força do vento.
Comportamento humano	Uso de cinto de segurança; uso de telemóvel; consumo de álcool; calendário
Condições de estrada	Redes de estrada; luminosidade; identificação da estrada; volume de trânsito.

3.2. Classificação Supervisionada

3.2.1. Métodos Clássicos

Uma grande quantidade de estudos procura classificar cada segmento de estrada num determinado momento em classes binárias (Haverá acidente, Não haverá Acidente). Em 2005, Chang [28] comparou o desempenho da ANN com o desempenho do modelo de regressão em 1338 acidentes. Acidentes estes que ocorreram entre 1997 e 1998 na autoestrada Nacional 1 de Taiwan. O estudo chegou à conclusão de que a ANN é um bom modelo para analisar acidentes de viação, por ter atingido 61.4% de exatidão. No mesmo ano e num conjunto de dados semelhante, por ser a mesma autoestrada, mas desta vez relativamente aos anos 2001-2002, com 1484 acidentes, Chang et al. [29] aplicaram também o modelo de árvores de regressão e classificação, ou seja, aquilo a que chamámos de árvores de decisão no capítulo de enquadramento teórico. A exatidão foi de cerca de 55%, demonstrando que as árvores de decisão são menos eficazes para a predição de acidentes do que as ANNs.

Começando pelos estudos com menor complexidade, tanto Muhammad et al. [30], em 2017 e Akomolafe et al. [31], em 2013, realizaram um estudo com o objetivo de prever o local, a causa e a hora do acidente apenas numa autoestrada. Muhammad et al. [30] analisaram 165 acidentes num período de 30 meses, desde janeiro de 2014 a junho de 2016, numa autoestrada entre Kano e Wudil, na Nigéria. Este estudo utilizou o software de mineração de dados WEKA e o algoritmo de árvore de

decisão Id3. As variáveis usadas foram apenas 4: Tipo de veículo, Hora do acidente, Causas do acidente e Localização do acidente. O modelo preditivo alcançou uma exatidão de 72.7%. Akomolafe et al. [31] criaram um modelo preditivo para uma autoestrada, em Lagos, Nigéria, com dados de 2002 e 2003, que englobam 148 acidentes de viação. Neste estudo as variáveis estudadas foram o tipo de veículo, a hora do dia, as condições atmosféricas, as causas do acidente e a localização. Neste trabalho foram usados dois algoritmos de seleção de características, *correlation based feature selection* e *consistency subset selection*. Tal como no trabalho anterior, também foi utilizado o software WEKA e para além do algoritmo de mineração de árvore de decisão Id3, também foi utilizado o *Function Tree*. Foi alcançada uma exatidão de cerca de 78%, usando a medida-F como medida de performance.

Em 2014, Olutayo et al. [32] aplicaram o modelo de árvore de decisão e ANN num conjunto de 417670 acidentes de 1995 a 2000, de uma amostra de acidentes do Estados Unidos da América. A árvore de decisão superou a ANN obtendo uma exatidão de cerca de 74%, no entanto o objetivo era classificar o tipo de severidade do acidente e não prever a ocorrência de acidentes. Já Shanti et al. [33] realizaram um estudo com o objetivo foi prever a forma de colisão no acidente, de forma a encontrar um padrão do tipo de acidentes que podem ocorrer com mais ou menos frequência de acordo com as características da estrada, condições meteorológicas, entre outras. Ao todo foram analisadas 27 variáveis numa amostra de 37248 acidentes. Foram considerados 9 tipos de colisão possíveis como variável de saída e foram comparados alguns algoritmos de mineração, entre os quais: árvores de decisão C4.5 e ID3, *naive bayes* e *Random Tree*. Foram utilizados alguns algoritmos de seleção de características, dos quais se destacou o *Feature Ranking*. O preditor com melhores resultados foi o *Random Tree*, no entanto o objetivo era também uma classificação e não uma regressão. Em 2016, Castro et al. [34], utilizaram uma base de dados de 451462 acidentes do Reino Unido de 2010 a 2012, sendo que desses, apenas 81690 desses acidentes foram incluídos no estudo. Foi utilizada a ferramenta WEKA e foram consideradas 7 variáveis de entrada, as quais foram: o tipo de estrada, as condições de luz, as condições meteorológicas, as condições da superfície da estrada, a manobra do veículo, o tipo de combustível da viatura, a idade do veículo e a severidade do acidente. Foram utilizados 3 algoritmos de mineração, os quais foram: a rede bayesiana, BayesNet, o algoritmo de árvore de decisão J48 e um algoritmo de rede neural *multi-layer perceptron*. Todos eles com uma variável de saída com 3 valores possíveis, que representam a severidade do acidente (fatal, grave ou normal). A medida de performance utilizada foi a precisão e severidade do acidente foi prevista pelos 3 diferentes algoritmos com uma precisão muito semelhante de cerca de 72%. No mesmo ano, Keshyap et al. [35] procuraram encontrar uma ligação entre as condições das estradas e a severidade do acidente. Em vez de árvores de decisão, foi utilizado o algoritmo de redes bayesianas, também através do software WEKA. Neste trabalho já foram incluídas 12 características, entre os quais: o estado do condutor, experiência do condutor, condições climatéricas, tipo de estrada, condições de luminosidade, condições do veículo, tipo de veículos incluídos no acidente, tipo de animais, severidade do acidente, utilização de cinto de segurança e a localização. Foram analisados 31698 acidentes provenientes de questionários feitos a pessoas que sofreram acidentes, desde 2003 a 2015. A utilização da seleção de características piorou a exatidão do modelo e o melhor resultado obtido foi de 89% sem qualquer algoritmo de seleção de características. Em 2019, Hussain et al. [36] realizaram

uma avaliação de diferentes métodos de mineração de dados clássicos em acidentes de viação, analisando literatura semelhante à referida no parágrafo anterior e chegando à conclusão que os algoritmos mais usados e com maior exatidão são o *Multi-layer Perceptron*, a árvore de decisão J48 e o *Naive Bayes*. No mesmo ano, Kumeda et al. [27] aplicaram 6 algoritmos de classificação clássicos, tais como o *Naive Bayes, Multi-layer Perceptron, Random Forest*, entre outros, de forma a encontrar os fatores que mais influenciam o acidente de viação. Já Almamlook et al. [37] também compararam 3 métodos de mineração de dados, os quais foram *Naive Bayes*, *Radom Forest* e *Logistic Regression* para encontrar os fatores que mais influenciam na severidade dos acidentes.

Apesar dos trabalhos acima descritos tratarem problemas de classificação e não de regressão, serão importantes para ter uma ideia do tipo de variáveis utilizadas. A maioria da literatura encontrada com aplicações de métodos clássicos não tinha como objetivo prever quantidades de acidentes, mas sim prever classes, como a severidade do acidente, entre outros. Todos os trabalhos acima mencionados aplicaram técnicas de mineração clássicas, a maioria deles num conjunto de dados de pequena escala (exemplo: em uma ou num pequeno número de estradas) com uma quantidade limitada de características. Também não abordaram as propriedades únicas dos dados que utilizaram, tais como a periocidade, as correlações e heterogeneidade derivadas dos locais. No entanto, este tipo de abordagem pode ser útil se for criado um modelo para cada estrada ou região.

3.2.2. Métodos de Aprendizagem Profunda

Alguns trabalhos mais recentes procuraram enfrentar os problemas na análise de acidentes de viação ao utilizar Modelos de Aprendizagem Profunda (do inglês: *Deep Learning Models*). Chen et al. [38] utilizaram dados de cerca de 1.6 milhões de registos de GPS e um histórico de registos de acidentes para construir um modelo que relaciona a mobilidade humana com o risco de acidente. Desta forma o modelo avalia o risco de acidente em tempo real através de uma classificação do risco de acidente para cada zona do mapa. Dados como a geo-localização do acidente e os níveis de mobilidade humana em tempo real mostraram-se essenciais. O autor refere também que existem muitos fatores que levarão a um acidente de trânsito, como o comportamento do motorista, o clima e as condições da estrada. Mas que, apesar de alguns estudos terem focado na correspondência entre acidente de trânsito e esses fatores, é muito difícil revelar a mudança dinâmica do risco de acidente apenas com esses fatores.

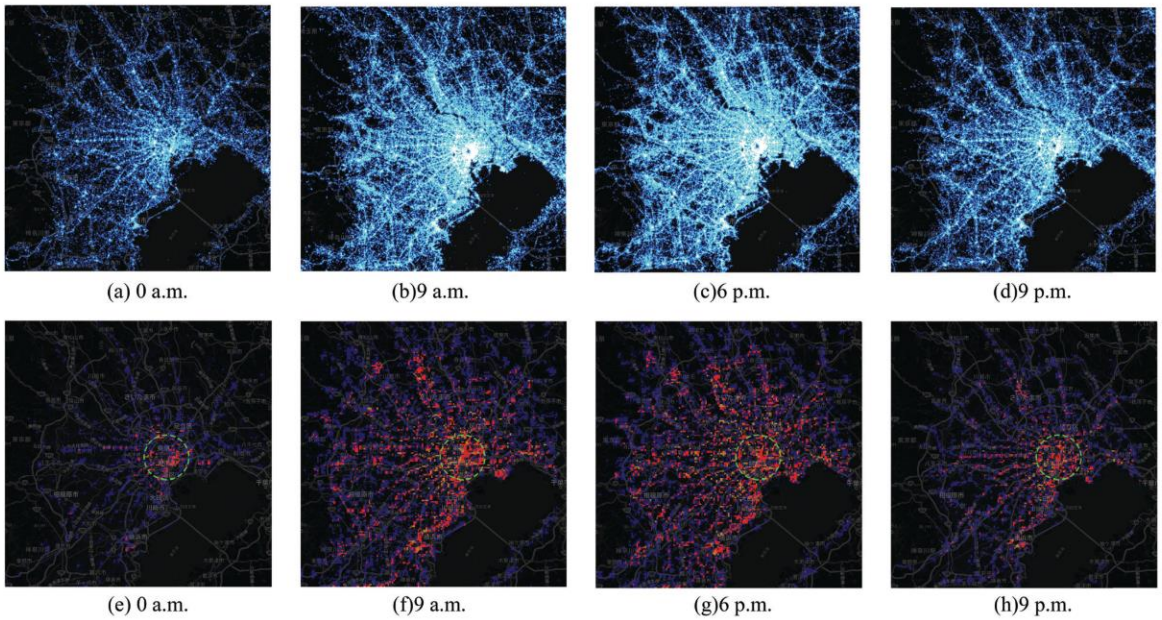

Figura 3 - Visualização dos resultados de simulações, retirado de [36] (figura 5). De (a) até (d) os níveis momentâneos de mobilidade humana, ou seja, as viaturas existentes nas estradas e a sua localização. De (e) até (h) o mapa de risco obtido pelo algoritmo de rede neural convolucional.

Ziakopoulos et al. [39] realizaram uma revisão de literatura relativa às diferentes abordagens que utilizam a heterogeneidade espacial dos acidentes na previsão de acidentes de viação. Daqui concluíram que a deteção de regiões problemáticas é uma característica crucial para a previsão de acidentes que é obtida através da análise da heterogeneidade espacial e que por isso seria vantajoso criar medidas de prevenção especificas para cada região.

Em 2018, Yuan et al. [40] realizaram um estudo de forma a conseguir prever o risco de acidente, de acordo com a hora, o local e o dia. Para isso foi utilizada uma abordagem de aprendizagem profunda que se baseia na heterogeneidade espacial e temporal dos dados, uma característica própria dos acidentes de viação. O estudo foi feito com dados do estado de Iowa, nos Estados Unidos da América e a amostra contém 375690 acidentes de 2006 a 2014. Para este estudo foram adicionadas bases de dados exteriores, tais como: dados relativos ao volume de transito, condições das estradas, dados de precipitação e temperatura ambiente de quatro bases de dados diferentes, ao longo de 8 anos. O algoritmo utilizado foi uma adaptação da rede neural convolucional de longa e curta memória e foi criado um software que cria um modelo preditivo para cada região do estado, porque foi concluído neste estudo que as principais causas de acidentes variam de região para região. O estado foi dividido em quadrados de 5km por 5km, criando assim as diferentes regiões, no entanto, para não se perder a ligação espacial dos acidentes, foi criado um grafo entre os acidentes , para desta forma se extrair informação relativa à proximidade entre acidentes. Cada entrada corresponde ao espaço temporal de um dia. Como resultado foram identificados 3 tipos de regiões que se assemelham entre si, que denominadas por: Região do tipo urbana, do tipo rural e do tipo mista. Para cada um dos tipos de região, as variáveis que se sobressaem como as principais causas de acidentes são diferentes. Nas regiões urbanas as variáveis que se sobressaíram foram a rede das estradas, o volume de trânsito, as condições das estradas e o calendário o que pode ser explicado pela grande densidade populacional,

que resulta em acidentes mais ligados à atividade humana. Nas regiões rurais e mistas, as variáveis que mais se sobressaíram foram a precipitação, a temperatura, a velocidade do vento e variáveis espaciais (localizações dos acidentes), o que pode ser explicado pela menor densidade populacional, que leva que os acidentes tenham uma ligação maior a causas naturais e não tanto à atividade humana [40].

Numa abordagem diferente, Yu et al. [41] defende que as típicas CNNs estão restritas a espaços euclidianos e por isso não são adequadas para dados em forma de grafo. Para abordar a predição de acidentes em forma de grafos foi proposta a utilização de redes convolucionais de grafos (GCN, do inglês: *Graph Convolution Network*) que podem ser categorizadas em dois tipos de métodos: espaciais e espectrais. Desta forma, no seu trabalho realizado em 2019, incorporaram as características espaciais e temporais na previsão de acidentes de viação. Para o estudo foram incluídas bases de dados relativas a ocorrências de acidentes, dados relativos ao fluxo de transito a partir de dados GPS de táxis, dados meteorológicos e dados relativos ao mapeamento das estradas, como o seu comprimento, identificação das estradas e existência cruzamentos ao longo das mesmas. Como modelo de predição, foi proposta uma nova Rede Convolucional de Grafo Espacial-Temporal Profundo que mostrou ter um grande sucesso comparativamente aos métodos tradicionais.

Resumindo, em relação aos métodos que usam redes neurais profundas, estes para além dos registos de acidentes usam também a localização dos mesmos e a quantidade de veículos a cada momento do dia (mobilidade humana). Assim é possível avaliar o risco em tempo real, ou num futuro próximo como em [36] e [39], respetivamente.

4. Metodologia

A metodologia proposta possui as etapas seguintes:

4.1. Seleção de Dados

Nesta primeira etapa do processo de descoberta de conhecimento em bases de dados, através da Junção Orientada, que significa que o analista deve entrevistar alguns especialistas da área, serão questionados alguns militares da área do trânsito de forma a selecionar as características relevantes numa participação de acidentes. Desta forma, ao excluir as características irrelevantes, é possível melhorar a precisão do algoritmo e também diminuir a complexidade do mesmo. Para além disso serão adicionados dados relativos a outras bases de dados como já referido anteriormente.

4.2. Pré-processamento

Os vários passos da etapa de pré-processamento estão ilustrados no fluxograma da Fig. 4.

