**UNIVERSIDADE PRESBITERIANA MACKENZIE**

Projeto Aplicado II:  
Análise de Turnover na TIVIT

**São Paulo**, Março de 2024

# Membros

|  | **TIA** | **Email** |
| --- | --- | --- |
| Elaine Di Fonzo Assumpção | *10407079* | 10407079@mackenzista.com.br |
| Felipe Gabriel de Avelar Teodoro | *10415788* | 10415788@mackenzista.com.br |
| Rafael José de Oliveira | 10415375 | 10415375@mackenzista.com.br |

# Sumário

[**Membros 2**](#_w9wmx4qzrx0y)

[**Sumário 3**](#_5e08rafcb02h)

[**Introdução 3**](#_wiueyvu8jr8c)

[**Premissas do Projeto 3**](#_npmwyd9efnub)

[**Objetivos e Metas 3**](#_wmstj77iy9je)

[**Cronograma 4**](#_dgt7uno2mlxe)

[**Github 4**](#_1i89txp4z2i)

**Linguagem Utilizada no Projeto...................................................................................4**

**Análise Exploratória de Dados (EDA)..........................................................................5**

**Tratamento da Base de Dados (Preparação e Treinamento) .................................11**

**Definição e Descrição das Bases Teóricas do Métodos.........................................11**

**Definição e Descrição do Cálculo da Acurácia........................................................12**

**Pré-Processamento e Treinamento...........................................................................12**

# Introdução

Nesse estudo, vamos analisar as principais causas de desligamentos voluntários (*turnover*) na base de funcionários da TIVIT. Analisaremos especialmente os desligamentos de funcionários qualificados com certificações importantes para a empresa.

# Premissas do Projeto

A TIVIT é, segundo suas [próprias palavras](https://www.linkedin.com/company/tivit/about/), "*uma multinacional brasileira e one-stop-shop de tecnologia, presente no mercado desde 1998 e com operações em dez países da América Latina. A companhia oferece serviços diversos como desenvolvimento de software ágil, nuvem híbrida e pública, cibersegurança, operações de alta complexidade, entre outros, que apoiam os clientes em sua jornada de transformação digital.*" Atualmente conta com um quadro de 5 mil colaboradores, atuando em 10 países da América do Sul com 50 unidades. Seus serviços oferecidos são Soluções Digitais, Soluções SAP, Mainframes e Serviços Transacionais, Cloud, Serviços Gerenciados, Sistemas Inteligentes, Cybersecurity, Digital Workplace, Dados e IA, Processos e Informações.

Por ser uma empresa de grande porte, a TIVIT tem muitas pessoas se juntando e deixando a companhia todos os meses. Procurar, contratar e treinar pessoas é um processo caro e demorado, que pode limitar o potencial de crescimento de uma empresa. Portanto, **propomos usar a Análise Exploratória de Dados (AED) e a aplicação de Aprendizado de Máquina (ML) como mecanismo para descobrir potenciais causas da saída voluntária de pessoas e o que fazer para retê-las**. Portanto, o projeto se tratará de uma **manipulação de texto.**

# Objetivos e Metas

Dentre os motivos principais, procuraremos compreender:

* Explorar e compreender os dados..
* Tratar e modelar a base de dados.
* Compreender se a coleta de dados atual pode ser otimizada para a geração de modelos mais precisos no futuro.
* Estimar quais colaboradores estão com maior probabilidade de deixarem a empresa através de um modelo supervisionado de machine learning.
* Testar a acurácia do modelo..

Ao longo do semestre iremos tratar (data cleaning), analisar e treinar um modelo supervisionado, apresentando nossas descobertas com o auxílio de bibliotecas de AED e Scikit-Learn na linguagem Python, em Jupyter.

