**Faculdade de Informática e Administração Paulista**

Engenharia de Software

Vitor da Silva Sant’ana RM99586

Lucca Alexandre Machado RM99700

Victor Augusto Wittner RM 98667

CHECKPOINT 3

Data Science and Statistical Computing

ESPW

Bela Vista

2024

1. INTRODUÇÃO

O dataset aborda o tema “Qualidade de vinho”, abordando os variantes tintos e brancos da marca Vinho Verde, de Portugal.

Os dados apresentam informações técnicas sobre a os resultados gerados por conta da composição dos vinhos, como a acidez cítrica, a acidez volátil, densidade, pH, sulfatos e outros pontos.

Sabe-se que estes fatores alteram o sabor dos vinhos, e, para bons apreciadores, isso é facilmente perceptível, necessitando de estudo sobre o assunto para que sejam desenvolvidos com maior eficiência.

1. ESCOPO

Esse projeto explora o conjunto de dados de vinhos, analisando correlações entre suas características químicas, para identificar padrões que possam impactar sua qualidade.

**Colunas do Dataset**

1. **fixed acidity**: Ácido não volátil no vinho.
2. **volatile acidity**: Ácido acético.
3. **citric acid**: Acidez cítrica.
4. **residual sugar**: Açúcar após fermentação.
5. **chlorides**: Nível de sal no vinho.
6. **free sulfur dioxide**: Proteção contra oxidação.
7. **total sulfur dioxide**: Total de SO₂ presente.
8. **density**: Densidade do vinho.
9. **pH**: Acidez medida em escala.
10. **sulphates**: Concentração de sulfatos.
11. **alcohol**: Percentual de álcool.
12. **quality**: Classificação do vinho (0-10).
13. OBJETIVO

O objetivo do projeto é realizar uma análise no *dataset*, observando padrões e buscando conclusões sobre as correlações entre as colunas e as informações dos vinhos, respondendo as seguintes perguntas:

1. **Qual é a relação entre a acidez fixa e o pH dos vinhos?**

-0.0683

Correlação baixa!

1. **Existe alguma correlação significativa entre a quantidade de acidez volátil e o total de dióxido de enxofre?**

0.0765

Correlação baixa! Existe, porém é tão baixa que se torna insignificante, aconteceu poucas vezes e por acaso.

1. **Os níveis de açúcar residual impactam o nível de cloretos nos vinhos?**

Não diretamente, pois, após análises, foi constatado que as duas variáveis não se influenciam. Entretanto, os sulfatos têm a função de adicionar sódio ao vinho, o que pode parecer palatavelmente maior caso o nível de açúcar seja menor, porém, quimicamente falando, não, as variáveis não se influenciam

1. **Os vinhos que têm a maior concentração de dióxido de enxofre livre têm o pH afetado?**

Sim, não em quantidades grandes, pois o sulfato não tem correlação alta com o pH, entretando, o sulfato é quimicamente um influenciador do pH, deixando o produto mais ácido

1. **Qual é o intervalo de densidade mais comum nos vinhos?**

0.9972, 0.9968 e 0.9976

1. **Existe alguma tendência entre a quantidade de sulfatos e a quantidade de dióxido de enxofre?**

0.0429

Existe, porém é tão baixa que se torna insignificante

1. **A acidez fixa afeta o nível de açúcar residual em diferentes faixas de vinhos?**

Não, além da correlação ser baixa, o açúcar não influencia na acidez fixa do produto, mas sim o ácido cítrico e o pH

1. **Quais características têm maior influência na densidade dos vinhos?**

Quase todos os químicos adicionados, mas, principalmente e em maioria, o álcool

1. **Qual é o impacto da adição de sulfatos sobre a acidez volátil e a acidez fixa?**

Imediatamente, nenhuma. Entretanto é necessário atentar-se com os outros aditivos para observar se não ocorre nenhuma outra alteração por conta de reações químicas não planejadas

1. **Quantos vinhos têm o teor alcoólico maior que 13.5?**

13 vinhos

1. ANÁLISE EXPLORATÓRIA

O dataset foi analisado, buscando informações que anulem a qualidade dos dados, afim de limpar e tratá-los para que os cálculos sejam feitos com dados bons.

