**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**

**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**PÓS-GRADUAÇÃO LATO SENSU EM CIÊNCIA DE DADOS E BIG DATA**

**MONICA TOLEDO DA COSTA SANTOS**

**Título: Uso de algoritmo de Machine Learning**

**para a classificação de vinhos**

**BELO HORIZONTE**

**ANO 2022**

**MONICA TOLEDO DA COSTA SANTOS**

**Título: Uso de algoritmo de Machine Learning**

**para a classificação de vinhos**

**Trabalho de Conclusão de Curso apresentando**

**ao Curso de Especialização em Ciência de**

**Dados e Big Data como requisito parcial à**

**obtenção do título de especialista**

**BELO HORIZONTE**

**ANO 2022**

**SUMÁRIO**

1. **Introdução..........................................................................................4**
   1. **Contextualização...............................................................................4**

**1.2 O problema proposto....................................................................... 4**

**2. Coleta dos dados..................................................................................5**

**3. Processamento/Tratamento de Dados...............................................6**

**4. Análise e exploração dos Dados.........................................................7**

**5. Criação de Modelos de Marchine Learning.....................................8**

**6. Apresentação dos Resultados..............................................................9**

**7. Links....................................................................................................10**

1. **INTRODUÇÃO**

**1.1. CONTEXTUALIZAÇÃO**

O vinho é uma bebida alcoólica feita pela fermentação do sumo da uva. Para especialistas, a bebida surgiu por acaso, talvez por um pouco de uvas amassadas e esquecidas em um recipiente que, após um tempo, sofreram o efeito da fermentação.

Basicamente, existem dois tipos de vinhos: branco e tinto. A diferenciação está nas frutas escolhidas para obter a bebida e no processo utilizado. Na fabricação do vinho tinto, são utilizadas uvas pretas ou tintas, e é necessário o contato do sumo com a casca da uva para atribuir cor e sabor à bebida. Já na fabricação do vinho branco, são utilizadas uvas brancas ou tintas, só que, nesse caso, a casca da uva não pode entrar em contato com a bebida.

Há inúmeros métodos para melhorar o aroma de vinhos, como maceração a frio, utilização de enzimas, processo de congelamento pré-fermentativo, utilização de inóculos (Substância com uma concentração de microrganismos para auxiliar o início da fermentação na preparação de alimentos ou bebidas fermentadas), diferentes e alguns casos inóculos mistos com o intuito de trazer características únicas na produção de vinho.

Obter um aroma e cor equilibrado e que vá de encontro ao paladar e aos olhos do consumidor ao final do processo de produção do vinho é um dos pontos chaves envolvidas no decorrer do processo.

**1.2. O PROBLEMA PROPOSTO**

O problema estudado, consiste na análise exploratória dos dados do dataset contendo 6.497 linhas **(amostras)** e 13 colunas **(características)** de vinhos classificados em 2 categorias: **tinto e branco.**

Para isso,suas variáveis categóricas nominais **tinto(‘Red’) e branco('White'),** serão substitúidas por valores binários **0 e 1.**

Será separado o conjunto de dados em duas partes, uma para **treino** e outra para **teste** e criar um algoritmo de **machine learning** para aprender a **classificar** vinhos com base nos dados de treino, e depois iremos verificar a performance com os dados de teste e apresentar o resultado.

1. **COLETA DE DADOS**

Para a coleta dos dados, foi usado a biblioteca **pandas** e para a codificação foi usado o **jupter notebook** e biblioteca do **python**.

**Bibliotecas usadas**

**• import pandas as pd**

**• import matplotlib.pyplot as plt**

**• from dash import Dash, html, dcc**

**• import plotly.express as px**

**• from mpl\_toolkits import mplot3d**

**• import scipy as sp**

**• import seaborn as sns**

**• import numpy as np**

**• %matplotlib inline**

**• plt.style.use("ggplot")**

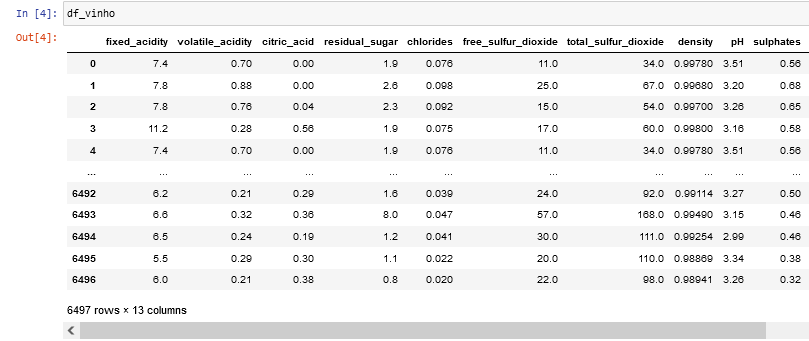
**DATASET**

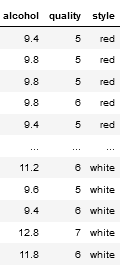
Os dataset é em formato **CSV**. Ele contém 6.497 linhas **(amostras)** e 13 colunas **(características)**.

