# Previsão de Emergêncis

Arthur Abrahão Santos Barbosa
*Universidade Federal de Pernambuco*
*Centro de Informática*
Pernambuco, Brasil
aasb2@cin.ufpe.br

Arthur Henrique Anibal da Costa
*Universidade Federal de Pernambuco*
*Centro de Informática*
Pernambuco, Brasil
ahac@cin.ufpe.br

Filipe Samuel da Silva
*Universidade Federal de Pernambuco*
*Centro de Informática*
Pernambuco, Brasil
fss8@cin.ufpe.br

Vinicius Bastos Moreira Principe
*Universidade Federal de Pernambuco*
*Centro de Informática*
Pernambuco, Brasil
vbmp@cin.ufpe.br

## I. Introdução

Acidentes de trânsito são inesperados e causam diversas perdas. Existem diversas variáveis que contribuem com a gravidade de um acidente. O intuito deste estudo é desenvolver um modelo que ajude na predição do chamado em casos de acidentes em vias terrestres.

O trabalho envolve também, a partir de uma base de dados, determinar um sistema de apoio à decisão, que a partir dos dados coletados, possui capacidade de apontar caminhos a serem seguidos. E, com base na experiência do domínio do problema, o stakeholder pode definir as decisões a serem tomadas.

Além disso, houve um avanço do conhecimento a partir da noção de IA explicável. E, com esse objetivo, usamos três métodos para, além de desenvolver um bom sistema, entender com propriedade o que está lá dentro, que determina o por quê e quais caminhos os algoritmos estão tomando

## II. Base de Dados

A base de dados vieram da junção das bases de acidentes de trânsito [1] ocorridos no Condado de Montgomery - Maryland, EUA e das informações dos motoristas envolvidos neste acidente [2].

Estas informações foram registradas pelo "Sistema automatizado de Relatórios de Acidentes da Polícia estadual de Maryland".

### *A. Escopo e Seleção dos Dados*

Como definido pela base de dados, o escopo é dado apenas pelos acidentes de trânsito que ocorreram no Condado de Montgomery. Nele, temos informações sobre estado da via, trânsito, clima e detalhes sobre o impacto (caso tenha ocorrido). Entretanto, grande parte desses dados são informações a posteriori, se fazendo necessário então a remoção dessas colunas.

### *B. Definição do Objetivo*

Com o conhecimento gerado a partir dos modelos treinados é possível aplicar a predição em diversos segmentos do mercado, como por exemplo: Seguradoras, empresas de locação de veículos, planos de saúde e outros prestadores de serviços relacionados a carro ou a saúde do condutor.

### *C. Pré Processamento dos Dados*

Inicialmente a ideia girava em torno de tentar predizer a gravidade do acidente, qual a chance de existirem vítimas graves (sendo esse o alvo binário). Neste sentido, foram encontradas algumas dificuldades.

Com a junção dos datasets, foram obtidas 77 colunas de atributos. Dentre essas colunas, 51 eram de valores que precisariam ser removidos (dados a posteriori) ou tratados, como por exemplo colunas com muitos valores nulos. Considerando também as restrições de captação de dados do veículo, foram incluídos apenas os atributos mais significativos e pertinentes para a análise, enquanto outras colunas foram agrupadas, para melhor organizar os dados, restando 26 features para serem analisadas.

A principal técnica utilizada para tratamento foi a criação de variáveis "dummies" [3]. Variáveis categóricas ou variáveis binárias foram alteradas para variáveis "flags". Um bom exemplo de uso dessa técnica é a coluna de local de impacto: variável categórica que indicava a posição do impacto no veículo. Possuía valores baseados na técnica militar que usa o relógio para indicar posição, ou seja, 9 horas, 1 horas, 6 horas. Agrupando esta variável, chegamos aos valores "colisão frontal", "colisão lateral", "colisão traseira" ou "outro". Assim, reduzimos de 15 (cima, baixo e desconhecido também eram valores) possíveis valores para apenas 4.