Portanto, o projeto se tratará de uma **manipulação de texto.**

# Cronograma

|  |  |
| --- | --- |
| **Atividade** | **Data** |
| **Análise exploratória** | **25/03** |
| **Tratamento e modelagem** | **13/04** |
| **Treinamento e Teste** | **29/04** |
| **Avaliação de acurácia** | **14/05** |

# Github

O link do arquivo encontra-se no [GitHub](https://github.com/rafaelmackenzie2024/projetoaplicado2.git).

# Linguagem Utilizada no Projeto

No projeto, utilizaremos a linguagem *Python,* devido a presença de bibliotecas como Pandas, Numpy, Matplotlib e Seaborn, que proporcionam utilização de estatísticas descritivas, geração de gráficos e modelagem dos dados de modo eficaz.

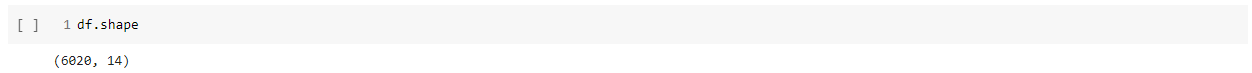
Além disso, na linguagem temos acesso ao SickitLean, que fornece um escopo de diversas ferramentas para a implementação e avaliação de algoritmos de aprendizado de máquina.

# Análise Exploratória de Dados (EDA)

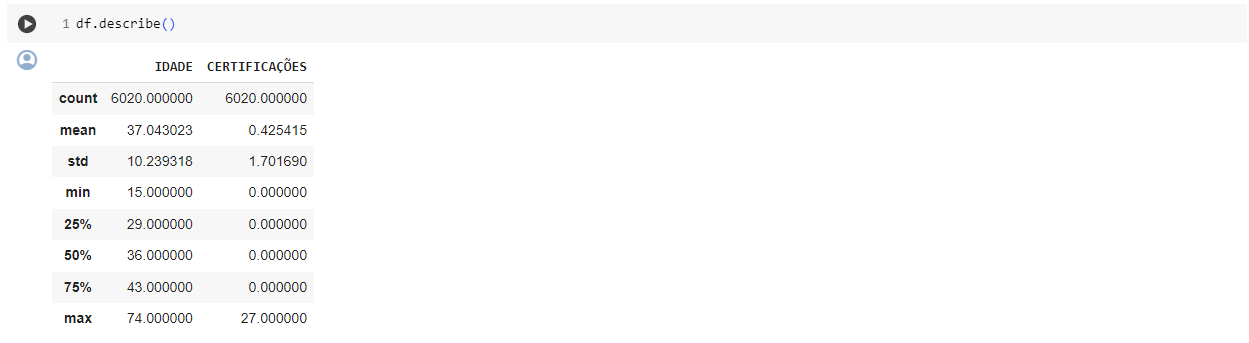
1 - Apresentação da base utilizada para criação das análises.



2 - Quantidade de linhas e colunas do Dataset



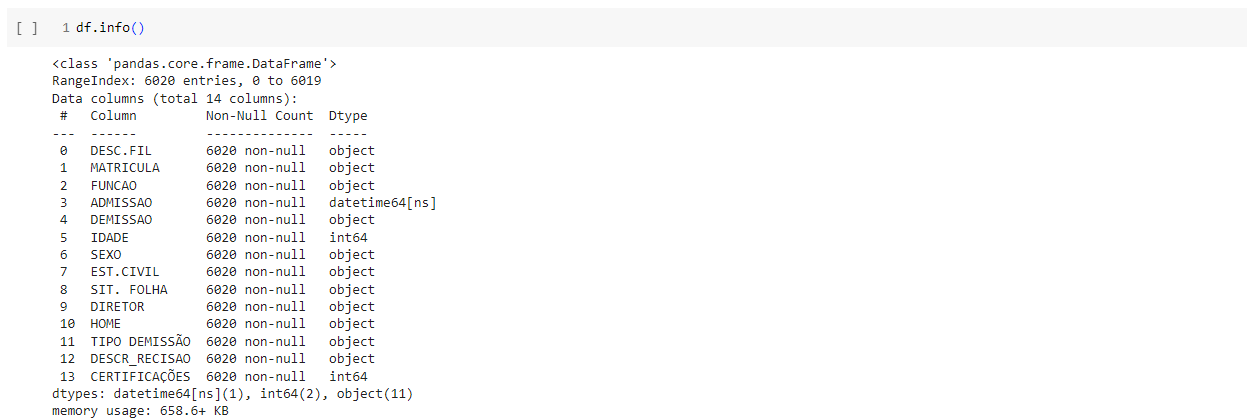
3 - Estatística Descritiva (Quantidade, média, desvio padrão, mínimo, quartis e máximo) dos campos Idade e Certificações do Dataset