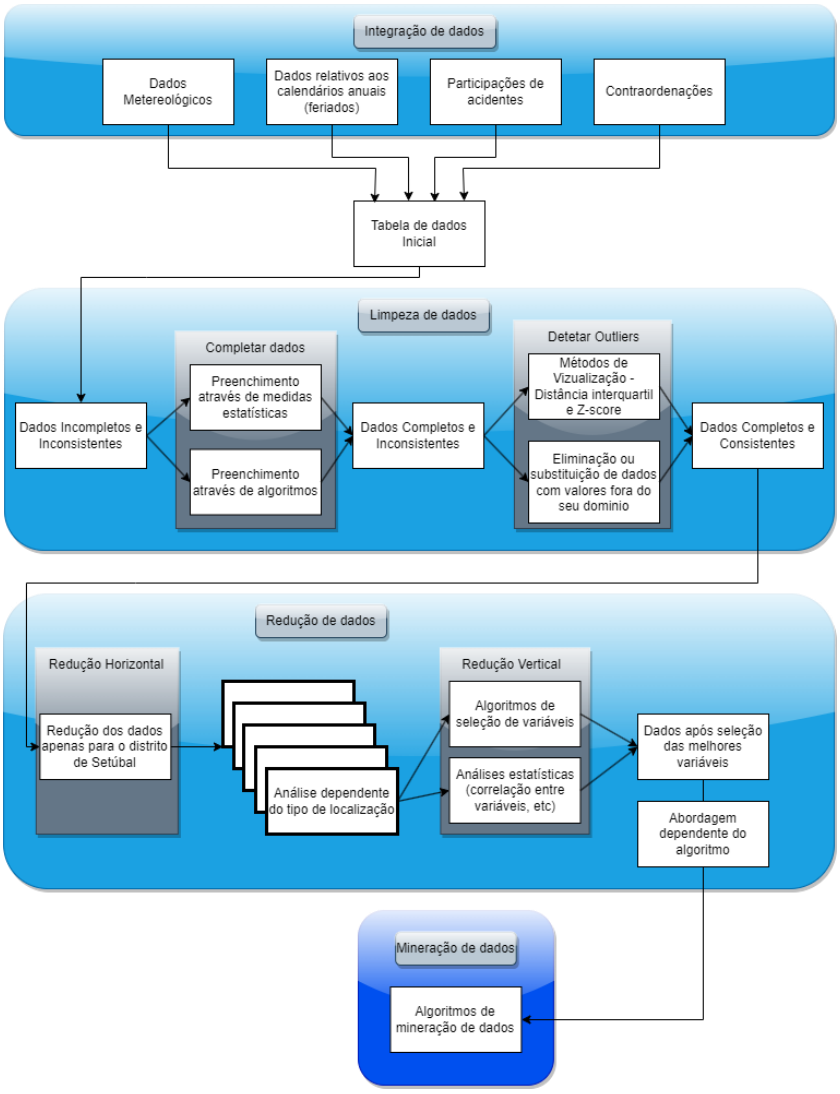

Figura 4 - Fluxograma das etapas de pré-processamento dos dados, a azul-claro: integração de dados, limpeza de dados e redução de dados. Que são prévias à etapa de mineração de dados, a azul-escuro.

Optou-se por realizar os vários passos na sequência proposta na Fig. 4. Inicialmente, na seleção de dados, foram agrupados os dados provenientes de diferentes fontes numa só tabela de dados.

Para a limpeza de dados, foram tratados os dados incompletos através do preenchimento com medidas estatísticas como a média e a moda, ou pela eliminação direta. Após tratar os dados incompletos, foram tratados os dados inconsistentes, retirando *outliers* por acidentes que tinham valores que não pertenciam ao domínio das variáveis, como localizações impossíveis, entre outros.

De seguida, na redução de dados, optou-se por realizar uma análise de características distinta para o tipo de região/estrada, devido a que para as diferentes regiões e estradas existiam diferentes fatores que podem influenciar o número de acidentes. A redução de dados foi feita através de uma análise de correlações entre as variáveis e através de algoritmos de seleção de características.

4.2.1. Integração dos dados

De forma a obter um modelo de classificação com melhor performance, foram integrados vários tipos de dados, entre os quais: as participações de acidente, contraordenações e crimes ocorridos nas estradas de Portugal Continental proveniente da base de dados da GNR; dados meteorológicos provenientes de estações de meteorologia que fornecem as suas bases de dados; dados relativos aos feriados anuais dos anos analisados na amostra;

Devido à relação temporal existente entre os acidentes, estes foram agrupados por hora do dia. Já em relação à heterogeneidade espacial, visto que nos dados fornecidos não foi possível obter a geo-localização dos acidentes e já que no estado da arte foi possível encontrar muitos trabalhos que fazem estudos individuais para uma certa estrada, os acidentes da base de dados da GNR foram divididos pelo concelho em que o acidente ocorreu e pela estrada em que ocorreu, no caso de ser uma autoestrada, itinerário ou estrada nacional. Assim, apesar de não ser o ideal devido às limitações dos dados fornecidos, poder-se-á realizar um estudo que relacione os acidentes com o concelho ou a estrada em que este ocorreu.

Por último, na integração de dados serão ainda integrados os dados fornecidos relativamente às contraordenações. Os dados que foram fornecidos estão representados na Tab. 7.

Tabela 7 - Dados fornecidos pela GNR relativamente ao número de fiscalizações e ao número de contraordenações no distrito de Setúbal agrupado por dia da semana

	Condutores Fiscalizados	Contraordenações Rodoviárias	Álcool-Testes	Álcool - Excessos	Percentagem De Contraordenações	Percentagem De Excessos De Álcool
1. Dom	2600	1423	727	31	0,54	0,042
2. Seg	3508	2118	1243	19	0,60	0,015
3. Ter	4005	2259	1061	16	0,56	0,015
5. Qua	3231	2237	1311	16	0,69	0,012
6. Qui	3281	2420	1165	17	0,73	0,014
7. Sex	3128	2192	958	12	0,70	0,012
8. Sáb	3477	1394	1082	24	0,40	0,022

Como é possível verificar apenas foram dadas informações relativas aos 3 anos em análise, sobre as fiscalizações realizadas, agrupadas em cada dia da semana. Assim decidiu-se que estes tipos de dados não seriam importantes já que a sua correlação com os dias da semana é total, logo o dia da semana contém a mesma informação para o modelo preditivo.

4.2.2. Limpeza dos dados

Já que para a maioria dos dados obtidos os valores não estavam incompletos nem apresentavam inconsistências, não foi necessário a utilização de algoritmos para completar dados em falta, nem se sentiu a necessidade de explorar extensivamente esta etapa. Para os poucos dados incompletos ou nulos encontrados, que continham valores de entrada em falta, estes dados foram eliminados diretamente ou, no caso do número de valores em falta não fosse muito elevado, o valor em falta foi substituído pela medida estatística mais indicada, ou seja, pela média no caso de valores numéricos e pela moda no caso de valores categóricos.

Já para dados com ruído ou inconsistentes foram utilizados métodos de visualização, ao criar, histogramas e diagramas de caixa para cada uma das características e ao observar se existem inconsistências nesses gráficos. Foram também eliminados dados com valores que não pertenciam ao domínio, como valores impossíveis para a localização.

4.2.3. Normalização e transformação dos dados

Esta etapa consistiu em trabalhar os dados de forma a obter melhores resultados na mineração. Foi feita uma codificação dos dados Categórica-numérica, que representa valores de características categóricas por valores numéricos. A técnica de codificação utilizada foi a *One-hot,* em que para uma certa variável com k categorias são criadas k-1 novas categorias, em que apenas uma dessas categorias pode ser igual a 1 para cada dado de entrada e uma das categorias é representada apenas por zeros de forma a prevenir problemas de Multicolinearidade. A maior parte das variáveis utilizadas são categóricas, exceto a variável alvo (o número de acidentes, que é uma variável numérica).

Após a codificação, foi feita uma normalização dos dados, em que se ajustam os valores de cada característica de forma que fiquem na mesma escala entre si, de forma que os algoritmos de mineração não considerarem valores mais ou menos elevados como mais ou menos importantes.

4.2.4. Redução dos dados

Como foi possível observar pela literatura, diferentes tipos de regiões apresentam diferentes tipos de características que influenciam os modelos. Assim, decidiu-se agrupar os dados a nível horizontal pela localização dos mesmos, em três conjuntos, dependendo dos limites de velocidade e características da estrada. Estes conjuntos são: **Autoestradas, Estradas Nacionais ou Itinerários e Municípios**.

Já para a redução de dados horizontal, ou seja, para a seleção das melhores características foram aplicados dois métodos:

1. Análise estatística, como por exemplo as correlações entre as diferentes características;

2. Algoritmos de seleção de características mencionados pela literatura, referidos na Tab. 5.

4.2.4.1. Redução através de análise estatística

Relativamente à análise estatística, optou-se por analisar a correlação entre as variáveis. Esta é uma relação estatística que envolve dependência entre duas variáveis. Frequentemente usamos o Coeficiente de Correlação de Pearson para calcular a correlação linear entre variáveis numéricas contínuas. No entanto, devemos usar uma métrica diferente para calcular a correlação entre variáveis categóricas, como é o caso do nosso conjunto de dados. A correlação de *V de Cramer* é usada para calcular a correlação entre variáveis categóricas nominais com mais de dois valores (não binárias) [9]. Devido às características dos nossos dados a métrica para o cálculo de correlações entre características categóricas foi o V de Cramer, definido como [42]:

$$\emptyset_c = \sqrt{\frac{X^2}{N(k-1)}} \quad ((1)$$

Onde:

- $\emptyset_c$ é o valor do V de Cramer;
- X^2 é o valor de chi-quadrado;
- *N* é o número de amostras;
- e *k* é o número de categorias da variável com o menor número de categorias.

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (2)$$

Onde:

- e_{ij} é o valor da frequência expectável
- o_{ij} é o valor da frequência observada de uma combinação de dois valores, um da variável *i*, outro da variável *j*;

$$e_{ij} = \frac{o_i \cdot o_j}{N} \quad (3)$$

Onde:

- e_{ij} é o valor da frequência expectável de uma combinação de dois valores, um da variável *i*, outro da variável *j*;
- o_i é a frequência marginal de um dos valores da variável *i*;
- o_j é a frequência marginal de um dos valores da variável *j*;
- N é número total de amostras

A interpretação de quão forte é a correlação entre duas variáveis categóricas foi feita a partir dos valores obtidos pelo V de Cramer. Esta interpretação foi feita através da Tab. 8.

Tabela 8 - Interpretação do valor de V de Cramer, adaptado de [43]

Valor de V de Cramer $\emptyset_c$	Interpretação
]0,25 ; 1,00]	Muito forte
]0,15 ; 0,25]	Forte
]0,10 ; 0,15]	Moderada
]0,05 ; 0,10]	Fraca
[0 ; 0,05]	Muito fraca

Já para a correlação entre características categóricas nominais e categóricas numéricos, o indicador utilizado foi o indicador de Kruskal Wallis. Este tem como objetivo verificar se existe uma diferença entre vários grupos independentes quando esses grupos não apresentam uma distribuição normal. No nosso caso o número de acidentes, quando agrupado pelas diferentes categorias das variáveis categóricas não apresenta uma distribuição normal, como é possível ver pelos exemplos dos histogramas das Figs. 5 e 6 que relacionam variáveis categóricas com a variável numérica do número de acidentes.

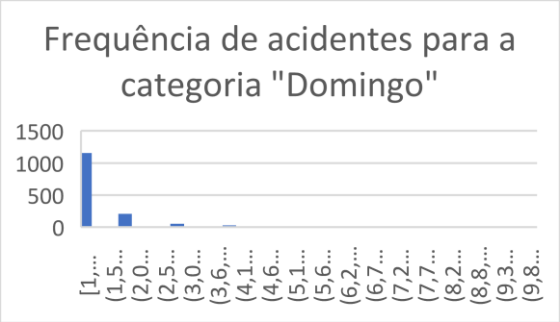

Figura 5 -Histograma com o número de acidentes no eixo X e com a frequência da respetiva quantidade de acidentes no eixo Y, para a categoria "Domingo"

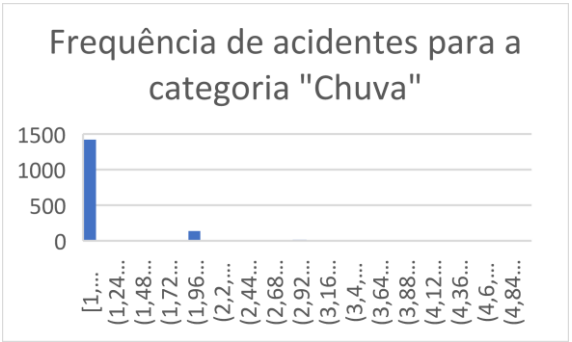

Figura 6 - Histograma com o número de acidentes no eixo X e com a frequência da respetiva quantidade de acidentes no eixo Y, para a categoria "Chuva"

Como foi explicado, o V de Cramer considera duas variáveis como independentes se o valor da frequência expectável, e_{ij} for igual ao valor da frequência observada, levando a que a probabilidade de duas das categorias ocorrerem seja dada pela multiplicação da probabilidade de cada uma. Já no

teste de Kruskal Wallis as variáveis serem independentes significa que a soma das classificações de todos os grupos/categorias tendam para o mesmo valor [9], [43], [44], [45].

Para ser possível obter um critério universal para eliminar características com valores de correlação baixos, é importante que todas as correlações calculadas sejam comparáveis. O Kruskal Wallis é equivalente ao chi-quadrado também utilizado no V de Cramer, pelo que os valores obtidos podem ser comparados entre as duas medidas. A expressão para o teste de Kruskal Wallis é dada por [44], [45]:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(r_{ij} - \bar{r}\right)^2} \quad (4)$$

- *N* é o número total de amostras ao longo de todos os grupos;
- *g* é o número de grupos;
- n_i é o número de amostras no grupo *i* ;
- r_{ij} é o valor da classificação da amostra *j* que pertence ao grupo *i* ;
- $\overline{r_i} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ é o valor médio da classificação de todas as observações *j* do grupo *i* ;
- $\bar{r} = \frac{1}{2}(N+1)$ é o valor médio do somatório de todas as classificações r_{ij}, ou seja, o valor esperado para a média de todos os grupos;

A interpretação do Kruskal Wallis é também dada pela Tab. 8 já que este equivale ao Chi-quadrado, tal como o V de Cramer. Podemos comparar facilmente a equação 4 do teste de Kruskal com o com a equação 2 e 3 do V de Cramer. O numerador da equação 4, $(\bar{r}_i - \bar{r})^2$, é equivalente ao numerador da equação 2, $(o_{ij} - e_{ij})^2$, pois ambos estão a medir a distância entre o valor expectável e o valor observado.

4.2.4.2. Redução através de algoritmos de seleção de características

Os algoritmos de seleção de características utilizados foram o RBA e o SBS.

Algoritmo de seleção RBA

No trabalho de Urbanowicz et al. [46] foi feita uma revisão extensa aos métodos de seleção de características baseados em RBA. Como um método de seleção de características, o RBA calcula uma estatística para cada característica que pode ser usada para estimar a qualidade ou a relevância do mesmo relativamente à característica alvo. Como valores de saída, o algoritmo retorna um vetor de pesos em que cada um desses pesos está associado a uma característica, ou seja, o vetor tem o mesmo tamanho que o número de características existentes. O peso atribuído a cada característica pode variar de -1 (pior) a +1 (melhor). O algoritmo de RBA original foi limitado a problemas de classificação binária. No entanto existem estratégias para estender o algoritmo para problemas de várias classes ou problemas de regressão.