Por recomendações do professor, evitamos deixar variáveis com valores ordenados, pois para o modelo pode passar a ideia de "prioridade" ou de "ordem", tornando a coluna enviesada. Sendo assim, usamos a técnica de variáveis "dummies" [3] e experimentamos usar a técnica "Frequency Encoding" [4] , onde codificamos uma variável categórica de forma que seu valor passe a ser representado pela frequência de vezes em que o seu valor aparece no conjunto de dados.

### D. Definição do Alvo

No primeiro momento, o alvo binário tratava-se de predizer se havia chance de ter vítimas graves ou não em um acidente. Entretanto, devido a dificuldade em obter bons resultados com o dataset, foi estudada a possibilidade de trocar para a coluna de dano ao veículo e o impacto que isso teria na ideia principal do projeto.

O alvo da classificação binária ficou então definido em relação ao dano do veículo, sendo então o objetivo descobrir se houve ou não dano significativo ao mesmo.

## III. Extração de Dados, Resultados e Discussão

Para extrair o conhecimento inserido na base de dados, foram utilizados três métodos: regressão logistica, arvore de decisão, e indução de regras

### A. Regressão Logistica

Após treinar o modelo de regressão logistica, foram analisadas as features com maior coeficiente beta, e que possuíssem maior significância de acordo com o p-valor, onde os coeficientes de maior módulo tem mais relevância ao definir a classe alvo.

As features com maior valor positivo tem maior contribuição para definir que houve dano significativo ao veículo, enquanto as de valores mais negativos possuem uma importância maior para definir se não houve dano significativo.

Se o carro está se movendo ou é particular há uma maior chance de possuir dano significativo após o acidente, enqunto se tiver algum não motorista participando do acidente (pedestre ou ciclista), se o carro estiver acelerando ou se a colisão for na mesma direção, a probabilidade de não haver dano significativo é bem menor.

TABLE I: características mais Relevantes de Acordo com A Regressão Logistica

| Feature | Beta | p-valor |
|---|---|---|
| is_moving | 1.259 | 0 |
| is_particular | 1.255 | 0 |
| RelatedNon-Motorist | -2.652 | 0 |
| is_accelerating | -1.212 | 0 |
| CT=SameDir | -1.175 | 0 |

### B. Árvore de Decisão

A árvore de decisão é um dos modos mais simples de visualisar o conhecimento presente em uma base de dados de modo compreenssível. As variáveis mais importantes para definir a classe alvo foi principalmente o tipo de colisão, sendo a colisão em ângulo a com maior probabilidade de causar dano ao veículo Outra características importante foi a divisão da estrada.


Fig. 1: Árvore de decisão gerada usando o Knime

### C. Indução de Regras de Classificação

A indução de regras vai trazer para nós um conjunto de regras em fórmula de E's e OU's, sendo utilizadas para classificar em uma determinada classe. Com o resultado da execução dos algoritmos indutores de regras. O primeiro o RIPPER (Repeated Incremental Pruning to Produce Error Reduction) que usa um processo iterativo de criação e refinamento de regras, com objetivo em melhor a precisão, sendo uma extensão do IREP, o segundo, o CN2(Classificação Numérica 2) também foi utilizado, sendo formulado em sua origem como uma extensão do primeiro. Após a execução iterativa dos algoritmos indutores de regra, obtivemos uma extensa lista de regras, algumas com muitas e outras com menos cláusulas. Também limitamos o número de cláusulas máximas que poderiam ser geradas.

TABLE II: Tabela de regras RIPPER Ruler Induction

| Regra: | Clausula 1 | 2 | 3 |
|---|---|---|---|
| 1 | is_particular=1 | is_moving=1 | is_slowing=0 |
| 2 | is_particular=1 | is_moving=1 | CollisionType=Angle |
| 3 | is_particular=1 | is_moving=1 | CollisionType=Outros |
| 4 | CollisionType=Angle | is_particular=1 | RD=TWO-WAY, DIVIDED |
| 5 | is_particular=1 | CollisionType=Outros | RoadCondition=0.0 |
| 6 | is_moving=1 | is_particular=1 | is_slowing=0 |
| 7 | Hit_Fixed_Object=1 | is_particular=1 | |
| 8 | CollisionType=Angle | TrafficControl=2 | is_particular=1 |
| 9 | is_moving=1 | is_particular=1 | RoadCondition=0.0 |
| 10 | CollisionType=Angle | Is_Junction=1 | is_particular=1 |