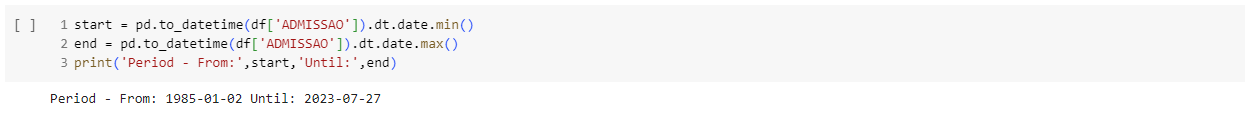
4 - Averiguação de valores nulos e ausentes de todos os campos do Dataset



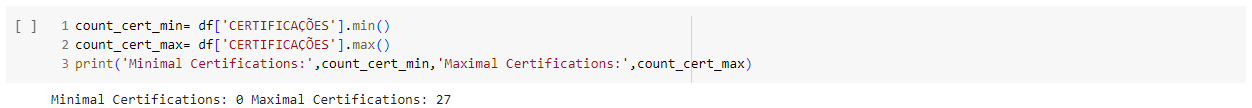
5 - Resumo de todos os campos do Dataset



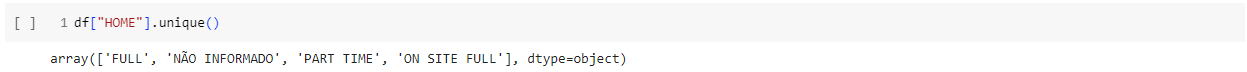
6 - Analise da primeira e última admissão



7 - Análise da menor e maior quantidade de certificações



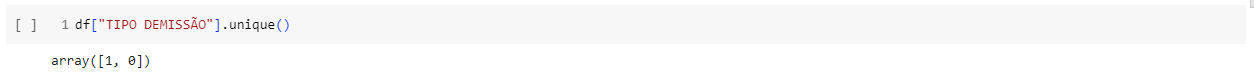
8 - Análise dos dados que compõem o campo "Home". Trata-se do colaboradores em regime de Home Office Full (Integral), Part Time (Hibrido) e On Site Full (Presencial)



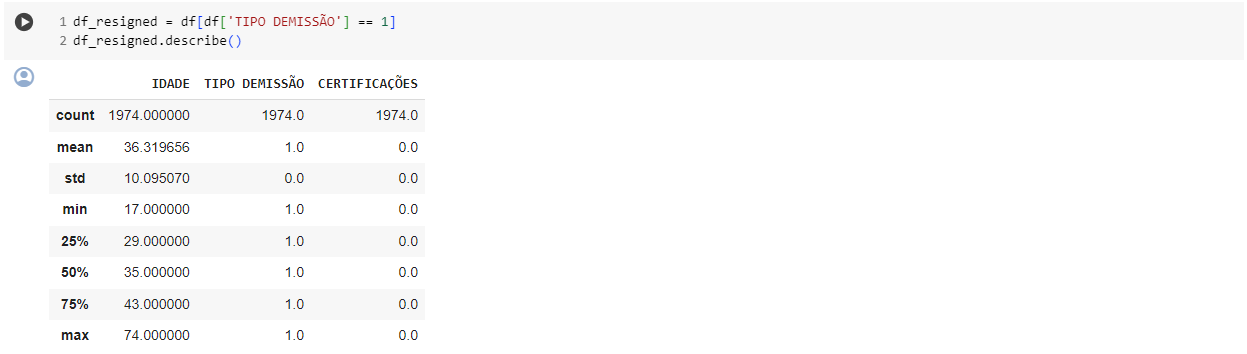
9 - Análise dos tipos de demissões, sendo, Voluntário (Quando o colaborador se demite da empresa), Provocado (Quando a empresa demite o colaborador e Experiência (Podendo ser Voluntario ou Provocado antes de completar os 90 dias)