Tabela 9 - Pseudocódigo para o algoritmo original de RBA, adaptado de [46]

Parâmetros:
n - número de amostras
A – número de características
m – número de amostras de treino, menor que n, que servirão para atualizar o vetor de pesos W
Algoritmo:
Inicializar vetor com peso de características a 0, W[A] = 0
de i = 1 **até** m **fazer**
    selecionar uma amostra alvo, R_i
    encontrar a amostra mais próxima 'certa', C, e a amostra mais próxima 'falhada', F
    de a = 1 **até** A **fazer**
        $W[a] = W[a] - diff(a, R_i, C)/m + diff(a, R_i, F)/m$
    fim de ciclo
fim de ciclo
retorna o vetor W com as pontuações para cada característica que estima a qualidade da respetiva característica

---

Este algoritmo original tinha como objetivo problemas de classificação binário. Conforme resumido pelo pseudocódigo descrito na Tab. 9, o algoritmo original RBA percorre **m** amostras de treino aleatórias, onde **m** é um parâmetro definido pelo utilizador. A cada ciclo, $\mathbf{R_i}$, uma das amostras de **m**, é definida como amostra alvo. Através desta amostra alvo o vetor de pesos das características, **W**, é atualizado com base nas diferenças entre a amostra alvo e duas amostras vizinhas mais próximas, mencionadas no pseudocódigo como **C** e **F**. Estas duas amostras vizinhas mais próximas, são as amostras com mais semelhança com a amostra alvo, ou seja, com o maior número de características com o mesmo valor, mas enquanto **C** corresponde à amostra mais próxima com a mesma classificação do que a amostra alvo, **F** corresponde à amostra mais próxima com uma classificação contrária à amostra alvo.

A última etapa do ciclo é um novo ciclo que percorre o vetor dos pesos das características, **W**, e atualiza o peso de uma característica **W[A]**, se o valor da característica difere entre **Ri** e **C**, ou entre **Ri** e **F**.

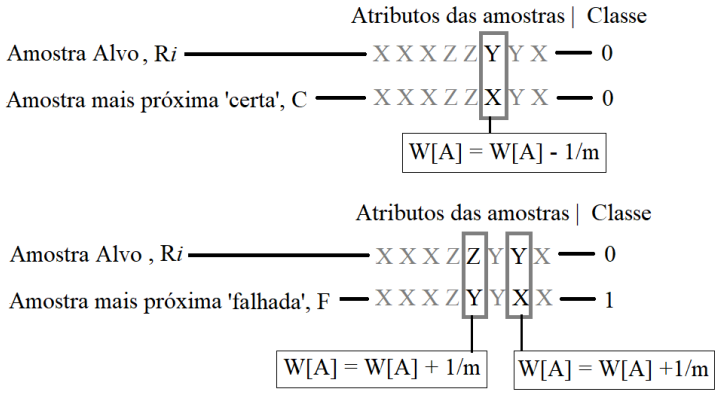

Figura 7 – Passo do algoritmo original RBA em que, para uma determinada amostra alvo Ri, é feita a atualização do vetor de pesos W[A] para as diferentes características que apresentem valores diferentes em relação às amostras mais próximas C e F. Neste exemplo, as características são discretas com valores possíveis de X, Y ou Z, e a classe alvo é binária com valores possíveis de 0 ou 1.

As características que têm um valor diferente entre **Ri** e **F** suportam a hipótese de que são características que acrescentam informação relativamente à classe alvo, de modo que a estimativa de qualidade da característica **W[A]** é aumentada. Por outro lado, características com diferenças entre **Ri** e **C** fornecem evidências do contrário, de modo que a estimativa de qualidade da característica **W[A]** é diminuída. A função *diff* no da Tab. 8 calcula a diferença no valor da característica **A** entre duas amostras **C** e **F**. Para **C** a equação *diff* e é definida da seguinte maneira:

$$diff(A, Ri, C) = \begin{cases} 0\ se, valor(A, Ri) = valor(A, C) \\ 1, caso\ contrário \end{cases} \quad (5)$$

Para a amostra **F**, bastaria substituir **F** no lugar de **C**. Assim, se para a mesma característica, ambas as amostras têm o mesmo valor, a essa característica é feita uma subtração ou soma ao peso da característica, dependendo de se tratar da amostra **C** ou **F**, respetivamente [46].

Para este trabalho, de forma a abordar o problema de regressão já que a variável alvo é o número de acidentes, o algoritmo utilizado foi a variante do RBA com aplicação em problemas de regressão, a que é dada o nome de RReliefF [47]. Em problemas de regressão a característica alvo é contínuo, portanto a comparação entre as classes das instâncias não pode ser utilizada. Para resolver este problema, tal como explicado por Sikonja et al. [47], em vez de tentar perceber se uma amostra tem a mesma classificação que outra amostra, de modo a determinar se devemos somar ou subtrair como demonstrado na Fig. 7, foi criada uma função probabilistica que indica qual a distância que a classificação de uma amostra está da classificação de uma outra amostra. Esta função é modelada a partir da distância relativa entre as classificações das duas amostras.

Algoritmo de seleção SFS e SBS

Relativamente ao segundo algoritmo de seleção de características utilizado, o SBS, este é um procedimento de pesquisa simples, com uma abordagem dependente do modelo, que como se viu no

enquadramento teórico, consiste em executar o algoritmo de mineração iterativamente ao mesmo tempo que se adiciona ou retira características, de forma a verificar quais as combinações de características que resultam em melhores predições. O SBS é uma variante do algoritmo SFS. O SFS inicia a partir de um conjunto vazio de características e gradualmente adiciona características selecionados por uma medida de performance, que mede quanto é que cada característica melhora ou piora um médodo de mineração. A cada iteração, a característica a ser incluída no conjunto de características é selecionada entre as características ainda disponíveis do conjunto. O algoritmo adiciona características com base numa métrica de desempenho de classificação/regressão definida pelo utilizador [48].

Tabela 10 – Pseudocódigo do algoritmo de seleção de características SFS, adaptado de [49]

Parâmetros:

X . número total de características

$S(X)$ – subconjunto de características pertencentes a X, com um número de características menor que X

J – medida de performance para um conjuntos de características (através de uma classificação/regressão) definida pelo utilizador

X' – subconjunto de características final

Algoritmo:

$X' = \acute{O}$ //conjunto vazio

Fazer enquanto existir melhorias na medida de performance J **ou** $X' = X$

$$x' = argmax\{ J(S(X' \cup x') \mid x \in X \setminus X'\}$$

$X' = X' \cup \{x'\}$

fim de ciclo

retorna X'

Como é possível verificar o algoritmo inicia com o subconjunto de características vazio ***X' = Ø*** e vai acrescentando iterativamente a característica que mais contribuição der para uma melhoria da medida de performance ***J*** utilizada, ***J(S(X' U x'))***. O algoritmo atinge a condição de paragem quando não existir melhorias na medida de performance ***J*** ao introduzir uma nova característica, ou quando o número de características no subconjunto for igual ao conjunto total de características. O SFS é amplamente utilizado pela sua simplicidade e velocidade [18].

Uma das variantes deste método é o Sequential Backward Selection (SBS) representado na Tab. 11. Existem ainda outras como o Sequential Forward Floating Selection (SFFS) e o Sequential Backward Floating Selection (SBFS). SFFS e SBFS, podem ser considerados como extensões dos algoritmos SFS e SBS [49]. No entanto, optou-se por utilizar o SBS pela sua simplicidade e o pseudocódigo representado na Tab. 11. Neste caso foi utilizada a rede neural para criar um modelo com os diferentes conjuntos de características e foi utilzada a medida de performance MAPE. Esta opção deve-se ao facto de a rede neural ter sido o algoritmo de mineração com melhor performance, dentro dos algoritmos de mineração aplicados.

Tabela 11 - Pseudocódigo de uma variante algoritmo de seleção de características SFS, o SBS. Adaptado de [49].

Parâmetros:
X . número total de características
$S(X)$ – subconjunto de características pertencentes a X, com um número de características menor que X
J – medida de performance para um conjuntos de características (através de uma classificação/regressão) definida pelo utilizador
X' – subconjunto de características final
Algoritmo:
$X' = X$ //conjunto completo
Fazer enquanto existir melhorias na medida de performance J **ou** $X' = \acute{O}$

$$x' = argmax\{ J(S(X' \setminus x') \mid x \in X'\}$$

$X' = X' \setminus \{x'\}$
fim de ciclo
retorna X'

Como podemos ver pela Tab. 11, o SBS é bastante semelhante ao SFS, no entanto tem uma abordagem simétrica, de cima para baixo, iniciando com todas as características no subconjunto ***X' = X***, e retirando iterativamente a característica que menos contribui para a performance da medida ***J,*** *em* ***J(S(X' U x'))***. O algoritmo atinge a condição de paragem quando não existir melhorias na medida de performance ***J*** ao retirar uma nova característica ou quando o número de características no subconjunto for igual ao vazio.

4.3. Mineração de dados

Como medida de performance para avaliar os diferentes algoritmos mineração, optou-se por utilizar o erro absoluto médio (do inglês: *Mean Absolute error, MAE*) que é uma medida de erro que soma o erro absoluto entre as observações e o valor obtido pelo modelo. É dado pela equação seguinte:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\overline{y_j} - y_j\right| \quad (6)$$

Optou-se por utilizar o MAE e não o erro médio quadrático (MSE, do inglês: *mean squared error*) devido à distribuição do número de acidentes. Visto que para o intervalo de tempo escolhido a quase totalidade do valor dos acidentes é baixo, não se quer penalizar o erro ao quadrado pois os valores altos correspondem a acontecimentos excecionais. Para além do MAE, optou-se por utilizar também o erro absoluto médio por percentagem (do inglês: *Mean Absolute Percentage Error*), que é dado pela seguinte equação:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (7)$$

Em que A_t é o valor real e F_t é o valor obtido através do modelo. Através deste modelo de erro podemos ter uma ideia da percentagem de erro média que existe para cada predição.

Sendo que o objetivo é descobrir o risco de acidente e não prever o valor exato de acidentes, os valores previstos e os valores reais serão agrupados em três grupos de risco: **baixo, médio, elevado.** A escolha do intervalo em que se inserem os valores de cada um destes agrupamentos será feita a partir da análise do **diagrama em caixa** da frequência de acidentes, em que é possível dividir os dados pelos respetivos quartis do diagrama em caixa. Após este agrupamento podemos obter uma medida de performance relacionada com classificação. Optou-se por utilizar a **exatidão**, referida no enquadramento teórico.

$$\frac{vp + vn}{vp + fp + vn + fn} \qquad (8)$$

A explicação desta equação encontra-se no capítulo anterior.

Para o modelo de mineração serão utilizados os seguintes algoritmos: **Regressão Linear, Regressão Lasso, Regressão Ridge, Regressão de árvore de Decisão, Regressão com K-vizinhos mais próximos** (KNN, do inglês: *K-nearest neighbors*) **e Regressão através de Rede neural**.

Segundo Alhamzawi et al. [50] o modelo padrão de regressão linear múltipla pode ser escrito da seguinte forma:

$$y = X.\beta + \epsilon \qquad (9)$$

Onde *X* é uma matrix de *n* por *k* em que *n* é o número de amostras e *k* o número de características para cada amostra. β é o vetor correspondente de coeficientes de regressão. O objetivo é descobrir o vetor β que resulte no vetor y mais próximo de y real. Assim, o objetivo é descobrir o mínimo de:

$$\min_{\beta}(y - X\beta)'(y - X\beta) \qquad (10)$$

O tamanho dos coeficientes aumenta exponencialmente com o aumento da complexidade do modelo. À medida que a complexidade do modelo aumenta, os modelos tendem a ajustar desvios ainda menores no conjunto de dados de treinamento que leva a um sobre ajustamento do modelo (do inglês: *overfitting*). Assim, ao colocar uma restrição na magnitude dos coeficientes pode reduzir a complexidade do modelo. Daí surge a regularização de Lasso [50]. Dada pela seguinte equação:

$$\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |B_j|, \qquad \lambda > 0 \qquad (11)$$

Para a regressão de Lasso, a otimização penaliza o somatório do valor absoluto dos coeficientes B_j, levando a que estes sejam mais pequenos e assim reduzindo a complexidade do modelo obtido. Já para a regressão de Ridge, esta penaliza o quadrado dos coeficientes e é dada pela seguinte equação:

$$\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} {\beta_j}^2, \qquad \lambda > 0 \tag{12}$$

Estes algoritmos são de simples aplicação e optou-se pela utilização dos mesmos apenas para comparação com modelos mais complexos.

As árvores de decisão conseguem resolver problemas não lineares porque dividem o espaço das características em vários subespaços através de condições. As árvores de decisão separam os dados através de condições nos diferentes nós intermédios. Os nós folha representam uma certa classe. A definição dos vários nós intermédios que contêm as condições que definem em que nó folha é que uma nova amostra será classificada, são definidos através do maior ganho de informação possível. A equação para o cálculo do ganho de informação é dada pela equação seguinte:

$$Ganho\ de\ Informação = Entropia(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropia(i)) \tag{13}$$

Em que *p* é o conjunto do nó atual e *i* são os diferentes subconjuntos obtidos para uma certa condição de separação do nó *p*.

Enquanto a entropia para um problema de classificação é dada por:

$$Entropia = \sum - p_i \log(p_i) \tag{14}$$

Para um problema de regressão em vez da entropia é utilizada a redução da variância:

$$Var = \frac{1}{n} \sum (y_i - \bar{y})^2 \tag{15}$$

E assim, o ganho de informação passa a ser dado por:

$$Ganho\ de\ Informação = Var(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Var(i)\right) \tag{16}$$

Para um problema de regressão em vez de ser atribuída uma classe à nova amostra, é atribuído o valor médio de todas as amostras do grupo de treino a qual corresponde o nó folha em que a nova amostra foi colocada [51].

Já o algoritmo KNN também é popularmente usado para problemas de regressão não linear. KNN assume que o novo ponto de dados é semelhante aos pontos de dados existentes. O novo ponto de dados é comparado às amostras existentes. Para o caso da regressão, o valor médio dos k vizinhos mais próximos é tomado como o valor previsto para a nova amostra. Os vizinhos nos modelos KNN recebem um peso específico que define sua contribuição para o valor médio. Os pesos podem ser definidos como iguais ou podem ser diferentes quando se quer ter em conta a que distância está cada um dos vizinhos mais próximos, de modo a alterar o seu peso na classificação da nova amostra. O

valor de K a ser utilizado pode ser procurado iterativamente [52]. Um exemplo para um valor de k=3 pode ser observado na Fig. 8.

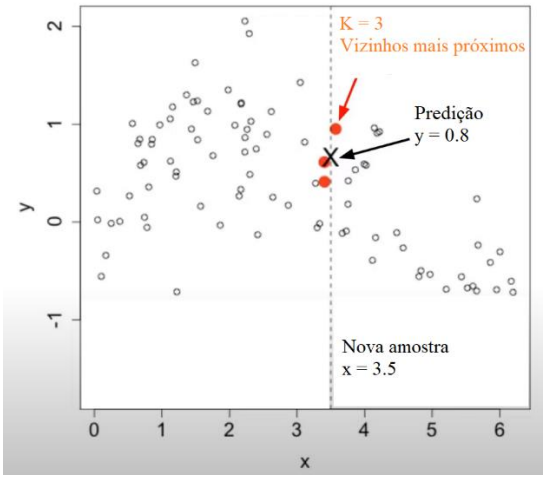

Figura 8 - Exemplo de aplicação de KNN para um problema de regressão, em que através dos 3 vizinhos mais próximos de x=3.5, é obtido o valor de y, através da média dos y's dos 3 vizinhos mais próximos.

