

Instituto Universitário de Lisboa

Escola de Tecnologias e Arquitetura

Relatório do desenvolvimento do poster científico sobre o *dataset* “*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 3](#_Toc182332189)

[Data Understanding 4](#_Toc182332190)

[Data preparation 5](#_Toc182332191)

[Escolha das Variáveis 6](#_Toc182332192)

[Tratamento de Variáveis Categóricas 7](#_Toc182332193)

[Modelagem 8](#_Toc182332194)

[Resultados 9](#_Toc182332195)

[Conclusões 9](#_Toc182332196)

# 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. Existe uma maior prevalência de acidentes consoante o dia da semana e período do dia?

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 publico, 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 cerca de 12316 registros, e 32 colunas, entre as quais, podemos referir, colunas sobre as Informações do acidente, tais como: hora, dia da semana, condições meteorológicas. Observamos ainda algumas categorias que representam o perfil e as características do tipo de motorista que está envolvido nos acidentes rodoviários, também estavam presentes variáveis referentes à categoria do veículo, informações da vitima e detalhes do acidente, que neste caso correspondiam às únicas 2 categorias numéricas que tínhamos no *dataset*, nomeadamente, o numero de veículos envolvidos, e o numero de vitimas. Assim sendo, o nosso *dataset* era composto por 30 variáveis categóricas, e apenas 2 numéricas.

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 analise 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, analisar a fundo as variáveis que têm muitas classes o que pode ser limitador no sentido de uma maior redundância e dificuldade em extração de informação, que pode tornar a modelagem mais complexa. 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, a maioria apenas com o ensino equivalente ao 9ºano, sendo que o intervalo de idades entre os 18-30, e os 30-50 é o mais prevalente, sendo importante realçar que a maioria dos acidentes envolve motoristas com experiência entre os 2 e os 10 anos, verificamos ainda que as áreas onde há maior taxa de acidentes corresponde em zonas de atividade económica, seguido por zonas residenciais, verificamos que a maior parte dos acidentes são de 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 essas variáveis e todas as outras que pouco acrescentavam para a nossa analise 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 9, 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 envolvidos, isto é, a idade e experiência de condução, e do contexto, nomeadamente, hora ou local do acidente. Descartamos também outras 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 ajustar e agrupar algumas classes, tornando o conjunto de dados mais eficiente para análises descritivas mais pertinentes e para reduzir a redundância. Assim, variáveis como *Type Of Vehicle* que continha mais de 17 classes, utilizamos funções de Python para agrupar diferentes tipos de veículos em classes mais amplas, reduzindo para 7 classes.

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 imprudentes, 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, uma vez que já tínhamos feito algumas analises bívariadas, queríamos sobretudo verificar quais as técnicas de *data mining* que poderíamos recorrer com o intuito de responder ao terceiro problema que tínhamos identificado ao iniciar este projeto, mais precisamente, entender quais as variáveis categóricas que mais poderiam fornecer insights, e consequentemente prever o comportamento da nossa variável targetde forma a aplicarmos os modelos adequados.

Nesse sentido, iniciamos a fase de modelação dos dados, de forma a analisar as correlações entre as variáveis categóricas do nosso conjunto de dados. Recorremos ao teste de Cramér’s V uma vez que consideramos este util para verificar a correlação entre variáveis categóricas, isto é, se a força de associação era alta ou baixa, sendo que um valor próximo de 0 sugere uma associação fraca, e valores próximos de 1, sugerem uma associação forte. O teste foi aplicado a todas as variáveis categóricas do *dataset*, e os resultados foram visualizados em uma matriz, onde cada célula representa o valor de Cramér's V entre duas variáveis. A análise dessa matriz revelou que a correlação entre a variável target e as variáveis independentes era, muito baixa. Os valores de Cramér’s V indicaram que a força da associação entre a variável target e outras variáveis, como o "*Day\_of\_week*", "*Age\_band\_of\_driver*" e "*Cause\_of\_accident*", variava entre 0 e 0.05, o que nos indicou que as relações entre essas variáveis e a severidade dos acidentes eram fracas.

Com base nesses 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, já que era esse o objetivo da nossa terceira questão que queríamos resolver com a exploração os dados. 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, especialmente, nomeadamente, idade e experiência do condutor, período do dia, área onde ocorreu, de forma a verificar a possibilidade de tendências que ajudassem a resolver a nossa terceira questão, e posteriormente, definir propostas de decisão suportada pelas evidencias extraídas através das tabelas.

# Resultados

# Conclusões

Tópicos a abordar:

1. A metodologia CRISP-DM orientou este projeto em etapas como compreensão do problema, entendimento e preparação dos dados, focando especialmente na análise descritiva para explorar os padrões de sinistralidade na Índia. Foram analisadas distribuições de frequência e gráficos que revelaram prevalências de acidentes em função de fatores como área, idade e causa do acidente. Embora não tenha sido utilizada modelagem avançada, a análise exploratória forneceu insights relevantes para identificar áreas críticas de risco, mesmo com as limitações de um escopo inicial.

-Aqui podemos falar sobre campanhas a adotar tais como: Se os gráficos de barras indicarem que os acidentes são mais frequentes em determinados dias da semana ou horários específicos, as autoridades podem direcionar campanhas de conscientização para esses períodos ou aumentar a fiscalização nesses horários críticos, especialmente em áreas onde há uma maior prevalência de sinistralidade. Tambem podemos falar de medidas a adotar a curto e a longo prazo!

Embora o conjunto de dados inclua informações como horário e condições climáticas, hora e dia, a ausência de dados sobre o local exato dos acidentes limita nossa capacidade de identificar pontos críticos de risco de forma precisa. A inclusão de uma variável com o sítio exato do acidente – como o nome da rua, o número do cruzamento, ou um ponto de referência próximo – facilitaria a implementação de ações preventivas específicas. Por exemplo, ao identificar cruzamentos ou trechos de ruas com alto índice de sinistralidade, medidas como melhoria da sinalização, instalação de semáforos, ou reforço na iluminação poderiam ser aplicadas para reduzir o risco de novos acidentes nessas áreas.

* Limitação do nosso dataset – não havia correlação entre as variaveis, e como tal não foi possível apalicar técnicas de data mining e machine learning, o que teria sido útil, já que iriamos conseguir obter insights/previsões e ter uma visão mais aprofundada de todas as fases da tomada de decisão de dados, desde à fase inicial de escolha do problema, exploração de dados, e técnicas de data mining para extrair informação relevante para a tomada de decisão.