Os dados foram obtidos na data 27/07/2022, no site**: https://www.kaggle.com/**

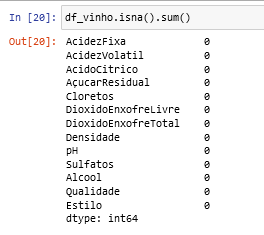
**3. PROCESSAMENTO/TRATAMENTO DE DADOS**

**Dataframe antes do tratamento dos dados**

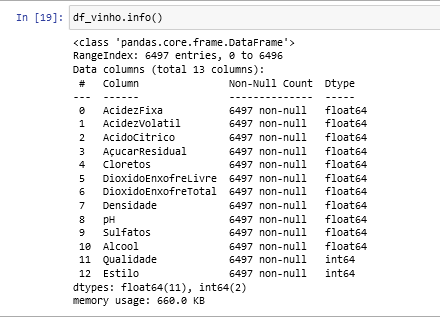




Para o tratamento dos dados, foram Analisados valores **missing**, verificando todas as linhas e colunas do dataframe em busca de valores faltantes ou nulos, onde não foram encontrado nenhum dado ausente. Foi usada a função isna().sum(), para detectar **valores ausentes** onde **'0'** significa que o dataframe não possui nenhum dado ausente.



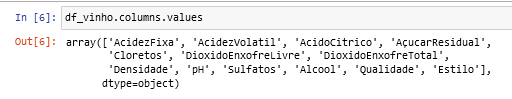
Foi usada a função info(), para saber informações e os tipos de variáveis do Dataframe. Possui 11 variáveis do tipo **‘float’** e 2 variáveis do tipo **‘int’**.



Foi usada a função rename(columns) para renomear o nome de cada coluna.

****

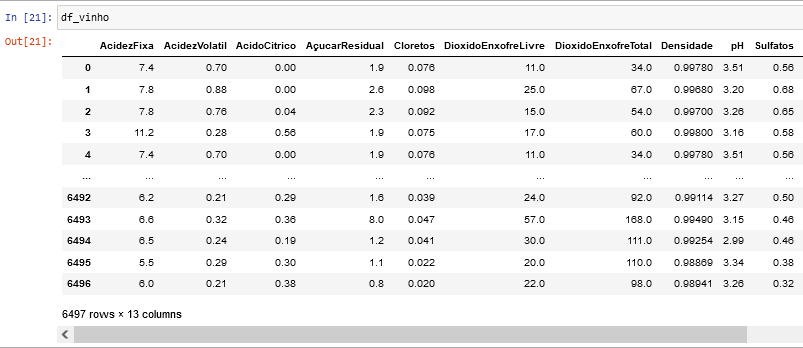
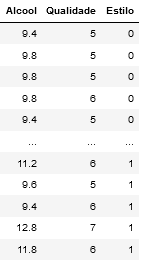
**Array** com nomes das colunas renomeadas, foram usadas todas e nenhuma foi descartada.

****

Foi feito o tratamento de uma coluna específica 'Estilo', usando a função replace() para substituir suas variáveis categóricas nominais ‘Red’ e 'White', por valores binários 0 e 1.

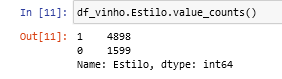
****

**Dataframe depois do tratamento dos dados**

** **

**4. ANÁLISE E EXPLORAÇÃO DOS DADOS**

Foi usada a função value\_counts(), para a contagem de amostras de cada valor binário da coluna **'Estilo'**.

