TECNOLOGIAS
E ARQUITETURA

Road Accident Severity in India

Tomada de Decisão Baseada em Dados

Docentes: Ana Rita Peixoto e Elsa Cardoso

Ano Letivo: 2024/2025

Grupo 5:

Pedro Conceição nº 129188

Ricardo Mororó nº 94562

Índice

Introdução

O presente trabalho visa explorar o dataset *"Road Accident Severity in India*", disponível na plataforma *Kaggle*, com o objetivo de desenvolver um poster científico. Para desenvolver e explorar os dados, baseamo-nos no modelo de Tomada de Decisão de Herbet Simon, na qual, este divide o processo de decisão em 3 fases: Inteligência, Design e escolha.

Assim, usando este modelo como referência, iniciamos o trabalho pela fase de Inteligência, onde nos dedicamos a identificar problemas que pudessem ser relevantes por meio de uma exploração inicial dos dados, e começamos a realizar diversas questões que pudessem vir a ser respondidas com a ajuda dos mesmos.

Com base numa análise inicial, definimos três questões/problemas centrais que queríamos obter resposta com os dados, sendo estes:

1. A causa do acidente está relacionada com a idade e experiência do condutor?
2. A hora e o local dos acidentes estão relacionados com a frequência de sua ocorrência?
3. A gravidade do acidente pode ser predita a partir dos dados disponíveis no dataframe?

Na fase de *design* do modelo de Simon, estas perguntas guiarão a escolha de técnicas analíticas para identificar relações significativas entre variáveis, ajudando a estruturar insights práticos para a tomada de decisão.

Após exploração e limpeza dos dados, o nosso objetivo será responder aos problemas identificados recorrendo também a técnicas de *data mining* e análises descritivas, permitindo a identificação de padrões e descobrir as causas principais dos acidentes relativamente às questões que procuramos responder.

Com base nesses resultados, procuramos que a nossa análise exploratória dos dados seja útil para a tomada de decisão e conduza a implementação de medidas que contribuam para a redução da sinistralidade rodoviária.

Quanto à ética dos dados, os mesmos estão disponíveis online, de forma transparente, sendo o seu acesso público, embora o nosso *dataset* não tenha informação confidencial sobre os indivíduos, consideramos que a responsabilidade no uso dos dados também implica interpretar e promover conclusões que sejam verdadeiramente representativas e possam beneficiar a sociedade de forma imparcial, isto é, evitando interpretações que possam reforçar preconceitos.

Data Understanding

Nesta fase, o nosso objetivo foi explorar e entender as diferentes variáveis que constituem o nosso *dataset*, de forma a perceber o impacto e a relevância das mesmas para os nossos objetivos previamente mencionados. Como tal, recorremos ao *Python*, e a ferramentas e bibliotecas como Pandas, para visualizar as variáveis existentes.

O nosso *dataset* contem 12316 registos, e 32 colunas. É composto por 30 variáveis categóricas, entre as quais, podemos referir, colunas sobre as informações do acidente, tais como: hora, dia da semana, condições meteorológicas.

Observou-se ainda variáveis qualitativas que representam o perfil do motorista e do veículo envolvido nos acidentes rodoviários, bem como de informações referentes a vítimas.

As duas variáveis quantitativas apresentadas no *dataset* dizem respeito ao número de veículos envolvidos, e o número de vitimas.

Em média, os acidentes envolvem dois veículos, sendo raros os acidentes com mais de 3. A maior parte dos acidentes (75%) envolve 1 ou 2 veículos. A variável *Number_of_vehicles_involved* apresenta um valor máximo de 7 veículos, o que é considerado um *outlier* devido à sua raridade.

Quanto à variável referente ao número de vítimas, a média é de 1,55, indicando que, em geral, os acidentes resultam em 1 ou 2 vítimas. No entanto a variável apresenta um valor máximo de 8 vítimas, o que também é considerado um *outlier*.

