|  |
| --- |
| Logo, company name  Description automatically generated |
| **VISUALIZAÇÃO DA INFORMAÇÃO**  **PROJETO FINAL** |
| Mestrado em Ciência dos Dados  2021/2022  Catarina Correia  Janice Dudensing  Lais Mendonça  Lissette Herrera Montero |

**Contents**

[1. Âmbito 4](#_Toc104634238)

[2. Objetivos 4](#_Toc104634239)

[3. Análise Exploratória de Dados | EDA 4](#_Toc104634240)

[3.1. Os dados do Dataset 4](#_Toc104634241)

[3.1.1. Variáveis discretas 5](#_Toc104634242)

[3.1.2. Variáveis contínuas 12](#_Toc104634243)

[3.1. Limpeza de dados 17](#_Toc104634244)

[3.1. Análise de Correlação 17](#_Toc104634245)

# Âmbito

Pretende-se com este projeto proceder a uma análise e descoberta de padrões em dados que nos permita tirar conclusões sobre um problema concreto recorrendo à visualização de informação.

Para o efeito foi disponibilizado o *dataset* *dadosRH.csv* que contém informação sobre os funcionários de uma organização. Nesse dataset estão dados devem ajudar a detetar padrões na razão de saída de funcionários da empresa.

# Objetivos

Será realizada uma análise dos dados que constam no *dataset* em diversas etapas, procurando, numa primeira fase, conhecer a estrutura dos dados e suas características principais, para posteriormente proceder à aplicação de técnicas de visualização e sumarização que permitirão atingir os seguintes objetivos:

* Análise e descoberta de padrões em dados, através da visualização de informação;
* compreender o que leva à saída ou não, dos funcionários da empresa;
* identificar padrões mais relevantes que determinam a saída dos funcionários da empresa.

# Análise Exploratória de Dados | EDA

A Análise de Dados Exploratórios | EDA consiste em compreender o conjunto de dados, resumindo as suas principais características, muitas vezes através de representação visual. A representação gráfica consiste na visualização através de histogramas, *bloxplots*, scatterplots, entre outros.

## Os dados do Dataset

Numa primeira fase procedemos à consulta da estrutura dos dados através da visualização do conjunto de dados.

Graphical user interface

Description automatically generated

Verificamos assim que o *dataset* é composto por *features* que caracterizam a situação de cada funcionário na empresa. Com base na designação de cada campo, procedeu-se à contextualização de cada variável (dedução), bem como a uma análise preliminar relativa à forma como esta informação poderá, eventualmente, influenciar a variável *target*: saiu.

Assim, em primeiro lugar, analisamos a variável objetivo ‘Saiu’, verificando que 23,8% dos colaboradores saíram da empresa. Podemos observar também que o problema tem duas classes e que o nosso *dataset* está desbalanceado: 23,8% pertence à classe 'Sim' e 76,2% à classe 'Não'.

Gráfico

Descrição gerada automaticamente

Na fase seguinte, analisamos a distribuição de colaboradores pelas duas classes relacionando com cada uma das features do dataset.

### Variáveis discretas

**Nível salarial** Classificação do nível salarial de cada funcionário (baixo, médio, alto).

Poderá ser uma variável explicativa para o problema levantado. Um funcionário com um nível salarial mais baixo, poderá ter maior probabilidade de abandonar a empresa.

Gráfico, Gráfico de explosão solar

Descrição gerada automaticamente

**Acidente de trabalho** Regista se o funcionário sofreu ou não algum acidente de trabalho. Um funcionário que tenha sofrido um acidente poderá ter maior probabilidade de abandonar a empresa?

Gráfico, Gráfico de explosão solar

Descrição gerada automaticamente

**Promovido5Anos** Indica se o funcionário foi promovido nos últimos 5 anos (1=sim; 0=não). Um funcionário que não tenha sido promovido nos últimos 5 anos poderá ter maior probabilidade de abandonar a empresa.

Gráfico, Gráfico de explosão solar

Descrição gerada automaticamente

Chart, sunburst chart

Description automatically generated

**Departamento** Departamento ao qual o funcionário está afeto. Existirão departamentos com maior número de saídas? Poderá este fenómeno estar relacionado com a chefia?

Gráfico

Descrição gerada automaticamente

**NumProjetos** Número de projetos nos quais o funcionário está envolvido. Um funcionário envolvido em muitos projetos pode significar que tem um nível de envolvimento com a empresa grande e que por isso não é provável que saia. No entanto, esta informação conjugada com o número de horas de trabalho, pode significar que o funcionário está com excesso de trabalho.

Gráfico

Descrição gerada automaticamente

**TempoServiço** Número de anos do funcionário ao serviço da empresa. Pode traduzir o nível de envolvimento ou o nível de saturação.

