ANÁLISE DA SEVERIDADE DAS LESÕES DOS PEÕES UTILIZANDO TÉCNICAS DE DATA MINING

Joana Sousa, João P. Dias[1], Conceição Amado[2] e Kenny Santos[1]

[1]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

email: joao.pereira.dias@tecnico.ulisboa.pt

[2]CEMAT, Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

Sumário

Os acidentes rodoviários que envolvem peões são uma problemática a nível mundial. Em Portugal a taxa de fatalidades por milhão de habitantes é muito elevada. Este estudo procura encontrar quais os fatores determinantes no aumento da severidade das lesões dos peões, de forma a minimizá-los. Foram utilizadas redes bayesianas, árvores de decisão e regressões logísticas. Ao longo da análise foram identificadas causas comuns aos diversos modelos: acidentes em zonas rurais, no distrito de Lisboa, em autoestradas, itinerários principais ou complementares e acidentes que envolvam veículos pesados. Estes fatores foram apontados como relevantes no aumento da gravidade das lesões dos peões.

Palavras-chave: árvores de decisão; data mining; peões; redes bayesianas; regressão logística ordinal, segurança rodoviária.

1 INTRODUÇÃO

As estatísticas de acidentes rodoviários na Europa mostram a necessidade de implementar mecanismos de análise e previsão de acidentes. Segundo a Organização Mundial de Saúde (OMS), estima-se que as lesões resultantes de acidentes rodoviários se tornem a sétima principal causa de morte até 2030 [1]. Em 2012 a Autoridade Nacional de Segurança Rodoviária (ANSR) registou em Portugal 29 867 acidentes com vítimas, dos quais resultaram 38 105 feridos. Para o mesmo período, registaram-se 5 245 peões vítimas em acidentes, dos quais 159 sofreram lesões fatais. Note-se que o número de peões mortos em Portugal correspondeu a 22% do total de mortes nas estradas no ano de 2012 [2]. Apesar de, em quase todos os estados membros da união europeia se ter verificado um declínio significativo na taxa de fatalidades por milhão de habitante, Portugal continua a ser um dos países com taxa mais elevada. Estes números levam a crer que, apesar de todas as estratégias aplicadas até agora, a severidade dos acidentes que envolvem peões em Portugal é um problema real que requer o desenvolvimento e implementação de medidas específicas de segurança rodoviária.

Diversos estudos têm analisado as causas que contribuem para a severidade das lesões nos acidentes, tais como, fatores relacionados com o condutor, características do veículo e condições da via, maioritariamente recorrendo a modelos de regressão logística ([3], [4], [5] e [6]) ou probit ordinais ([7], [8] e [9]). Os modelos de regressão logística e probit pertencem à família dos modelos lineares generalizados, os quais assentam sobre hipóteses pré definidas na determinação das relações entre as variáveis explicativas e a variável resposta como se verifica em [10] e [11]. De forma a evitar estimações incorretas ou dificuldade em lidar com as características próprias deste tipo de modelos alguns estudos utilizaram outras técnicas, tais como, árvores de decisão ou, menos comum, redes bayesianas e redes neuronais.

Os modelos de árvore de decisão foram adotados, por exemplo no estudo [10], essencialmente por não assumir à priori uma relação pré-definida entre as variáveis dependentes e independentes. Neste estudo, o objetivo foi estabelecer uma relação entre a severidade das lesões e as características do acidente. Chang e Wang [8], identificaram a categoria do veículo como a característica que influência a severidade do acidente. Para além disso, verificou-se também que os peões, os ciclistas e os motociclistas são aqueles que correm maiores riscos de sofrerem uma lesão. Por outro lado, as redes bayesianas foram aplicadas em [11] para classificar acidentes rodoviários de acordo com a sua severidade. No presente trabalho, foram construídas três redes bayesianas, uma para cada nível de severidade (ligeira, grave e fatal) e, foi possível concluir que, o tipo de acidente, a idade do

condutor do veículo, as condições de luminosidade e o número de vítimas são os fatores que estão fortemente associados a fatalidades e lesões graves.

Apesar de ser possível encontrar diversos estudos a apontar as vantagens de cada um destes métodos (regressões, árvores de decisão e redes bayesianas) são raros aqueles que de facto comparam os métodos entre eles para este tipo de base de dados. Um dos poucos estudos que apresenta uma comparação entre redes bayesianas e modelos de regressão é o apresentado em [12]. Este estudo concluiu que as redes bayesianas eram o modelo mais adequado para prever a severidade dos acidentes.

Contrariamente ao que acontece com a severidade global de um acidente, a severidade das lesões dos peões é um assunto muito menos estudado. Os poucos estudos que se focaram nessa questão recorreram maioritariamente aos modelos de regressão logística [13] e [14].

Este trabalho pretende identificar e entender os fatores que contribuem para o aumento da severidade das lesões dos peões, focando-se num histórico de acidentes que resultaram em peões feridos. Para isso, a ANSR disponibilizou a base de dados com informação detalhada dos condutores e peões envolvidos em acidentes. O objetivo é determinar quais os fatores que realmente influenciam a severidade das lesões dos peões. Nesta análise serão utilizadas diversas técnicas de *Data Mining*. Por último, serão comparados os resultados obtidos com as diferentes técnicas de forma a determinar qual delas dará o melhor contributo para este problema.