****

Foi usada a função describe(), para informar de forma simples e rápida, um conjunto de estatísticas descritivas das variáveis:

**•**A quantidade de valores**(count)**

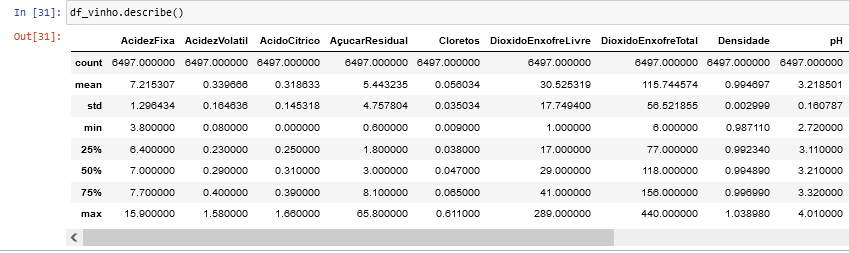
**•** A média**(mean)**

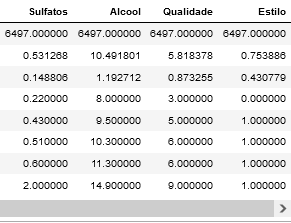
**•** O desvio padrão**(std)**

**•** O valor mínimo**(min)**

**•** Os quartis da distribuição**(25%, 50%, 75%)**

**•** O valor máximo**(max)**

****

****

Foi feito um algoritmo para calcular a porcentagem e a distribuição de cada elemento( valor binário 0 e 1) da coluna específica ‘Estilo’.



**Passo a passo para o cálculo:**

**Soma** os dois valores **4898 + 1599 =6497(total)**

Em seguida pega cada valor e **divide** pelo **6497 (total)**

Logo após, **multiplica** o resultado por **100** e terá o (**resultado final)**.

**‘white’, 1 4898 / 6497= 0,7538864091118978 \* 100 = (75,38864091118978)**

**‘red’, 0 1599 / 6497=0,2461135908881022 \* 100 = (24,61135908881022)**

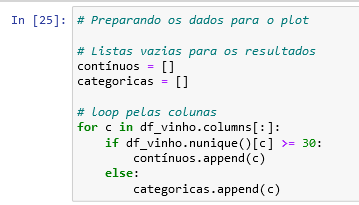
**(resultado final) (resultado final)**

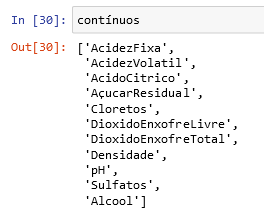
De acordo com o (**resultado final)**, ele nos mostra que a **categoria vinho branco** (‘white’) ou seja, **o valor binário** (1), possui a maior porcentagem.

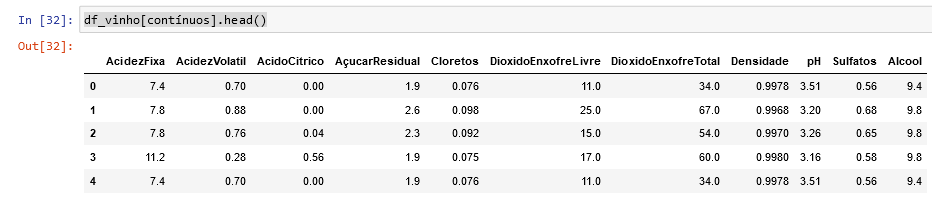
**Gráfico de pizza com a porcentagem**



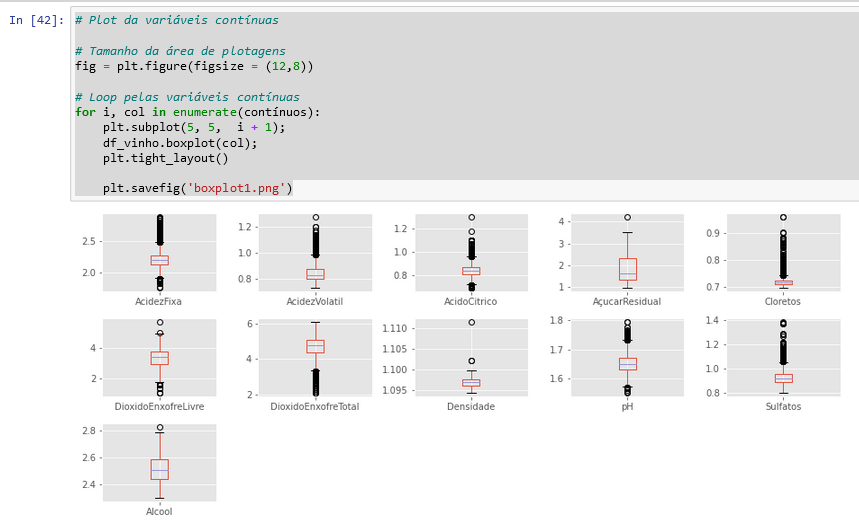
**Separando as variáveis em Contínuos e Categóricas**

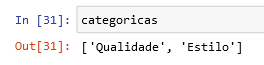


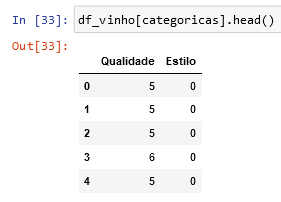




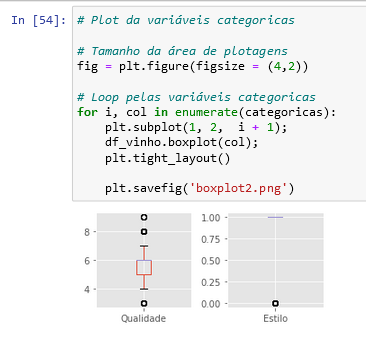
**Gráficos com os Quartis das variáveis contínuos**





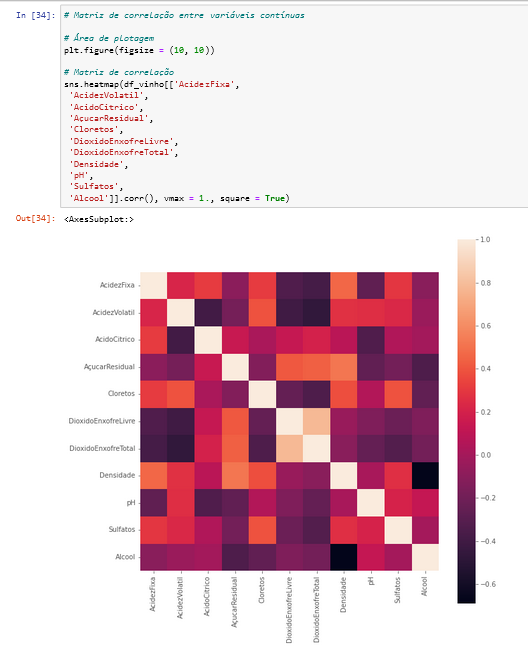


**Gráficos com os Quartis das variáveis categoricas**

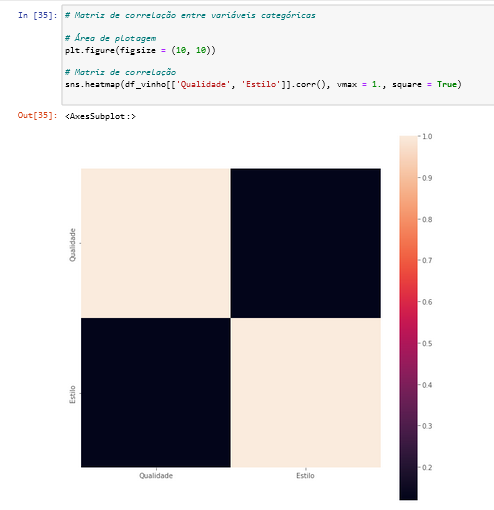


**Matriz de Correlação**

**A matriz de correlação** mostra os valores de correlação de **Pearson**, que medem o grau de relação linear entre cada par de itens ou variáveis. Os valores de correlação podem cair entre -1 e +1. Entretanto, na prática, os itens geralmente têm correlações positivas.

**Matriz de correlação entre variáveis contínuas** 

**Matriz de correlação entre variáveis categoricas**



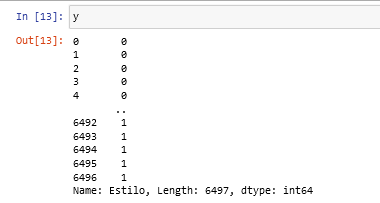
.

**5. CRIAÇÃO DE MODELOS DE MACHINE LEARNING**

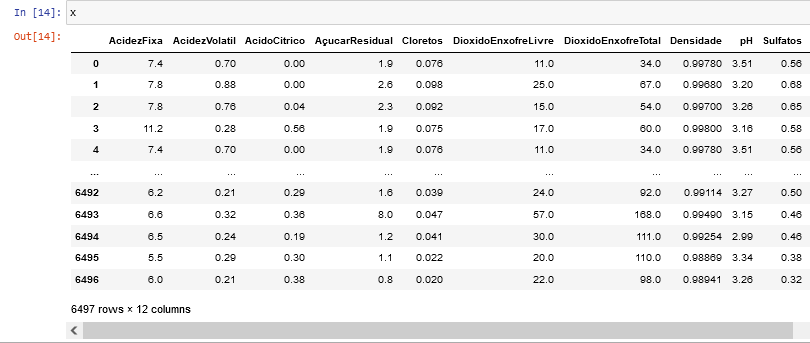
**Separando as variáveis em 'Preditoras' e 'Alvo'**

****

**Onde 'y' é a variável 'Alvo'**

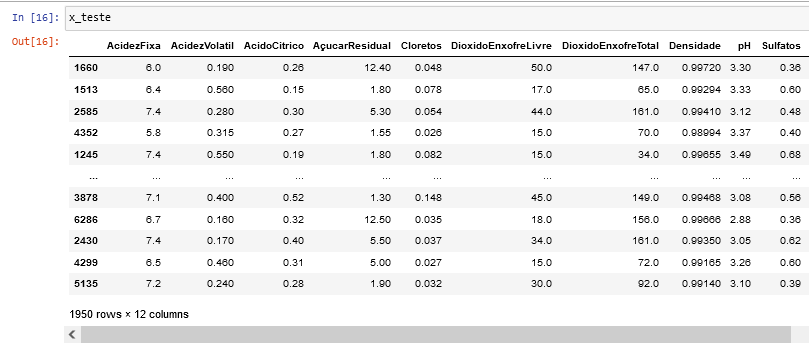
****

**Onde 'x' é a variável 'Preditora'**



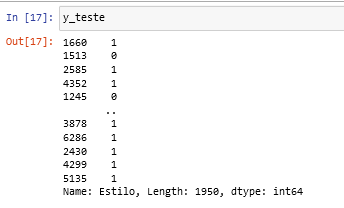
****

Sobre as variáveis **'Preditoras'**, elas são para **prever** algo.





Sobre a variável **'Alvo'**, ela é usada para estimar valores discretos como(valores binários 0 e 1, sim e não, verdadeiro e falso e etc).

****

**Dados de Treino e Teste**

**Dados de Treino:** São os dados que serão apresentados ao algoritmo de machine learning para criação do modelo. Estes dados costumam representar cerca de 70% da totalidade dos dados.

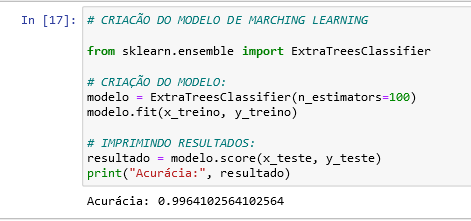
**Dados de Teste:** Serão apresentados ao modelo após a sua criação, simulando previsões reais que o modelo realizará, permitindo assim que o desempenho real seja verificado. Estes dados costumam representar cerca de 30% da totalidade dos dados.

O **scikit**-**learn** é uma biblioteca da linguagem **Python** desenvolvida especificamente para aplicação prática de machine learning.

**Criando os conjuntos de dados de 'Treino' e 'Teste'**

****

**ExtraTreesClassifier (Modelo de Classificação)**



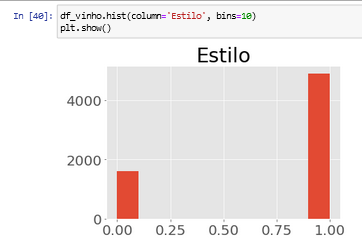
**6. APRESENTAÇÃO DOS RESULTADOS**

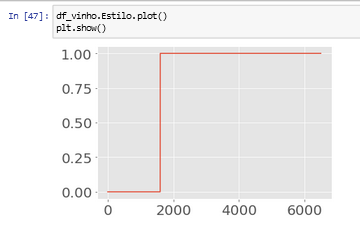
Foi feito um algoritmo para calcular a porcentagem de cada elemento( valor binário 0 e 1), onde o resultado mostra que o valor binário 1 da coluna em análise, obteve a maior porcentagem de 75.4% .

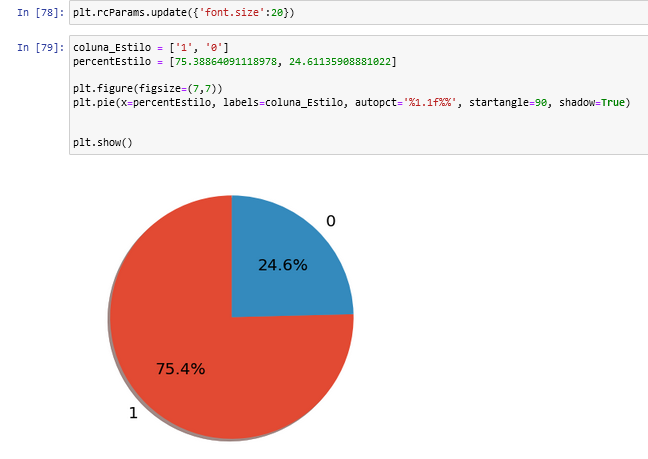
Em aplicação de algoritmo de Marching Learning, foi criado um conjunto de dados de treino e teste. Em seguida foi criado um modelo de classificação, onde a sua **acurácia** obteve o resultado de 0.9964102564102564.