As Redes Neurais Artificiais são estruturas que contêm nós. A sua arquitetura pode ser dividida em três tipos de camadas: de entrada, ocultas e de saída. A camada de entrada tem um número de nós correspondente às variáveis de treino. O número de camadas ocultas pode ser maior que um e pode ter um número arbitrário de nós. Por último, a última camada, de saída, tem o número de classes ou valores possíveis para a variável que queremos prever. Cada camada é constituída por um certo número de nós. e, para as redes tradicionais, a ligação entre nós é feita de camada em camada.

Ao treinar Redes Neurais Artificiais existem duas fases principais: propagação para frente e propagação para trás. Propagação para a frente é o processo de multiplicar pesos com cada variável de entrada no nó e adicioná-los. Supondo que para um certo nó existem D ligações de outros nós da camada anterior. Supondo que x representa o vetor de nós conectados a um certo nó da camada seguinte, w representa o vetor de pesos para cada um dos valores de x então o valor do nó da camada seguinte é dado pela seguinte equação [53].

$$z(x) = w.x = \sum_{i=1}^{D} w_i x_i \qquad (17)$$

Após este cálculo o valor de saída do nó z(x) passa por aquilo a que se dá o nome de uma função de ativação, que pode ser diferente dependendo do objetivo, como por exemplo ativação linear, por sigmoide, tangente hiperbólica, max-pooling, entre outros. Sendo g a função de ativação, esta é dada pela seguinte equação [53].

$$g(z(x)) \qquad (18)$$

Estas equações são aplicadas nó a nó até à camada de saída, sendo que, tradicionalmente, todos os nós de uma camada estão ligados aos nós da camada seguinte. Já propagação para trás é o processo de atualização dos pesos no modelo. A propagação para trás requer uma função/algoritmo

de otimização e uma função de perda, em que a função/algoritmo de otimização otimiza os pesos w de forma a reduzir o valor da função de perda [53].

Tendo como objetivo a aplicação de uma classificação supervisionada, inicialmente é realizada a divisão dos dados em grupos de treino e grupos testes com os dados de 2019 a 2021. O conjunto de dados fornecido contém características/variáveis maioritariamente categóricas. O classificador será capaz de indicar a probabilidade de ocorrer acidentes num certo distrito ou estrada, de acordo com as seguintes entradas: data, hora, fatores atmosféricos, calendário (feriados), rua/distrito. No entanto, o objetivo é que, por parte do utilizador, este apenas necessite de inserir as seguintes entradas: data, hora do dia, tipo de estrada/ distrito. Com esta entrada por parte do utilizador, será possível obter os restantes dados necessários através de uma aplicação. Por exemplo, ao obter a localidade e a data para as quais se pretende prever o a probabilidade/risco de acidente, o objetivo é que o programa consiga obter os dados meteorológicos relativos a essa data e nessa localidade. O programa também deverá verificar se a data corresponde a um feriado. A obtenção destes dados que não são fornecidos pelo utilizador será possível através da criação do serviço Web. Os pedidos GET que correspondem a um método HTTP[1], frequentemente utilizado em serviços Web RESTful, permitem obter recursos de outras páginas Web, como por exemplo a previsão de dados meteorológicos, entre outros.

4.4. Criação de um serviço Web[2] através de uma arquitetura REST[3]

O esquema para o sistema proposto está ilustrado na Fig. 4. Como é possível ver na figura, serão realizadas as várias etapas do processo de KDD referidas nos pontos anteriores. Após essas etapas o conhecimento obtido pelos algoritmos de mineração é, por fim, comunicado com o cliente através do serviço Web contruído a partir de uma arquitetura REST.

[1] Os métodos HTTP mais usados são: POST, GET, PUT, PATCH e DELETE. Correspondem às operações de criação, leitura, atualização e exclusão, respetivamente.

[2] Um serviço Web é um software desenhado para suportar interação entre duas ou mais máquinas numa rede de internet.[55]

[3] REST (representational state transfer) é uma arquitetura que estabelece regras na construção de um serviço Web. Existem alguns tipos de serviço Web, no entanto, a arquitetura REST baseia-se na troca de recursos (páginas HTML, imagens, vídeos, documentos e outros ficheiros). [55]

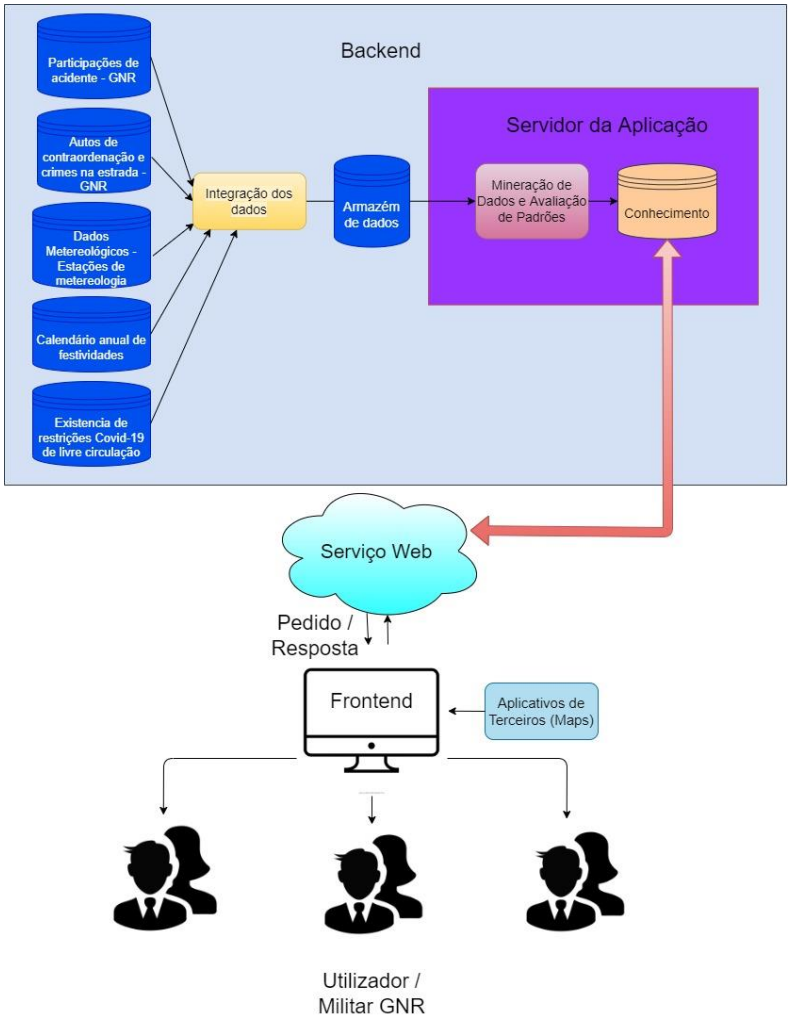

Figura 9- Esquema para o sistema de descoberta de conhecimento em bases de dados proposto, com a utilização de um serviço Web.

Integrando na página do cliente aplicativos como o Maps, será possível indicar o risco de acidente para a localização e data selecionada pelo utilizador através de um mapa das estradas com níveis de cor correspondentes aos níveis de risco de existir um novo acidente. Desta forma o utilizador, que será um militar da GNR, poderá decidir se é necessário realizar uma operação de fiscalização na data e na localização selecionada.

5. Resultados e Discussão

Neste capítulo são apresentados e analisados os resultados produzidos pela metodologia apresentada no capítulo anterior. Nas tarefas em que foram apresentadas mais do que uma técnica, estas são alvo de comparação, de forma a eleger a técnica que mais se adequa para a realização da tarefa em causa.

O presente capítulo está dividido em 3 Secções, iniciando-se pela apresentação da base de dados dos acidentes fornecida pela GNR para o distrito de Setúbal. Na secção seguinte apresenta-se um estudo comparativo das diferentes técnicas utilizadas para as tarefas de seleção de características e uma análise de correlação entre as diferentes características. Depois, é avaliado o desempenho dos diferentes algoritmos de mineração de dados, de forma a selecionar o modelo com menor erro na medida de performance. Ainda na última secção é efetuada a avaliação do desempenho geral da metodologia proposta.

5.1. Base de Dados

Relativamente à seleção de dados, na Tab. 12 estão indicadas todas as características selecionadas das participações de acidentes. Devido às limitações de tempo para a realização de entrevistas com especialistas da área, como mencionado no enquadramento teórico, optou-se por selecionar as variáveis de acordo com o estudo de J. Costa et. Al. [56]. Assim obtiveram-se as seguintes variáveis:

Tabela 12 – Características selecionadas da base de dados da GNR relativa às participações de acidentes.

Característica	Tipo de dado	Valores Possíveis
Identificação do acidente	Numérico	
Data	Data	
Hora	Hora	
Tipo de Local	Booleano	Dentro ou fora de localidade.
Localização	Acidente - Localização	Nome da localidade ou identificação da estrada. (Montijo, A2, EN-252, Almada, entre outros)
Tipo de acidente	Acidente - tipo acidente	Só com danos materiais; com vítimas
Dia da semana	Acidente - dia da semana	Segunda-feira; Terça-feira; Quarta-feira; Quinta-feira; Sexta-feira; Sábado; Domingo
Feriado	Booleano	
Álcool	Numérico	Percentagem de veículos fiscalizados com álcool
Contraordenação	Numérico	Percentagem de veículos fiscalizados que cometeram contraordenação
Fatores atmosféricos	Acidente - fatores atmosféricos	Bom tempo; Chuva; Vento forte; Nevoeiro; Neve; Nuvem de fumo; Granizo

A recolha dos dados feita pelo utilizador pode muitas vezes conter erros. De forma a conhecer melhor os dados e a verificar a sua qualidade pode ser feita uma comparação com algumas informações encontradas na literatura relativas a cada uma das características dos nossos dados. Para

algumas das características da Tab. 12, foi analisada a sua distribuição, através de diferentes tipos de gráficos.

Grupo Hora

Devido à relação temporal existente entre os acidentes, estes foram agrupados por hora do dia, tal como no trabalho de Ren et al. [54], em que o agrupamento de acordo com o estilo de vida dos Chineses foi feito nos diferentes grupos: 00:00–06:59 (meia-noite até ao amanhecer), 07:00–08:59 (transito matinal), 09:00–11:59 (horas de trabalho parte da manhã), 12:00–13:59 (pausa de almoço), 14:00–16:59 (horas de trabalho durante a tarde), 17:00–19:59 (pico de trânsito após trabalho), e 20:00–23:59 (horário noturno).

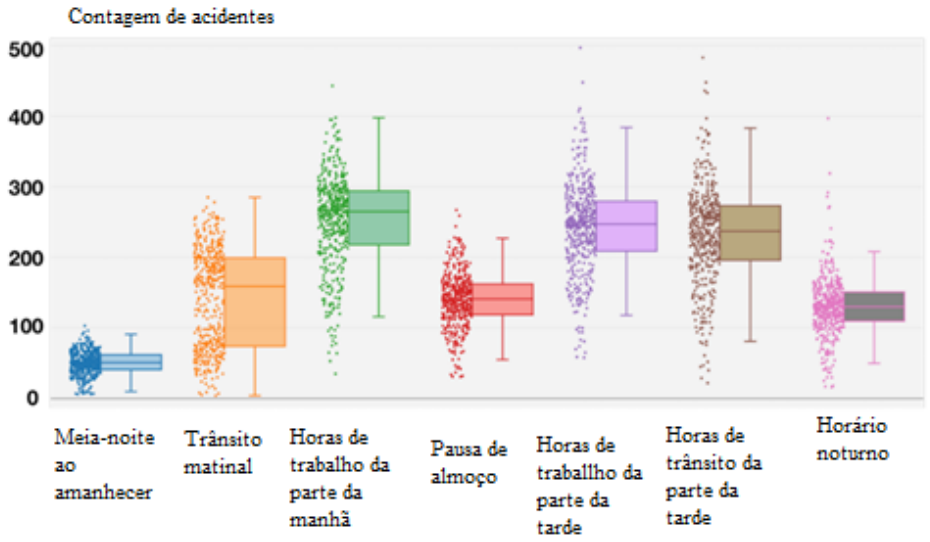

Figura 10- Diagrama de dispersão e diagrama em caixa da frequência de acidentes ocorridos nos diferentes intervalos de tempo em Beijing, nos diferentes dias do intervalo de tempo analisado. Adaptado de Ren et al. [54]

Como seria de esperar, a hora com maior afluência de trânsito, quotidianamente designada por hora de ponta, é, sempre, a altura do dia com mais trânsito. No entanto, é possível verificar que, em Beijing, os intervalos com maior número de acidentes nem sempre são nas horas de maior trânsito. Como se vê na Fig. 10, nas horas de trabalho de manhã, de tarde e nas horas de maior trânsito, apenas da parte da tarde, é quando existem mais acidentes.

O mesmo estudo pode ser aplicado aos nossos dados, segundo [57], os períodos entre as 8h e as 10.30h da manhã é bastante congestionado. Mais tarde, o intervalo entre as 17 e as 20h também tem uma afluência maior de viaturas. Assim, os intervalos foram definidos de forma semelhante a Ren et al. [54], mas adaptado ao estilo de vida português, nos seguintes: 00:00–07:59 (meia-noite até ao amanhecer), 08:00–09:59 (transito matinal), 10:00–12:29 (horas de trabalho parte da manhã), 12:30–13:59 (pausa de almoço), 14:00–16:59 (horas de trabalho durante a tarde), 17:00–19:59 (pico de trânsito após trabalho), e 20:00–23:59 (horário noturno).

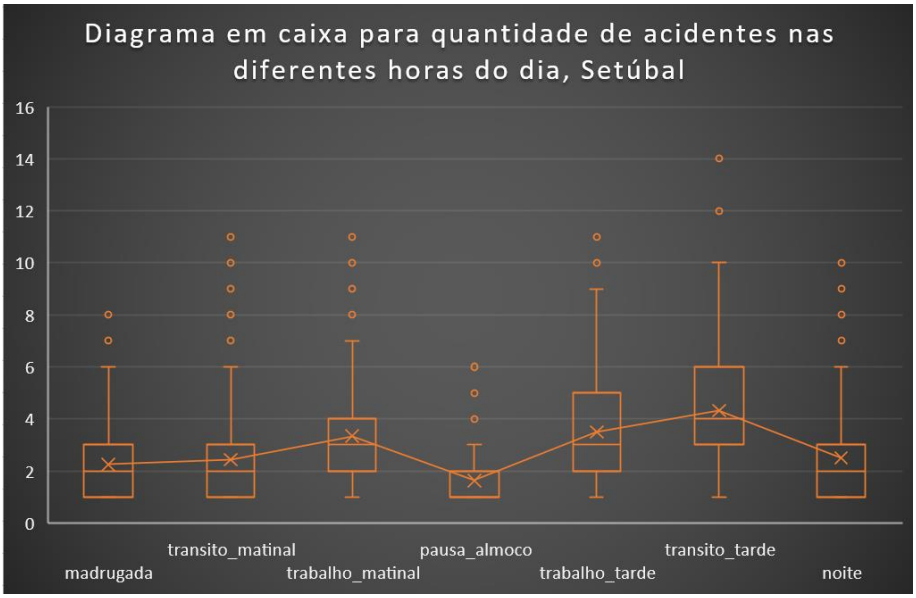

Figura 11 - Diagrama em caixa da frequência de acidentes ocorridos nos diferentes intervalos de tempo em Setúbal, para o intervalo de tempo analisado 2019-2021 (dados fornecidos pela GNR).

