Rio, 08/07/2025

Relatório Machine Learning

Alunos:

Brenno Cezário de Oliveira Pereira

Paulo José Gomes de Almeida

1- Descrição do Problema

Os dados vêm por meio de uma planilha .csv, contendo dados de diversas amostras de vinhos, branco ou vermelho.

Por meio de técnicas de treinamento de dados, o objetivo é conseguir prever a Qualidade, que é nosso alvo (target), de um vinho.

2- Justificativa e Descrição da Ferramenta Escolhida

Escolhemos a linguagem de programação Python por termos certa familiaridade por conta de outros trabalhos, e foi utilizado as bibliotecas scikit-learn, PyTorch, NumPy, Pandas, Matplotlib, Seaborn e skorch.

Usos de cada biblioteca no projeto:

torch - deep learning, numpy - operações numéricas, pandas - ler csv, matplotlib e seaborn - criar gráficos, skorch - GridSearchCV.

Os parâmetros utilizados para projeto foram:

Divisão dos Dados: Treino (75%) e Teste (25%)

Nós da camada oculta: 512

Número de Épocas de treino: 150

Taxa de Aprendizagem: 0.001

Função de Perda: Entropia cruzada

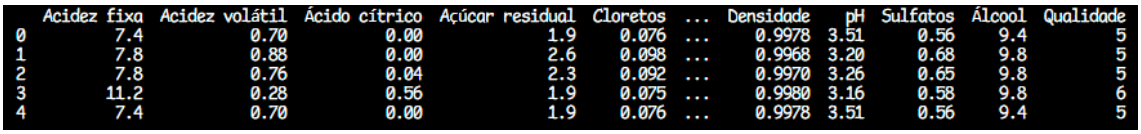
Otimizador: Adam

3- Descrição do Dataset Escolhido

São duas planilhas, um de vinho branco e um de vinho vermelho.

Fonte: Wine Quality UCI Machine Learning

Segue as 5 primeiras linhas da planilha de vinho vermelho:



	Acidez fixa	Acidez volátil	Ácido cítrico	Açúcar residual	Cloretos	...	Densidade	pH	Sulfatos	Álcool	Qualidade
0	7.4	0.70	0.00	1.9	0.076	...	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	...	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	...	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	...	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	...	0.9978	3.51	0.56	9.4	5

Acidez fixa Acidez volátil Ácido cítrico Açúcar residual Cloretos ... Densidade pH Sulfatos Álcool Qualidade

0 7.4 0.70 0.00 1.9 0.076 ... 0.9978 3.51 0.56 9.4 5

1 7.8 0.88 0.00 2.6 0.098 ... 0.9968 3.20 0.68 9.8 5

2 7.8 0.76 0.04 2.3 0.092 ... 0.9970 3.26 0.65 9.8 5

3 11.2 0.28 0.56 1.9 0.075 ... 0.9980 3.16 0.58 9.8 6

4 7.4 0.70 0.00 1.9 0.076 ... 0.9978 3.51 0.56 9.4 5

Legenda de qualidade do vinho:

3 e 4 = Ruim

5 e 6 = Médio

7 e 8 = Bom

9 e 10 = Excelente

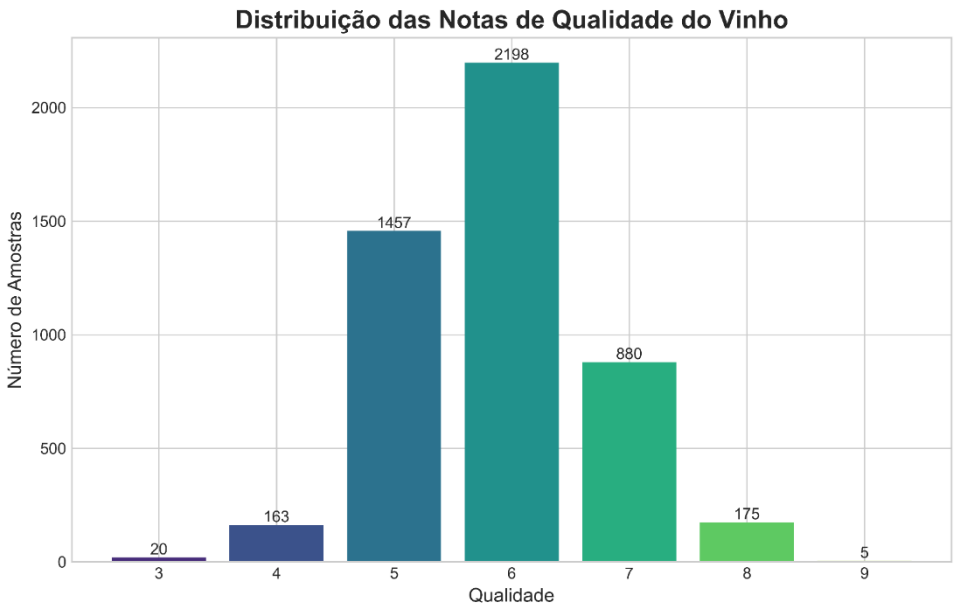
São 1600 amostras para cada tipo de vinho, com as seguintes características (colunas):