Após analisarmos as variáveis numéricas, decidimos verificar o número de classes presentes nas variáveis categóricas, e qual a informação contida nessas classes que seja relevante, isto é, que nos permita, mais tarde, extrair e retirar conclusões relativamente aos objetivos deste projeto.

Assim sendo, após verificarmos quais as classes presentes nas variáveis categóricas, conduzimos a uma análise univariada, de forma a extrair informação e obter uma visão mais detalhada do nosso *dataset*, tal como, verificar as classes modais, distribuições, frequências relativas e globais.

Para uma melhor compreensão desta fase que é a analise das variáveis, recorremos a algumas bibliotecas do *Python* como *Seaborn* e *Matplotlib* para gerar gráficos de barras de forma a ter uma visualização mais detalhada, o que nos permitiu identificar padrões, detetar desequilíbrios entre as classes, e, ainda, entender quais variáveis agregam pouca informação aos problemas que queremos resolver com os dados.

Através dos gráficos de barras, observámos padrões interessantes sobre os acidentes de trânsito na Índia, aos quais podemos destacar: mais de 90% dos acidentes rodoviários envolvem cidadãos do sexo masculino.

Ainda, a maioria dos condutores tinha apenas o ensino equivalente ao 9ºano, sendo que o intervalo de idades entre os 18-30, e os 30-50 é o mais prevalente. A maioria dos acidentes envolve motoristas com experiência entre os 2 e os 10 anos, e as áreas onde há maior taxa de acidentes correspondem a zonas de atividade económica, seguido por zonas residenciais.

Verificamos que a maior parte dos acidentes envolvem veículos automóveis, seguido de caminhões, e que a maioria dos acidentes ocorre em condições climatéricas consideradas normais.

Data preparation

Neste estágio, após termos uma compreensão mais profunda das variáveis e da sua distribuição categórica, iniciamos um processo de limpeza dos dados, com o objetivo de criar um conjunto de dados mais relevante e enriquecedor de forma a ser pertinente para os problemas que queremos resolver.

Escolha das Variáveis

Certas variáveis categóricas apresentaram classes modais com frequências muito altas, como as variáveis '*Frequency of Defect of Vehicle'* e '*Frequency of Owner of Vehicle'*, entre outras.

Isso limitou a capacidade de extrair insights relevantes, uma vez que a concentração de dados em algumas categorias reduz a variabilidade informativa, e verificamos que nenhuma destas variáveis seria útil para responder aos nossos problemas.

Como resultado, decidimos filtrar as variáveis que pouco acrescentavam para a nossa análise nesta fase, selecionando aquelas que, de forma mais eficaz, poderiam ajudar a compreender e responder às questões definidas inicialmente, garantindo uma análise mais robusta e focada.

As variáveis numéricas, como o número de veículos envolvidos e número de vítimas, não ofereceriam um aumento substancial no poder preditivo para as questões que desejávamos explorar, já que as variáveis numéricas, embora sejam de certa forma informativas, não ajudam diretamente a responder às questões propostas, e como tal, não seriam relevantes para a analise.

Durante esta fase, optamos por reduzir o número de variáveis categóricas de 30 para 10, de modo a alinhar o nosso conjunto de dados com as questões que pretendemos explorar. Foram selecionadas variáveis que abordam o perfil do condutor, o local do acidente e as causas atribuídas ao mesmo.

A variável *"Accident_severity*" foi escolhida como *target* para a aplicação de técnicas de *Machine Learning*, pois na altura, consideramos tratar-se da variável mais relevante para entender e prever a gravidade dos acidentes com base nas características dos condutores envolvidos.

Descartamos também variáveis por terem classes com frequências muito altas, ou por serem demasiado redundantes, e de certa forma, serem irrelevantes para os dados que queríamos explorar, já que não seria possível extrair informação pertinente.