Como podemos verificar pela comparação dos diagramas em caixa da Fig. 10 e 11, apesar da frequência de acidentes de Setúbal ser muito mais reduzida do que o número de acidentes em Beijing, os três grupos de horas onde existem mais acidentes são idênticos. Neste caso os grupos de hora de trabalho matinal, trabalho de tarde e trânsito de tarde.

Podemos assim verificar que os nossos dados são coerentes com os dados encontrados na literatura relativamente às horas de trânsito com maior frequência de acidentes em Beijing.

Tipo de local

O tipo de local, isto é, se o acidente ocorreu dentro ou fora de uma localidade, foi agrupado por mês, ano e tipo de local.

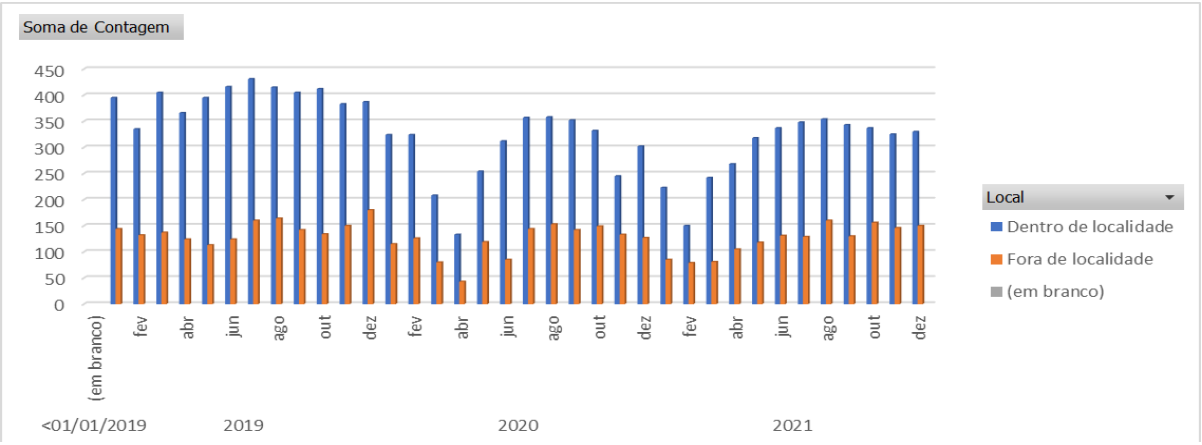

Figura 12 - Agrupamento de acidentes de acordo com a data e se aconteceu fora ou dentro de uma localidade

Como é possível verificar na Fig. 12, o número de acidentes que ocorre dentro das localidades é constantemente superior ao número de acidentes que ocorre fora de localidades, como também foi

indicado pelos estudos mencionados na revisão de literatura. Também é possível verificar que em Setúbal o maior número de acidentes ocorre sempre no mês de julho, que corresponde ao mês com maior quantidade de acidentes segundo [58], que justifica este facto com alguns fatores como: a existência de mais ciclistas, peões e motociclistas nas estradas durante o verão, a existência de maior número de obras nas estradas por ser um período mais favorável para as mesmas e por ser uma época do ano mais propensa a avarias nas viaturas, devido ao calor.

Dia da semana

Relativamente ao dia da semana os dados foram agrupados por dia da semana e por ano, como se pode verificar na Fig. 13.

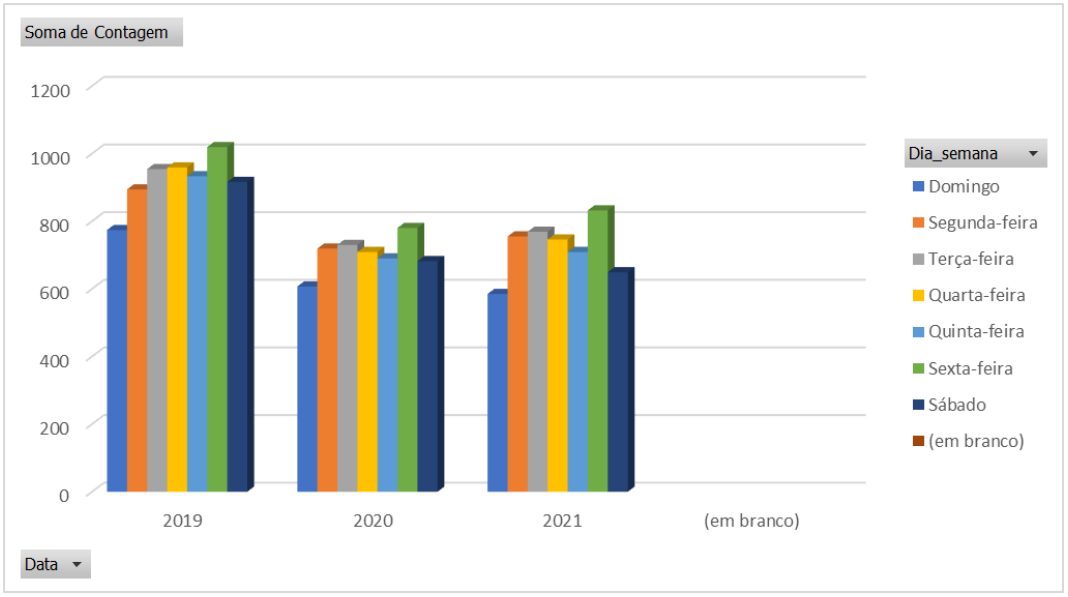

Figura 13 - Agrupamento do número de acidentes por dia da semana e por ano

Podemos verificar através da análise da Fig. 13 os dias da semana em que ocorrem mais acidentes ao longo dos três anos no distrito de Setúbal. O dia da semana em que ocorreram mais acidentes foi sempre à sexta-feira e com menos acidentes, sábado e domingo. Visto que a sexta-feira é o dia da semana em que frequentemente existe mais congestionamento [57], esta observação da Fig. 13 prova que os dados são coerentes, pois confirma que o volume de trânsito é um fator influenciador do número de acidentes.

Fatores atmosféricos

Os fatores atmosféricos foram agrupados por mês, ano e condição atmosférica presente no momento do acidente.

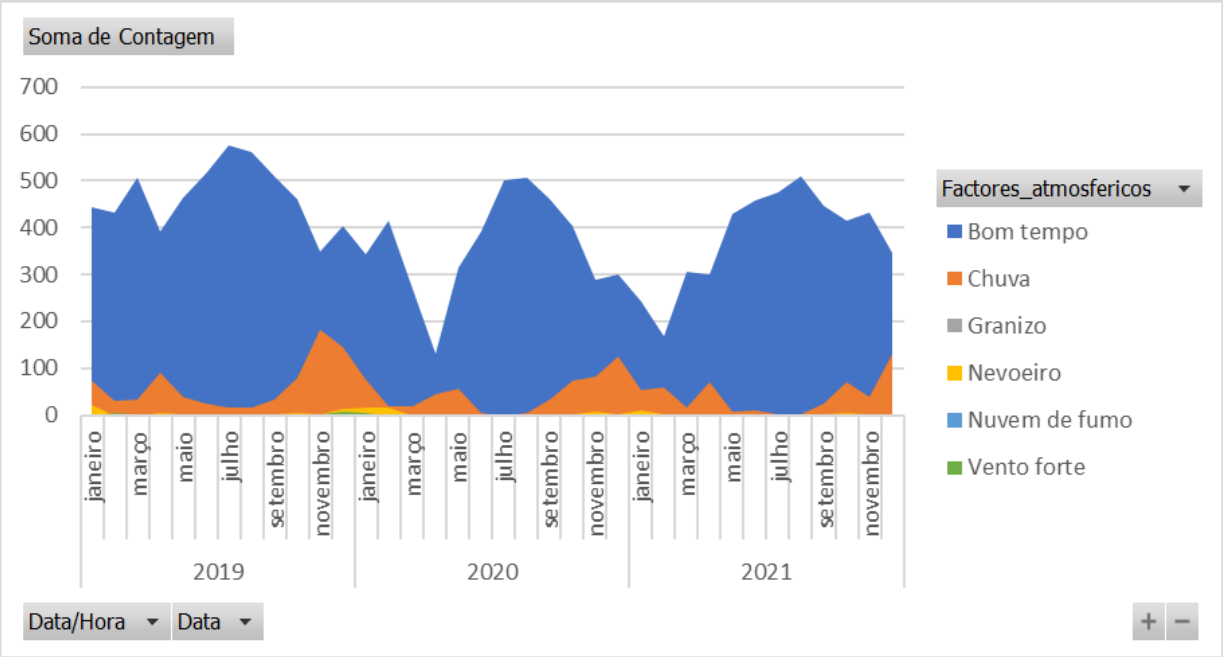

Figura 14 - Agrupamento do número de acidentes por mês, ano e tipo de condição atmosférica. Gráfico acumulativo.

Como se pode observar na Fig. 14, utilizou-se um gráfico com acumulação de acidentes por fator atmosférico. O maior número de acidentes ocorre quando está bom tempo pois este é o fator atmosférico que existe na maior parte do tempo. Em segundo lugar, o maior número de acidentes ocorre quando está chuva e de seguida quando está nevoeiro. No entanto isto não significa que o bom tempo seja a condição atmosférica que mais influência a ocorrência acidentes, pois deve ser considerada a probabilidade de cada uma das condições atmosféricas existir. Considerando a probabilidade de chover como $P(C)$ e a probabilidade de haver acidente como $P(A)$, o gráfico da Fig. 14 dá-nos a probabilidade de chover dado que houve acidente, ou seja, $P(C|A)$.

Pretende-se comparar a probabilidade de haver acidente sabendo que choveu $P(A|C)$ com a probabilidade de haver acidente sabendo que está bom tempo $P(A|B)$. Para realizar esta comparação utilizou-se o mês de dezembro. Assim, $P(C|A)$ para o mês de dezembro é dado por:

$$P(C|A) = \frac{144+126+132}{404+300+346} = 0.38 \qquad P(B|A) = 1 - P(C|A) = 0.62 \tag{19}$$

Como pretendemos confirmar que $P(A|B) < P(A|C)$, simplificou-se o cálculo da probabilidade de estar bom tempo, por excesso, como todos os dias em que não chove, desconsiderando o nevoeiro, granizo e vento forte.

Para obter a probabilidade de chover, podemos verificar através de informação obtida pelo *Weather Spark* que os níveis de precipitação para o distrito de Setúbal são as indicadas na Fig. 15 e na Tab. 13.

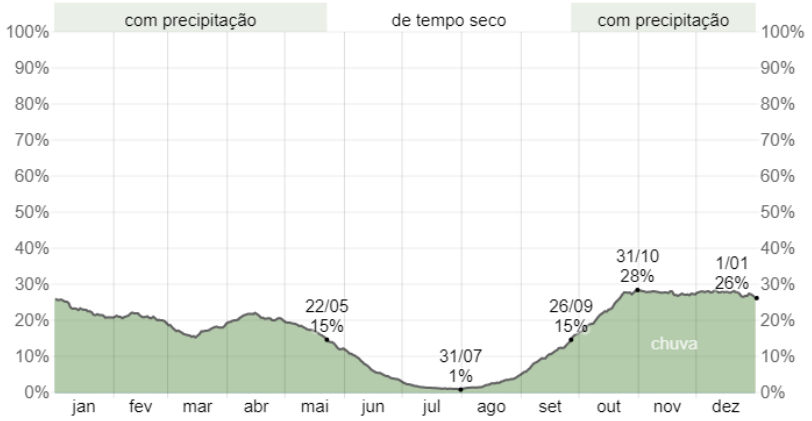

Figura 15 - Percentagem de dias em que vários tipos de precipitação são observados, retirado de https://pt.weatherspark.com/y/32195/Clima-caracter%C3%ADstico-em-Set%C3%BAbal-Portugal-durante-o-ano#Sections-Precipitation

É possível verificar através da Fig. 15 que os maiores níveis de precipitação acontecem entre os meses de outubro, novembro e dezembro.

Tabela 13 - Número médio de dias por mês em que se observa precipitação no distrito de Setúbal, retirado de https://pt.weatherspark.com/y/32195/Clima-caracter%C3%ADstico-em-Set%C3%BAbal-Portugal-durante-o-ano#Sections-Precipitation

Mês	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Média de Dias de Chuva	7,1	5,8	5,3	6,2	5,0	2,0	0,4	0,8	3,3	7,3	8,3	8,5

Continuando a análise para o mês de dezembro, a probabilidade de chover é dada por:

$$P(C) = \frac{8,5}{31} = 0,27 \qquad P(B) = 1 - P(C) = 0,73 \tag{20}$$

:

Dado as probabilidades que já temos, para obter $P(A|C)$, usando o teorema de Bayes, temos que:

$$P(A|C) = \frac{P(C|A).P(A)}{P(C)} = \frac{0,38.P(A)}{0,27} = 1,4.P(A) \tag{21}$$

Comparando esta probabilidade com a probabilidade de haver acidente dado que está bom tempo:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} = \frac{0,62.P(A)}{0,73} = 0,85.P(A) \tag{22}$$

E daqui se conclui que:

$$P(A|B) = 0{,}85.P(A) < 1{,}4.P(A) = P(A|C) \quad (23)$$

Logo, pelos nossos dados, a probabilidade de haver acidente dado que chove é maior do que a probabilidade de haver acidente dado que está bom tempo.

Conclui-se, para este exemplo, que tal como a literatura indica, a chuva é um fator atmosférico que leva a um maior número de acidentes do que o bom tempo. Assim, para este exemplo, os nossos dados estão também coerentes com a literatura. Confirmou-se também que para os restantes meses o mesmo se verifica ao aplicar o mesmo raciocínio.

5.2. Transformação de dados

A característica "Hora", foi agrupada, nos seguintes intervalos: 00:00–07:59 (meia-noite até ao amanhecer), 08:00–09:59 (transito matinal), 10:00–12:29 (horas de trabalho parte da manhã), 12:30–13:59 (pausa de almoço), 14:00–16:59 (horas de trabalho durante a tarde), 17:00–19:59 (pico de trânsito após trabalho), e 20:00–23:59 (horário noturno).

Já para a característica "Data", como indicado na literatura que os acidentes têm uma ligação temporal forte de cerca de 15 em 15 dias, foi considerado o mês em que o acidente ocorreu e os acidentes de cada mês foram agrupados pelo dia da semana em que ocorreram.

Por fim, ao agrupar os acidentes nos intervalos de tempo definidos acima e ao mesmo tempo pelas restantes variáveis, foi possível obter a contagem de acidentes para as diferentes possibilidades de dados de entrada.

Para a codificação de características categóricos foi usada a codificação distribuída (do inglês: *One Hot Encoding*) já mencionada anteriormente. Esta transforma características do conjunto de dados em vetores binários. É usada para transformar características categóricas onde as categorias não têm uma hierarquia/ordenação. Para evitar que os modelos interpretem essas características como sendo numéricos, pois os valores numéricos seguem uma ordem, são criadas variáveis binárias. É criada uma variável binária para cada categoria da variável categórica.

5.3. Limpeza de dados

Em relação à limpeza de dados, iniciando pelo desafio dos dados incompletos, foi utilizada a eliminação direta para eliminar a identificação do acidente por ser uma característica que não acrescenta informação relevante para a predição do acidente. Algumas variáveis também foram removidas, como por exemplo, as variáveis relativas ao álcool e às contraordenações, já que os valores fornecidos foram apenas para o dia da semana, a variável do dia da semana contém a mesma informação que estas duas e, por isso, estas deixam de ser relevantes para o modelo preditivo. Outra variável removida foi o tipo de acidente, porque é impossível ter como dado de entrada para uma predição de acidente se este terá ou não feridos, antes do acidente ocorrer. A eliminação direta foi também utilizada para remover acidentes onde não existisse informação relativamente à localização do mesmo. As restantes variáveis encontravam-se completas para todas as instâncias.

