Logo, company name

Description automatically generated**VISUALIZAÇÃO DA INFORMAÇÃO**

# 

**Mestrado em Ciência dos Dados**

**2021/2022**

|  |  |
| --- | --- |
| Catarina Correia | Janice XXXXX |
| Laís Mendonça | Lissete Herrera Montero |

Índice

[1. Âmbito 2](#_Toc104905780)

[2. Objetivos 2](#_Toc104905781)

[3. Análise Exploratória de Dados | EDA 2](#_Toc104905782)

[3.1. Os dados do Dataset 2](#_Toc104905783)

[3.1.1. Variáveis discretas 3](#_Toc104905784)

[3.1.2. Variáveis contínuas 5](#_Toc104905785)

[3.1. Análise de Correlação 10](#_Toc104905786)

[4. Classificação 13](#_Toc104905787)

[5. Dashboard 14](#_Toc104905788)

# 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 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 de 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 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'.

Chart

Description automatically generatedChart, bar chart

Description automatically generated

Na fase seguinte, analisamos a distribuição de colaboradores pelas duas classes (saiu | não saiu) relacionando-as 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. | Chart  Description automatically generated |
| **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? |  |
| **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? |  |
| **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. |  |
| **TempoServiço** Número de anos do funcionário ao serviço da empresa. Pode traduzir o nível de envolvimento ou o nível de saturação. |  |

### 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.

Chart, bar chart

Description automatically generated

Juntando a esta análise a variável Nível Salarial, podemos tirar outras conclusões dobre a relação entre os níveis salariais e satisfação dos funcionários. Pela imagem abaixo, poderemos talvez concluir que os funcionários com níveis salariais mais **altos, são os que menos satisfeitos estão.**

Chart, bar chart

Description automatically generated

**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.

Para avaliação desta variável contínua, utilizou-se um histograma combinando com violinplot e scatterplot. **Conclusão**.

Chart, bar chart

Description automatically generated

Conjugando a variável *Última Avaliação* com o *Nível Salarial*:

Chart, bar chart

Description automatically generated

Conjugando a variável *Última Avaliação* com a variável *Satisfação*:

Chart

Description automatically generated

**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.

Chart, bar chart

Description automatically generated

Conjugando a variável *Horas Mensais* com o *Nível Salarial*:

Chart

Description automatically generated

Conjugando a variável *Horas Mensais* com a variável *Satisfação*:

Graphical user interface

Description automatically generated

Análise conjunta das variáveis *Número de Projetos* e *Satisfação*:

Table

Description automatically generated

Análise conjunta das variáveis *Satisfação*, *Última Avaliação* e *Horas Mensais*:

A picture containing diagram

Description automatically generated

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

## Análise de Correlação

Para a avaliação de possíveis correlações lineares entre as variáveis do *dataset*, criámos 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

Foi também possível descobrir que existe uma correlação entre as variáveis *UltimaAvaliacao*, *NumProjetos* e *HorasMensais*. De seguida, será feita uma visualização com

estas três variáveis.

A picture containing calendar

Description automatically generated

Scatter chart

Description automatically generated

# Classificação

Com o objetivo futuro de definição de um modelo de classificação que, com base nos dados recolhidos, seja capaz de determinar o perfil de novos funcionários a integrar na empresa, ou outros não existentes na amostra de dados fornecida, aoplicaram-se seis algoritmos de classificação e compararam-se os respetivos resultados de *accuracy*. Estes resultados são representados nas seguintes figuras:

|  |  |
| --- | --- |
| Graphical user interface, text, application, table  Description automatically generated |  |

Chart, waterfall chart

Description automatically generated

# Dashboard

Layout geral do Dashboard:

Graphical user interface

Description automatically generated

Para a criação do Dasboard, foram criadas antecipadamente três listas que guardam o nome das variáveis que vão ser utilizadas no Dashboard:

Text, calendar

Description automatically generated with medium confidence

Criamos também um *dataframe* que resume as *features* do *dataset* pela média de cada valor, agrupado pela variável objetivo ‘Saiu’:

Text

Description automatically generated with medium confidence

Este *Dataframe* será apresentado no *header* do Dashboard.

