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

Academia Militar
Lisbon, Portugal
12721218david@gmail.com [1],
jose.silva@academiamilitar.pt [2]
David Dias [1], José Silvestre Silva [2]

Instituto Superior Técnico
Lisbon, Portugal
alex@isr.tecnico.ulisboa.pt [3]
Alexandre José Malheiro Bernardino [3]

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. Foram testados tanto métodos clássicos como métodos de aprendizagem profunda. Os melhores resultados foram obtidos para o algoritmo de rede neural.

Palavras-Chave: predição de risco; acidentes de viação; classificação supervisionada; métodos clássicos; redes neurais profundas.

I. 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. Visto que sempre que existe um acidente os dados relativos ao mesmo são guardados na base de dados da GNR, faz com que exista uma base de dados na qual é possível descobrir padrões e criar conhecimento. As técnicas de mineração de dados têm vindo a evoluir e têm vindo a ser aplicadas em cada vez mais em problemas do mundo real. Os métodos de mineração de dados podem ser utilizados na base de dados fornecida 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, 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.

Os dados disponibilizados pela GNR correspondem 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 serão necessárias técnicas para colmatar estes problemas. Para além disso, será a primeira vez que serão aplicados algoritmos de mineração de dados neste conjunto de dados, pelo que poder-se-á concluir que a medidas de erro sejam demasiado altas para que esta ferramenta seja viável e que sejam necessários outro tipo de dados para atingir melhores desempenhos.

Este trabalho tem como objetivo desenvolver uma ferramenta de auxílio ao policiamento guiado por informações, relativamente à área do trânsito. Para isso, serão testados vários algoritmos de mineração de dados que já demonstraram ter sucesso quando aplicados a diferentes tipos de conjuntos de dados. Estas ferramentas serão 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.

II. Enquadramento Teórico

A. Descoberta de Conhecimento em Bases de Dados

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], [3].

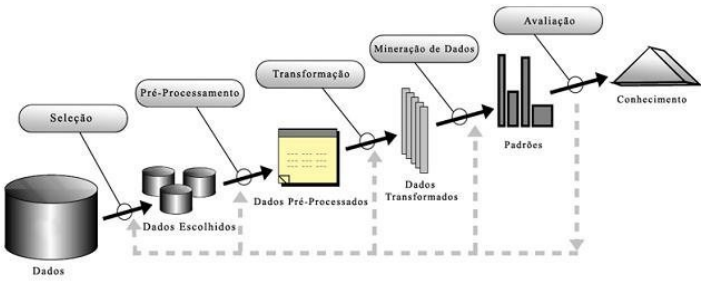

Figura 1 - Etapas que compõe o processo KDD, adaptado de [3]

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.
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.

B. *Mineração de dados*

A mineração de dados é frequentemente dividida em grupos principais. A aprendizagem supervisionada, não supervisionada, por reforço, entre outras. A aprendizagem supervisionada 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], [6]–[9].

Em [7] 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 usadas com mais frequência nesses 84 artigos. Outros algoritmos de aprendizagem supervisionada utilizados com mais frequência são o K-vizinhos mais próximos (KNN, do inglês: *K-neirest neighbors*) [6]–[8], [10] e a rede neural artificial (ANN, do inglês: *Artificial Neural Network*) [8], [11], [12].

C. *Seleção de atributos*

Para a seleção de atributos é importante analisar a correlação entre as diferentes variáveis e a variável alvo. 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) [13]. Devido às características dos nossos dados a métrica para o cálculo de correlações entre atributos será o V de Cramer [14].

A interpretação de quão forte é a correlação entre duas variáveis categóricas nominais a partir dos valores obtidos pelo V de Cramer é dada pela Tab. 1.

Tabela 1 - Interpretação do valor de V de Cramer, adaptado de [15]

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 atributos categóricos nominais e categóricos numéricos, um bom indicador é o teste de Kruskal Wallis. Este tem como objetivo verificar se existe uma diferença entre vários grupos independentes quando esses grupos não apresentam uma distribuição normal. Neste caso os grupos não necessitam de ter qualquer tipo de distribuição.

Como foi explicado, o V de Cramer considera duas variáveis como independentes se o valor da frequência expectável 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 tendem para o mesmo valor [13], [15]–[17].

Para ser possível obter um critério universal para eliminar atributos com valores de correlação baixos, é importante que todas as correlações calculadas sejam comparáveis. Se não for possível comparar pode-se utilizar modelos de avaliação diferentes. 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 [16], [17]:

Além da análise de correlação foram utilizados dois algoritmos de seleção de atributos, o Relief-based feature selection (RBA) e o Sequential Backward Selection (SBS).

D. *Medidas de desempenho*

Como medida de desempenho 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.

Para além do MAE, optou-se por utilizar também o erro absoluto médio por percentagem (do inglês: *Mean Absolute Percentage Error*).