5.4. Redução de dados

5.4.1. Análise estatística

Iniciando pela análise da correlação entre as variáveis, como foi visto na metodologia, foram propostas duas medidas de correlação. Uma para pares de variáveis categóricas nominais e numéricas (teste de Kruskal Wallis) e outra para pares de variáveis categóricas nominais com categóricas nominais (V de Cramer). Assim, foi possível obter os diferentes valores de correlação da Fig. 16. Relembrando os intervalos da Tab. 8, iremos apenas focar os resultados dos seguintes intervalos:]0.10 ; 0.15] que equivale a uma correlação moderada,]0.15 ; 0.25] que equivale a uma correlação forte e]0.25 ; 1.00] que equivale a uma correlação muito forte. Fazendo uma análise da matriz da Fig. 16 da esquerda para a direita na sua triangular inferior (já que a matriz é simétrica pela sua diagonal principal), podemos tirar algumas conclusões.

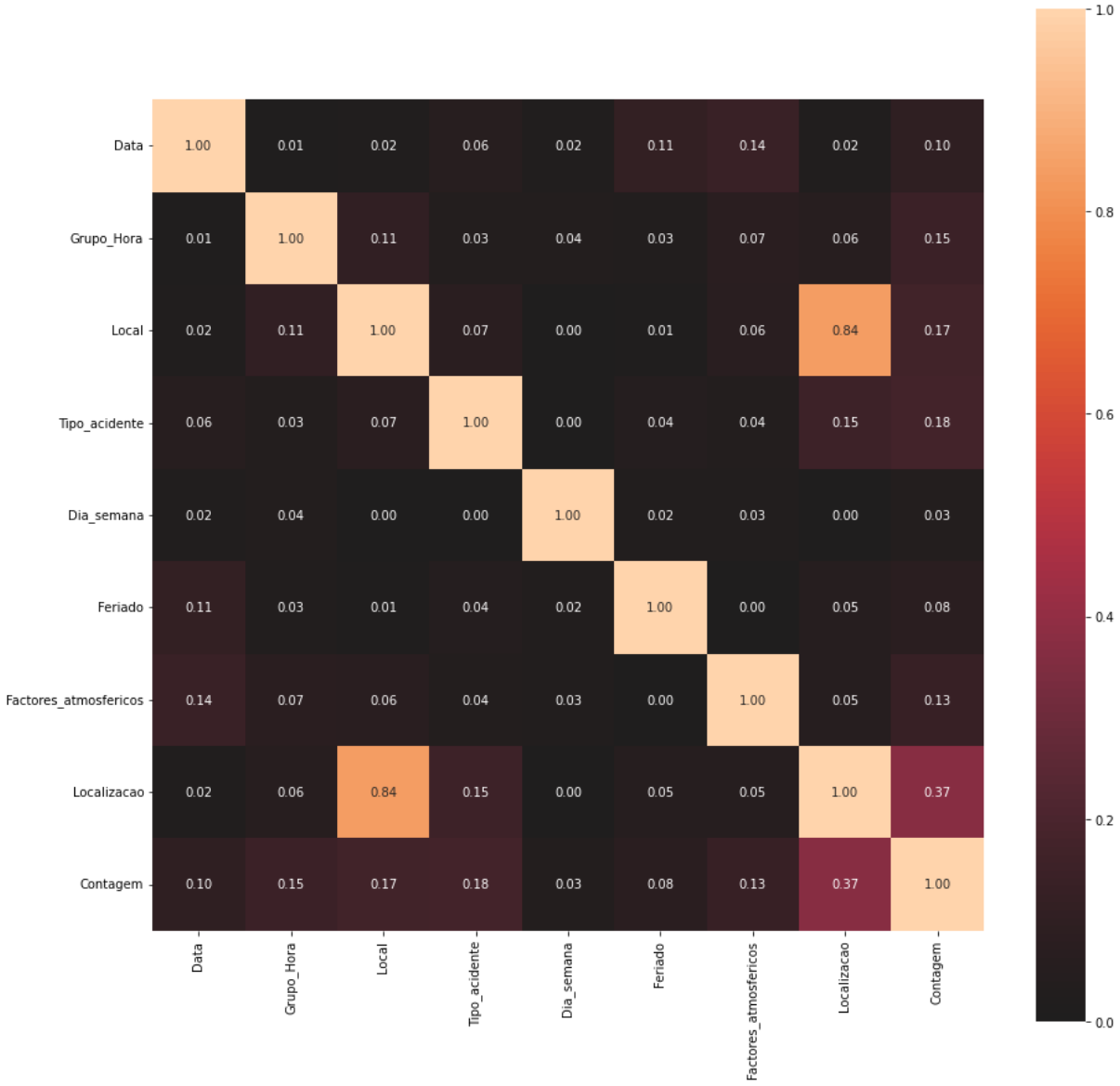

Figura 16 - Correlações para os diferentes pares de variáveis através das medidas de V de Cramer e do teste de Kruskal Wallis, dependendo do tipo de pares de variáveis. Os valores estão contidos no intervalo [0;1]. A variável "Contagem" corresponde à nossa variável alvo, ou seja, à contagem de acidentes.

1. **Data**: a variável Data tem uma correlação moderada com a variável Feriado (apesar de a data de alguns feriados mudar de ano para ano, outros mantêm uma data constante de ano para ano). A variável Data tem também uma correlação moderada com os fatores atmosféricos (pela Fig. 16 podemos ver que, em média, para certos meses existem fatores atmosféricos que são mais comuns, como a chuva nos meses de inverno).
2. **Grupo_Hora**: O grupo da hora do dia em que os acidentes ocorrem tem uma correlação moderada com a contagem de acidentes. Como já foi verificado na Fig. 5 a frequência de acidentes varia ligeiramente, de acordo com a hora do dia.
3. **Local**: A variável local, que indica se o acidente ocorreu dentro ou fora de uma localidade, tem uma correlação muito forte com a variável Localização (já que a variável Localização está dividida por concelhos ou pelo nome da estrada). Significa que é possível obter a informação quanto ao tipo de local apenas a partir da variável localização. Por esse motivo, podemos desprezar a variável local, já que a informação dessa variável fica contida na localização.
4. **Tipo de acidente**: apesar de não ser possível utilizar esta variável para o modelo de aprendizagem pois não se saberá se o acidente terá feridos ou apenas danos antes do acidente ocorrer, é interessante ver que existe uma correlação moderada/forte entre o tipo de acidente e a localização, ou seja, o facto do acidente ocorrer numa certa localização tornará mais provável que ocorram ou não feridos no acidente. Podemos verificar na Fig. 17 que por exemplo municípios como Sines, Montijo, entre outros, têm maioritariamente acidentes apenas com danos quando comparado por exemplo com a Estrada Nacional 10 (EN10);

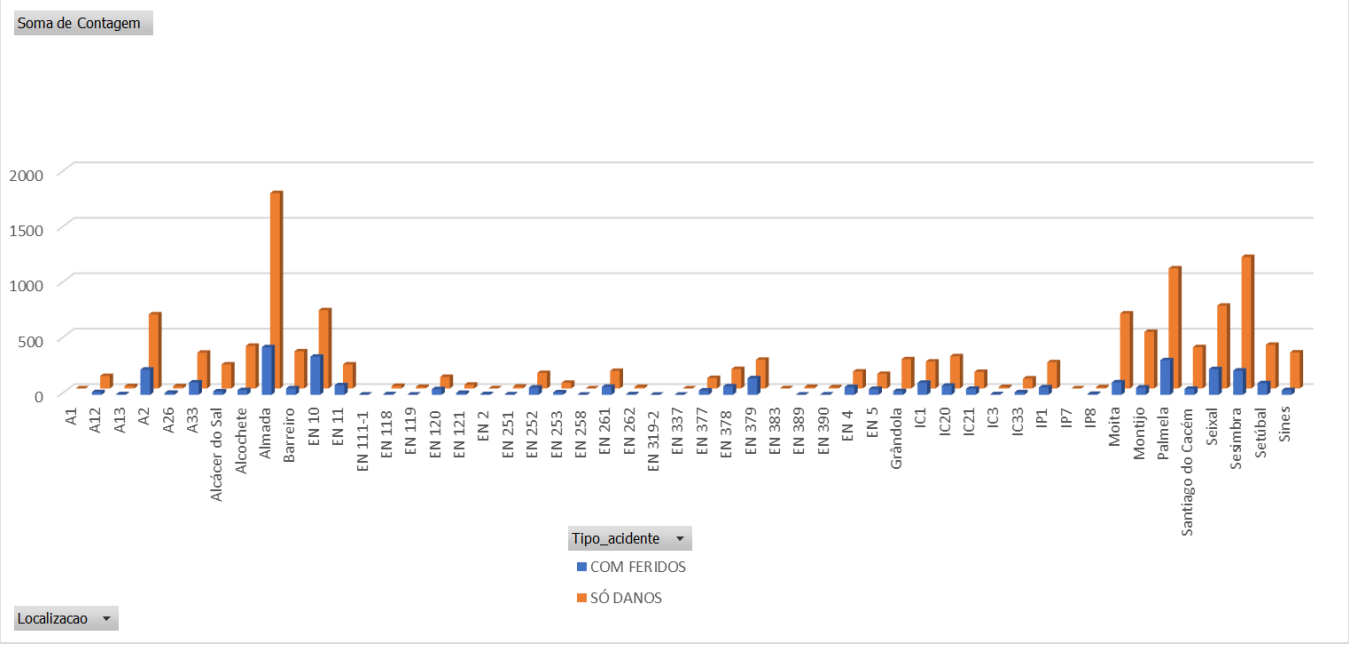

Figura 17 - Acidentes agrupados por localização e por tipo de acidente (apenas com danos ou com feridos)

5. **Dia_semana:** Não apresenta correlações com nenhuma variável;
6. **Feriado:** Apresenta apenas correlações com a data como já mencionado no ponto 1;
7. **Fatores atmosféricos:** Para além da correlação com a data, já mencionada, apresenta uma correlação moderada com a contagem de acidentes, ou seja, tem influência na variável alvo.
8. **Localização**: para além das correlações com o local e o tipo de acidente, já referidas, apresenta também uma forte correlação com a contagem de acidentes, ou seja, tem influência na variável alvo.

5.4.2. Algoritmos de seleção de características

Para os algoritmos de seleção de características RBA e SBS, optou-se por considerar apenas os resultados obtidos para as autoestradas, já que foi concluído que apenas para a autoestradas é que é possível obter um modelo credível para a predição de acidentes devido às percentagens de erro obtidas para cada modelo. As diferentes categorias das variáveis foram agrupadas por relevância. Os resultados obtidos na Tab. 14 são diferentes para os dois algoritmos utilizados, o que é natural visto que as abordagens dos mesmos são diferentes. No entanto, pode-se ver que existem várias características em que os algoritmos consideraram que estas têm a mesma relevância.

Tabela 14 - Relevância das características para a criação de modelos preditivos, obtida a partir do algoritmo RBA e SBS, para os acidentes ocorridos nas autoestradas

Autoestradas	Muito relevantes	Neutros	Pouco Relevantes
RBA	'abril', 'agosto', 'fevereiro', 'março', 'novembro', 'outubro', 'setembro', 'madrugada', 'noite', 'trabalho_matinal', 'transito_tarde', 'Quarta-feira', 'Sexta-feira', 'Sábado', 'Bom tempo', 'Chuva',	'Granizo', 'Vento forte'	'dezembro', 'janeiro', 'julho', 'junho', 'maio', 'pausa_almoco', 'trabalho_tarde', 'transito_matinal', 'Domingo', 'Quinta-feira', 'Segunda-feira', 'Terça-feira', 'Não-feriado', 'Sim-feriado', 'Nevoeiro'
SBS	'fevereiro', 'chuva', 'granizo', 'nevoeiro', 'julho', 'junho', 'maio', 'agosto', 'janeiro', 'março', 'trabalho_matinal', 'transito_tarde', 'quinta-feira', 'segunda-feira', 'sexta-feira', 'sábado', terça-feira'	'abril' 'novembro', 'outubro', 'madrugada', 'noite', 'pausa_almoço', 'Não-feriado', 'Sim-feriado'	'dezembro', 'abril', 'vento forte', 'domingo', 'quarta-feira', 'bom tempo'
Concordância	'Chuva', 'trabalho_matinal', transito_tarde', 'Sexta-feira', 'Sábado', 'agosto', 'fevereiro'		'domingo'

Para o algoritmo de RBA foi usada a variante do RBA com aplicação em problemas de regressão, a que é dada o nome de RReliefF. O número de amostras aleatórias foram 200, dos 1005 acidentes em autoestradas, sendo que cada amostra foi comparada com os seus 4 vizinhos mais próximos já que o algoritmo de KNN obteve melhores resultados para 4 vizinhos. Já para o algoritmo SBS, este usa um algoritmo de mineração para medir a performance ao retirar cada uma das características. O algoritmo escolhido foi aquele em que se obteve melhores resultados para a

mineração, ou seja, a rede neural. E a medida de performance foi a mesma que se utilizou na avaliação dos algoritmos de mineração, o MAPE. Ao retirar iterativamente cada uma das características conseguimos perceber quais delas é que melhoram ou pioram a performance da rede neural.

Através da Tab. 14 conseguimos entender que existem certas variáveis que são mais relevantes. A credibilidade destes algoritmos é difícil de medir já que estes dados não são sintéticos e não sabemos à priori quais são realmente as variáveis mais importantes para cada caso. Apesar dos algoritmos medirem a importância de cada variável de forma diferente, pode-se observar que existem várias variáveis em que ambos os algoritmos estão em concordância e que influenciam ou não os acidentes, como podemos ver na última linha da Tab. 14. Daqui se pode concluir que as variáveis que mais influenciam a ocorrência de acidentes em autoestradas, segundo os algoritmos utilizados são: o fator atmosférico chuva, os grupos de horas das 10:00–12:29 e 17:00–19:59, os dias da semana sexta-feira e sábado, e os meses de agosto e fevereiro.

5.5. Algoritmos de Mineração

Os diferentes algoritmos foram testados com vários parâmetros que estão representados na Tab. 15.

Tabela 15 - Parâmetros treinados nos diferentes algoritmos aplicados e respetivos valores testados.

Algoritmo	Parâmetro a ser escolhido	Valores testado
KNN	Número de Vizinhos, k	De 0 a 20
Lasso	Alpha	0,0001; 0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 0,8; 1; 10 ;100
Ridge	Alpha	0,01; 0,1; 1; 0; 100; 1000; 3000; 5000
Rede Neural (7xNxNxNx1)	Tamanho da amostra de treino	3; 5; 7; 8; 10; 15; 20
	Número de períodos de treino	2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12
	N = Número de nós por camada oculta	40; 60; 90; 120; 140; 160; 180
	Número de camadas ocultas	3

Inicialmente optou-se por utilizar o conjunto de dados inicial, sem dividir o mesmo pelos diferentes tipos de localização. No agrupamento de acidentes desta primeira experiência, para o cálculo da exatidão, optou-se por classificar o risco de acidentes nos intervalos da Tab. 16.

Tabela 16 - Intervalos utilizados para a definição da classe de risco da frequência de acidentes.