Tratamento de Variáveis Categóricas

De forma a melhorar e otimizar as variáveis categóricas para análise, decidimos agrupar algumas classes, tornando o conjunto de dados mais eficiente para análises mais pertinentes e para reduzir a redundância.

Assim, a variável *Type Of Vehicle* que continha mais de 17 classes passou a ter 7 classes, utilizando-se funções de Python para agrupar diferentes tipos de veículos em classes mais amplas.

Para a variável que representa a causa do acidente, também realizamos um agrupamento, mapeando categorias através do *Python*, como "*Risky driving*" em que juntamos comportamentos de risco, como consumo de álcool, consumo de drogas e excesso de velocidade. Adicionalmente, criamos categorias como "*Directional changes*," "*Priority violation*," "*Reckless driving*," e "*Other"* para organizar as causas de forma mais simplificada e coerente.

Para a variável Time, que correspondia à hora de ocorrência do acidente, recorremos aos métodos acima, e agrupamos em períodos do dia, de forma a colocar em evidência quais são os horários onde ocorrem mais acidentes.

Para a variável que corresponde à área onde aconteceu o acidente, também reduzimos as 17 classes iniciais e agrupamos em 5 categorias, de forma a posteriormente extrair mais informação sobre a predominância dos acidentes, que nos permitisse estabelecer relações e retirar conclusões mais enriquecidas.

Após a criação de um novo *dataframe* com as variáveis fulcrais e com as classes devidamente tratadas e agrupadas, fomos verificar a proporção de valores nulos nas colunas.

Relativamente aos valores omissos, optamos por imputar estes pela classe modal, justificado pelos próprios valores observados nos dados, ou seja, verificamos que havia uma percentagem de valores nulo relativamente baixa, sendo a mais alta, a variável que correspondia à experiência do condutor, que ainda assim, não ultrapassava os 7% dos valores omissos.

Para além disso, conseguimos assim preservar a representatividade dos dados, já que imputar pela moda mantém o peso da categoria dominante, uma vez que os valores omissos eram relativamente baixos.

Modelagem

Nesta fase, nosso objetivo é verificar técnicas com o intuito de responder ao terceiro problema, ou seja, tentar prever o comportamento da variável *target* consoante as variáveis categóricas com maior influência para a sua previsão.

Nesse sentido, recorremos à métrica de correlação V de Cramer, uma vez que a mesma é aplicável para verificar a correlação entre variáveis categóricas.

Com base nos resultados, concluímos que devido à fraca correlação observada, não fazia sentido explorarmos técnicas de *data mining*, já que este tipo de técnicas, como algoritmos de m*achine learning*, geralmente exigem associações mais fortes entre as variáveis para fornecer resultados significativos e confiáveis.

Devido a baixíssima correlação entre as variáveis do *dataframe*, optamos por uma análise com enfoque apenas nos acidentes cuja severidade tinha sido fatal.

Para isso, recorremos à filtragem do *dataset*, isolando os acidentes classificados como "*Fatal injury*". A partir deste novo subconjunto de dados, fomos capazes de focar-nos apenas nos dados de acidentes que tinham sido fatais, recorrendo a analises de gráficos para extrair o máximo de informação pertinente que conseguíssemos.

Ao aplicarmos novamente a métrica do V de Cramer no novo *dataset,* verificou-se uma melhor correlação entre as variáveis independentes, o que nos ajudou a resolver a terceira questão inicialmente proposta de forma suportada pelos dados.

Resultados e Discussões

1. **A causa do acidente está relacionada com a idade e experiência do condutor?**

A principal causa de acidentes são as mudanças de direção da viatura (mudança de faixa, ultrapassagem, p.ex.), seguida pelos comportamentos imprudentes do condutor (distanciamento, estacionamento inapropriado, p.ex.). Condutores com experiência de condução entre 5 e 10 anos, e com idade entre os 18 e 50 anos, são os mais envolvidos nessas causas.