2 MODELOS ESTATÍSTICOS

Neste capítulo faz-se uma descrição resumida dos modelos estatísticos das redes bayesianas à regressão logística ordinal, passando pelas árvores de decisão.

2.1 Redes Bayesianas

Uma rede bayesiana pode ser vista como uma representação gráfica de uma distribuição de probabilidade de um conjunto de k variáveis aleatórias $\{X_1, \dots, X_k\}$. Esta pode ser descrita pelo par $\mathcal{B} = (\boldsymbol{X}, \mathcal{G})$ composto por uma variável aleatória $\boldsymbol{X} = \{X_1, \dots, X_k\}$ e um grafo acíclico dirigido (DAG) $\mathcal{G}$. Este grafo é composto por um conjunto de k vértices (os quais representam cada uma das variáveis aleatórias X_i, $i = 1, \dots, k$) e um conjunto de arcos entre esses mesmos vértices, os quais representam as dependências diretas entre as variáveis. Para cada variável X_i, com predecessores $Y_1, \dots, Y_n$ está associada uma tabela de probabilidade condicional $\mathrm{P}(X_i|Y_1, \dots, Y_n)$. No caso da variável X_i não ter predecessor, a tabela de probabilidade condicional reduz-se à probabilidade $\mathrm{P}(X_i)$. Considere $\prod_{X_i}$, o conjunto de variáveis aleatórias que são os predecessores imediatos das variáveis X_i. De acordo com $\mathcal{G}$, a distribuição de probabilidade conjunta de $\mathcal{B}$ é dada por:

$$\mathrm{P}_{\mathcal{B}}(X) = \mathrm{P}(X_1 = x_1, \dots, X_k = x_k) = \prod_{i=1}^{k} \mathrm{P}_{\mathcal{B}}\left(x_i|\prod_{X_i}\right) \tag{1}$$

onde x_i é a realização de X_i. Por outras palavras, a distribuição de probabilidade conjunta da rede pode ser fatorizada e, ser vista como o produto da distribuição de probabilidade de cada nó e dos seus predecessores.

Para construir uma rede bayesiana é necessária especificar a relação de dependência condicional (ou independência) entre as variáveis e a distribuição de probabilidade condicional. A estrutura da rede pode ser definida *a priori* ou através de métodos de aprendizagem de estruturas. Os algoritmos que aprendem a estrutura de uma rede bayesiana podem ser classificados como: algoritmos *constraint-based*, algoritmos *score-based* e algoritmos híbridos.

Os algoritmos *constraint-based* usam testes de independência condicional no conjunto de dados e procuram uma rede consistente com o observado. Os algoritmos *score-based* classificam as estruturas da rede de acordo com a capacidade de ajustamento desta. Numa primeira fase, o algoritmo quantifica o ajustamento da rede aos dados, tendo em conta o critério estabelecido. Em segundo lugar, define um sistema de pesquisa de forma a encontrar uma estrutura que maximize o critério escolhido. Ambos os passos são aplicados iterativamente até que uma nova iteração não apresenta melhorias ao resultado, ou seja, não é encontrada estrutura que seja melhor do que qualquer uma já obtida. Como função de *score* podem considera-se dois tipos de função: classificação bayesiana, tais como o K2, o Bayesian Dirichlet (BD) e as suas variantes (BDe e BDeu); classificação de informação

teórica, tais como, o Log-Likelihood (LL), AIC ou BIC. Para o segundo passo de um algoritmo *score-based* é necessário definir um algoritmo de procura através da rede, o qual pode ser um algoritmo clássico de procura como o *hill-climbing*, *tabu search* ou até mesmo o *simulated annealing*.

2.2 Árvores de Decisão

Existem diversos tipos de algoritmos de árvores de decisão, um dos mais utilizados é o CART (árvore de regressão e classificação). Uma árvore CART é uma árvore binária, ou seja, é uma árvore construída através da divisão sucessiva de cada um dos nós em dois. O primeiro passo neste modelo é encontrar a variável que é a melhor divisão do conjunto, considerada posteriormente a raiz da árvore. E, de seguida, encontrar a melhor divisão em dois dessa mesmo variável. Os caminhos construídos entre a raiz e as folhas da árvore são considerados as regras de classificação. Para um dado nó t, a divisão s é selecionada de forma a maximizar o critério de divisão definido $\Delta G(s,t)$. O critério mais utilizado é o coeficiente de Gini, um dos critérios que permite utilizar dados categóricos.

Árvores de inferência condicionada são um outro algoritmo desenvolvido em [18] para regressão e classificação que usa partições binárias recursivas tal como o algoritmo CART. Este novo algoritmo baseia a construção da estrutura da árvore em propriedades estatísticas, a qual pode ser dividida em três etapas:

1. Testar a hipótese nula de independência entre cada variável exploratória e a variável resposta. Este passo termina se a hipótese nula não for rejeitada, caso contrário, seleciona a variável com o valor-p mais elevado para o teste da hipótese nula parcial.
2. Aplicar a divisão binária à variável selecionada.
3. Iterar o passo 1 e o passo 2 recursivamente para os dois elementos obtidos após cada divisão, até que a hipótese nula de independência não possa ser rejeitada.