Gráfico

Descrição gerada automaticamente

### Variáveis contínuas

Para análise das variáveis contínuas, foi utilizado o histograma marginal que permite na mesma representação visualizar a distribuição dos dados através de um histograma e do correspondente bloxplot.

**Satisfação** Nível de satisfação do funcionário, medido, provavelmente, através de inquérito de satisfação realizado anteriormente. Poderá traduzir a intenção de o funcionário se manter na empresa ou de sair.

**Gráfico, Gráfico de barras

Descrição gerada automaticamente**

Juntando a esta análise a variável Nível Salarial, podemos tirar outras conclusões sobre a relação entre os níveis salariais e satisfação dos funcionários. Os funcionários que saíram apresentam salários mais baixos, com uma variância alta entre eles. Os que saíram e tinham salários altos são outliars.

**Gráfico

Descrição gerada automaticamente com confiança média**

**Última avaliação** Resultado do último processo de avaliação de desempenho do funcionário. Um bom desempenho, poderá, eventualmente, indicar que o funcionário se mantém na empresa.

Gráfico

Descrição gerada automaticamente

Juntando a esta análise a variável Nível Salarial, podemos complementar a análise:

Gráfico

Descrição gerada automaticamente

Na visualização abaixo podemos ver padrões naqueles que saíram em relação a última avaliação e satisfação: Aqueles com última avaliação e satisfação baixos saíram; os que tiveram boa avaliação e satisfação altos também saíram.

Gráfico

Descrição gerada automaticamente

Por fim, ao comparar a variável última avaliação com horas mensais, vemos que os bem avaliados que tinham mais horas trabalhadas saíram.

Gráfico

Descrição gerada automaticamente

**HorasMensais** Total de horas mensais de trabalho do funcionário. Um funcionário com excesso de horas de trabalho poderá querer sair da empresa por esse motivo. A variância de horas trabalhadas entre aqueles que saíram é alta.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Analisando em conjunto as horas trabalhadas e o nível salarial vemos que aqueles que saíram estão nos extremos (muitas horas trabalhadas ou poucas, não estão na média assim como aqueles que não saíram).

Gráfico, Histograma

Descrição gerada automaticamente

Analisando em conjunto as horas trabalhadas e o satisfação vemos padrões e uma alta incidência de saídas entre os que trabalharam mais de 250 horas.

Uma imagem contendo Gráfico

Descrição gerada automaticamente

Análise conjunta variáveis 'Satisfação', 'UltimaAvaliacao' e 'HorasMensais':

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Nesta representação gráfica, podemos observar regiões bem marcadas dos funcionários que saem da empresa. Por exemplo, funcionários com elevada satisfação e muitas horas mensais e última avaliação com valores altos.

## Análise de Correlação

Para a avaliação de possíveis correlações lineares entre as variáveis do *dataset*, criamos a seguinte matriz através da qual concluímos que existe uma relação negativa expressiva entre a Satisfação e a variável objetivo.

Chart, treemap chart

Description automatically generated

Concluímos também que a variável que outras variáveis se relacionam com a saída de funcionários de forma mais ligeira: acidentes de trabalho, número de projetos e tempo de serviço.

Por outro lado, verificámos que existe correlação entre a variável Última Avaliação e as variáveis Satisfação e Horas Mensais.

A picture containing calendar

Description automatically generated

Scatter chart

Description automatically generated

# Classificação

## Análise de Clusters

Fizemos uma análise de cluster considerando as variáveis satisfação e última avaliação:

Gráfico

Descrição gerada automaticamente

## Classificação

Separamos o dataset entre conjunto de teste e treino com a proporção de 70 e 30 para testar os classificadores: KNN, LDA, QDA, Logistic Regression e Naive Bayes. Fizemos os testes com o sem cross validation para comparar resultados:

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Comparação entre a acurácia dos modelos mostra que o KNN é um bom classificador para esse dataset:

Gráfico, Gráfico de cascata

Descrição gerada automaticamente com confiança média

# DASHBOARD

Utilizando a biblioteca Dash criamos um dashboard com as visualizações que fizemos para facilitar a análise das diversas variáveis.

No primeiro quadro é possível escolher entre as variáveis e os tipos de gráfico.

Gráfico

Descrição gerada automaticamente

A segunda visualização conta com 3 dimensões que podem ser cruzadas. Além disso é possível escolher os tipos de gráficos.

Interface gráfica do usuário

Descrição gerada automaticamente

A última visualização permite o cruzamento de 4 variáveis e a mudança de cores do gráfico

Uma imagem contendo Linha do tempo

Descrição gerada automaticamente