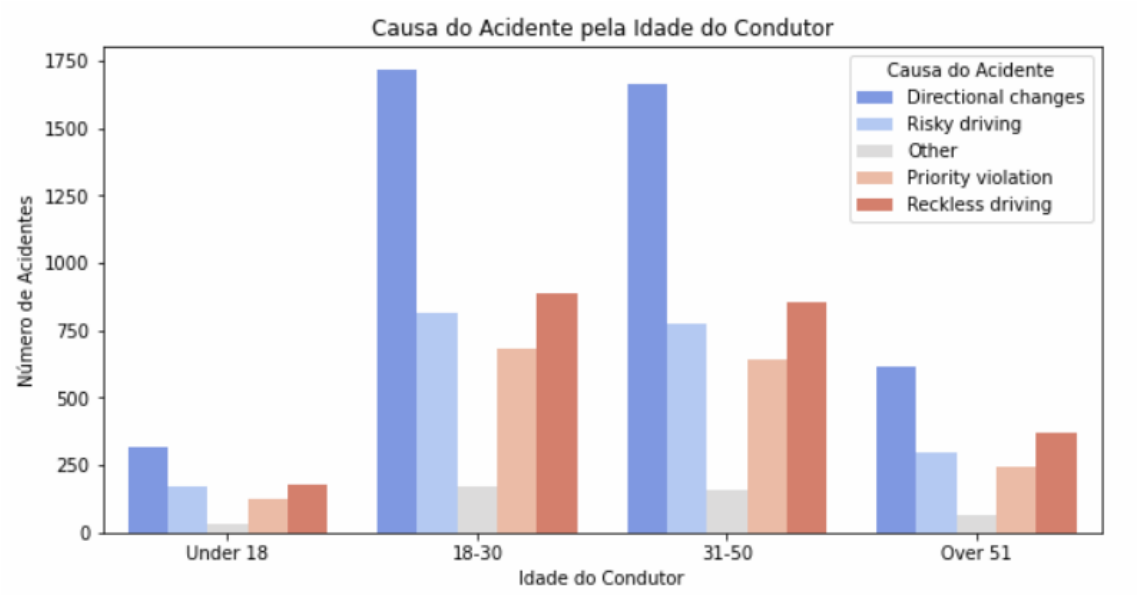

Gráfico 1 – Causa do Acidente pela Idade do Condutor

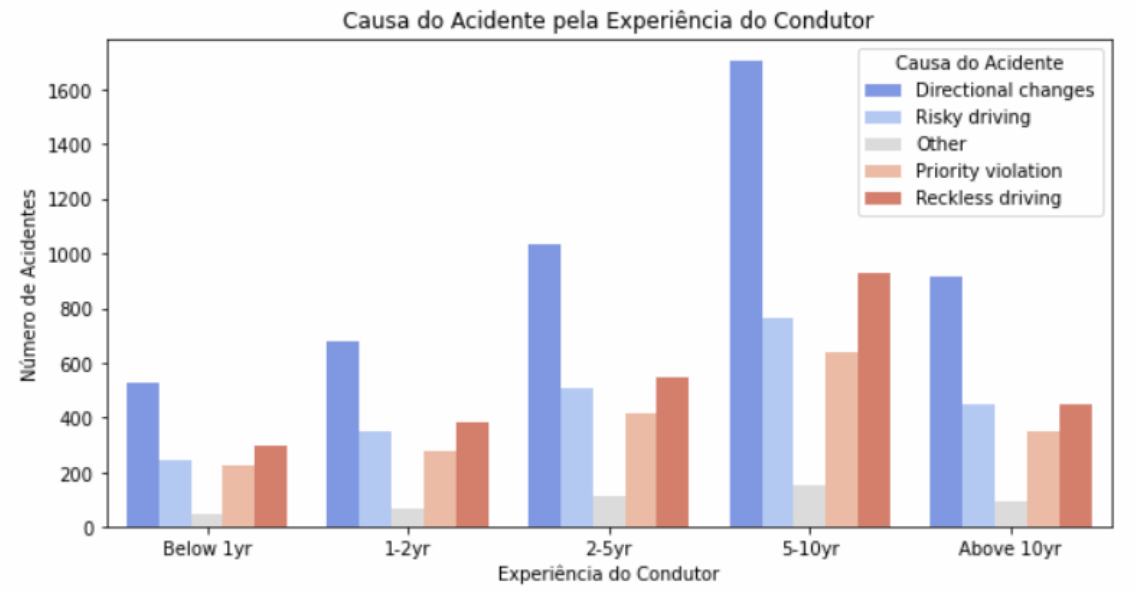

Gráfico 2 – Causa do Acidente com base na Experiência do Condutor

2. **A hora e o local dos acidentes estão relacionados com a frequência de sua ocorrência?**

A maioria dos acidentes ocorre no período entre as 14h e as 19h (*Evening*) em áreas relacionadas a negócios, como escritórios e zonas industriais.