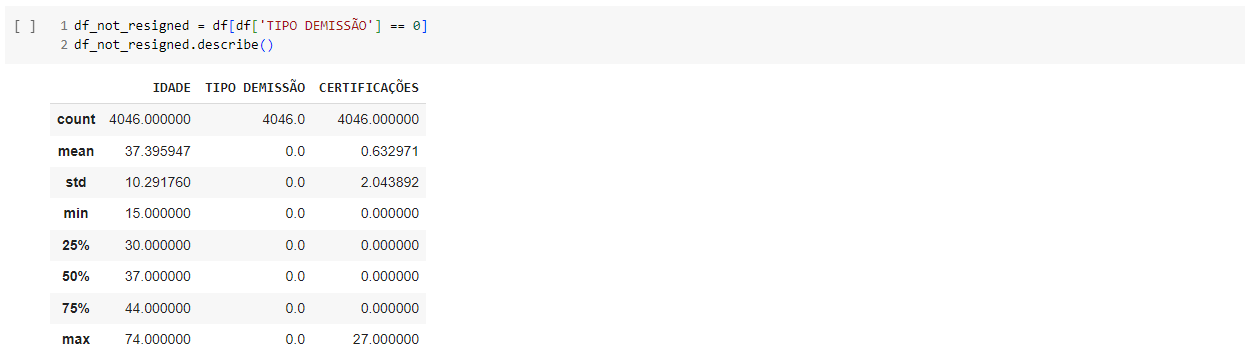


10 - Averiguação se existe mais um campo "Tipo de Demissão" no Dataset



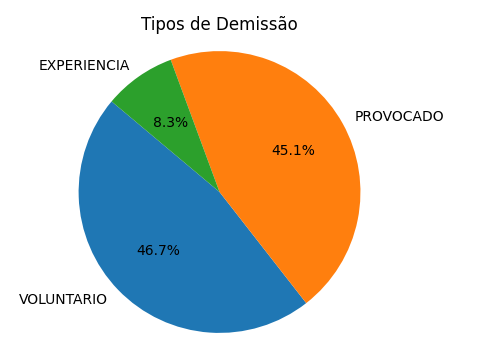
11 - Estatística Descritiva (Quantidade, média, desvio padrão, mínimo, quartis e máximo) dos campos Tipo de Demissão e Certificações do Dataset





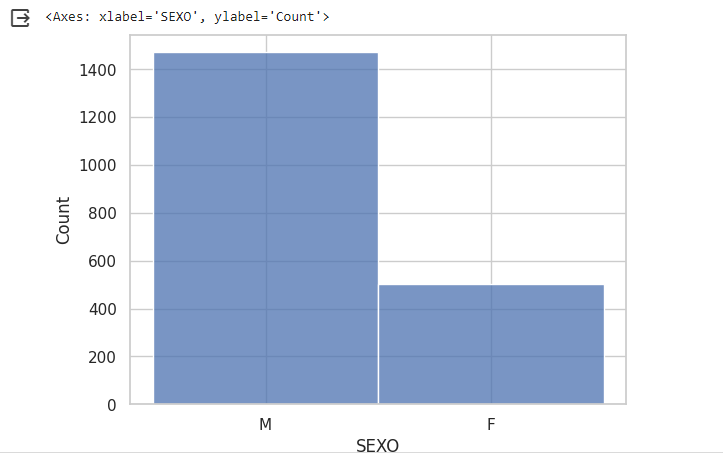
12 - Total de demissões por tipos Voluntário (Quando o colaborador se demite da empresa), Provocado (Quando a empresa demite o colaborador e Experiência (Podendo ser Voluntario ou Provocado antes de completar os 90 dias) do período de Jan22 até Jul23. O maior volume de demissões ocorre pela saída voluntária do colaborador