**Acuracidade** é a precisão e exatidão de dados e informações, quando há ausência de erros ou equívocos.

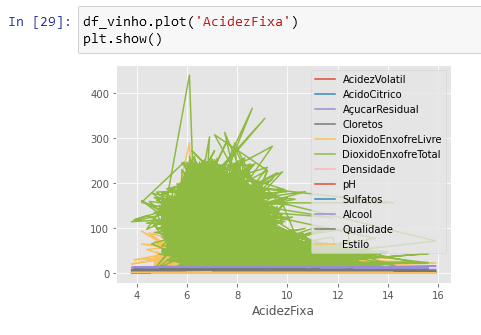
**Gráficos da coluna que está sendo analisada**

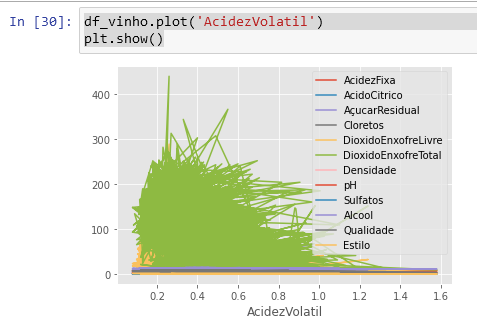


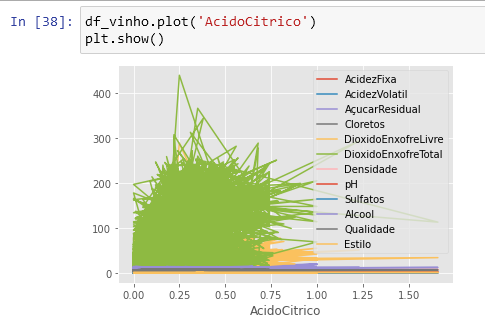


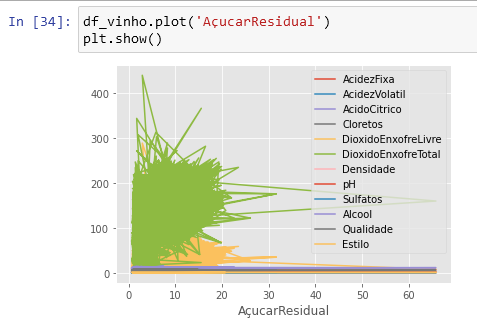


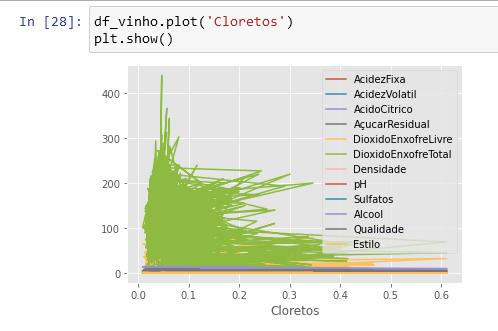
**Gráficos de cada coluna**

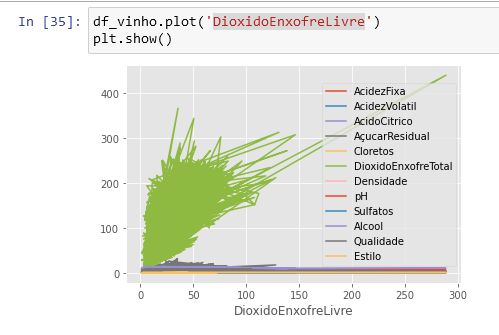


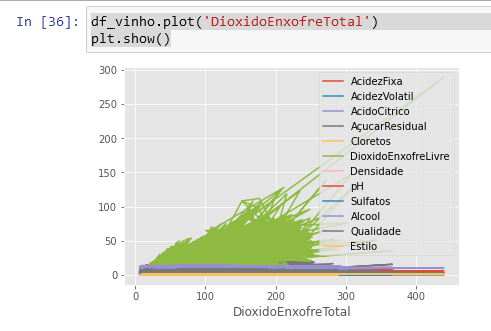


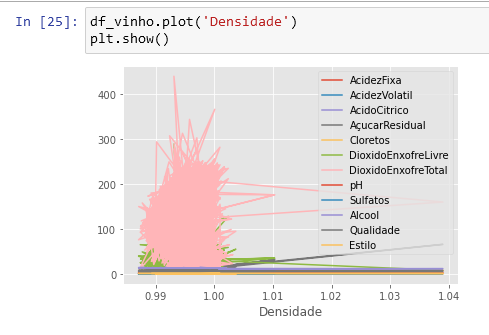


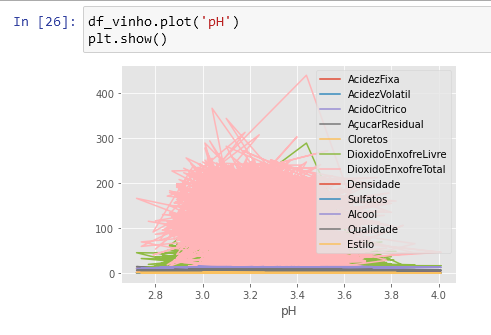