| Regra: | Cobertura | Confiança | Lift |
|---|---|---|---|
| 1 | 0,409 | 0,591 | 1,420 |
| 2 | 0,141 | 0,675 | 1,620 |
| 3 | 0,175 | 0,659 | 1,582 |
| 4 | 0,107 | 0,682 | 1,638 |
| 5 | 0,190 | 0,638 | 1,532 |
| 6 | 0,078 | 0,620 | 1,489 |
| 7 | 0,128 | 0,744 | 1,786 |
| 8 | 0,124 | 0,674 | 1,619 |
| 9 | 0,169 | 0,708 | 1,700 |
| 10 | 0,144 | 0,675 | 1,621 |

Tabela: CN2 Rule Induction

| Regra | Clausula 1 | 2 | 3 |
|---|---|---|---|
| 1 | CollisionType==Angle | TrafficControl¿=2.0 | is_slow |
| 2 | CollisionType==Outros | Is_Junction!=0 | Surfac |
| 3 | Hit_Fixed_Object!=0 | SecondHarmfulEvent!=OBJECT RELATED | is_slow |
| 4 | CollisionType==Angle | RoadDivision==TWO-WAY, DIVIDED | Traffic |
| 5 | CollisionType==Angle | TrafficControl¿=2.0 | Numb |
| 6 | CollisionType==Angle | WeatherGroup==Clear | Light= |

| Regra | Cobertura | Confiança | Lift |
|---|---|---|---|
| 1 | 0,106 | 0,659 | 1,582 |
| 2 | 0,103 | 0,578 | 1,388 |
| 3 | 0,075 | 0,656 | 1,575 |
| 4 | 0,092 | 0,679 | 1,630 |
| 5 | 0,066 | 0,683 | 1,640 |
| 6 | 0,096 | 0,588 | 1,413 |

## IV. Avaliação de Performance

Para avaliar a performance do regressor logistico, foi utilizada a métrica AUC_ROC que tem como valor ideal 1, o valor obtido foi 0.788, mostrando que o regressor consegue distinguir bem entre os valores possíveis da decisão binária.

## V. Conclusão

Neste trabalho foi analisado quão significante foi o dano ao veìculo após o acidente, técnicas como árvore de decisão e regressão logistica permitiram entender melhor a relação das características da base de dados com a variáveis alvos, enquanto a indução de regras permitiu que relações que envolvem mais de uma característica fossem encontradas.

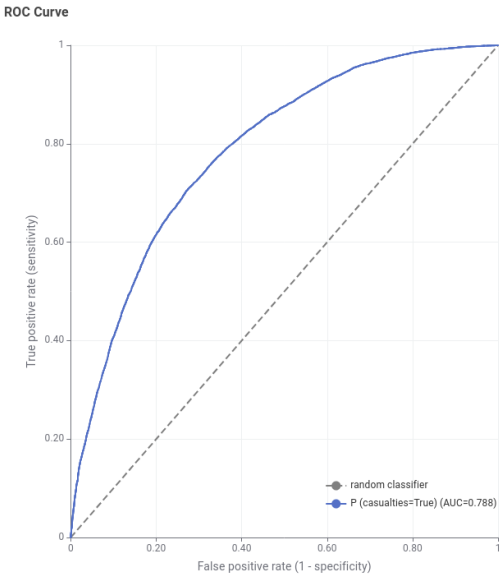


Fig. 2: Curva ROC Para o Regressor Logistico

## References


[1] M. Montgomery County. (2024) Crash reporting - incidents data. [Online]. Available: https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Incidents-Data/bhju-22kf/about_data

[2] ——. (2024) Crash reporting - drivers data. [Online]. Available: https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632/about_data

[3] S. Date. What are dummy variables and how to use them in a regression model. [Online]. Available: https://timeseriesreasoning.com/contents/dummy-variables-in-a-regression-model/

[4] M. Halford. (2018, Oct.) Target encoding done the right way. [Online]. Available: https://maxhalford.github.io/blog/target-encoding/

[5] W. H. Organization. (2024, Dec.) Road traffic injuries. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries