

**PROJETO:** PREVENDO DIABETES EM MULHERES

**DISCENTE:** GUILHERME ALVES ()

**DISCENTE:** VINICIUS DE SOUZA SANTOS(BI3008061)

**DOCENTE:** DR. MURILO VARGES DA SILVA

**DISCIPLINA:** MDAEL8 – MINERAÇÃO DE DADOS

Sumário

[1 Objetivo 2](#_Toc129295815)

[1.1 Sobre a base de dados 2](#_Toc129295816)

[2 Pré-processamento 4](#_Toc129295817)

[3 Normalização dos Dados 10](#_Toc129295818)

[Referências Bibliográficas 11](#_Toc129295819)

# 1 Objetivo

Estimular o aluno a aplicar os conhecimentos apresentados no decorrer da disciplina em problemas reais de mineração de dados, utilizando as técnicas de seleção, pré-processamento e transformação de dados, técnicas de visualização de dados, análise descritiva, análise de grupos, classificação e estimação/regressão.

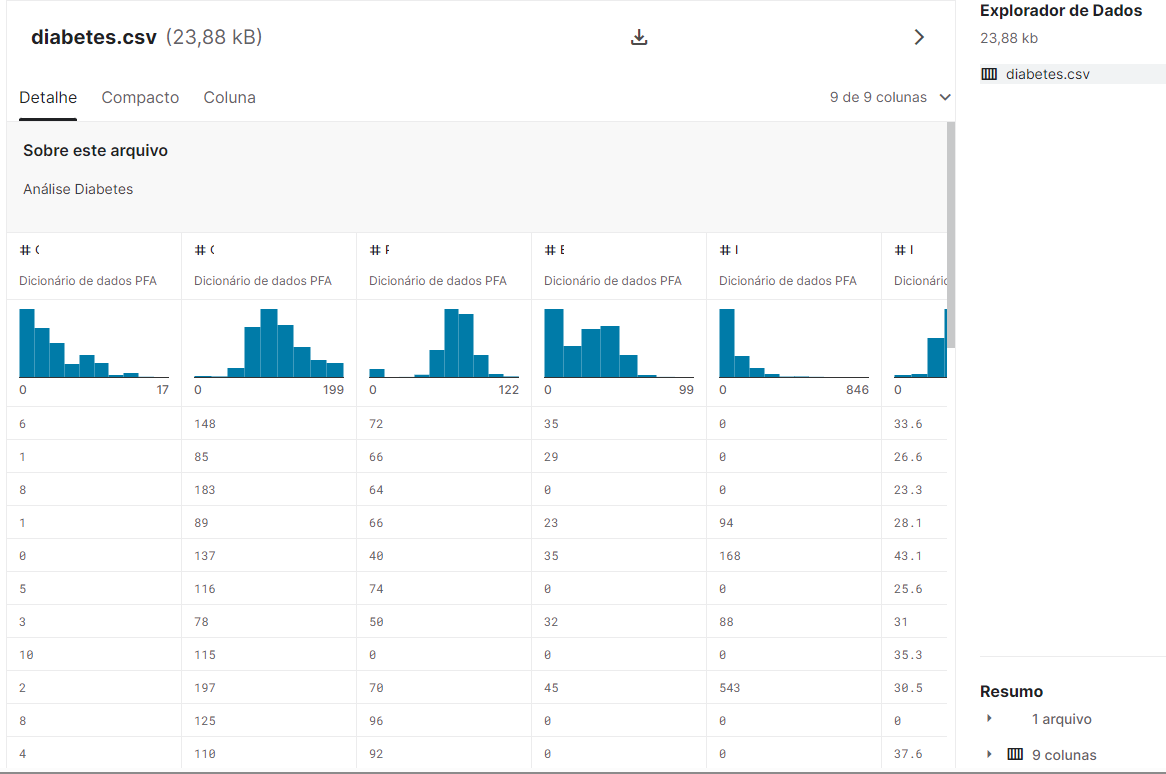
## 1.1 Sobre a base de dados

Este conjunto de dados é originalmente do Instituto Nacional de Diabetes e Doenças Digestivas e Renais. O objetivo do conjunto de dados é prever de forma diagnóstica se um paciente tem diabetes, com base em certas medidas de diagnóstico incluídas no conjunto de dados. Várias restrições foram colocadas na seleção dessas instâncias de um banco de dados maior. Em particular, todos os pacientes aqui são mulheres com pelo menos 21 anos de idade, descendentes dos índios Pima.2 A partir dos dados do arquivo (.csv) Podemos encontrar várias variáveis, algumas delas são independentes (várias variáveis reditivas médicas) e apenas uma variável dependente de destino (resultado).

# 2 Pré-processamento

Nessa primeira etapa do projeto foi feito o pré-processamento dos dados da base de prevenção a diabetes, essa base de dados foi disponibilizada pelo site da Kaggle nela foi abstraída informações relevantes como: Número de gestações, insulina, IMC, espessura da pele entre outros, com essa base de dados vamos fazer a preparação desses dados, conforme observado na figura 1 apresentando a base de dados analisada.

Figura 1: Base de dados de prevenção de diabete em mulheres



fonte: CHAUHAN, 2022

Na aula foi apresentado através de um repositório do professor vários algoritmos de mineração de dados e para esse capítulo usamos apenas o do diretório de processamento.

