**Centro Universitário Carlos Drummond de Andrade**

Andrew de Souza Almeida
Arthur Morais Silva
Carlos Alberto Vieira de Sousa
Matheus Serrate Neiva Gonçalves
Robson do Amaral Rosa

# Análise de dados: Base Adult census

São Paulo - SP
2021

**Centro Universitário Carlos Drummond de Andrade**

# Análise de dados: Base Adult census

Relatório Técnico-Científico apresentado na disciplina de Projeto Integrador para o curso de Análise e desenvolvimento de sistemas da Centro Universitário Carlos Drummond de Andrade (UNIDRUMMOND).

São Paulo - SP
2021

Almeida, Andrew; Silva, Arthur; Sousa, Carlos; Gonçalves, Matheus; Rosa, Robson. **Análise de dados: Base Adult census.** Relatório Técnico-Científico. Análise e desenvolvimento de sistemas – **Centro Universitário Carlos Drummond de Andrade**. Tutor: Me Eduardo Palhares Junior. Polo Ponte Rasa, 2021.

## RESUMO

Este projeto foi criado com a finalidade de ler, estudar e analisar os resultados obtidos por uma análise de dados com base na *Base adult census* e identificar quantas pessoas possuem uma renda maior que $50.000,00 Dólares, por meio de filtragem de informações e utilizando dados válidos perante ao que foi decidido no processo de criação do projeto. Tudo sendo feito com código com a linguagem Python e uma ferramenta especifica para a análise de dados e criação de matrizes de correlação de dados.

**PALAVRAS-CHAVE:** Adult; Dados; Matrizes; Python;

## LISTA DE ILUSTRAÇÕES

**FIGURA 1**– IMPORTAÇÃO DAS BIBLIOTECAS..... 10
**FIGURA 2**– COMUNICAÇÃO COM O GOOGLE DRIVE.....10
**FIGURA 3**– VARIÁVEL INCOME.....11
**FIGURA 4**– TABELA DO MODELO NEAREST NEIGHBOR.....12
**FIGURA 5**– TABELA DO MODELO NAIVE BAYES.....13
**FIGURA 6**– TABELA DO MODELO ÁRVORE DE DECISÃO.....13
**FIGURA 7**– TABELA DO MODELO REGRESSÃO LOGÍSTICA.....14
**FIGURA 8**– TABELA DO MODELO SVC.....14
**FIGURA 9**– TABELA DO MODELO FLORESTA ALEATÓRIA.....15

## SUMÁRIO

**1. INTRODUÇÃO.....6**

**2. DESENVOLVIMENTO.....8**

**2.1 OBJETIVOS**.....8
**2.2. JUSTIFICATIVA E DELIMITAÇÃO DO PROBLEMA**.....8
**2. 3. FUNDAMENTAÇÃO TEÓRICA**.....8
**2.4. APLICAÇÃO DAS DISCIPLINAS ESTUDADAS NO PROJETO INTEGRADOR**.....9
**2.5. METODOLOGIA**.....9
**2.6. O PROJETO**.....10

**3. RESULTADOS.....12**

**3.1. SOLUÇÃO INICIAL**.....12
**3.2. SOLUÇÃO FINAL**.....12

**4. CONSIDERAÇÕES FINAIS.....12**

**REFERÊNCIAS.....14**

**ANEXOS.....15**

## 1. INTRODUÇÃO

O controle dos dados é essencial no momento atual. Nome, idade, status, documento, nível de educação, salário, ocupação etc. Quanto mais completo for o cadastro melhor. Os datastes (ou conjuntos de dados) são o principal insumo dos processos de análise de dados. Eles são representados por dados tabulares em formato de planilha onde as linhas são os registros dos acontecimentos e as colunas são as características desses acontecimentos.

É usado para investigar a dependência entre várias variáveis ao mesmo tempo e para destacar as variáveis mais correlacionadas em uma tabela de dados. Neste visual, os coeficientes de correlação são coloridos de acordo com o valor. A matriz de correlação também pode ser reordenada de acordo com o grau de associação entre as variáveis ou agrupada usando algoritmo de agrupamento hierárquico. O uso deste visual é muito simples e intuitivo.