De início, foi verificado se existia algum valor nulo no *dataset*, usando a função ***isnull().any(),*** e o resultado foi de nenhum valor nulo no *dataset*.

Logo após isso, foram gerados gráficos de violinos, que são boxplots mais aprofundados, pois apresentam a densidade dos dados também, para todas as colunas do *dataset,* e alguns gráficos chamaram atenção para valores altos, sendo eles:

Gráfico, Histograma

Descrição gerada automaticamente Gráfico, Histograma

Descrição gerada automaticamente

Gráfico

Descrição gerada automaticamente Gráfico

Descrição gerada automaticamente

Esses gráficos geram dúvida pois o valor máximo encontrado é bem mais alto do que a densidade, que é onde a maioria dos dados se encontra, porém há alguns assuntos a retratar:

* **A escala de representação do gráfico**: a parte visual difere, mas os valores numéricos apenas estão bem distribuídos no eixo y, como no caso de *chlorides*, que altera em apenas 0.45, aproximadamente, o mesmo acontece em *sulphates* e *total sulfur dioxide*, onde a diferença comparada à distribuição dos dados não é tão diferente.
* **Quantidade**: quanto à coluna *residual sugar*, o valor que poderia ser exorbitante é acompanhado de outros valores quase únicos, que não se repetem muito, entre 10 e 15.5, por isso não se enquadram na densidade, mas nada de *outliers,* é apenas uma concentração de açúcar maior. Informação obtida por ***data[data['residual sugar'] > 10]***
* Quanto aos outros gráficos, não foram apresentados dados extremos duvidosos ou algo do tipo.

Para finalizar a busca por outliers ou valores estranhos, foi feita uma comparação entre a

média e a mediana das colunas, com o seguinte código:

***for i in data.columns:***

***print(f"Coluna {i} -- média: {data[i].mean()} -- mediana: {data[i].median()}")***

Nenhum valor de média obtido foi grandemente diferente da média, o que, além de

comprovar que não há *outliers* de fato, mostra que a distribuição dos dados está apresentando

uniformidade.

Foi usada agora a função ***sns.kdeplot()*** para plotar histogramas de linhas contínuas,

função da biblioteca Seaborn e, colocando-o na mesma linha, comparamos com o ***violinplot,*** para

verificar se os dados condizem nos dois gráficos.

1. ESTATÍSTICA

Ao analisar a correlação do *dataset,* algumas colunas chamaram atenção por correlação positiva relativamente alta, sendo elas:

* *total sulfur dioxide x free sulfur dioxide* (68%)
* *density x fixed acidity* (67%)
* *citric acid x fixed acidity* (67%)
* *alcohol x quality* (48%)

Plotagem dos gráficos das colunas referidas:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Uma imagem contendo Gráfico

Descrição gerada automaticamente

* **O gráfico 1 (vermelho)** mostra uma grande interação entre os dados, pois está com a cor mais concentrada, principalmente na parte inferior dos dados, onde tanto o dióxido livre quanto o total estão em quantidades menores.
* **O gráfico 2 (verde)** mostra grande concentração na parte central e muita uniformização dos dados, mostrando correspondência entre densidade e acidez fixa.
* **O gráfico 3 (marrom)** indica grande concentração e bastante uniformidade, mostrando que os dados estão distribuidos quase total e uniformemente, em vários aspectos do gráfico, ou seja, o ácido cítrico implica na acidez fixa claramente.
* **O gráfico 4 (azul)** é de compreensão um pouco mais complexa, pois ele agrupa valores que não são livremente expressos, pois se tratam de escalas, tanto de nível de álcool, quando de avaliação de qualidade, são números com apenas uma casa decimal após a vírgula no máximo, ainda assim a coluna de qualidade apresenta apenas valores inteiros.

Com isso, pode-se concluir que, quando a quantidade de álcool está entre 9 e 13, a qualidade é majoritamente mediana, o que mostra a inflência do álcool na avaliação do vinho.