2.3 Regressão Logística Ordinal

Um modelo de regressão logística ordinal é uma técnica estatística utilizada para analisar a relação entre as diferentes classes ordinais da variável resposta e uma ou mais variáveis explicativas. Seja Y a variável de resposta ordinal com m categorias. Considere que apenas é possível observar Y em pontos de corte específicos de Y^*, o que pode ser visto como as regiões de corte das diferentes categorias da variável. Formalmente, ${Y_i}^*$ é definido como uma variável aleatória continua com $i = 1, \dots, n$ e π_{ic} a probabilidade de i pertencer à categoria c com $c = 1, \dots, m$. Assim, os pontos de corte $\gamma_{i,1}, \dots, \gamma_{i,m-1}$ da distribuição de ${Y_i}^*$para cada i podem ser definidos por:

$$\begin{aligned} \mathrm{P}({Y_i}^* \le \gamma_{i1}) &= \pi_{i1} \\ \mathrm{P}\left(\gamma_{i,c-1} \le {Y_i}^* \le \gamma_{ic}\right) &= \pi_{ic}, \quad c = 2, \dots, m-1 \\ \mathrm{P}\left({Y_i}^* > \gamma_{i,m-1}\right) &= \pi_{im} \end{aligned} \tag{2}$$

Considere as probabilidades cumulativas ${\pi_{ik}}^* = \sum_{c=1}^{k} \pi_{ic}$, com $0 < \pi_{ic} < 1,\ k = 1, \dots, m$ e ${\pi_{i,c}}^* > {\pi_{i,c-1}}^*$. O modelo logit cumulativo com probabilidades proporcionais é dado por:

$$\begin{aligned} \log\left(\frac{\mathrm{P}(Y^* \le \gamma_{ic})}{\mathrm{P}(Y^* > \gamma_{ic})}\right) &= \log\left(\frac{\pi_{i,1} + \dots + \pi_{i,c}}{\pi_{i,c+1} + \dots + \pi_{i,m}}\right) \\ &= \beta_{0,c} - \beta_{1,c}x_{i,1} - \dots - \beta_{k-1,c}x_{i,k-1} \end{aligned} \tag{3}$$

onde $x_1, \dots, x_{k-1}$são as $k-1$ variáveis explicativas e $\beta_{0,1} < \dots < \beta_{0,m}$ de forma a garantir que $\pi^*_{i,c} \ge \pi^*_{i,c-1}$. O preditor linear $\boldsymbol{x}^{\boldsymbol{T}}\beta_c$ é restrito, assim, $\beta_{0,c}$ depende de c, mas os efeitos de todas as outras variáveis explicativas permanecem constante ao longo de todas as categorias da variável resposta (probabilidades proporcionais).

3 SELEÇÃO DE MODELOS

Nesta secção apresentam-se as métricas utilizadas para a seleção do modelo que melhor se ajusta aos dados. Conjuntamente com esta seleção efetuou-se uma análise de seleção de variáveis, por forma a encontrar as características que poderiam ser mais relevantes para o ajuste dos modelos. Com esse objetivo, várias técnicas

foram usadas, nomeadamente critérios de informação mútua, de mínima redundância máxima relevância (mRMR) [15], assim como métodos de seleção passo a passo. Para além disso, e uma vez que as categorias de severidade dos peões não estavam igualmente representadas na base de dados foi necessário utilizar técnicas para o equilíbrio dos dados. Para mais detalhes sobre estas dois tópicos ver [17].

3.1 Métricas de desempenho

Podem ser ajustados um sem número de modelos ao mesmo conjunto de dados. Diferentes algoritmos podem ser aplicados, ou até o mesmo algoritmo pode ser aplicado com variações nos parâmetros que o definem. Em problemas reais não é possível selecionar o modelo perfeito, assim, é necessário conseguir identificar qual o melhor modelo para cada caso. Existem diversas medidas que podem ser utilizadas para avaliar o desempenho de cada modelo, por norma baseadas nas medidas que constituem a matriz de confusão. Uma matriz de confusão é um caso especial de uma tabela de contingência, que permite visualizar o desempenho de um modelo. O Quadro 1 representa uma matriz de confusão para um modelo cuja variável resposta apresenta três classes distintas.

Quadro 1. Exemplo de uma matriz de confusão para uma variável resposta com três classes.

		Classe 1	Classe 2	Classe 3
Modelo	Classe 1	TP_1	FN_{12}	FN_{13}
	Classe 2	FN_{21}	TP_2	FN_{23}
	Classe 3	FN_{31}	FN_{31}	TP_3

No Quadro 1, TP_i descreve os verdadeiros positivos da i-ésima classe, ou seja, as observações que o modelo prevê pertencerem à classe i e, que de facto, pertencem a essa mesma classe. Por outro lado, FN_{ij}, descreve os falsos negativos para a i-ésima classe. Neste caso, o modelo prevê que as observações pertencem à classe i quando na verdade estas pertencem à classe j. O Quadro 2 sumaria as principais medidas que podem ser estimadas com base nos verdadeiros positivos e falsos positivos apresentados numa matriz de confusão.

Quadro 2. Medidas de desempenho, baseadas em probabilidades, onde ϕ denota a classificação no modelo, c^+ as observações classificadas como classe i que pertencem à classe i e c^- as observações classificadas como classe i que de facto pertencem a uma outra classe da variável resposta.

