**Uma imagem com preto, escuridão

Descrição gerada automaticamente**

Licenciatura em Ciência de Dados

Projeto Aplicado em Ciência de Dados II

**Base de Dados: Acidentes Rodoviários**

Uma imagem com veículo, Veículo terrestre, ar livre, roda

Descrição gerada automaticamente

Afonso Gião Santana Oliveira e Silva Nº 105208 | Turma: CDC2  
Ana Reis Abreu Nº 98625 | Turma: CDPL1  
Nº | Turma: CD1  
Nº | Turma: CD1  
Tomás Francisco Ribeiro Nº 105220 | Turma: CDC2

ISCTE – IUL | dezembro de 2023

Docente ………

1. **Introdução**
   1. **Motivação**

O trânsito nas estradas representa um aspeto crucial do dia a dia de muitos indivíduos, atravessando diferentes partes de todo o território nacional. Incluídos neste contexto estão elementos críticos a considerar, como os acidentes que frequentemente ocorrem nas nossas vias.

Os acidentes de trânsito afetam significativamente a vida dos cidadãos, incluindo desde os condutores, passageiros a pedestres. Por isso, é vital promover a conscientização da população sobre como adotar comportamentos responsáveis, tanto por parte da comunidade em geral quanto das respetivas autoridades, para reduzir a frequência e o impacto desses eventos. Isto envolve analisar os diversos tipos de acidentes ocorridos no país e suas causas principais, visando identificar padrões e implementar estratégias eficazes para sensibilizar a comunidade.

Posto isto, para avançar com este projeto, optou-se por utilizar como referência os dados disponibilizados pela Autoridade Nacional de Segurança Rodoviária (ANSR). Este é um dataset em que os dados incluem informações detalhadas sobre acidentes de trânsito ocorridos em várias áreas de Portugal Continental, cobrindo o período dos anos de 2010 a 2019.

* 1. **Objetivo de Estudo**

Neste projeto foi usada a metodologia *Cross-industry Standard Process for Data Mining*, ou CRISP DM1. Como o nome sugere, esta metodologia é normalmente usada na extração de dados, por isso será a base do projeto e do relatório. Foram realizadas análises, com suporte na base de dados, na tentativa de chegar a conclusões importantes para definir os nossos objetivos de estudo.

Assim após esta análise e discussão de ideias, o grupo decidiu incidir o tema do projeto em algumas questões a serem respondidas ao longo deste relatório. A primeira questão levantada foi “que impacto têm as épocas festivas nos acidentes rodoviários?”. Em que queremos perceber em datas como o Natal, Ano Novo, Carnaval, entre outras épocas festivas o que acontece nas estradas português., sendo que nestas datas existe sempre um maior tráfego rodoviário. Por sua vez, a segunda questão levantada como objeto de estudo focou-se em conseguirmos obter algo que conseguisse caracterizar um acidente como grave, isto através de um modelo de classificação, de forma a conseguirmos perceber quais as variáveis que podiam influenciar mais nessa classificação.

O projeto é então baseado no tratamento e processamento de dados e a sua análise exploratória, trabalhados no R com linguagem *Python*, com a finalidade de obter respostas para as questões debatidas e propostas a estudo pelo grupo. As questões formuladas são:

* Que impacto têm as épocas festivas nos acidentes rodoviários?
* O que caracteriza um acidente como grave? (modelo de classificação)

1. **Que impacto têm as épocas festivas nos acidentes rodoviários?**

**2.1. Metodologia**

A nossa base de dados tem por fonte a Autoridade Nacional de Segurança Rodoviária com acidentes de vários anos, desde 2010 ao ano de 2019. Iniciámos com uma base de dados com 309874 registos. Existiam originalmente 43 variáveis, algumas das principais como por exemplo data e hora do acidente, a sua localização, número de feridos, condições climáticas, estacão do ano, entre outras. Verificou-se que existiam vários datasets repartidos em anos, foram então unidos todos os anos e lidos em um único dataframe ‘acidentes’. Feita esta junção o dataset passou a ter 41 variáveis, em que foram eliminadas duas novas variáveis ‘latitude\_gps’ e a ‘longitude\_gps’ que não foram consideradas pertinentes para o estudo.