Em que At é o valor real e Ft 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. Após este agrupamento podemos obter uma medida de desempenho relacionada com classificação. Optou-se por utilizar a exatidão, que é dada por:

$$Exatidão = \frac{vp+vn}{vp+fp+vn+fn} \quad (1)$$

Em geral, a exatidão mede o rácio de predições corretas no número total de instâncias avaliadas. v_p é o número de verdadeiros positivos, v_n é o número de verdadeiros negativos, f_p o número de falsos positivos e f_n o número de falsos negativos [18].

E. Trabalhos Relacionados

Os trabalhos relacionados mencionados nesta secção estão dentro da classificação supervisionada e dividem-se em métodos clássicos e métodos de aprendizagem profunda.

Relativamente aos métodos clássicos, em 2016, Castro et al. [19], 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 e um algoritmo de rede neural. Todos eles com uma variável de saída com 3 valores possíveis, que representam a severidade do acidente (fatal, grave ou normal). A medida de desempenho 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. [20] procuraram encontrar uma ligação entre as condições das estradas e a severidade do acidente. Neste trabalho já foram incluídos 12 atributos, 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. O melhor resultado obtido foi de 89%.

Em 2019, Hussain et al. [21] 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 e o Naive Bayes.

Apesar dos trabalhos acima mencionados 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 atributos.

Relativamente aos 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. [22] 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. Em 2018, Yuan et al. [23] 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.

III. RESULTADOS E DISCUSSÃO

´Nesta secção são apresentados e analisados os resultados produzidos pela metodologia. 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.

A. Base de Dados

Relativamente à seleção de dados, na Tab. 2 estão indicados todos os atributos selecionados.

Tabela 2 - Atributos selecionados da base de dados da GNR

Atributo	Tipo de dado
Identificação do acidente	Numérico
Data	Data
Hora	Hora
Tipo de Local	Booleano
Localização	Acidente - Localização
Tipo de acidente	Acidente - tipo acidente
Dia da semana	Acidente - dia da semana
Feriado	Booleano
Álcool	Numérico
Contraordenação	Numérico
Factores atmosféricos	Acidente - fatores atmosféricos

Relativamente ao Grupo da Hora do acidente, o maior número de acidentes ocorre nos intervalos hora de trabalho matinal, trabalho de tarde e trânsito de tarde.

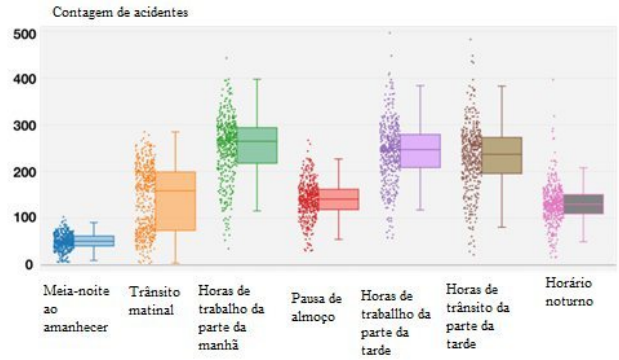

Figura 2 - Diagrama de dispersão e diagrama em caixa da frequência de acidentes ocorridos nos diferentes intervalos de tempo em Beijing. Retirado de Ren et al. [24]

Comparando a Fig. 2 e 3, podemos 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.

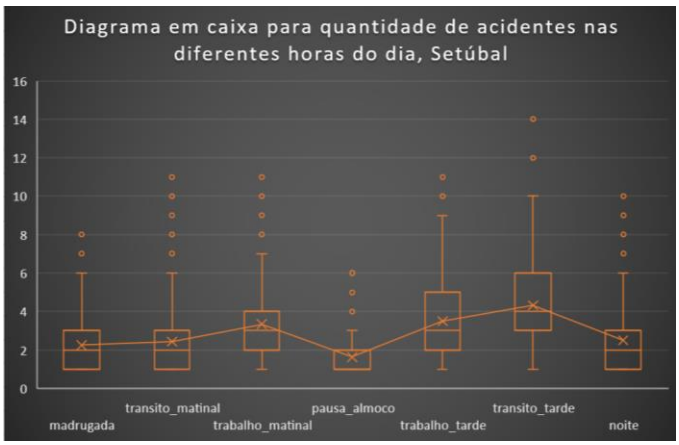

Figura 3 - Diagrama em caixa da frequência de acidentes ocorridos nos diferentes intervalos de tempo em Setúbal, entre 2019-2021.

O intervalo de tempo para os agrupamentos de acidentes em Setúbal foi realizado com base nos horários de trabalho e trânsito mais comuns em Portugal.