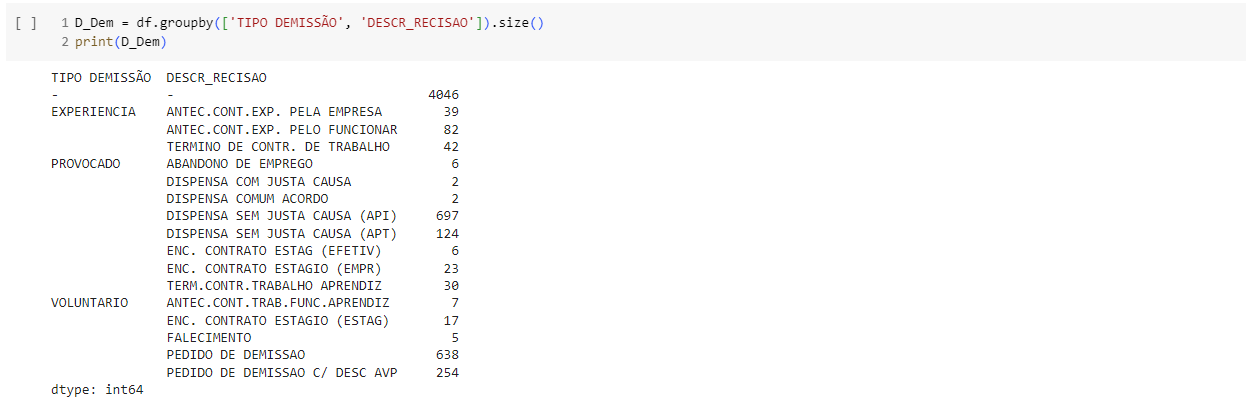


13 - Análise por sexo. A maior quantidade de demitidos e do sexo masculino, mas é influência da concentração de colaboradores homens ser maior.