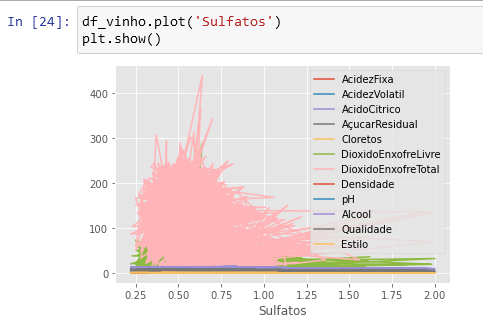


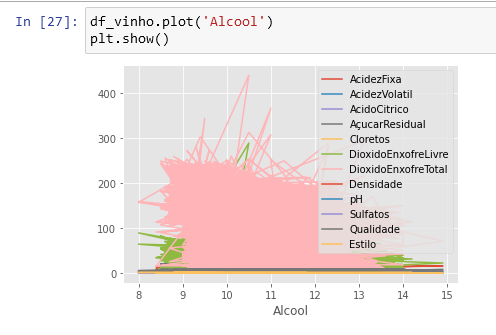


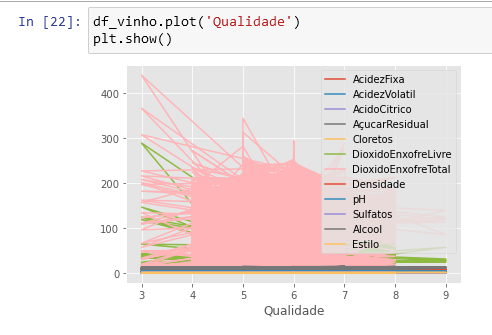


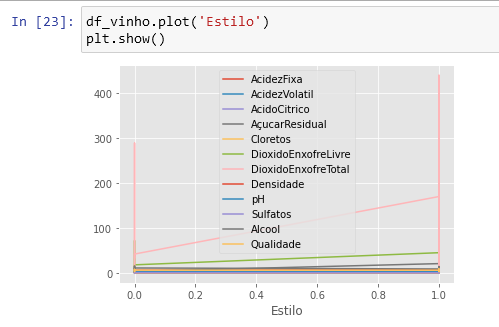




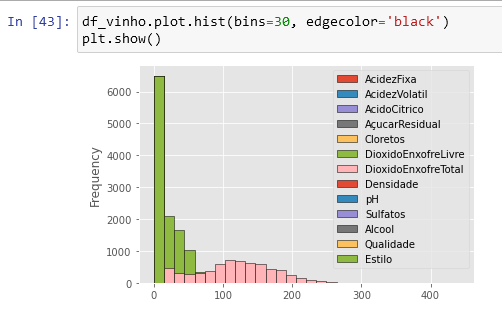


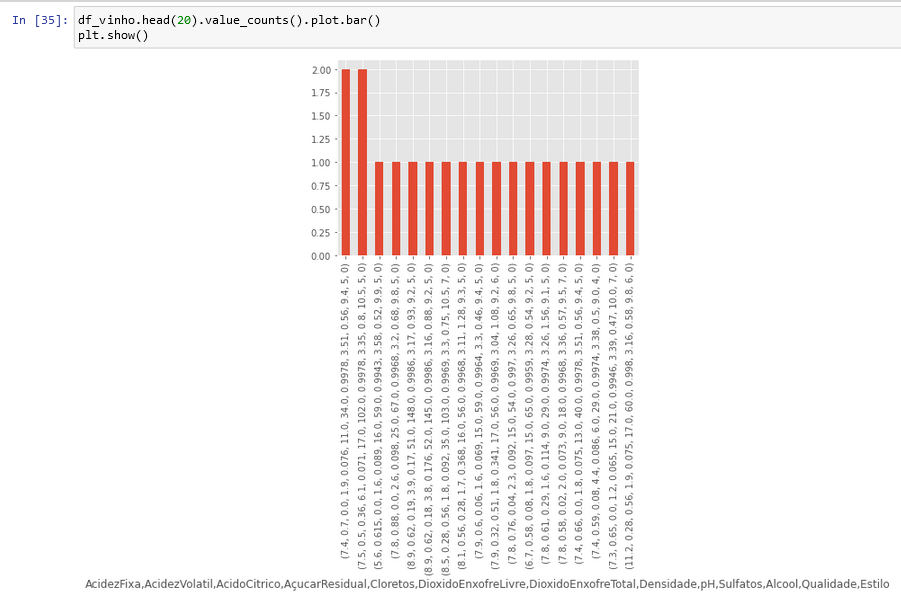






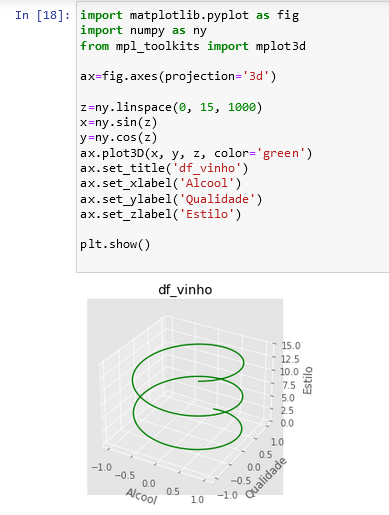
**Gráficos do Dataframe com todas as colunas**



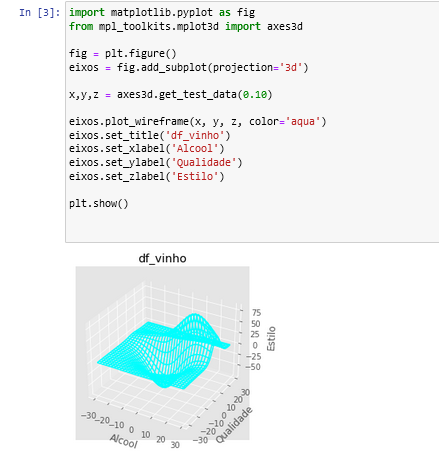


**Modelos de Gráficos 3D**

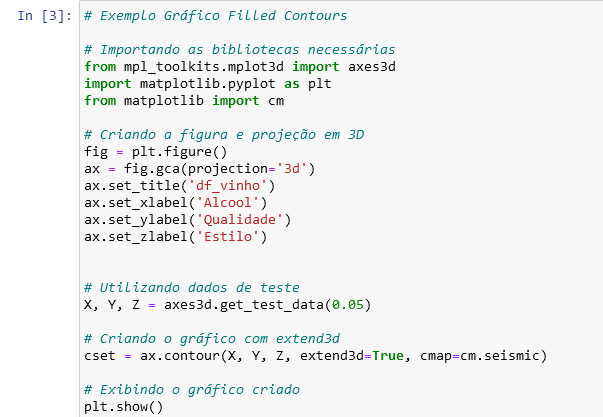
**Gráfico de linha de dispersão no espaço 3D**

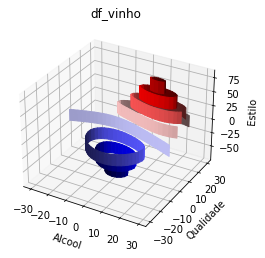


**Gráfico de Superfície**



**Gráfico de contornos preenchidos**





**7. LINKS**

[**https://brasilescola.uol.com.br/saude/vinho.htm**](https://brasilescola.uol.com.br/saude/vinho.htm)

[**https://www.divinho.com.br/blog/qual-a-diferenca-vinho-tinto-vinho-branco/**](https://www.divinho.com.br/blog/qual-a-diferenca-vinho-tinto-vinho-branco/)

[**https://pt.wikipedia.org/wiki/Vinho#O\_processo\_de\_produção\_e\_fermentação**](https://pt.wikipedia.org/wiki/Vinho#O_processo_de_produção_e_fermentação)

**https://www.kaggle.com/**