As principais estruturas utilizadas no python são as listas de dados, que armazena em sequência em cada valor um índice correspondente, os dicionários que são coleção de itens desordenados que possui uma diferença maior comparada a outro tipo de coleção e tuplas que são estruturas próximas as listas que tem uma característica diferente, onde os elementos inseridos não podem ser alterados.

O Google Colaboratory foi a ferramenta escolhida para a codificação do código, esta que nada mais é que um serviço em nuvem gratuito da google onde pode ser usado principalmente para aprendizado e estudo do machine learning.

Seu sistema de divisões de instruções e separação entre texto e código é algo que facilita o entendimento tornando o ambiente “limpo” para a leitura do código. Uma vantagem em se usar esta ferramenta do google é a rápida e fácil comunicação com o Google Drive, podendo assim acessar qualquer base de dados desejada que estiver lá. A linguagem de programação usada nesta ferramenta é o Python, linguagem esta que é considerada uma ótima opção para os códigos de análise de dados e inteligência artificial, pois sua fácil escrita e suas bibliotecas dedicadas para análise de dados facilitam a codificação. Na linguagem python , temos vários tipos de estruturas de dados que podemos utilizar, pode ser a solução de problemas ocasionais e outras situações durante o processo de desenvolvimento.

# 2. DESENVOLVIMENTO

## 2.1 Objetivos

O objetivo deste projeto é análise e discretização de uma base de dados, a Adult census income, uma base de dados feita com informações de mais de 30 mil pessoas com fins de identificar quem possuí uma renda maior a $ 50.000,00 dólares. Assim usando Python e estrutura de dados para poder simplificar as informações obtidas, formando matrizes de correlação, gráficos comparativos para a análise do capital e trazendo os resultados em um número médio.

## 2.2. Justificativa e delimitação do problema

O problema da nossa solução é o algoritmo usado para gerar o modelo, porém o grande vilão é o seu próprio conjunto de dados que podem possuir muitos atributos com valores faltantes, (outliers) e escalas de valores contra dizentes e por fim nenhum modelo será capaz de trabalhar com esses dados e gerar resultados de qualidade. Com o código feito no projeto todos os dados que seriam incapazes de serem avaliados foram retirados, sendo assim tornando possível discretizar os valores antes em palavras por números, sendo assim tornando mais fácil para a criação das matrizes de correlação dos resultados, e do projeto em um todo.

## 2. 3. Fundamentação teórica

Machine learning é uma área da ciência da computação que apresenta como significado ser o “aprendizado da máquina”, o qual evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial. Desse modo, estuda meios para que máquinas posam fazer tarefas que seriam executadas por pessoas, bem como, é uma programação usada nos computadores, formada por regras previamente definidas que permitem que os computadores tomem decisões com base nos dados disponíveis. Nesse sentido, a base do funcionamento são os algoritmos, que são sequências definidas e instruções

que vão ser seguidas pelo computador de acordo com programações feitas, o computador tem habilidade para tomar decisões que podem resolver problemas ou impulsionar publicações na internet, por exemplo.

As aulas do professor Eduardo Palhares, conhecimento da linguagem python que nos direcionou para a realização do relatório, utilizando todas as metodologias proposta na disciplina.

### 2.4. Aplicação das disciplinas estudadas no Projeto Integrador

Estrutura de dados, análise de dados, Python, lógica de programação e programação orientada a objetos foram os principais assuntos nesse semestre, e foram a base destas matérias que fizeram surgir esse projeto. Neste caso, por conta da ferramenta usada para a codificação do código a programação orientada a objetos não foi tão aproveitada, porém Python e a lógica de programação foram os principais temas utilizados e aplicados no projeto sendo eles e vistos em aula.

Saber de sintaxe e léxica básica foi essencial na codificação do programa seja ele fácil ou não. Saber interpretar um possível erro e saber que pode ser por uma simples instrução digitada errada ou a ordem das mesmas foi de suma importância.

Um dos temas das aulas utilizado no desenvolvimento do projeto foi o Dicionário, utilizado para a discretização da base de dados mapeando as variáveis desejáveis, assim facilitando a leitura dos dados, deixando-os apenas em dois estados, o estado “1” e em “0”.