* **Tabela “Acidentes”**

|  |  |  |
| --- | --- | --- |
| **Nome da Variável** | **Descrição** | **Tipo de Variável** |
| **ID.ACIDENTE** | Identificação do acidente | Quantitativa contínua |
| **COND\_ADERÊNCIA** | Condição de aderência da estrada | Qualitativa nominal |
| **TIPOS\_VIAS** | Tipo de via do acidente | Qualitativa nominal |
| **FATORES\_ATMOSFERICOS** | Fatores atmosféricos que ocorrem durante o acidente | Qualitativa nominal |
| **DATAHORA** | Data e Hora quando ocorreu o acidente | Quantitativa contínua |
| **NATUREZA** | Tipo de acidente | Qualitativa nominal |

O DataSet continha 12767321 NA´s que correspondiam a 9% do dataset inicial, estes foram tratados com base no manual de preenchimento da beav. Foi assim realizada uma análise detalhada dos valores ausentes em várias colunas. Foram removidas linhas com valores NA em colunas específicas como: **velocidade\_local, velocidade\_geral, características\_tecnicas1, marca\_via, luminosidade,** entre outras que não consideramos relevantes tratar de outra forma os valores omissos sem ser eliminando os mesmos que pouco influenciavam na análise do nosso objetivo ou eram colunas com um alto teor de NA´s. Substituímos os valores NA em outras variáveis com valores predeterminados, refetindo sempre a interpretação do manual de preenchimento, estas colunas foram: **pov\_proxima, nome\_arruamento, km, factores\_atmosfericos, obras\_arte, sinais, sinais\_luminosos, sentidos**. Em maior parte destas variáveis foram substituídos os valores nulos por “Outro” para permitir a inclusão do registo na análise evitando distorções na mesma e para não perder o restante da informação.

Para tratamento de outliers foram utilizados boxplots para as variáveis **‘num\_feridos\_ligeiros\_a\_30\_dias’,’num\_feridos\_graves\_a\_30\_dias’** e **‘num\_mortos\_a\_30\_dias’**. Usamos ainda *bloxplot.stats* para termos uma análise estatística mais detalhada do boxplot, incluindo assim a identificação dos valores de outliers que foram então removidos. Ao analisarmos variáveis como estas conseguimos analisar a relação entre a gravidade dos acidentes e fatores como a luminosidade, estado da via e perceber então o porquê de inúmeros acidentes.

Criámos gráficos de linhas para observar as tendências nos nossos dados neste caso da ocorrência de acidentes de 2010 a 2019. Bem como na distribuição de acidentes ao longo de um ano, e um mês específicos, e até em épocas festivas importantes com o Natal, Ano Novo e Carnaval, revelando assim tendências sazonais ou relacionadas a eventos.

* + 1. **Tratamento de variáveis**

Foi feita uma padronização dos nomes das colunas onde utilizamos a função **make\_clean\_names** durante a leitura dos dados de cada ano (de 2010 a 2019) com a função **read\_excel**. Esta função transforma os nomes das colunas para um formato consistente e fácil de usar no R. Converte todos os caracteres para minúsculo, substitui espaços e caracteres especiais por *underscores* (\_) ainda garante que os nomes das colunas são únicos e legíveis.

De seguida convertemos algumas variáveis quanto ao seu tipo de dados:

* Data e Hora:

A coluna datahora é convertida para o formato de data e hora com a função **as\_datetime**. Esta conversão é crucial para análises temporais, permitindo assim manipularmos e analisarmos as datas e os horários dos acidentes de uma forma mais eficaz.

* Conversão de Variáveis *Character* para *Factor*:

Foram convertidas também as variáveis do tipo *character* no dataframe acidentes para o tipo factor usando a função **mutate\_if** do pacote *dplyr* juntamente com **is.character** e **as.factor**. A conversão para factor é importante para variáveis categóricas, pois assim otimizamos o armazenamento de dados e facilitamos a realização de certas análises estatísticas e visualizações.

* 1. **Análise Exploratória**

Como ponto de partida, e de forma a observarmos melhor de forma global o número de acidentes fizemos assim uma análise geral através de um gráfico de linhas. O gráfico resultante mostra a tendência de ocorrências de acidentes rodoviários ao longo do tempo, de 2010 a 2019. As variações na linha mostram como o número de acidentes mudou ao longo dos anos. O que nos é útil para identificar padrões temporais, como aumentos ou diminuições na frequência de acidentes ao longo do período estudado. Obtivemos o seguinte resultado: (Gráfico 1)

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

Gráfico 1 - Ocorrência de Acidentes 2010-2019

Verificámos que o padrão de tendência de acidentes de ano para ano é bastante similar, desde o ano 2011 que o número de acidentes tem vindo a aumentar de ano para ano.   
( explicar possíveis causas + pessoas c carta, + carros, + circulação, + acidentes)  
  
Posteriormente, quisemos analisar o padrão do número de acidentes num só ano, por sua vez num só mes