Algumas tiveram correlação negativa relativamente alta, sendo elas:

* *fixed acidity x pH* (68%)
* *citric acid x pH* (54%)
* *alcohol x density (*49%)
* *volatile acidity x citric acid* (55%)

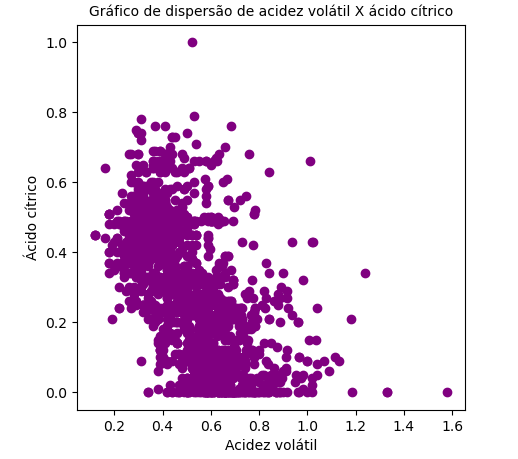
Plotagem dos gráficos das colunas referidas:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente 

Estes gráficos se relacionam inversamente, ou seja, quando uma variável aumenta, a outra diminui

* **O gráfico 1 (azul)** mostra grande interligação dos dados, especialmente na parte esquerdo-central do gráfico, que mostra que o pH influencia na acidez fixa, principalmente entre 3.0 e 3.6 de potencial de hidrogênio.
* **O gráfico 2 (laranja)** mostra grande concentração na parte esquerdo-central, mostrando correspondência entre o ácido cítrico e o pH.
* **O gráfico 3 (vermelho)** indica grande concentração na parte central do gráfico, mas não muita uniformização, então o álcool infere na densidade, porém ainda assim há casos em que ele pode não ser o agente principal para alteração da mesma.
* **O gráfico 4 (roxo)** mostra grande concentração majoritamente na parte inferior esquerda do gráfico, o que mostra grande inferência do ácido cítrico na acidez volátil do produto.

Isso implica que, ao comparar os dados, muito provavelmente, o ácido cítrico, a quantidade de álcool, e os dióxidos de enxofre são agentes principais para a alteração de indicadores como a densidade, o pH, as acidezes e a qualidade.

Para confirmar as hipóteses apresentadas acima, foi criado um modelo de testes de hipótese, feito como uma função, que recebe dois parâmetros (*coluna1* e *coluna2*), e realiza o teste de hipótese entre as duas.

O teste de hipótese funciona comparando duas variáveis, e, primeiramente assume a hipótese nula como certa até que se seja provado o contrário. A hipótese nula afirma que não há relação entre as duas variáveis, e a rejeição é feita quando o valor “p”, obtido pela função ***stats.chi2\_contingency(tabela\_de\_contigencia\_das\_colunas).***

O que significa cada valor obtido pela função?

* **Valor p**: É a chance de que os resultados que foram apresentados acontecessem por acaso. Se for baixo, significa que o resultado não aconteceu por sorte, foi estatisticamente influenciado.
* **Graus de liberdade**: É o número de opções que temos para fazer os cálculos.
* **Chi²:** Serve para ver se as contagens que foram observadas são diferentes do que era esperado.
* **Frequências esperadas**: São os números que deveriam aparecer se tudo estivesse normal.

Com isso, pode-se obter o resultado esperado – todas as hipóteses nulas foram rejeitadas, ou seja, a relação entre todas elas é real e foi confirmada pelo teste de hipótese.

1. MACHINE LEARNING

Conforme realizados os processos de análise de correlação dos dados e os testes de hipótese, foram realizados alguns comparativos com aprendizado de máquina.

Utilizando a biblioteca ***Scikit-learn,*** os dados foram treinados e testados, tanto em acurácia quanto em previsão deles, porém, antes disso, foi necessário criar os modelos desejados, sendo eles:

* **Regressão linear –** analisa a influência de uma variável na outra, minimizando a soma dos erros (valores reais – valores previstos);
* **Classificação (Regressão logística)** **–** prevê a probabilidade de um evento acontecer, com base na função logística;
* **Árvore de decisão –** modelo quedetermina as decisões com caminhos em formatos de árvores, cada nó representa uma pergunta e decisão com base nos dados;
* **Random Forest –** modelo que utiliza várias árvores de decisão aleatoriamente para treinar os dados.

Após a decisão dos modelos que serão utilizados, o *dataset* foi separado em colunas categóricas e colunas alvo (X e y), onde só então serão treinados.

A coluna alvo foi ***quality,*** enquanto todas as outras restantes foram definidas como colunas categóricas, pois são colunas contendo dados que alteram o resultado da qualidade de avaliação dos usuários.