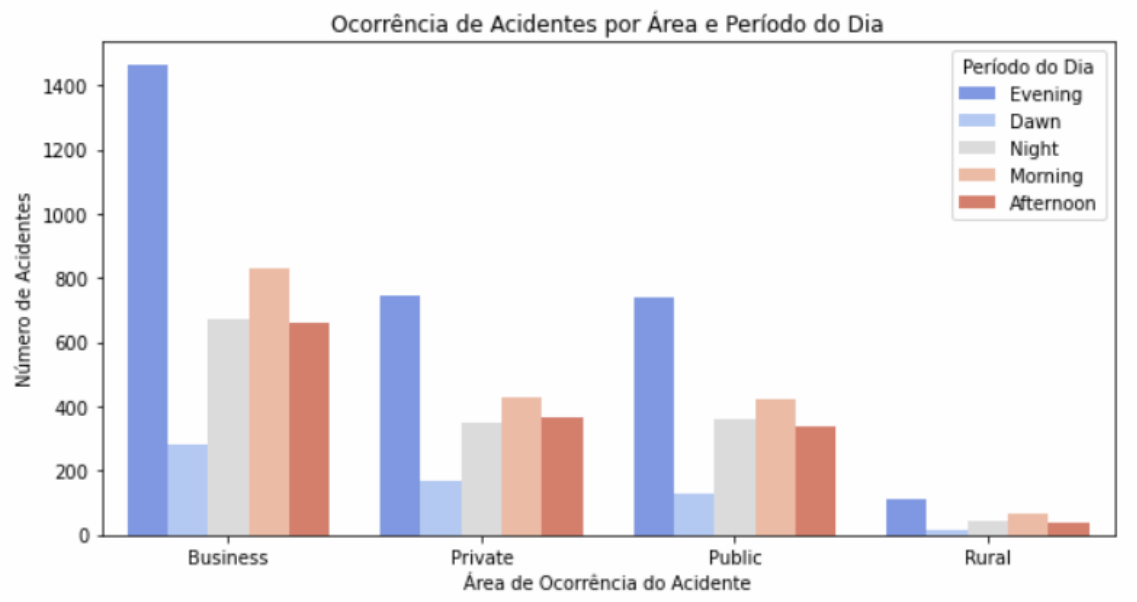

Gráfico 3 – Frequência de Acidentes por Área e Período do Dia

3. **A gravidade do acidente pode ser predita a partir dos dados disponíveis no dataframe?**

Os acidentes fatais apresentam correlação com o dia da semana e o período do dia em que acontecem, com maior predominância nos dias de fim de semana e nos períodos da tarde e noturnos.