Exatidão	Precisão
$p(\phi(\boldsymbol{X}) \neq C)$	$p(C = c^+ \mid \phi(\boldsymbol{X}) = c^+)$
Sensibilidade	Especificidade
$p(\phi(\boldsymbol{X}) = c^+ \mid C = c^+)$	$p(\phi(\boldsymbol{X}) = c^- \mid C = c^-)$

Sem perda de generalidade, serão utilizados os termos exatidão, precisão, sensibilidade e especificidade como as medidas estimadas em vez das suas verdadeiras probabilidades. Exatidão avalia de forma geral o desempenho do modelo, estimando a quantidade de observações que o modelo classificou correta ou incorretamente. A precisão representa a fração de observações dos dados que foram classificadas como classe i que, de facto, pertencem à classe i. A sensibilidade mede a classe positiva de observações corretamente identificadas. Num problema com três categorias, irão existir três classes positivas (uma por cada categoria). Por outro lado, a especificidade mede a fração de classes negativas corretamente identificadas. Note que, por exemplo, quando se calcula a sensibilidade para a classe 1 da matriz de confusão do Quadro 2, c^+ representa a classe 1. Por outro lado, quando é calculada a especificidade para a classe 1, c^- representa as restantes classes (classe 2 e classe 3 num problema com três categorias). Uma outra medida igualmente utilizada é a média harmónica da precisão e da sensibilidade, usualmente designada de F-score e dada por:

$$F = 2 \times \frac{precisão \times sensibilidade}{precisão + sensibilidade} \tag{4}$$

Para analisar o desempenho dos modelos, é utilizado igualmente a curva ROC. A curva ROC é o gráfico que ilustra o desempenho dos modelos ao representar os verdadeiros positivos (sensibilidade) e os falsos positivos

(1-especificidade). A área sob a curva ROC, também chamada de AUC, é outra medida amplamente utilizada, pois permite avaliar a capacidade do modelo para distinguir as diferentes categorias de resposta. No problema multi-classe, existe uma especificidade e sensibilidade para cada categoria da variável resposta. Ou seja, existe uma curva ROC e uma AUC para cada uma dessas classes. Com base nesse facto, o estudo [19] sugere a extensão da definição da AUC para um problema multi-classe.

4 CONJUNTO DE DADOS/BASE DE DADOS

O conjunto de dados diz respeito aos acidentes rodoviários que ocorreram em Portugal em 2011-2012 e envolveram peões. Esta base de dados foi disponibilizada pela ANSR e é composta por informação detalhada de todos os acidentes registados com vítimas. A escolha deste período foi devida a terem sido realizadas anteriormente análises detalhadas com métodos de regressão linear, para este período.

Cada observação diz respeito a um peão ferido e contém informação que pode ser dividida em três tipos: informação do acidente (incluí condições da via, tipo de veículo, etc), dados sobre o condutor e dados relativos ao peão. Neste conjunto de dados, corresponde às vítimas a 30 dias em que é utilizada a definição de fatalidade adotada na convecção de Viena de 1968, a qual considera que uma fatalidade é toda a vítima que faleceu até 30 dias após o acidente. Por outro lado, um ferido grave é todo o peão que necessitar de assistência médica nas primeiras 24h após o acidente, mas que não faleceu nos primeiros 30 dias após o mesmo. Informação mais detalhada encontra-se no relatório anual da ANSR [2].

O conjunto de dados é composto por 8 431 peões feridos, dos quais 7 479 sofreram ferimentos ligeiros, 666 foram gravemente feridos e 286 faleceram. A variável resposta é designada de "Lesões Peões" e é composta por três categorias, "leve", "grave" e "fatal" (variável categórica ordinal, com ordenação de acordo com a gravidade). Uma análise primária à base de dados permite concluir que o conjunto de dados é extremamente não equilibrado face ao número de observações em cada uma dessas categorias. De forma a lidar com esta situação, o conjunto foi equilibrado através do método SMOTE [16]. Optou-se também por dividir o conjunto de dados iniciais num conjunto de treino (composto por 75% das observações iniciais) e num conjunto de teste com as restantes 25%, de forma a poder ser avaliado o desempenho de cada um dos modelos.

Em todo o conjunto de dados existem 24 variáveis das quais três são contínuas, duas delas relacionadas com a idade (do condutor e do peão) e uma outra relacionada com o teor de álcool no sangue do condutor. Uma análise detalhada a estas variáveis mostra que a severidade das lesões dos peões varia entre classes de idades e diferentes níveis de álcool. Assim, decidimos categorizar estas três variáveis e reorganizar diversas classes de outras variáveis categóricas.

Como referido, foi efetuada a seleção de variáveis com o objetivo de determinar quais das 24 variáveis deveriam ser incluídas no modelo. A seleção foi feita com base nas medidas de correlação, na informação mútua, no critério mRMR e na regressão passo a passo. A seleção foi aplicada ao conjunto original e ao conjunto após a categorização de variáveis e reajuste de categorias. Após análise aos resultados de todas as abordagens foi necessário definir critérios que permitissem obter um conjunto final. O primeiro critério foi escolher as variáveis selecionadas por todos os critérios, e estas foram o "Tipo Via" e "Idade Peão". O segundo critério foi a escolha das variáveis selecionadas pelo menos por um conjunto de dados em cada critério e, nesse caso, foram selecionadas as seguintes: "Região", "Localização", "Luminosidade", "Idade Condutor", "Lesões Condutor", "Categoria Veículo"," Álcool" e "Ações Peão". O conjunto final de variáveis selecionadas são as apresentadas no Quadro 3.

