**Uma imagem com preto, escuridão

Descrição gerada automaticamente**

Projeto Aplicado em Ciência de Dados II

Data Set: Acidentes de Viação em Portugal

**Uma imagem com veículo, Veículo terrestre, ar livre, céu

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 ………

**Introdução**

Este trabalho foi realizado no âmbito do projeto de ciência de dados II. A base de dados disponibilizada pela Autoridade Nacional de Segurança Rodoviária foi escolhida para concluir o trabalho. Para concluir este trabalho, foi utilizado um conjunto diversificado de habilidades, todas elas relacionadas à exploração de dados e extração de informações deles. Ao trabalhar com dados reais, o nosso objetivo com este trabalho foi conseguir melhorar o nosso conhecimento de ciência de dados trabalhando com informação real e retirar informação que possa ser útil para o nosso dia a dia.

**Metodologia**

A metodologia *Cross-industry Standard Process for Data Mining*, ou CRISP DM1, será a metodologia que vamos utilizar neste trabalho. Como o nome sugere, esta metodologia é normalmente usada na extração de dados, por isso será a base do projeto e do relatório. Os seis estágios principais do CRISP-DM são compreensão / identificação do problema em causa, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação.

A primeira fase do trabalho corresponde também à primeria fase do CRISP-DM, o *Business Understanding*, que será descrita no início do trabalho e será onde iremos descrever a realidade descrita pelo conjunto de dados e as ferramentas utilizadas. A seguinte fase corresponde ao *Data Understanding* e este será um processo de constante aprendizagem ao longo do trabalho. Inicialmente será feita uma descrição, exploração e verificação dos dados que serão utilizados, de seguida iremos formular os nossos objetivos para o projeto e para cada objetivo haverá uma exploração mais focada e precisa. Após esta exploração mais detalhada os dados serão trabalhados seguindo o *Data Preparation* necessário para cada objetivo, a partir dos dados trabalhados iremos passar para o *Modeling* onde iremos realizar estatísticas acerca dos dados. A fase final do trabalho será a *Evaluation* onde com base nos nossos resultados iremos tentar chegar a respostas para os objetivos propostos.

**Data Understanding**

O nosso DataSet tem por fonte a Autoridade Nacional de Segurança Rodoviária com acidentes de vários anos, desde 2010 ao ano de 2019. Iniciámos com um dataset com ???? registos. A nossa base de dados tem N 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. O DataSet era repartido em anos que foi unido e lido em um único dataframe ‘acidentes’.

O DataSet continha XXX NA´s que 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 porque 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.

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.