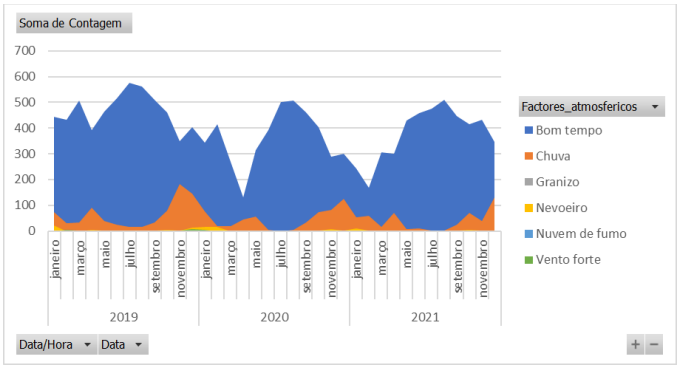

Figura 4 - Número de acidentes agrupados por ano, mês e tipo de condição atmosférica

Relativamente aos fatores atmosféricos foi possível dividir os acidentes pelas diferentes condições atmosféricas em que ocorreram. Considerando a probabilidade de chover como P(C) e a probabilidade de haver acidente como P(A), o gráfico da Fig.4 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:

$$P(C|A) = \frac{144+126+132}{404+300+346} = 0.38 \quad (10)$$

Através do número médio de dias de chuva para o mês de dezembro em Setúbal, obtemos então que:

$$P(C) = \frac{8,5}{31} = 0,27 \quad (9)$$

Pelo teorema de Bayes, obtém-se que:

$$P(A|C) = \frac{P(C|A).P(A)}{P(C)} \quad (10)$$

Utilizando o mesmo raciocínio para o bom tempo, conclui-se que a probabilidade de existir acidente sabendo que está a chover é maior que a probabilidade de existir acidente, sabendo que está bom tempo:

$$P(A|B) = 0,85.P(A) < 1,4.P(A) = P(A|C) \quad (11)$$

B. *Seleção de atributos*

Os resultados obtidos para RBA e SBS foram analisados apenas para as autoestradas, pois concluiu-se que somente para este tipo de localização foi possível obter bons resultados.

Tabela 3 - Relevância de atributos obtida através dos algoritmos de RBA e SBS, para acidentes ocorridos em autoestradas

Autoestradas	Atributos considerados relevantes por ambos os algoritmos	Atributos considerados irrelevantes por ambos os algoritmos

RBA & SBS	'Chuva', 'trabalho_matinal', 'transito_tarde', 'Sexta-feira', 'Sábado', 'agosto', 'fevereiro'	'domingo'

Para o algoritmo RBA, foi utilizada a variante RBA com aplicação em problemas de regressão, que recebe o nome de RReliefF. O número de amostras aleatórias foi de 200, de 1005 acidentes em rodovias, e cada amostra foi comparada com seus 4 vizinhos mais próximos, pois o algoritmo KNN obteve melhores resultados para 4 vizinhos. Quanto ao algoritmo SBS, este usa um algoritmo de mineração para medir o desempenho removendo cada um dos recursos. O algoritmo escolhido foi aquele que obteve os melhores resultados para mineração de dados, ou seja, a rede neural. E a medida de desempenho foi a mesma utilizada na avaliação de algoritmos de mineração, o MAPE. Ao remover iterativamente cada uma das características, podemos ver quais melhoram ou pioram o desempenho da rede neural.

Embora os algoritmos meçam a importância de cada variável de forma diferente, existem diversas variáveis em que ambos os algoritmos concordam que essas variáveis influenciam ou não os acidentes, como podemos ver na Tab. 3. A partir disso pode-se concluir que as variáveis que mais influenciam na ocorrência de acidentes nas rodovias, de acordo com os algoritmos utilizados são: o fator atmosférico chuva, os grupos de horas das 10:00 às 12:29 e das 17:00 às 19:59, os dias da semana sexta e sábado, e os meses de agosto e fevereiro.

Foi possível obter os diferentes valores de correlação da Fig.5 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).

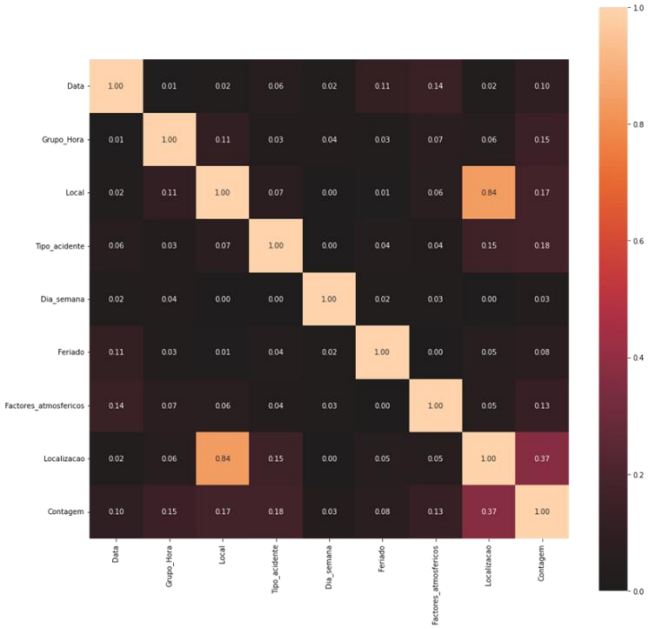

Figura 5 - Correlações de V de Cramer e do teste de Kruskal Wallis dependendo do tipo de pares de variáveis