Quadro 3. Conjunto final das variáveis selecionadas.

Conjunto final de variáveis		
Tipo Via	Distrito	Localização
Luminosidade	Lesões Condutor	Categoria Veículo
Álcool	Idade Peão	Ação Peão

5 RESULTADOS

Neste capítulo apresentam-se os resultados agrupados por modelo estatístico.

5.1 Redes bayesianas

Foi utilizado o pacote do R *bnlearn* no processo de encontrar a melhor rede bayesiana para este problema. Foram utilizados dois tipos de algoritmos de aprendizagem para construir as redes: *hill-climbing* e *tabu search*. Foram aplicadas as funções *score* para variáveis categóricas (AIC, BIC, BDe, mBDe, K2 e loglik). As diferenças entre os dois algoritmos mostraram-se mínimas, como tal, apenas serão analisados os resultados do algoritmo *hill-climbing*. Analisando as medidas de desempenho para os modelos obtidos foi possível verificar que os *scores* mBDE e BDe apresentam os valores mais elevados de exatidão, sensibilidade e especificidade. No entanto, os modelos aprendidos através destes dois *scores* não são capazes de encontrar uma exata identificação das causas das lesões graves e lesões fatais nos peões. Ponderando todas as medidas de desempenho dos restantes *scores*, foi escolhido como melhor modelo o construído com o *score* loglik. A escolha recaiu sobre esta medida, uma vez que, neste problema em particular, considera-se mais relevante identificar corretamente todas as lesões graves e fatais nos peões (rácio verdadeiros positivos) do que identificar todos os casos em geral (exatidão). Em suma, o modelo escolhido foi aprendido com o algoritmo *hill-climbing*, com o *score* loglik e, a estimação dos seus parâmetros foi feita através do método de máxima verosimilhança.

Uma vez que especificadas as ligações entre os nós da rede, o próximo passo é quantificar as relações entre esses mesmos nós. A quantificação das relações é feita estabelecendo a distribuição de probabilidade condicional de cada nó. Neste problema em questão, todas as variáveis são discretas e, como tal, são representadas numa tabela de probabilidade condicional (CPT). Cada nó tem uma CPT e, para analisá-la é necessário considerar todas as combinações de valores dos nós predecessores. Neste estudo, o objetivo é encontrar as causas para a severidade das lesões dos peões, assim, o nó ” Lesões Peões” e os seus predecessores, são os de maior interesse. Em particular, na rede selecionada para proceder à análise deste nó é necessário ter em consideração todos os outros nós, uma vez que todos eles são seus predecessores. Assim sendo, a CPT para este nó é constituída pelas estimativas das probabilidades condicionais para todas as combinações de valores de todos os outros nós, o que resulta em centenas de tabelas, tornando-se incomportável analisá-las individualmente. Apresenta-se de seguida, algumas dessas tabelas com os valores considerados mais relevantes. O Quadro 4 representa a probabilidade estimada para as lesões dos peões para o tipo de via condicionada pelas seguintes características: a localização ser em áreas urbanas, as condições de luminosidade serem luz do dia, localizar-se no distritos de Lisboa, o condutor ficar ileso, o veículo ser do tipo ligeiro, o teor de álcool no sangue do condutor ser inferior a 0.2 g/L, o peão ter menos de 20 anos e este ter realizado uma ação incorreta do ponto de vista do código da estrada. De forma a simplificar, e eliminar informação irrelevante, os Quadro 4 e 5 centram-se nos arruamentos e nos restantes tipos de vias.

Quadro 4. Probabilidade condicional estimada para cada tipo de via por tipo de lesão do peão no distrito de Lisboa

	TIPOS DE VIA	
	Arruamentos	Outros
Lesões leves	0.6667	0.0000
Lesões graves	0.3333	0.5714
Lesões fatais	0.0000	0.4286

A análise ao Quadro 4 mostra que no caso dos arruamentos a probabilidade estimada mais elevada diz respeito às lesões leves. Para todo o outro tipo de vias são as lesões graves e as lesões fatais que apresentam probabilidades estimadas mais elevadas. Considere-se agora o

Quadro 5, o qual apresenta o mesmo caso que o quadro anterior relativo a vias do distrito do Porto.

Quadro 5. Probabilidade condicional estimada para cada tipo de via por tipo de lesão do peão, no distrito do Porto.

	TIPOS DE VIA	
	Arruamentos	Outros
Lesões leves	0.7333	0.0000
Lesões graves	0.2667	0.9999
Lesões fatais	0.0000	0.0000

Comparando os resultados de ambas os quadros, as diferenças mais significativas encontram-se na categoria de outros tipos de vias. A probabilidade estimada para a morte de um peão decresce de 0.4286 para praticamente 0.000 quando comparamos o distrito de Lisboa com o do Porto. No entanto, a probabilidade estimada mais elevada continua a pertencer às lesões graves. No que diz respeito ao tipo de vias arruamentos não existe praticamente alteração.

5.2 Árvores de decisão

A segunda abordagem passou por ajustar uma árvore de decisão ao conjunto de dados. Foram testados dois tipos de árvores: uma árvore de decisão simples (DT) e uma árvore de decisão condicional (CT). Neste caso a árvore DT é uma árvore construída com base no algoritmo CART, enquanto as CT são árvores de inferência condicional.