Classificação	Intervalos de frequência de acidentes (Nº de acidentes por Grupo de hora)
Baixo Risco	<1,5
Médio Risco	>1,5 ; <2,5
Alto Risco	>2,5

Esta opção foi tomada com base no diagrama em caixa da Fig. 18, pois como é possível observar a média do número de acidentes por grupo de hora é de cerca de 1.5 e através dos diferentes quartis, chegou-se à Tab. 16.

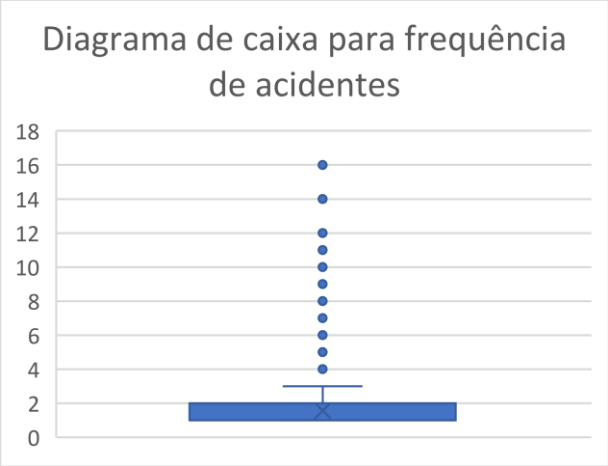

Figura 18 - Diagrama em Caixa para a frequência de acidentes na tabela de dados inicial

Em todos os algoritmos os parâmetros escolhidos foram aqueles que levaram a uma melhor performance dos algoritmos. Os melhores modelos obtidos para cada algoritmo segundo as métricas utilizadas, relativamente à primeira experiência, estão representados na Tab. 17.

Tabela 17 - Valores obtidos nas diferentes métricas de performance para os algoritmos utilizados, através da tabela de dados inicial.

Tabela de dados inicial			
Algoritmo	**MAE (distância)**	**100-MAPE (%)**	**Exatidão (Classes)**
Regressão KNN	0,77	2,1%	57%
Regressão Linear	0,65	2,4%	59%
Regressão Lasso	0,62	5,5%	61%
Regressão Ridge	0,62	5,4%	61%
Regressão Árvore de Decisão	0,65	10,8%	63%
Regressão Rede Neural	0,49	55,1%	88%

Como segunda experiência, optou-se por agrupar os dados pelos conjuntos já referidos anteriormente: Autoestradas; Estradas Nacionais ou Itinerários; e Municípios. Optou-se por dividir as classes pelos intervalos definidos pela Tab. 16. O objetivo de manter o intervalo de classificação é

porque apenas se pretende entender se o modelo de regressão obtido ao criar três modelos individuais para cada tipo de localização melhora ou piora as métricas de performance. Os acidentes ocorridos em Autoestradas representam 9,3% do total de acidentes, os acidentes ocorridos em Itinerários ou estradas nacionais representam 30% do total de acidentes e os acidentes ocorridos fora das duas anteriores, em ruas de um Município/Concelho, representam 60,7% do total de acidentes.

Começando pelo conjunto de dados relativos à autoestrada, os melhores modelos obtidos para cada algoritmo segundo as métricas utilizadas, relativamente à segunda experiência, estão representados nas tabelas 18, 19 e 20.

Tabela 18 - Valores obtidos nas diferentes métricas de performance para os algoritmos utilizados, através da tabela de dados relativa aos acidentes ocorridos em autoestradas.

Tabela de dados de Autoestradas - 9,3% do total de acidentes			
Algoritmo	**MAE (distância)**	**100-MAPE (%)**	**Exatidão**
Regressão KNN	0,74	5,9%	56%
Regressão Linear	0,63	5,2%	57%
Regressão Lasso	0,60	8,3%	54%
Regressão Ridge	0,61	8,1%	52%
Regressão Árvore de Decisão	0,69	12,2%	56%
Regressão Rede Neural	0,57	56,4%	89%

Tabela 19 - Valores obtidos nas diferentes métricas de performance para os algoritmos utilizados, através da tabela de dados relativa aos acidentes ocorridos em Itinerários ou estradas nacionais.

Tabela de dados de Itinerários - 30% do total de acidentes			
Algoritmo	**MAE (distância)**	**100-MAPE (%)**	**Exatidão**
Regressão KNN	0,30	60,3%	81%
Regressão Linear	0,27	64,1%	86%
Regressão Lasso	0,28	63,4%	86%
Regressão Ridge	0,28	63,3%	80%
Regressão Árvore de Decisão	0,31	61,4%	76%
Regressão Rede Neural	0,55	50,6%	87%

Tabela 20 - Valores obtidos nas diferentes métricas de performance para os algoritmos utilizados, através da tabela de dados relativa aos acidentes ocorridos dentro de Municípios.

Tabela de dados de Municípios - 60,7% do total de acidentes			
Algoritmo	**MAE (distância)**	**100-MAPE (%)**	**Exatidão**
Regressão KNN	0,93	-35,4%	48%
Regressão Linear	0,85	-34,3%	50%
Regressão Lasso	0,80	-29,6%	51%
Regressão Ridge	0,79	-29,7%	50%
Regressão Árvore de Decisão	0,91	-30,6%	55%
Regressão Rede Neural	0,52	-4,3%	88%

De forma a comparar a experiência 1 e 2, criou-se a Tab. 21 onde se utilizou a média ponderada da experiência 2. Neste caso, como o modelo em que se obteve melhores resultados foi a regressão através de uma rede neural, apenas se comparou os resultados obtidos para esse algoritmo de mineração.

Tabela 21 Comparação de resultados entre a experiência 1 e 2, para o melhor modelo obtido.

Comparação entre a experiência 1 e 2			
Regressão Rede Neural	**MAE (distância)**	**100-MAPE (%)**	**Exatidão**
Modelo geral	0,49	55,1%	88%
Modelos individuais (média ponderada)	0,54	16,9%	88%

O resultado obtido para a média dos modelos individuais deve-se ao facto de que ao dividir os dados, o modelo obtido para os municípios apresentou um erro muito elevado 104,3% (correspondente ao -4,3% da tabela 20). Assim, para uma melhor visualização individual dos dados, criou-se a Tab. 22.

Tabela 22 - Resumo dos resultados obtidos pelos diferentes algoritmos para as duas experiências realizadas

Resumo dos melhores resultados obtidos			
Algoritmo (Regressão Rede Neural)	**MAE (distância)**	**100-MAPE (%)**	**Exatidão**
Modelo Geral	0.49	55.1%	88%
Autoestradas (9,3% do total de acidentes)	0.57	56.4%	89%
Itinerários ou Estradas Nacionais(30% do total de acidentes)	0.55	50.6%	87%
Municípios (60,7% do total de acidentes)	0.52	-4.3%	88%

Os parâmetros ideais para as redes neurais obtidas para os diferentes conjuntos de dados da Tab. 22, estão representados na Tab. 23.

Tabela 23 - Parâmetros das diferentes redes neurais obtidas para os diferentes conjuntos de dados.

Resumo dos melhores resultados obtidos			
Algoritmo (Regressão Rede Neural)	**Nº de nós por camadas**	**Nº de períodos**	**Tamanho da amostra de treino da rede por período**
Modelo Geral	120	4	7
Autoestradas (9,3% do total de acidentes)	160	3	20
Itinerários ou Estradas Nacionais(30% do total de acidentes)	140	2	8
Municípios (60,7% do total de acidentes)	120	3	8

Como se pode observar, ao dividir o conjunto de dados pelo tipo de localização obteve-se um melhor MAPE e exatidão para o modelo das autoestradas. Para o modelo obtido para os **municípios** obtém-se um erro muito elevado de aproximadamente 104,3% (correspondente ao -4,3% da tabela 21). Isto pode ser justificado pelo elevado número de exceções (ou seja, grupos de hora em que existiram muitos acidentes e que não é comum acontecer, pela análise do diagrama em caixa) e por essas exceções terem valores muito elevados (valores como 16 acidentes, num total de 4 horas, para um certo município), que leva a erros percentuais maiores para os valores mais altos. Como confirmação, ao medir o erro percentual para valores da variável alvo iguais a 1 e diferentes de 1, obteve-se um erro de 30% e de 197%, respetivamente. Como podemos observar ao comparar os diagramas em caixa dos três conjuntos de dados, na Fig. 19. No entanto, não leva a que a exatidão seja baixa o que significa que a predição para os valores mais comuns não tenha erros tão elevados.

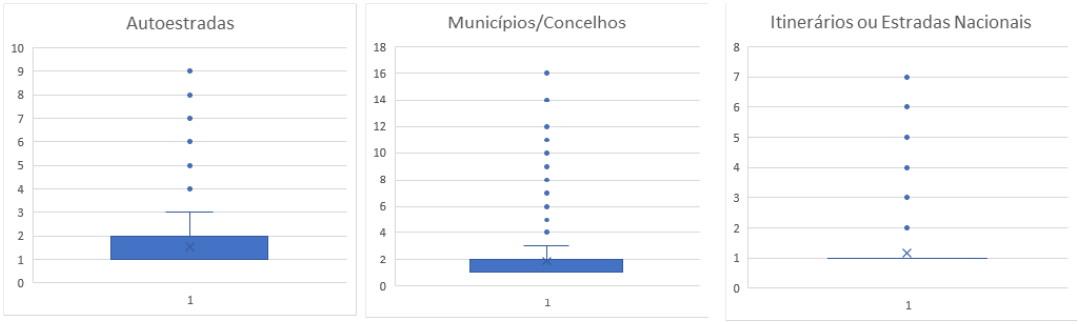

Figura 19 - Diagramas em caixa para os valores da frequência de acidentes por grupo de hora nas diferentes localizações (autoestradas, municípios, estradas nacionais ou itinerários)

Pela Fig. 19 podemos entender que apesar do modelo relativo às **autoestradas** ter o menor número de amostras de treino e maior variância no valor da frequência de acidentes, é aquele em que é possível obter um modelo com melhor MAPE e melhor exatidão.

Podemos perceber também que o motivo para o MAPE obtido para o modelo do conjunto de dados de **itinerários ou estradas nacionais** não afetar a exatidão é porque, para o intervalo de tempo escolhido, na maioria dos casos só existe 1 acidente, logo, mesmo com um erro de 0.5, para a classificação esse valor será classificado corretamente.

Daqui se conclui que apenas o modelo obtido para as autoestradas pode ter utilidade, já que para . Para o modelo obtido para itinerários ou estradas nacionais, seria necessário ter uma frequência maior de acidentes, ou obter mais dados como o nível de mobilidade humana ou a geo-localização dos acidentes, para perceber se existem locais com grande concentração de acidentes, e assim, ter mais variáveis de forma a criar modelos com abordagens semelhantes às encontradas na literatura e assim obter melhores resultados. O mesmo se aplica para os acidentes nos municípios. Para este tipo de acidentes é normal existirem mais exceções devido à área ser maior e a haver mais fatores como peões, ciclistas, maior número de cruzamentos, entre muitas variáveis que não existem para as autoestradas.

Tabela 24 - Dados relativos ao número de acidentes com feridos ou mortos e ao número de feridos e mortos por acidente.

Tipo de Localização	Nº de feridos/ mortos	Nº de acidentes com feridos/ mortos	Nº de acidentes total	Percentagem de acidentes com feridos	Feridos/ mortos por acidente com feridos	Feridos por acidente
Autoestradas	651	385	1531	25,1%	1,7	0,43
Concelho/ Município	2105	1731	9988	17,3%	1,2	0,21
Itinerários ou Estradas Nacionais	2074	1459	4892	29,8%	1,42	0,42
Todas as localizações	4830	3575	16411	21,8%	1,35	0,3

Apesar das limitações dos dados conseguiu-se obter um modelo para as autoestradas com um erro um erro de 11% para a precisão. A autoestrada, apesar de ser a localização onde existe menor quantidade de acidentes para o distrito de Setúbal, é a localização com maior concentração de acidentes por área quando comparado aos municípios e maior concentração de acidente por autoestrada, quando comparado com a concentração de acidentes em itinerários ou estradas nacionais, como se pode verificar pelo gráfico da Fig. 17. A Autoestrada é o tipo de localização onde existem mais feridos e mortos por acidente, que pode ser observado na Tab. 24. A autoestrada é também a localização onde é possível que a GNR faça uma fiscalização com mais efeito, já que para os municípios existe uma grande quantidade de estradas e uma vasta área onde o acidente pode ocorrer, e nos itinerários não existem grandes concentrações de acidentes no tempo, como se pode verificar pelo diagrama em caixa da Fig. 19. Isto permite que o modelo tenha mais informação para encontrar os parâmetros certos.

Na Tab. 25 é possível observar um exemplo de 3 conjuntos de dados de entrada e a sua respetiva predição, tanto em quantidade numérica, com um MAPE de 56%, como em classe de risco, com uma exatidão de 89%.

Tabela 25 - Exemplo de 3 amostras de entrada que se pretende prever a quantidade de acidentes e o respetivo risco de acidente associado a essa previsão.

Mês	Dia da semana	Grupo de horas do dia	Feriado	Tipo de local	Condição atmosférica	Localização	Regressão (56%)	Classificação (89%)
Setembro	Quarta-feira	Trabalho matinal (10:00–12:30)	Não	Fora de localidade	Bom tempo	A2	0.7	Pouco risco
Setembro	Quinta-feira	Transito matinal (08:00–10:00)	Sim	Fora de localidade	Bom tempo	A26	2.3	Médio risco
Setembro	Quinta-feira	Transito tarde (17:00–20:00)	Sim	Fora de localidade	Chuva	A26	7.4	Muito risco

Como se pode observar pela Tab. 25, o modelo obtido permite realizar uma predição do risco de acidente através das variáveis: Mês, Dia da semana, Grupo de horas do dia, Feriado, Tipo de local, Condição atmosférica e Localização com uma exatidão de 89%.

6. Conclusão

Os acidentes rodoviários causam várias mortes por ano e têm como consequência danos económicos e físicos para as vítimas e para o Estado. Devido à existência de uma base de dados de acidentes na qual é possível descobrir padrões e criar conhecimento optou-se por aplicar as técnicas de mineração de dados à mesma. Foram feitas diferentes análises aos dados, pois, por se tratar de um problema real foi necessário conhecer os dados fornecidos a partir de diferentes abordagens. Foram também realizadas diferentes experiências e mesmo com a limitação do conjunto de dados, tentou-se adaptar a metodologia à metodologia aplicada pelos trabalhos relacionados que apresentaram bons resultados.

Nesta dissertação foi proposto o tema predição do risco de acidente rodoviário através de métodos de mineração de dados. Foram disponibilizados dados de participações de acidentes pela GNR, dos acidentes ocorridos em Setúbal de 2019 a 2021. Tem como objetivo criar um modelo que consiga realizar uma predição com baixo erro. O sistema desenvolvido é constituído por 3 etapas principais: (i) seleção e recolha dos dados, (ii) pré-processamento, (iii) algoritmos de mineração.

Neste trabalho, várias técnicas foram implementadas, sendo que a recolha e seleção dos dados e o pré-processamento levaram mais tempo do que o esperado devido a se tratar de um problema com dados reais.

Inicialmente, através da análise dos dados, foi possível concluir que a maior concentração de acidentes ocorre durante o intervalo de tempo de 17:00-20:00. Foi possível concluir também que para os nossos dados, a chuva é o fator atmosférico com maior probabilidade de ocorrer acidente. Conclui-se também que o dia da semana em que ocorrem mais acidentes é na sexta-feira. Daqui provou-se que os dados têm alguma credibilidade já que estas conclusões são coerentes com a literatura.