### 2.5. Metodologia

O projeto consiste de uma análise de código em Python transcrito diretamente na plataforma colaboratory, todo o código foi escrito com o acompanhamento de um vídeo produzido pelo professor Eduardo Palhares Junior. A metodologia foi transcrever os códigos e analisa-los assim entendendo o que está sendo codificado e as estruturas de dados utilizadas. Pesquisas na área também foram feitas para o melhor entendimento, porém a maior parte de todos o conteúdo utilizado foi abstraído em aula.

### 2.6. O projeto

Em relação a codificação primeiro importamos as bibliotecas:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
from google.colab import drive
```

Figura 1 - Importação das bibliotecas

O NumPy é uma biblioteca para a linguagem Python com funções para se trabalhar com computação numérica. Pandas é uma biblioteca e os dois principais objetos são as series e os DataFrames. Uma Serie é uma matriz unidimensional que contém uma sequência de valores que apresentam uma indexação (que podem ser numéricos inteiros ou rótulos. DataFrame é uma estrutura de dados tabular, semelhante a planilha de dados do Excel, em que tanto as linhas quanto as colunas apresentam rótulos. Matplotlib é uma biblioteca para a visualização de dados em Python, O PyPlot é um módulo do Matplotlib para criação de gráficos.

Após a importação das bibliotecas é preciso importar a tabela desejada, no caso usamos uma tabela encontrada no Kaggle. Para a importação devemos fazer a comunicação com o Google Drive, onde deve estar a tabela em formato .csv e efetuar algumas linhas de código.

```
drive.mount("/content/drive")
```

```
pd.set_option('display.max_columns', 20)

pd.set_option('display.width', 1000)
```

```
df = pd.read_csv("/content/drive/MyDrive/adult.csv",1,",")
data = [df]
print ("Inicial dataset \n\n", df.head(10), "\n")
```

Figura 2 - Comunicação com o Google Drive

Precisamos transformar as palavras em números (valores binários), trocamos todas os valores da tabela em números discretos para poder calcular.

```
income_map = {'<=50K':1,'>50K':0}
df ['income'] = df['income'].map(income_map).astype(int)
```

Figura 3 – Variável income

É feita essa conversão com todas os tópicos da tabela, pois todas as informações, porém alguns dos tópicos da tabela possuem dados onde não podemos utilizar por serem informações faltantes, sigilosas ou não explicativas, nesses casos são removidos quaisquer dados dessas naturezas.

# 3. RESULTADOS

## 3.1. Solução inicial

Mediante a análise dos dados coletados e das soluções desenvolvidas através dos processos de construção e desenvolvimento de cada uma das etapas, foram implementados e testados alguns modelos de treinamento. Os resultados de cada um deles foi avaliado e a partir da avaliação, classificados como bons ou ruins.

O primeiro modelo de treinamento usado para classificação da base de dados foi o “Nearest Neighbor”, o que apresentou uma acurácia de cerca de 70,2%, porém este modelo encontrou grande dificuldade para classificar as pessoas que ganham, mais do que 50 mil por ano, tendo uma acurácia de 23%.

```
Accuracy score:
0.7029506022764946
---------------------------------------
Confusion Matrix:
[[ 393 1898]
 [ 790 5968]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.33      0.17      0.23      2291
           1       0.76      0.88      0.82      6758

    accuracy                           0.70      9049
   macro avg       0.55      0.53      0.52      9049
weighted avg       0.65      0.70      0.67      9049
```

Figura 4 – Tabela do Modelo Nearest Neighbor

O segundo modelo de treinamento usado foi “Naive Bayes”, onde foi possível ter uma acurácia geral de 77%, porém ainda possui pouca precisão (31%) na classificação de pessoas que ganham mais do que 50 mil dólares.

```
Accuracy score:
0.7735661399049619
---------------------------------------
Confusion Matrix:
[[ 461 1830]
 [ 219 6539]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.68      0.20      0.31      2291
           1       0.78      0.97      0.86      6758

    accuracy                           0.77      9049
   macro avg       0.73      0.58      0.59      9049
weighted avg       0.76      0.77      0.72      9049
```

Figura 5 – Tabela do Modelo Naive Bayes

O terceiro modelo de treinamento foi a Árvore de Decisão, o que conseguiu aumentar a acurácia geral para 77%, porém a acurácia na classificação de pessoas que ganham acima de 50 mil por ano ainda está abaixo do nível satisfatório, com cerca de 56%.