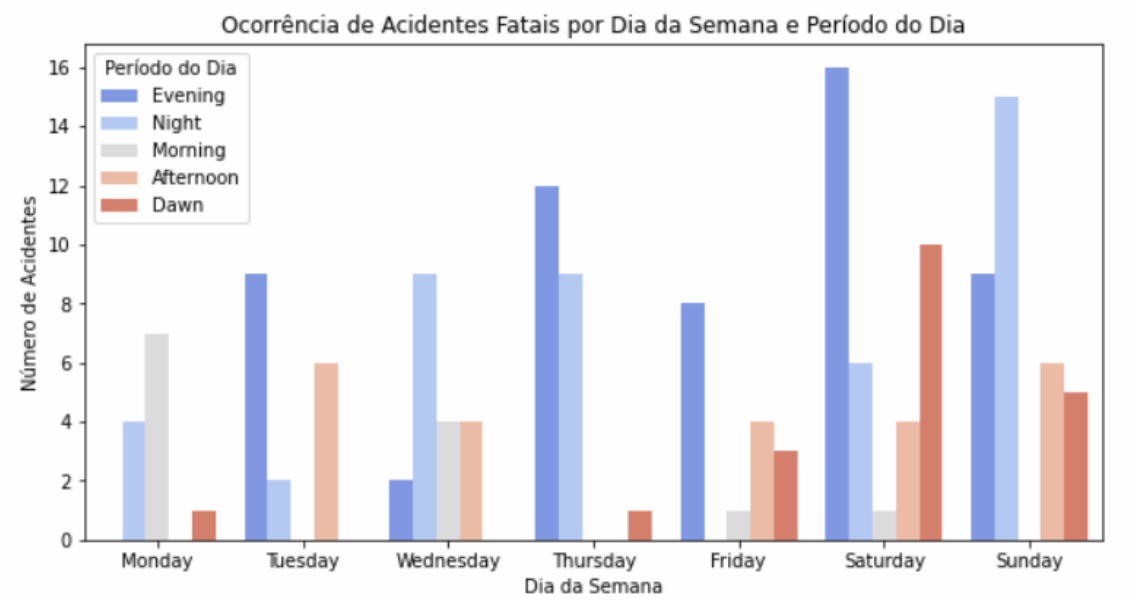

Gráfico 4 – Frequência de Acidentes Fatais por Dia da Semana e Período do Dia

As autoridades rodoviárias devem investir em reforço de sinalização e iluminação das vias públicas, principalmente nas áreas relacionadas a atividades económicas, e com maior ênfase para proibir a mudança de faixa nas zonas indicadas como de maior ocorrência de acidentes por este motivo. Também se recomenda o reforço de fiscalização policial noturna aos fins de semana e realização de campanhas de conscientização.

Os condutores devem tomar em atenção a práticas de condução segura com observância das normas de trânsito, e conduzir com maior atenção nos locais e dias da semana em que se observam mais acidentes e com maior gravidade. Especialmente os que têm menos de 10 anos de experiência e menos de 50 anos de idade.

Conclusões

- Embora o conjunto de dados inclua informações valiosas relativamente aos acidentes rodoviários na India, como horário, área do acidente, dia da semana, tipo de veículo, perfil do condutor, sentimos que faltaram algumas variáveis que poderiam ser de grande utilidade.
- A inclusão de uma variável com o sítio exato do acidente – como o nome da rua, número do cruzamento, ou um ponto de referência próximo, facilitaria a implementação de ações preventivas específicas para esses locais. Por exemplo, direcionadas com maior precisão para reduzir o risco de novos acidentes nessas áreas especificas.
- A análise exploratória revelou uma fraca correlação entre as variáveis observadas, o que limitou a viabilidade de técnicas avançadas de *data mining*, que poderiam ter oferecido insights adicionais.
- Esse cenário reduziu a possibilidade de observar um ciclo de análise completo e de aprofundar a tomada de decisão. Uma vez que com correlações mais fortes ou variáveis adicionais, seria possível realizar uma modelagem mais detalhada, permitindo visualizar tendências e identificar segmentos de alto risco com maior clareza, contribuindo para aumentar a riqueza da análise, conduzindo a recomendações mais precisas e abrangentes para mitigar a sinistralidade.