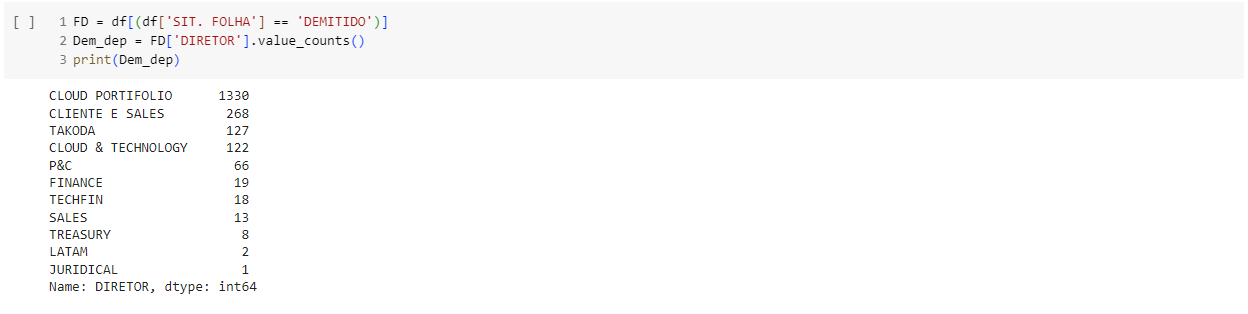


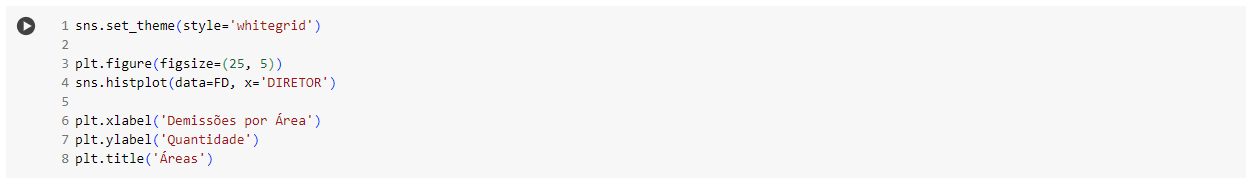


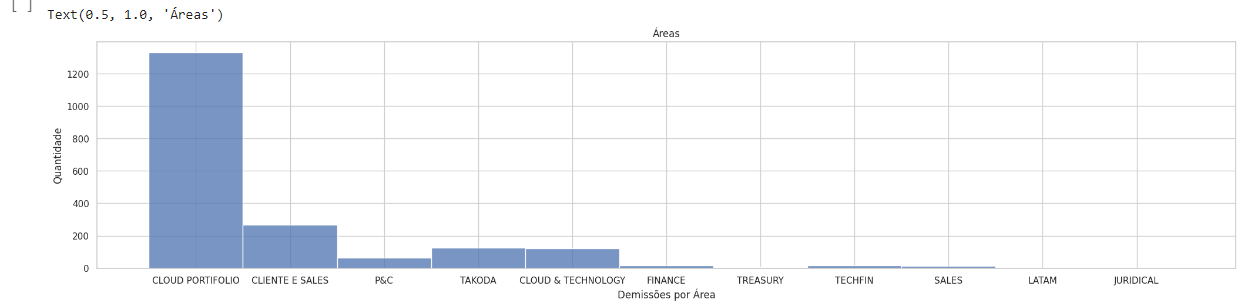
14 - Detalhamento das principais causas de demissões do período de Jan22 até Jul23, agrupadas por tipos (Voluntário, Provocado e Experiência). Em Experiência o maior volume está na antecipação por parte do funcionário, ou seja, voluntário. Em provocado o maior volume Dispensa sem justa causa e em Voluntario Pedido de Demissão



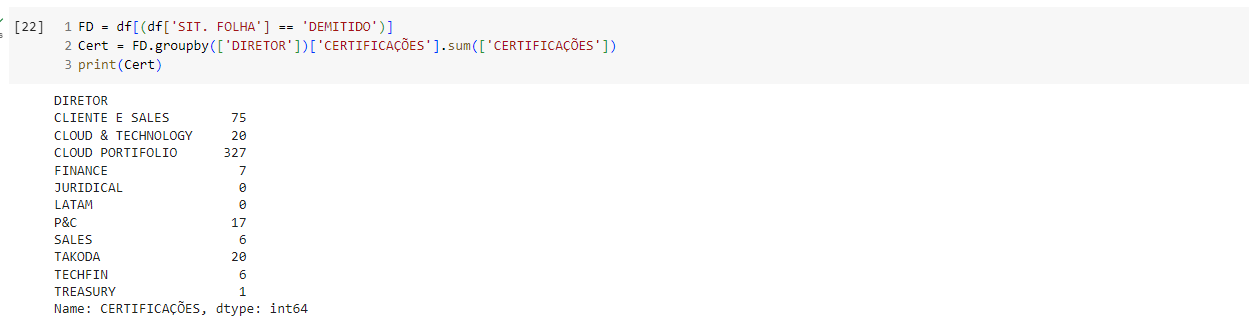
15 - Análise por departamento da empresa, considerando o mesmo período de Jan22 a Jul23, o maior volume está na área de Cloud Portfólio que também possui a maior quantidade de colaboradores e em Cliente e Sales, esta última por sua vez trata-se da área comercial e de relevância para o negócio.