Para fazer o pré-processamento usamos o algoritmo DataCleaning.py.

import pandas as pd

import numpy as np

def main():

# Faz a leitura do arquivo

names = ['Número Gestações','Glucose','pressao Arterial','Expessura da Pele','Insulina','IMC','Função Pedigree Diabete','Idade','Resultado']

features = ['Número Gestações','Glucose','pressao Arterial','Expessura da Pele','Insulina','IMC','Função Pedigree Diabete','Idade','Resultado']

output\_file = '0-Datasets/diabetesClear.data'

input\_file = '0-Datasets/diabetes.data'

df = pd.read\_csv(input\_file, # Nome do arquivo com dados

names = names, # Nome das colunas

usecols = features, # Define as colunas que serão utilizadas

na\_values='?') # Define que ? será considerado valores ausentes

df\_original = df.copy()

# Imprime as 20 primeiras linhas do arquivo

print("PRIMEIRAS 20 LINHAS\n")

print(df.head(20))

print("\n")

# Imprime informações sobre dos dados

print("INFORMAÇÕES GERAIS DOS DADOS\n")

print(df.info())

print("\n")

# Imprime uma analise descritiva sobre dos dados

print("DESCRIÇÃO DOS DADOS\n")

print(df.describe())

print("\n")

# Imprime a quantidade de valores faltantes por coluna

print("VALORES FALTANTES\n")

print(df.isnull().sum())

print("\n")

columns\_missing\_value = df.columns[df.isnull().any()]

print(columns\_missing\_value)

method = 'mode' # number or median or mean or mode

for c in columns\_missing\_value:

UpdateMissingValues(df, c)

print(df.describe())

print("\n")

print(df.head(15))

print(df\_original.head(15))

print("\n")

# Salva arquivo com o tratamento para dados faltantes

df.to\_csv(output\_file, header=False, index=False)

def UpdateMissingValues(df, column, method="mode", number=0):

if method == 'number':

# Substituindo valores ausentes por um número

df[column].fillna(number, inplace=True)

elif method == 'median':

# Substituindo valores ausentes pela mediana

median = df['Density'].median()

df[column].fillna(median, inplace=True)

elif method == 'mean':

# Substituindo valores ausentes pela média

mean = df[column].mean()

df[column].fillna(mean, inplace=True)

elif method == 'mode':

# Substituindo valores ausentes pela moda

mode = df[column].mode()[0]

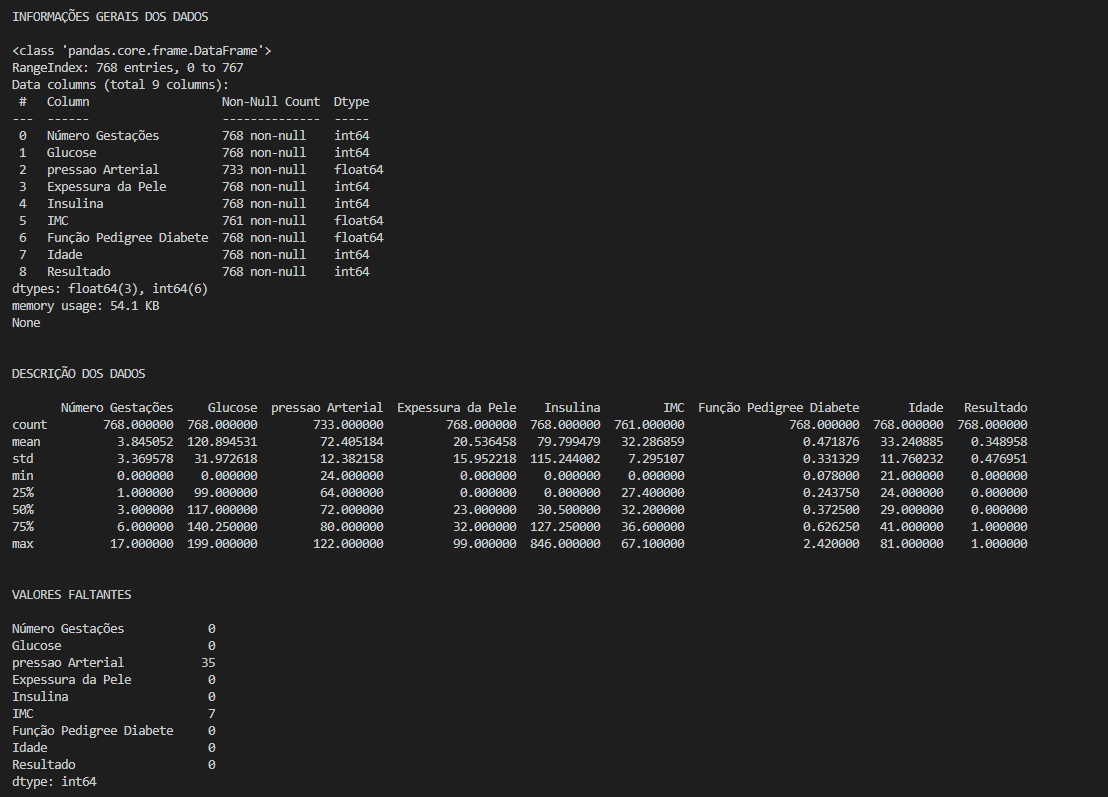
df[column].fillna(mode, inplace=True)

if \_\_name\_\_ == "\_\_main\_\_":

main()

Ao adaptar esse algoritmo a nossa base de dados obtivemos as seguintes saídas.

Figura 2: saída do DataCleaning do projeto.



fonte: vscode, 2023

# 3 Normalização dos Dados

# Referências Bibliográficas

CHAUHAN, Aman. Predict Diabetes. Kaggle.com. Disponível em: <https://www.kaggle.com/datasets/whenamancodes/predict-diabities>. Acesso em: 22 fev. 2023.