Acidez fixa, Acidez volátil, Ácido cítrico, Açúcar residual, Cloretos, Dióxido de enxofre livre, Dióxido de enxofre total, Densidade, pH, Sulfatos, Álcool e Qualidade.

É removido do eixo X o atributo de qualidade e colocado em Y. Qualidade é o nosso alvo, target, e o atributo alvo fica em Y. Se deixássemos em X, por exemplo, os dados seriam treinados junto com a qualidade, meio que "trapaceando" nos testes.

As amostras não são balanceadas, então temos muito mais vinhos de qualidade normal do que excelentes ou ruins, dificultando o treino.

Segue diversos gráficos que descrevem o Dataset:



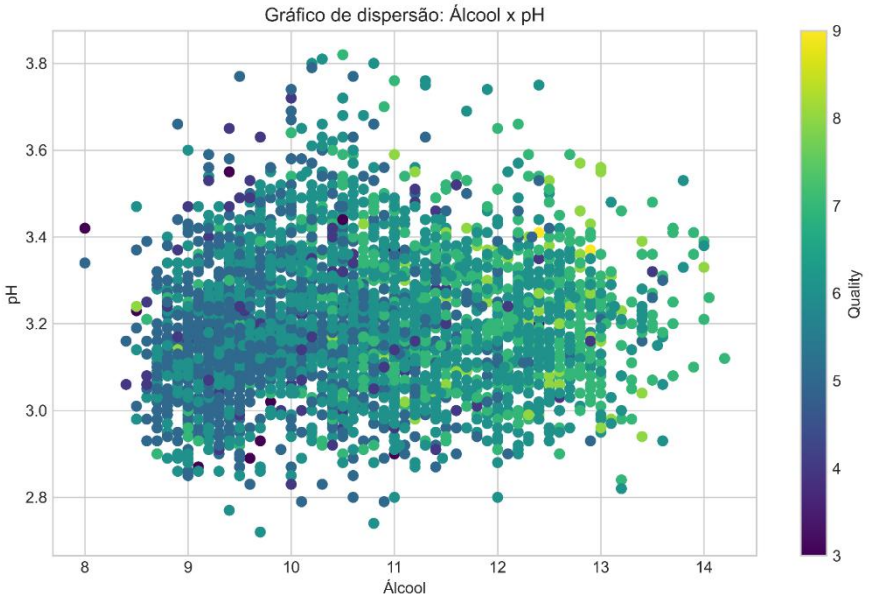
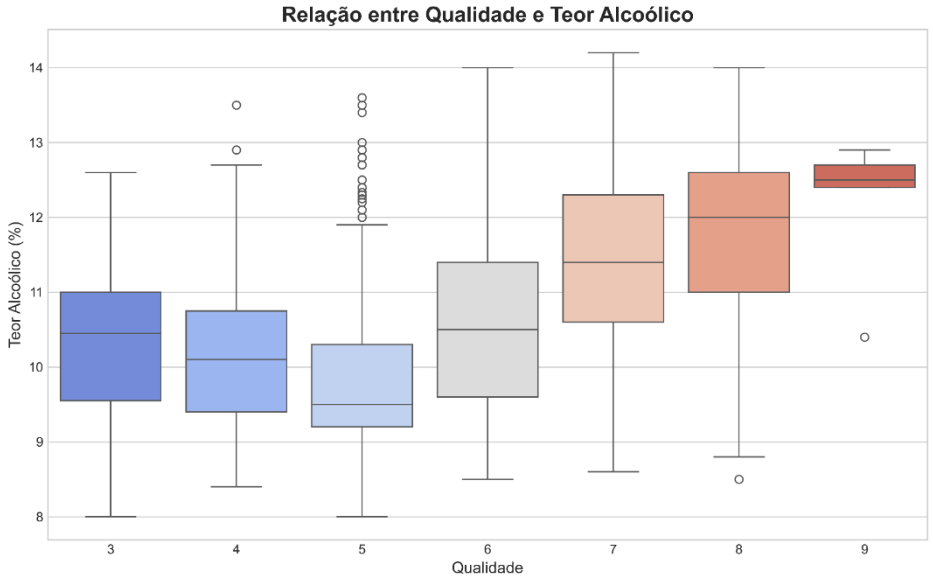
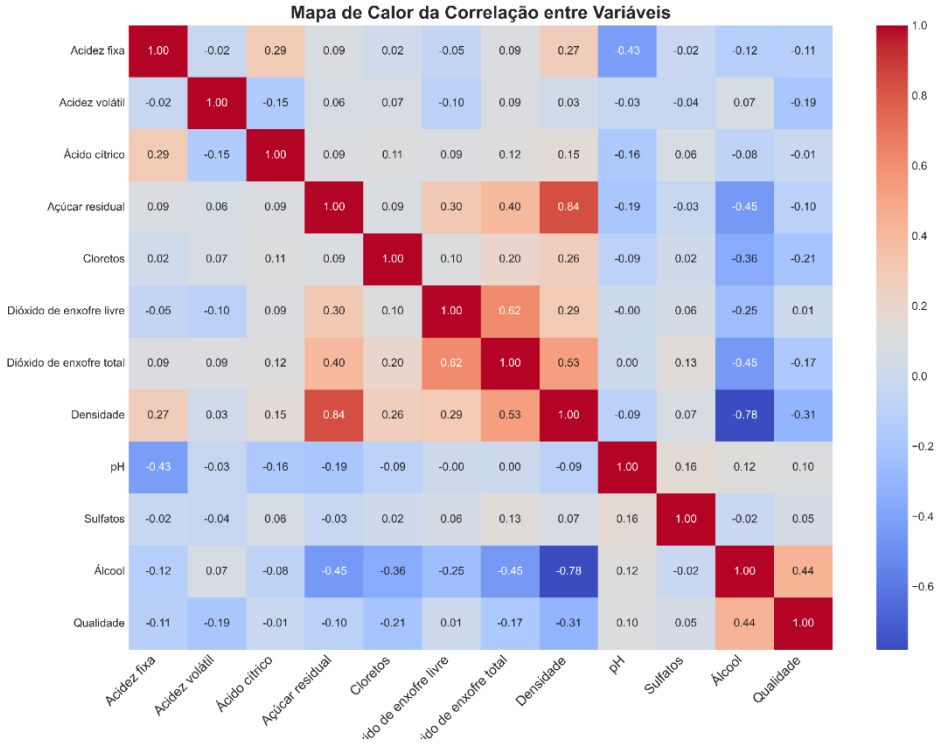