```
Accuracy score:
0.7768814233616974
---------------------------------------
Confusion Matrix:
[[1263 1028]
 [ 991 5767]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.56      0.55      0.56      2291
           1       0.85      0.85      0.85      6758

    accuracy                           0.78      9049
   macro avg       0.70      0.70      0.70      9049
weighted avg       0.78      0.78      0.78      9049
```

Figura 6 – Tabela do Modelo Árvore de Decisão

O próximo modelo avaliado foi a Regressão Logística, porém seu desempenho piorou com relação a Árvore de Decisões, apresentando uma acurácia geral de 46%.

```
Accuracy score:
0.4661288540170185
---------------------------------------
Confusion Matrix:
[[1807  484]
 [4347 2411]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.29      0.79      0.43      2291
           1       0.83      0.36      0.50      6758

    accuracy                           0.47      9049
   macro avg       0.56      0.57      0.46      9049
weighted avg       0.70      0.47      0.48      9049
```

Figura 7 – Tabela do Modelo Regressão Logística

Outro modelo que também foi testado foi o SVC, mas embora a sua acurácia geral tenha sido de 74%, o tempo de processamento é muito grande e a sua precisão em classificar as pessoas que ganham acima de 50 mil dólares é praticamente nula (0.5%).

```
Accuracy score:
0.7483699856337717
---------------------------------------
Confusion Matrix:
[[  54 2237]
 [  40 6718]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.57      0.02      0.05      2291
           1       0.75      0.99      0.86      6758

    accuracy                           0.75      9049
   macro avg       0.66      0.51      0.45      9049
weighted avg       0.71      0.75      0.65      9049
```

Figura 8 – Tabela do Modelo SVC

### 3.2. Solução Final

Baseado nos testes efetuados com os modelos anteriores, foi determinado que o modelo que apresentou melhor desempenho foi “Floresta Aleatória", pois além de alcançar uma

acurácia geral de 82%, tanto sua acurácia para pessoas que ganham acima de 50 mil (62%) quanto para pessoas que ganham abaixo de 50 mil (88%) estão acima do nível satisfatório.

```
Accuracy score:
0.8188750138136811
---------------------------------------
Confusion Matrix:
[[1331  960]
 [ 679 6079]]
---------------------------------------
Classification Matrix:
              precision    recall  f1-score   support

           0       0.66      0.58      0.62      2291
           1       0.86      0.90      0.88      6758

    accuracy                           0.82      9049
   macro avg       0.76      0.74      0.75      9049
weighted avg       0.81      0.82      0.81      9049
```

Figura 9 – Tabela do Modelo Floresta Aleatória

## 4. CONSIDERAÇÕES FINAIS

Com base na proposta apresentada pela universidade a respeito do escopo que o Projeto Integrador deveria possuir, o grupo foi capaz de alcançar os pontos propostos, sendo capaz de compreender o problema apresentado, identificar onde e como cada conteúdo aprendido ao longo do semestre poderia ser aplicado, identificar os problemas que surgiram ao longo do desenvolvimento do trabalho e como eles poderiam ser solucionados e implementar uma solução consistente, visando sanar o problema identificado.

O Centro Universitário Carlos Drummond de Andrade juntamente com o seu corpo docente, foram de grande ajuda que o grupo conseguisse alcançar tal objetivo, através das aulas ministradas pelos professores, disponibilidade de material para consulta e ajuda em especial do Professor Eduardo Palhares.

**REFERÊNCIAS**

ABNT – Associação Brasileira de Normas Técnicas. **NBR 14724**: Informação e documentação. Trabalhos Acadêmicos - Apresentação**.** Rio de Janeiro: ABNT, 200

**ANEXOS**

https://colab.research.google.com/drive/1dp4xStlWf36FAKQWhfiBXrOzDG6bmH3v#scrollTo=m9t-PRTv3VTl

https://www.kaggle.com/uciml/adult-census-income

https://www.youtube.com/watch?v=HaQDQvHmOtc&list=PLA2RZdy3Z15oK3JFWYmIUTZjQ13O5r6qW