Então, usando a função *train\_test\_split*, foram criadas as variáveis de treino e teste categóricas (***X\_train***, ***X\_test***), e as de treino e teste alvo (***y\_train***, ***y\_test***). As variáveis de treino serão usadas para treinar o modelo e então realizar os testes com as variáveis de teste.

Com os dados treinados com seu respectivo modelo, em sua respectiva função, foi gerada a predição (resultados previstos pelo modelo de *Machine Learning)* e a acurácia, métrica que avalia a qualidade do modelo em relação aos resultados gerados.

Todos os modelos apresentaram o *score* (acurácia) menor que 70%, sendo que o modelo que gerou o *score* mais alto foi o de ***Random Forest*** com 67.5% de acerto, com uma predição de qualidade de vinho média em 5.6.

Gráficos resultados do processo de *Machine Learning*:

Regressão Linear

Para a regressão linear não foi possível fazer um gráfico de matriz de confusão, pois o mesmo informa apenas sobre modelos de classificação, e não de valores contínuos como o da regressão linear.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

A linha azul retrata significa uma ocorrência normal dos dados, e, na regressão linear, os dados previstos são apresentados como os pontos em vermelho.

Pode-se afirmar que a concentração dos dados se encontra na parte medial da linha de comparação a números reais.

Considerando:

* Erro absoluto médio: 0.51
* Erro absoluto quadrático: 0.41

Classificação

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Calendário

Descrição gerada automaticamente

A linha azul retrata significa uma ocorrência normal dos dados, e, na regressão logística, os dados previstos são apresentados como os pontos em vermelho.

Pode-se afirmar que os dados se concentram ao redor da parte medial da linha, afirmando ainda que a previsão conseguiu acertar 3 pontos, onde em 2 há grande concentração de dados.

Considerando:

* Erro absoluto médio: 0.49
* Erro absoluto quadrático: 0.56

141: Modelo acertou predição positiva

54: Modelo errou predição negativa

79: Modelo errou predição positiva

119: Modelo acertou predição negativa

Random Forest

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Calendário

Descrição gerada automaticamente

A linha azul retrata significa uma ocorrência normal dos dados, e, na *Random Forest*, os dados previstos são apresentados como os pontos em vermelho.

Pode-se afirmar que os dados se concentram ao redor da parte medial da linha, afirmando ainda que a previsão conseguiu acertar 3 pontos, onde em todos há grande concentração de dados.

Considerando:

* Erro absoluto médio: 0.35
* Erro absoluto quadrático: 0.39

Gráfico com menor nível de erros!

150: Modelo acertou predição positiva

43: Modelo errou predição negativa

46: Modelo errou predição positiva

140: Modelo acertou predição negativa

Árvore de decisão

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Calendário, Gráfico de dispersão

Descrição gerada automaticamente

A linha azul retrata significa uma ocorrência normal dos dados, e, na árvore de decisão, os dados previstos são apresentados como os pontos em vermelho.

Pode-se afirmar que os dados se concentram ao redor da parte medial da linha, com grande distância, afirmando ainda que a previsão conseguiu acertar 5 pontos, onde em 3 deles há grande concentração de dados.

Considerando:

* Erro absoluto médio: 0.50
* Erro absoluto quadrático: 0.64

124: Modelo acertou predição positiva

56: Modelo errou predição negativa

51: Modelo errou predição positiva

116: Modelo acertou predição negativa

1. HALLWAY TESTING

**Sugestões de melhoria (criadas por pessoas de fora do trabalho):**

* Buscar as produtoras dos vinhos
* Aumentar, de alguma forma, a acurácia dos modelos
* Gerar insights mais incisivos, falando mais detalhadamente sobre os vinhos

1. CONCLUSÃO

Com a finalização do projeto, foi possível concluir que algumas variáveis do projeto se relacionam entre si, e mudam o valor uma da outra, a depender de sua influência. Com os testes de hipótese foi possível confirmar essa informação e anular todas as hipóteses nulas.

Com a técnica de *Machine Learning,* foi possível treinar o modelo e criar algumas previsões, que tiveram um pouco mais de 65% de acurácia, podendo gerar alguns insights positivos sobre os vinhos, sua qualidade e o que a influencia.