Pela Fig. 5 podemos observar que para a variável "*Contagem*" que representa a contagem de acidentes, as variáveis com maior correlação são a hora do dia, o tipo de local, a localização e os fatores atmosféricos. O tipo de acidente, que representa a severidade do acidente, foi considerado apenas para se verificar se havia correlação entre a severidade do acidente e o número de acidentes ocorridos, que se confirma. No entanto, não é útil para o modelo preditivo, já que não é uma informação que se consiga obter à priori do acidente.

C. Algoritmos de mineração

Devido à importância da localização dos acidentes, após as experiências iniciais, optou-se por agrupar os dados pela sua localização: Autoestradas; Estradas Nacionais ou Itinerários; e Municípios. Para além disso, visto que se pretende obter um risco de acidentes e não um valor exato de acidentes (o valor obtido pela regressão), optou-se por dividir os valores em intervalos que correspondem a classes de risco, definidos pela Tab. 4.

Tabela 4 - Intervalos correspondentes a classes de risco.

Classificação	Intervalos de frequência de acidentes
Baixo Risco	<1.5
Médio Risco	>1.5 ; < 2.5
Alto Risco	>2.5

Esta opção foi tomada com base nos diagramas em caixa da Fig. 6. Como se pode observar, a variância dos valores é maior para o conjunto de dados relativo às autoestradas e aos municípios.

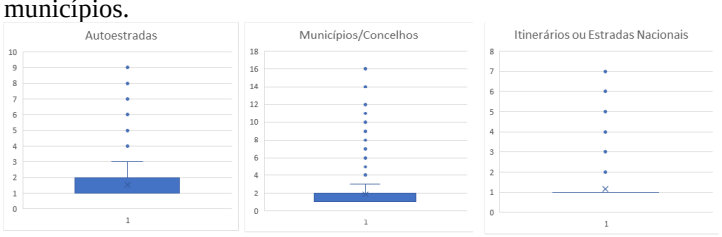

Figura 6 - Diagramas em caixa para os valores da frequência de acidentes nas diferentes localizações

No conjunto de dados de itinerários ou estradas nacionais, para o intervalo de tempo escolhido, na maioria dos casos só existe 1 acidente, por isso seria necessário ter uma frequência maior de acidentes para o modelo ser útil.

Tabela 5 - 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%

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 e por essas exceções terem valores muito elevados, 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 e o facto de a exatidão se manter alta deve-se a que a divisão em classes tem intervalos grandes, quando a variância dos acidentes para este tipo de localização é pequena.

IV. CONCLUSÕES

Nesta dissertação, foi proposto o tema da previsão do risco de acidentes rodoviários através de métodos de mineração de dados. Foram disponibilizados pela GNR dados sobre os relatórios de acidentes, dos acidentes ocorridos em Setúbal de 2019 a 2021. Este trabalho visa criar um modelo que consiga fazer uma previsão com baixo erro.

Apesar das limitações dos dados conseguiu-se obter um bom modelo para as 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 com maior concentração de acidente por autoestrada, quando comparado com a concentração de acidentes em itinerários ou estradas nacionais.

A percentagem de erro na regressão foi de 44%, no entanto não se pretendeu obter o número exato de acidentes, pelo que os resultados foram agrupados em 3 classes de risco, de acordo com o diagrama de caixa obtido para a frequência de acidentes. Assim, obteve-se uma percentagem de erro de apenas 11%.

Desta forma é possível a criação de um aplicativo em que o militar, através de dados de entrada relativos ao mês, dia da semana, grupo de hora do dia, se é feriado ou não, se é dentro de uma localidade ou não, a previsão atmosférica para esse dia e local e, por último, a autoestrada que quer prever o risco, possa verificar o risco de acidente para este conjunto de entrada.

REFERÊNCIAS BIBLIOGRÁFICA

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

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

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

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

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] L. A. Belanche and F. F. González, "Review and Evaluation of Feature Selection Algorithms in Synthetic Problems," no. December 2013, 2011, [Online]. Available: http://arxiv.org/abs/1101.2320.

[19] 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.

[20] 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]. Available: https://www.irjet.net/archives/V3/i10/IRJET-V3I1041.pdf.

[21] 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.

[22] 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.

[23] 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.

[24] 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