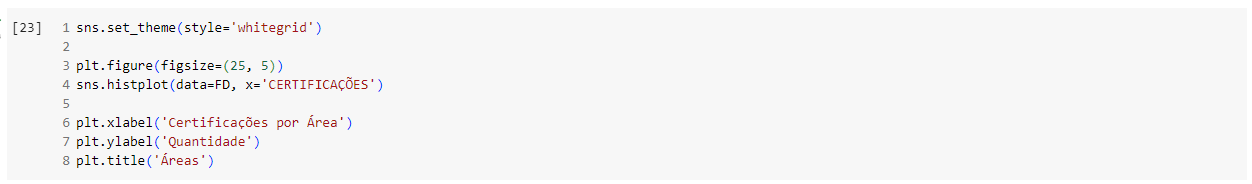


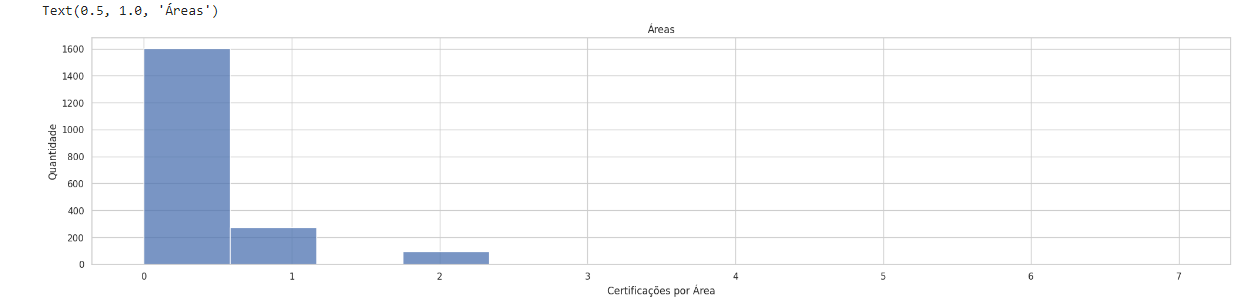




16 - Analise de Certificações por área par entender se os demitidos tiveram investimento em certificações pela empresa.







# Tratamento da Base de Dados(Preparação e Treinamento)



# Definição e Descrição das Bases Teóricas dos Métodos

No pré-processamento e treinamento, foi realizado a speração das fetures(x) e da label(y).

Por conseguinte, transformamos as variáveis categóricas e numéricas, através da função pd.get\_dummies(), proporcionando que nossa base de dados seja compreendida pelo algoritmo de classificação.

Devido a discrepância de valores da label, balanceamos os dados para que o treinamento dos algoritmos tenham o menor viés possível na classificação.

Normalizamos os dados para otimizar o aprendizado das features, com a função MinMaxScaler() e separamos os dados nos conjuntos de treino e teste com a função train\_test\_split(). Utilizamos a SEED fixa 1234 para replicarmos os resultados nos posteriores testes do modelo.

O modelo escolhido para a classificação, será a árvore de decisão. A árvore de decisão é um modelo de classificação supervisionado. A árvore de decisão, funcionando caminhos de decisão e o resultado de cada um deles.

# Definição e Descrição do Cálculo da Acurácia

O cálculo de acurácia, será realizado por meio do método *score* do scikit learn, que realiza a previsão com as bases de teste, verificando qual a porcentagem de acertos nesse conjunto de dados. Para complementar a análise, verificaremos a matriz de confusão, de modo que analisaremos especificamente a quantidade de true positives(TP), true negatives(TN), false positives(FP) e false negatives(FN).

Dessa forma, poderemos verificar a eficácia do modelo em prever a label de forma detalhada.

Gráfico

Descrição gerada automaticamente

Mapa de calor da matriz de confusão para avaliar o desempenho de um classificador, mostrando como as previsões do modelo se comparam às classes verdadeiras.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

# REFERÊNCIAS

GRUS,Jeol. Título: Data Science do zero: primeiras regras com o Python. Edição. SÃO PAULO, Alta Books, 27 de junho de 2016.