Ambas as abordagens apresentaram valores de medidas de desempenho bastantes idênticos, com uma ligeira vantagem para a CT. No entanto, os valores de sensibilidade, especificidade e F-score da DT são mais equilibrados para as três categorias da variável resposta. Por esta razão, e tendo em conta a dimensão da CT e consequentemente a dificuldade de visualização dos seus resultados optou-se por analisar os resultados da DT.

O critério de divisão utilizado para construir a DT foi o *Gini Gain*. Existem cinco divisores primários da árvore e são eles: a localização do acidente, a categoria do veículo, a idade do peão, o distrito e o tipo de via. Estas são as variáveis que terão impacto na análise à severidade das lesões dos peões em acidentes rodoviários.

A primeira divisão da árvore é a da variável localização do acidente, o que significa que esta é a melhor variável para avaliar a severidade das lesões dos peões. A árvore indica que os acidentes que ocorrem em zonas rurais têm tendência a provocar lesões mais graves nos peões (61%). Enquanto nas áreas urbanas os peões têm mais tendência a sofrer lesões leves (53%). Em relação à categoria do veículo a árvore mostra que quando existem veículos pesados envolvidos no acidente os peões têm 55% de tendência de sofrerem lesões fatais contra 28% lesões graves e apenas 17% de lesões leves. A variável tipo de via, quando condicionada à idade do peão, à categoria do veículo e à localização do acidente é dividida em dois nós pela árvore. No nó esquerdo encontram-se os arruamentos que apresentam 71% de probabilidade de os peões sofrerem lesões ligeiras. No nó direito encontram-se todos os outros tipos de via onde os peões têm 53% de probabilidade de sofrerem lesões graves contra 38% de lesões ligeiras. O último fator é o distrito o qual a árvore divide no nó esquerdo Lisboa, Braga e Viana do Castelo e para o nó direito os restantes distritos. Neste caso é nas regiões do nó esquerdo que os peões apresentam uma maior possibilidade de virem a sofrer lesões fatais (45%) face a 39% de lesões ligeiras.

5.3 Regressão logística ordinal

A última abordagem para ajuste de um modelo ao conjunto de dados foi a utilização de um modelo de regressão logística multinominal ordinal uma vez que a variável resposta apresenta três categorias. O teste do rácio de verosimilhança para a significância total das 10 variáveis selecionadas garante que pelo menos uma delas é significante para explicar a severidade das lesões dos peões. De seguida foram analisados os resultados para cada uma das variáveis. Nos quadros 6 a 9 s encontram-se os valores das *odds* (chances), valor p, limite inferior (LI) e limite superior (LS) dos intervalos de confiança a 95% do modelo ajustado. As *odds* representam o aumento ou a diminuição na probabilidade de as lesões pertencerem a cada uma das categorias da variável resposta.

Quadro 6. Estimativa das *odds* para a variável: Localização.

	odds	valor-p	LI	LS
Zonas urbanas	-	-	-	-
Zonas rural	1.81	0.00	1.54	2.13

Em relação à variável localização do acidente, os resultados indicam que, quando os acidentes ocorreram em zonas rurais existe um aumento de 81% das *odds* de ter de lesões mais severas quando comparado com acidentes que ocorreram em áreas urbanas.

Quadro 7. Estimativa das *odds* para a variável: Tipo de Via.

	odds	valor-p	LI	LS
Arruamentos	-	-	-	-
Estradas Nacionais (EN)	1.60	0.00	1.25	2.04
Autoestradas, Itinerários Principais (IP) e Complementares (IC)	2.98	0.00	2.31	3.55
Outros tipos de vias	1.89	0.00	1.65	2.17

Os resultados do Quadro 7 mostram que quando se compara acidentes que ocorreram em arruamentos com todos os outros tipos de vias todos eles apresentam um aumento das *odds*. Acidentes que ocorrem em estradas nacionais, têm um aumento de 60%, acidentes que ocorrem em autoestradas, IP's e IC's têm um aumento de 198% na chance de resultarem em lesões graves ou fatais.

Quadro 8 – Estimativa das *odds* para a variável: Categoria do Veículo.

	odds	valor-p	LI	LS
Veículo Ligeiro	-	-	-	-
Veículo Pesado	4.88	0.00	4.16	5.73
Veículo motorizado de duas rodas	0.99	0.91	0.76	1.28
Outros tipos de veículos	0.57	0.00	0.41	0.78

Comparando com veículos ligeiros, os veículos pesados apresentam um aumento de 388% nas *odds* dos peões sofrerem acidentes onde resultem lesões mais gravosas.

Quadro 9. Estimativa das *odds* para a variável: Álcool do Condutor.

	odds	valor-p	LI	LS
≤ 0.2	-	-	-	-
]0.2,0.5[	1.55	0.06	0.98	2.42
[0.5,1.0[	1.32	0.01	1.06	1.64
≥ 1	1.71	0.00	1.20	2.43

Em relação ao teor de álcool no sangue do condutor, quando se compara todos os níveis à classe dos valores inferiores a 0.2 g/L, verifica-se um aumento significativo nas *odds* de sofrer lesões mais graves e/ou fatais. No entanto, e ao contrário do que seria de esperar, as *odds* não aumentam diretamente com o aumento do nível de álcool. Para valores entre 0.5 g/L e 1.0g/L a *odd* é inferior à classe dos valores entre 0.2 g/L e 0.5 g/L.