Gráfico de dispersão: Álcool x pH
pH
Álcool
Quality
3.8
3.6
3.4
3.2
3.0
2.8
8
9
10
11
12
13
14
9
8
7
6
5
4
3


Relação entre Qualidade e Teor Alcoólico
Teor Alcoólico (%)
Qualidade
14
13
12
11
10
9
8
3
4
5
6
7
8
9



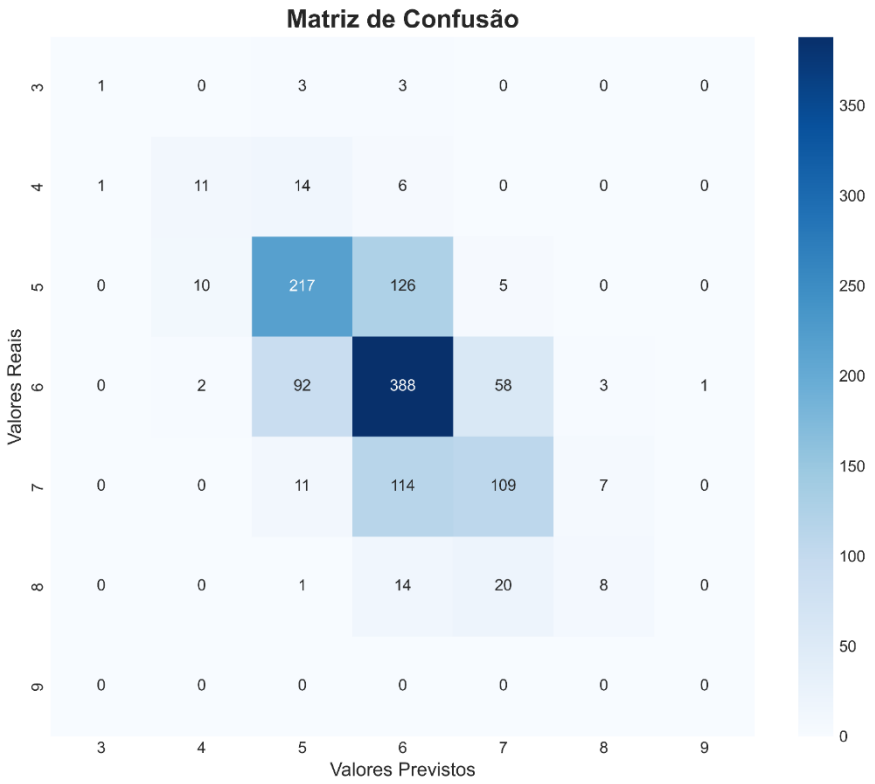
4- Apresentação dos resultados

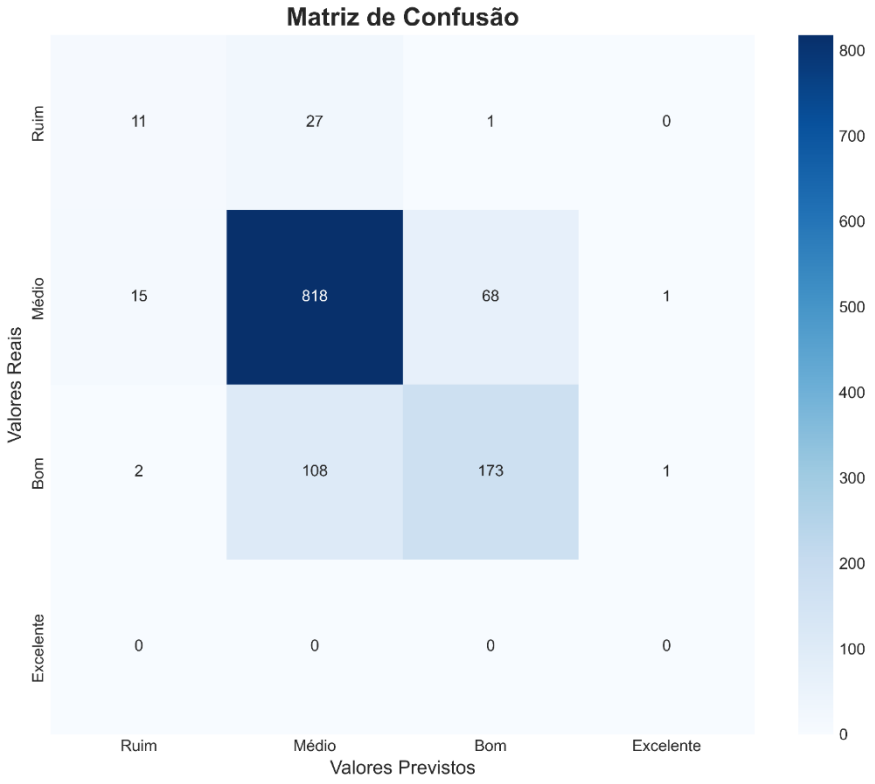
Sem Agrupamento

```
Acurácia no teste: 59.59%
```

Com Agrupamento

```
Acurácia no teste: 82.29%
```


Matriz de Confusão
Valores Reais
Valores Previstos
3 4 5 6 7 8 9
1 0 3 3 0 0 0
1 11 14 6 0 0 0
0 10 217 126 5 0 0
0 2 92 388 58 3 1
0 0 11 114 109 7 0
0 0 1 14 20 8 0
0 0 0 0 0 0 0
350 300 250 200 150 100 50 0


Matriz de Confusão
Valores Reais
Valores Previstos
Ruim Médio Bom Excelente
11 27 1 0
15 818 68 1
2 108 173 1
0 0 0 0
800 700 600 500 400 300 200 100 0

5- Discussão dos resultados

Ao terminar de desenvolver o projeto, verificamos uma baixa acurácia e um dos fatores se deve à grande quantidade de amostras desbalanceadas.

Utilizando a técnica de agrupamento (clustering) conseguimos aumentar consideravelmente a acurácia, indo de ≈60% à ≈80%.

Testando o modelo com um vinho de teste com os atributos, o observamos como tendo qualidade 5 (Médio):

```
 AF  AV  AC  AR    Clo    DEL   DET   Den  pH  Sulf  Alc
6.3,0.45,0.1,1.2,0.03335,15.5,21.0,0.9946,3.39,0.47,10.0
```

AF AV AC AR Clo DEL DET Den pH Sulf Alc

6.3,0.45,0.1,1.2,0.03335,15.5,21.0,0.9946,3.39,0.47,10.0

Legenda siglas:

AF - Acidez fixa

AV - Acidez volátil

AC - Ácido cítrico

AR - Açúcar residual

Clo - Cloretos

DEL - Dióxido de enxofre livre

DET - Dióxido de enxofre total

Den - Densidade

Sulf - Sulfatos

Alc - Álcool