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

Contents

[1. Âmbito 4](#_heading=h.30j0zll)

[2. Objetivos 4](#_heading=h.1fob9te)

[3. Análise Exploratória de Dados | EDA 4](#_heading=h.3znysh7)

[3.1. Os dados do Dataset 4](#_heading=h.2et92p0)

[3.1. Limpeza de dados 7](#_heading=h.tyjcwt)

[3.1. Análise de Correlação 7](#_heading=h.3dy6vkm)

# 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, e que inclui diversas variáveis que podem em conjunto ajudar a detetar padrões que determinam a razão de saída 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 e um primeiro 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 á forma como esta informação poderá, eventualmente, influenciar a variável *target*: saiu.

Assim:

| **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. | Chart, bar chart  Description automatically generated |
| --- | --- |
| **Acidente de trabalho** Regista se o funcionário sofreu ou não algum acidente de trabalho (0 ou 1).Um funcionário que tenha sofrido um acidente poderá ter maior probabilidade de abandonar a empresa? | Chart, bar chart  Description automatically generated |
| **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? |  |
| **Departamento** Departamento ao qual o funcionário está afeto. Existirão departamentos com maior número de saídas? Poderá este fenómeno estar relacionado com a chefia? |  |
| **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. |  |
| **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. |  |
| **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. |  |
| **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. |  |
| **TempoServiço** Número de anos do funcionário ao serviço da empresa. Pode traduzir o nível de envolvimento ou o nível de saturação. |  |
| **Saiu** Variável target – assume valor 0 (não) ou 1 (sim). |  |

## Limpeza de dados

Procuramos registos duplicados e colunas irrelevantes para a análise pretendida e concluímos que a coluna ID não é necessária e que existem 3008 registos duplicados (cerca de 20%). Concluímos ainda que não existem registos nulos em nenhuma das variáveis.

O dataset fica assim com um total de 11.991 registos. Características dos dados:

Table

Description automatically generated with low confidence

## Análise de Correlação

Para a avaliação de possíveis correlações lineares entre as variáveis do *dataset*, criamos a seguinte matriz através da qual concluímos que existem relações mais expressivas entre a

Graphical user interface

Description automatically generated

Numa breve análise concluímos que a variável que mais se relaciona com a saída de funcionários é o tempo de serviço, e verificamos ainda que o nível de satisfação tem uma relação negativa com a variável que representa o número de funcionários que saíram.

Treemap chart

Description automatically generated with medium confidence