Analisando os resultados das condições de iluminação verifica-se que existe um aumento das *odds* para lesões mais graves quando os acidentes não ocorrem à luz do dia. É de salientar ainda que, tal como seria de esperar, acidentes que ocorrem durante a noite sem iluminação ou com iluminação insuficiente têm um aumento de 176% nas *odds* de resultarem lesões mais gravosas.

Considerando a variável distrito e, utilizando Lisboa como base de comparação para todas as outras categorias, verificou-se uma diminuição generalizadas das *odds* para o aumento da severidade das lesões dos peões em todos os outros distritos. Note-se que foi no distrito do Porto que se verificou a descida mais acentuada.

Por último foi analisada a variável idade do peão, considerando a categoria dos peões com idade inferior a 20 anos a base de comparação. Nesse caso, verificou-se que existe um aumento das *odds* para um agravamento da severidade das lesões dos peões quando estes têm entre 20 e 29 anos (77%) e um aumento ainda mais significativo quando têm entre 40 e 49 anos (209%).

6 DISCUSSÃO

Ao longo desta análise foram estudadas três abordagens para encontrar um modelo que melhor explicasse a relação entre a severidade da lesão de peões em acidentes rodoviários com as diversas características dos acidentes. Daqui em diante, para simplificar a análise vamos designar o melhor modelo das redes bayesianas por BN, a melhor árvore de decisão por DT e o modelo de regressão logística ordinal por LR. Com base nas medidas de desempenho não é claro qual dos modelos é a melhor escolha para esta base de dados. Contudo, deve ser tido em conta que, foi o modelo LR que obteve globalmente os melhores resultados para todas as medidas de

desempenho (com exceção do AUC). Por outro lado, foram os modelos LR e DT que obtiveram os valores mais reduzidos de f-score para as lesões graves e lesões fatais. No entanto, de um modo geral foi o modelo BN que obteve os valores mais equilibrados para todas as classes de lesões de peões. Para além disso os resultados da exatidão e da AUC são considerados bastante satisfatórios para o modelo BN, em detrimento dos restantes. Note-se que, apesar de este não ser um problema clássico de classificação, estes modelos têm que ser capazes de identificar as razões por de trás da severidade das lesões dos peões corretamente. Isto significa que, a taxa de verdadeiros positivos e sensibilidade para cada categoria têm uma maior relevância do que a exatidão global. Assim, a rede bayesiana foi considerada o modelo que melhor se ajustava a estes dados.

De uma forma global, o modelo DT identifica o tipo de via, o distrito, a localização do acidente, a categoria do veículo e a idade do peão como os fatores críticos para a análise da severidade das lesões dos peões. De acordo com o modelo, peões com idades compreendidas entre 20 e 29 anos e 40 e 49 anos têm uma maior tendência para sofrer lesões mais gravosas. Em relação ao tipo de via, é nas estradas nacionais, autoestradas, IP's e IC's que os acidentes provocam lesões mais graves e fatais nos peões. Este modelo, em particular, indica que os veículos pesados estão fortemente associados a lesões mais graves nos peões. Por último, a localização do acidente é outra das variáveis críticas, e esta estima uma probabilidade de 61% dos acidentes que ocorrem em zonas rurais resultarem em lesões graves para os peões.

No modelo LR foi a categoria do veículo que apresentou o maior *odd*, mais especificamente, os veículos pesados apresentam um aumento de 388% na odd para um agravamento da severidade das lesões dos peões quando comparados com veículos ligeiros. Para além disso, todos os distritos, quando comparados com Lisboa, apresentam uma diminuição nas odds para o agravamento da severidade das lesões dos peões. Outros fatores relacionados com as condições da via foram considerados igualmente relevantes, como por exemplo, as autoestradas, IP's e IC's que verificam um aumento de 198% das *odds* quando comparados com os arruamentos. Esta situação pode ser explicada pelo facto de os limites de velocidade nestas vias serem muito mais elevados do que nos arruamentos. Por outro lado, os acidentes que ocorrem durante a noite sem iluminação adequada apresentam um aumento de 173% nas *odds* para agravamento da severidade das lesões dos peões quando comparados com acidentes que ocorreram à luz do dia. Acidentes em zonas rurais são outro dos fatores que apresentam um aumento das *odds* quando comparado com acidentes que ocorreram em zonas urbanas. Por último, para peões com idades compreendidas entre os 40 e 49 anos verificou-se um enorme aumento de 209% nas *odds* de agravamento da severidade das lesões dos peões quando comparado com peões com idades inferiores a 20 anos.

Os resultados do modelo BN não são tão simples de analisar como os anteriores, uma vez que apresenta centenas de tabelas de estimativas de probabilidade condicional para caracterizar todas as variáveis. No entanto, a análise à rede mostrou que, a localização do acidente, as condições de luminosidade, as lesões do condutor, a categoria do veículo, o teor de álcool no sangue do condutor, a idade do peão, o tipo de ação do peão, o distrito e o tipo de estrada têm influência direta na severidade das lesões sofridas pelo peão. Para além disso as probabilidades condicionais obtidas nas tabelas permitem sustentar todos os resultados e conclusões obtidos com os modelos DT e CT.