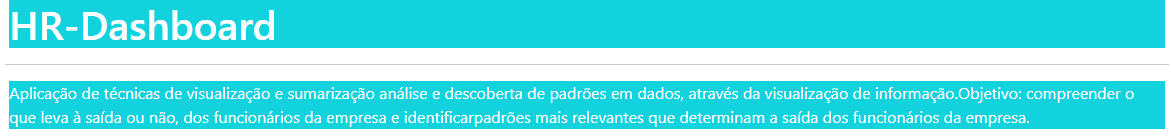
O *Dashboard* encontra-se dividido em 3 áreas:

1 | Visualização da distribuição de cada variável

2 | Visualização da distribuição de cada variável por tipo de gráfico

3 | Visualização de Gráfico com 3 dimensões de escolha e tipo de objeto

A página dispõe também de um heade com uma brave descrição dos objetivos do Dashboard:



**SECÇÃO 1**

Na seção 1, é disponibilizada a tabela que sintetiza as médias de valores de cada variável quantitativa, agrupadas pela variável objetivo. É possível escolher na combobox o tipo de gráfico que se pretende visualizar (barras ou pie):

Chart, bar chart

Description automatically generated with medium confidence

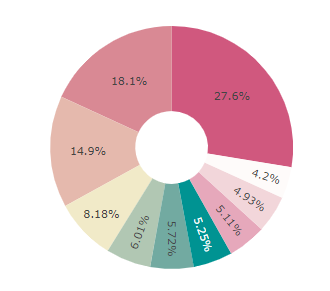
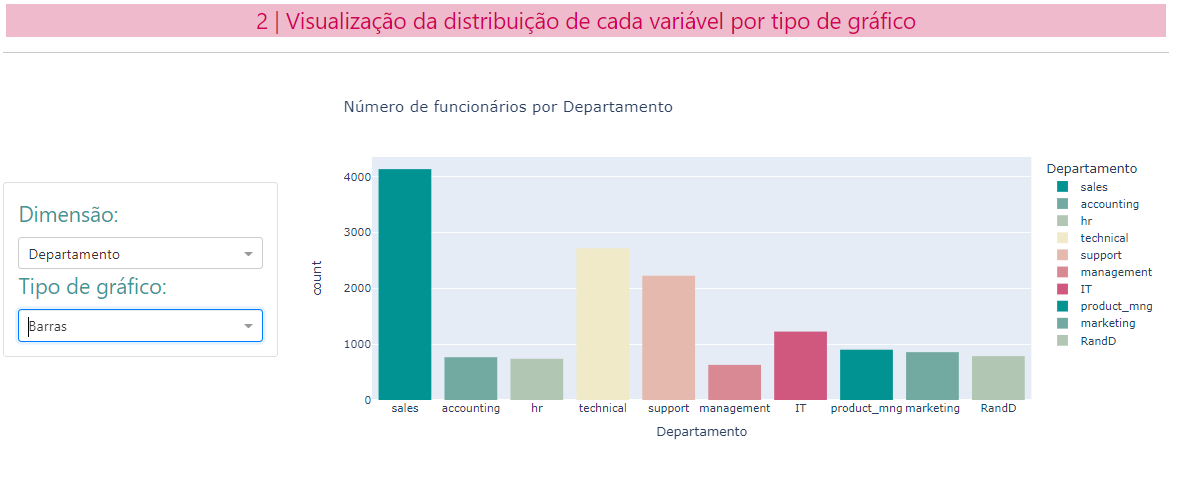
Escolhendo a opção *Pie*:

Chart, bubble chart

Description automatically generated

**SECÇÃO 2**

Nesta secção, é possível selecionar, para cada variável discreta, o tipo de gráfico que se pretende visualizar. (barras ou pie):



**SECÇÃO 3**

Nesta secção, podemos escolher três dimensões do dataset e o tipo de subplot que pretendemos visualizar no 2º gráfico (boxplot ou violin). Os gráficos mostram a relação entre as duas variáveis discretas selecionadas e a variável objetivo ‘Saiu’.