Através de uma análise da correlação entre as diferentes variáveis foi possível concluir que a localização do acidente é a variável que mais influência na frequência de acidentes, no conjunto de varáveis existentes. A partir dessa conclusão e da ideia de que para diferentes tipos de localizações existem diferentes fatores que influenciam o acidente, estes foram agrupados pelo tipo de localização em que se encontravam. Por este motivo foi necessário criar diferentes modelos para cada conjunto. Para além da localização, a correlação entre variáveis também indicou outras variáveis que mais influenciam na frequência de acidentes tais como a hora do dia, o fator atmosférico e se o acidente ocorreu dentro ou fora de uma localidade. Após a divisão do conjunto de dados nos três tipos de localização (autoestradas, estradas nacionais ou itinerários e municípios), foi possível, através dos algoritmos de seleção de variáveis, perceber quais as variáveis que mais influenciam cada tipo de localização.

O problema de mineração foi abordado como um problema de regressão já que a variável alvo era a frequência de acidentes no intervalo de tempo definido. Os algoritmos de mineração testados foram o KNN, a regressão linear simples, Lasso e Ridge, a árvore de decisão para regressão e a rede neural tradicional. Tanto para o conjunto de dados inicial como para os conjuntos de dados separados pela localização nos seguintes conjuntos: autoestradas, estadias nacionais ou itinerários e municípios.

O melhor resultado foi obtido através da rede neural. No entanto, para cada conjunto obteve-se diferentes modelos, com diferentes arquiteturas (número de nós, períodos de treino, etc). O melhor resultado foi obtido para o conjunto de dados das autoestradas. A autoestrada, apesar de ser a localização onde existe menor quantidade de acidentes para o distrito de Setúbal, é a localização com maior concentração de acidentes por área quando comparado aos municípios e maior concentração de acidente por autoestrada, quando comparado com a concentração de acidentes em itinerários ou estradas nacionais. Para além disso é a Autoestrada a localização onde existem mais feridos e mortos por acidente. A autoestrada é também a localização onde é possível que a GNR faça uma fiscalização com mais efeito, já que para os municípios existe uma grande quantidade de estradas e uma vasta área onde o acidente pode ocorrer, e nos itinerários não existem grandes concentrações de acidentes no tempo. Isto permite que o modelo tenha mais informação para encontrar os parâmetros certos. A percentagem de erro na regressão foi de 44%, no entanto não se pretendia obter o número exato de acidentes, por isso agrupou-se os resultados em 3 classes de risco, de acordo com o diagrama em caixa obtido para a frequência de acidentes. Obteve-se assim uma percentagem de erro de apenas 11%.

6.1. Trabalho futuro

Como trabalho futuro, o primeiro passo seria melhorar a recolha de dados de forma a garantir que a geo-localização dos acidentes era adquirida. Dessa forma seria possível optar por abordagens mais complexas como as que se mencionaram nos estudos de abordagem de aprendizagem profunda. Outra variável importante de obter seria os níveis de mobilidade humana, que seriam possíveis de obter através de cooperações com aplicativos como a google Maps, Waze, ou apenas através da velocidade como táxis da Uber ou outras companhias se deslocam, como foi o caso de alguns dos estudos mencionados da literatura. Ainda na obtenção de dados, seria também importante garantir a credibilidade dos mesmos ao criar mais parâmetros obrigatório quando a participação de acidentes é inserida na base de dados pelo militar, e também sensibilizar os militares para a importância de inserir os dados da forma mais coerente possível com a realidade do acidente.

Após garantir a obtenção destes dados, seria possível utilizar as abordagens mencionadas e dessa forma obter percentagens de erro ainda mais baixas e também a localização das zonas com maior risco de acidente.

Por último, a nível de aplicação é possível a criação de um serviço Web em que o cliente, possa verificar o risco de acidente para este conjunto de entrada. Para a GNR, o objetivo seria dar mais informação sobre as zonas onde o risco de acidente é maior de forma a direcionar as suas patrulhas e ações de fiscalização para essas zonas. No entanto, não foi possível no tempo disponível criar esta ferramenta.

Bibliografia

[1] R. Goldshmidt, E. Passos, E. Bezerra, "*Data mining : conceitos, técnicas, algoritmos, orientações e aplicações"*, Elviser Ed. Rio de Janeiro, 2015.

[2] U. Fayyad, P. Smyth, G. Piatetsky-Shapiro, "Knowledge Discovery and Data Mining: Towards a Unifying Framework" *American Association for Artificial Intelligence*., 1996, pp. 82-88 [online] Acedido em: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

[3] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, "Mining association rules in graphs based on frequent cohesive itemsets," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[4] S. Agarwal, *Data mining: Data mining concepts and techniques*, Elsevier I. Estados Unidos da America, 2014.

[5] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.

[6] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *Found. Trends Mach. Learn.*, vol. 9, no. 1, pp. 1–118, 2016, doi: 10.1561/2200000055.

[7] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1275–1289, 2019, doi: 10.1016/j.jksuci.2019.06.012.

[8] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, *A review of feature selection methods on synthetic data*, vol. 34, no. 3. 2013.

[9] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 2, pp. 289–303, 2015, doi: 10.1109/TVCG.2014.2350494.

[10] L. Massaron, J. P. Mueller, *Deep Learning for Dummies*, New Jersey, For Dummies, 2019.

[11] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and Unsupervised Learning for Data Science*, no. January. 2020.

[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms: Second Edition*, Jonh Wiley., vol. 2. New Jersey, 2014.

[13] P. C. Sen, M. Hajra, and M. Ghosh, *Emerging Technology in Modelling and Graphics*, vol. 937. Springer Singapore, 2020.

[14] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms", *Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016*, pp. 1310–1315, 2016.

[15] R. Indrakumari, T. Poongodi, and K. Singh, *Introduction to Deep Learning*, Springer I. Croatia, 2021.

[16] L. Massaron, J. P. Mueller, *Deep Learning for Dummies*, New Jersey, For Dummies, 2019.

[17] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[18] L. A. Belanche and F. F. González, "Review and Evaluation of Feature Selection Algorithms in

Synthetic Problems," Universitat Politécnica de Catalunya, Barcelona, Spain, 2011, doi: https://doi.org/10.48550/arXiv.1101.2320, [Online]. Available: http://arxiv.org/abs/1101.2320.

[19] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 55, pp. 444–459, 2015, doi: 10.1016/j.trc.2015.03.015.

[20] D. Eisenberg, "The mixed effects of precipitation on traffic crashes," *Accid. Anal. Prev.*, vol. 36, no. 4, pp. 637–647, 2004, doi: 10.1016/S0001-4575(03)00085-X.

[21] R. B. Hayat *et al.*, "Explaining the road accident risk : Weather effects," *Accid. Anal. Prev.*, vol. 1, no. 60, pp. 456–465, 2013.

[22] J. D. Tamerius, X. Zhou, R. Mantilla, and T. Greenfield-Huitt, "Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions," *Weather. Clim. Soc.*, vol. 8, no. 4, pp. 399–407, 2016, doi: 10.1175/WCAS-D-16-0009.1.

[23] J. D. Febres, S. García-Herrero, S. Herrera, J. M. Gutiérrez, J. R. López-García, and M. A. Mariscal, "Influence of seat-belt use on the severity of injury in traffic accidents," *Eur. Transp. Res. Rev.*, vol. 12, no. 1, 2020, doi: 10.1186/s12544-020-0401-5.

[24] G. Musile, N. Pigaiani, D. Sorio, M. Colombari, F. Bortolotti, and F. Tagliaro, "Alcohol-associated traffic injuries in Verona territory: A nine-year survey," *Med. Sci. Law*, vol. 61, no. 1_suppl, pp. 7–13, 2021, doi: 10.1177/0025802420937577.

[25] L. M. Martín-delosReyes, V. Martínez-Ruiz, M. Rivera-Izquierdo, E. Jiménez-Mejías, and P. Lardelli-Claret, "Is driving without a valid license associated with an increased risk of causing a road crash?," *Accid. Anal. Prev.*, vol. 149, no. November 2020, pp. 1–7, 2021, doi: 10.1016/j.aap.2020.105872.

[26] Y. Song, S. Kou, and C. Wang, "Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach," *J. Safety Res.*, vol. 76, pp. 64–72, 2021, doi: 10.1016/j.jsr.2020.11.006.

[27] B. Kumeda, F. Zhang, F. Zhou, S. Hussain, A. Almasri, and M. Assefa, "Classification of road traffic accident data using machine learning Algorithms," in *2019 IEEE 11th International Conference on Communication Software and Networks, ICCSN 2019*, 2019, pp. 682–687, doi: 10.1109/ICCSN.2019.8905362.

[28] L. Y. Chang, "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network," *Saf. Sci.*, vol. 43, no. 8, pp. 541–557, 2005, doi: 10.1016/j.ssci.2005.04.004.

[29] L. Y. Chang and W. C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *J. Safety Res.*, vol. 36, no. 4, pp. 365–375, 2005, doi: 10.1016/j.jsr.2005.06.013.

[30] L. J. Muhammad *et al.*, "Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents, its Prone Locations and Time along Kano –Wudil Highway," *Int. J. Database Theory Appl.*, vol. 10, no. 1, pp. 197–206, 2017, doi: 10.14257/ijdta.2017.10.1.18.

[31] D. T. Akomolafe and A. Olutayo, "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways," *Am. J. Database Theory Appl.*, vol. 1, no. 3, pp. 26–38, 2013, doi: 10.5923/j.database.20120103.01.

[32] O. V.A and E. A.A, "Traffic Accident Analysis Using Decision Trees and Neural Networks", *Int. J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 22–28, 2014, doi: 10.5815/ijitcs.2014.02.03.

[33] S. Shanti, D. R. G. Ramani, R. G. Shanthi, S., & Ramani, S. Shanti, and D. R. G. Ramani, "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms", *Int. J. Comput. Appl.*, vol. 35, no. 12, pp. 975–8887, 2011.

[34] Y. Castro and Y. J. Kim, “Data mining on road safety: Factor assessment on vehicle accidents using classification models”, *Int. J. Crashworthiness*, vol. 21, no. 2, pp. 104–111, 2016, doi: 10.1080/13588265.2015.1122278.

[35] J. Kashyap, A. Chandra, and P. Singh, “Mining Road Traffic Accident Data to Improve Safety on Road-related Factors for Classification and Prediction of Accident Severity,” *Int. Res. J. Eng. Technol.*, vol. 10, pp. 2395–56, 2016, [Online]. Disponível em: https://www.irjet.net/archives/V3/i10/IRJET-V3I1041.pdf.

[36] S. Hussain, L. J. Muhammad, F. S. Ishaq, A. Yakubu, and I. A. Mohammed, “Performance evaluation of various data mining algorithms on road traffic accident dataset”, *Smart Innov. Syst. Technol.*, vol. 106, pp. 67–78, 2019, doi: 10.1007/978-981-13-1742-2_7.

[37] R. E. Almamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, “Comparison of machine learning algorithms for predicting traffic accident severity”, in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*, 2019, pp. 272–276, doi: 10.1109/JEEIT.2019.8717393.

[38] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, “Learning deep representation from big and heterogeneous data for traffic accident inference”, *30th AAAI Conf. Artif. Intell. AAAI 2016*, pp. 338–344, 2016.

[39] A. Ziakopoulos and G. Yannis, “A review of spatial approaches in road safety”, *Accid. Anal. Prev.*, vol. 135, no. July 2019, p. 105323, 2020, doi: 10.1016/j.aap.2019.105323.

[40] Z. Yuan, X. Zhou, and T. Yang, “Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data”, *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 18, pp. 984–992, Jul. 2018, doi: 10.1145/3219819.3219922.

[41] L. Yu, B. Du, X. Hu, L. Sun, L. Han, and W. Lv, “Deep spatio-temporal graph convolutional network for traffic accident prediction”, *Neurocomputing*, vol. 423, pp. 135–147, 2021, doi: 10.1016/j.neucom.2020.09.043.

[42] A. Bhattacharya and D. B. Dunson, “Simplex factor models for multivariate unordered categorical data”, *J. Am. Stat. Assoc.*, vol. 107, no. 497, pp. 362–377, 2012, doi: 10.1080/01621459.2011.646934.

[43] H. Akoglu, “User’s guide to correlation coefficients”, *Turkish J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, 2018, doi: 10.1016/j.tjem.2018.08.001.

[44] S. Jun, "The Microbiome in Health and Disease", *Volume 171 in the Progress in Molecular Biology and Translational Science*, Elsevier Science, 2020, pp. 309-450.

[45] A. C. Leon, “Descriptive and Inferential Statistics” *Compr. Clin. Psychol.*, New York, USA, vol. 3, pp. 243–285, 1998, doi: 10.1016/b0080-4270(73)00264-9.

[46] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *J. Biomed. Inform.*, vol. 85, no. June, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.

[47] Robnik-Šikonja, M., Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning 53, pp. 23–69 (2003). https://doi.org/10.1023/A:1025667309714

[48] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, “Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network,” *IECON Proc. (Industrial Electron. Conf.*, no. December, pp. 2845–2850, 2010, doi: 10.1109/IECON.2010.5675075.

[49] L. C. Molina, L. Belanche and A. Nebot, "Feature selection algorithms: a survey and experimental evaluation," *in 2002 IEEE International Conference on Data Mining*, 2002, pp.

306-313, doi: 10.1109/ICDM.2002.1183917.

[50] R. Alhamzawi and H. T. M. Ali, “The Bayesian adaptive lasso regression,” *Math. Biosci.*, vol. 303, pp. 75–82, 2018, doi: 10.1016/j.mbs.2018.06.004.

[51] R. Timofeev, “Classification and Regression Trees (CART) Theory and Applications”, CASE - Center of Applied Statistics and Economics, Humboldt University, Berlin, 2004.

[52] S. Barrash, Y. Shen, and G. B. Giannakis, “SCALABLE AND ADAPTIVE KNN FOR REGRESSION OVER GRAPHS", University of Minnesota, Minneapolis , USA, 2019.

[53] R. HECHT-NIELSEN, *Theory of the Backpropagation Neural Network Based on ‘‘nonindent’’ by Robert Hecht-Nielsen, which appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989. © 1989 IEEE.*, no. June 1989. Academic Press, Inc., 1992, pp.65-93, ISBN: 978-0-12-741252-8. doi: https://doi.org/10.1016/B978-0-12-741252-8.50010-8

[54] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, “A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction,” *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, vol. 2018-Novem, pp. 3346–3351, 2018, doi: 10.1109/ITSC.2018.8569437.

[55] S. Vinoski, "RESTful Web Services Development Checklist," in IEEE Internet Computing, vol. 12, no. 6, pp. 96-95, Nov.-Dec. 2008, doi: 10.1109/MIC.2008.130.

[56] J. Costa, E. Freitas, P. Pereira, M. Jacques, "Acidentes Rodoviários das Estradas Nacionais de Portugal", C-TAC - Comunicações a Conferências Nacionais, 2011. [Online]. URL: https://hdl.handle.net/1822/15483, Disponível em: https://repositorium.sdum.uminho.pt

[57] Seguropordias (2022). *O congestionamento nas estradas da cidade do Porto*. Disponível em: https://seguropordias.pt/blog/tr%C3%A2nsito-porto-portugal

[58] The Littleton Law Office of Bahr, Kreidle & Flicker (2022), *4 Reasons You Are More Likely to Be in a Car Accident in the Summer*. Disponível em*:* https://littletonlawyers.com/4-reasons-you-are-more-likely-to-be-in-a-car-accident-in-the-summer/