7 CONCLUSÕES

Em suma, todos os modelos finais permitiram encontrar fatores comuns que foram considerados relevantes para o problema em estudo. De facto, a categoria do veículo, o distrito, o tipo de estrada e a localização do acidente foram indicados por todas as abordagens como estando estritamente associados a lesões graves ou fatais dos peões. Mais ainda, o teor de álcool no sangue do condutor e as condições de luminosidade foram considerados relevantes nos modelos BN e LR.

Algumas das conclusões do estudo foram consideradas críticas e deveriam ser tomadas como prioritárias em ações e planos de prevenção que promovam a segurança dos peões na estrada. A deficiência na iluminação das vias de circulação está fortemente associada a lesões mais graves. Campanhas de prevenção para potenciar o uso de material refletor por parte dos peões, quando circulam à noite nas bermas das vias, poderão trazer benefícios para a sua segurança. O material refletor não serve apenas para reduzir a probabilidade de acontecer um acidente, mas serve sim, também, para reduzir a sua severidade, pois se o peão for avistado mais cedo a velocidade de impacto pode ser reduzida. O teor de álcool no sangue do condutor é outro fator a ter em consideração e, neste

caso, deveria ser ponderada a aplicação de multas mais rigorosas de forma a desencorajar os condutores a conduzirem com qualquer nível de álcool.

A redução da velocidade de circulação dos veículos é crucial para a redução da severidade das lesões. Todas as medidas de acalmia de tráfego que reduzam as velocidades de circulação, são benéficas para a segurança dos peões.

Em conclusão, uma base de dados com informação mais detalhada sobre a exposição do peão ao perigo, tais como, distância percorrida, número de passadeiras nas proximidades, número de peões envolvidos no acidente e número de veículos em intersecções poderiam ser fatores a ter em conta numa análise futura. No entanto, obter este tipo de dados é um enorme desafio. Acresce ainda o facto de este tipo de estudo, baseado no registo realizado pelas forças de segurança, não contem informação sobre a dinâmica do acidente e sobre as causas que conduziram à severidade das lesões, isto é, velocidade real de impacto, peso do veículo ou até mesmo o movimento do peão no momento do impacto. Dados deste género apenas são possíveis de recolher através de testes e uma reconstituição cientifica do acidente. Outro fator que não estava presente nesta base de dados e, que se mostrou relevante noutros estudos foi o teor de álcool no sangue do peão.

Por último, e uma vez que foi a rede bayesiana o modelo considerado com melhores resultados sugere-se o uso de outro tipo de modelos de redes, tais como, as redes neuronais. No caso específico das redes bayesianas podem ser igualmente testados outro tipo de algoritmos de aprendizagem, tais como o *simulated annealing* ou algoritmos genéticos.

8 AGRADECIMENTOS

Este trabalho foi suportado pela FCT, pelo IDMEC, sobre o LAETA, UID/EMS/50022/2019.

Agradece-se à ANSR (Autoridade Nacional para a Segurança Rodoviária) por facultar a base de dados dos atropelamentos.

9 REFERÊNCIAS

1. World Health Organization (2015). Global status report on road safety 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/
2. Autoridade Nacional de Segurança Rodoviária - Portugal National Road Safety Authority (2015). Relatório anual - vítimas a 30 dias.
3. Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention, 34(06):729–741.
4. Bédard, M., Guyatt, G. H., Stones, M. J., and Hirdes, J. P. (2002). The independent contribution of driver, crash and vehicle characteristics to driver fatalities. Accident Analysis & Prevention, 34(06):717–727.
5. Yau, K. K. W., Lo, H. P., and Fung, S. H. H. (2006). Multiple-vehicle traffic accidents in hong kong. Accident Analysis & Prevention, 38(06):1157–1161.
6. Milton, J. C., Shankar, V. N., and Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis & Prevention, 40(01):260–266.
7. Kockelman, K. M. and Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. Accident Analysis & Prevention, 34(03):313–321.
8. Zajac, S. S. and Ivan, J. N. (2003). Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural connecticut. Accident Analysis & Prevention, 35(03):369–379.
9. Clifton, K. J., Burnier, C. V., and Akar, G. (2009). Severity of injury resulting from pedestrian-vehicle crashes: What can we learn from examining the built environment? Transportation Research Part D: Transport and Environment, 14(06):425–436.
10. Chang, L.-Y. and Wang, H.-W. (2006). Analysis of traffic injury severity: An application of nonparametric classification tree techniques. Accident Analysis & Prevention, 38(05):1019–1027

11. Õna, J., Mujalli, R. O., and Calvo, F. J. (2011). Analysis of traffic accident injury severity on spanish rural highways using bayesian networks. Accident Analysis & Prevention, 43(01):402–411.

12. Zong, F., Xu, H., and Zhang, H. (2013). Prediction for traffic accident severity: Comparing the Bayesian network and regression models. Mathematical Problems in Engineering, 2013.

13. Sze, N. N. and Wong, S. C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis & Prevention, 39(6):1267–1278.

14. Kim, J.-K., Ulfarsson, G. F., Shankar, V. N., and Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. Accident Analysis & Prevention, 40(5):1695–1702.

15. Peng, H., Long, F., and Ding, C. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 03(02):185–205.

16. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357.

17. Sousa, J. Analysis of pedestrian injury severity using Data Mining techniques. Lisboa: IST – Universidade de Lisboa, 2017. Tese de Mestrado.

18. Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3):651–674.

19. Hand, D.J., Till, R.J. (2001). A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems", Machine Learning, 45: 171–186.