UNIVERSIDADE REGIONAL INTEGRADA DO ALTO URUGUAI E DAS MISSÕES

CIÊNCIA DA COMPUTAÇÃO

TÓPICOS ESPECIAIS EM COMPUTAÇÃO I

SOFTWARE PARA ANÁLISE E MODELAGEM DE DADOS PARA UM DATASET DE QUALIDADE DE VINHOS

Alexandre S. Castro

091881@aluno.uricer.edu.br

Resumo

Este trabalho apresenta uma análise e modelagem de dados utilizando o dataset de qualidade do vinho. O processo começa com o carregamento dos dados, seguido pela visualização da distribuição das classes de qualidade do vinho e pela análise de correlação entre as variáveis. Em seguida, um modelo de Support Vector Machine (SVM) é treinado para prever a qualidade do vinho com base em suas características químicas. A performance do modelo é avaliada por meio de métricas como acurácia, relatório de classificação e matriz de confusão, proporcionando uma visão detalhada sobre a eficácia do modelo na previsão da qualidade do vinho.

Abstract

This work presents a data analysis and modeling approach using the wine quality dataset. The process begins with loading the data, followed by visualizing the distribution of wine quality classes and analyzing the correlation between variables. Next, a Support Vector Machine (SVM) model is trained to predict wine quality based on its chemical characteristics. The model's performance is evaluated through metrics such as accuracy, classification report, and confusion matrix, providing a detailed overview of the model's effectiveness in predicting wine quality.

1. INTRODUÇÃO

Este trabalho aplica técnicas de análise de dados e aprendizado de máquina para prever a qualidade do vinho com base em suas características químicas, utilizando o dataset de qualidade do vinho. O modelo escolhido para a tarefa é o Support Vector Machine (SVM), que será avaliado por meio de métricas como acurácia, relatório de classificação e matriz de confusão. O objetivo é explorar as correlações entre as variáveis e avaliar a eficácia do modelo na previsão da qualidade do vinho.

2. ÁNALISES REALIZADAS

Neste trabalho, diversas análises exploratórias e quantitativas foram realizadas com o objetivo de entender melhor as características do dataset e a relação entre as variáveis, antes de treinar o modelo de aprendizado de máquina.

2.1. Análise de Correlação entre as Variáveis

A primeira análise consistiu em observar a distribuição das classes de qualidade do vinho, representadas pela variável quality. Utilizou-se um gráfico de barras (countplot) para visualizar a quantidade de vinhos em cada nível de qualidade, permitindo uma rápida avaliação da distribuição dos dados. Esta etapa é importante para entender se as classes estão balanceadas ou se há alguma classe mais predominante, o que pode influenciar a performance do modelo de classificação.

2.2. Análise de Correlação entre as Variáveis

A seguir, foi realizada uma análise de correlação entre as variáveis independentes (características químicas do vinho). Para isso, a coluna quality, que é a variável alvo, foi removida e foi calculada a matriz de correlação entre as outras variáveis numéricas do dataset. Essa análise permite identificar como as diferentes características do vinho se relacionam entre si e pode fornecer dados sobre quais variáveis podem ter maior influência na previsão da qualidade. Para a visualização, foi utilizado um heatmap, facilitando a interpretação das correlações de forma gráfica.

2.3. Divisão entre Dados de Treinamento e Teste

Para treinar o modelo e avaliá-lo de forma confiável, os dados foram divididos em duas partes: um conjunto de treinamento (70% dos dados) e um conjunto de teste (30% dos dados). A divisão foi feita utilizando a função train_test_split, garantindo que os dados de teste não fossem utilizados durante o treinamento, o que permite uma avaliação imparcial da performance do modelo.

2.4. Avaliação da Performance do Modelo

Após treinar o modelo Support Vector Machine (SVM), a performance do modelo foi avaliada com base nas previsões feitas sobre o conjunto de teste. Foram utilizadas as métricas de acurácia, relatório de classificação e matriz de confusão para avaliar a precisão das previsões. A acurácia fornece uma visão geral da taxa de acerto, enquanto o relatório de classificação oferece detalhes sobre precisão, recall e f1-score para cada classe. A matriz de confusão, por

sua vez, detalha as previsões corretas e incorretas em uma tabela, permitindo uma avaliação mais granular da performance do modelo.

3. RESULTADOS

Nesta seção, são apresentados os resultados obtidos a partir da aplicação do modelo **Support Vector Machine (SVM)** no dataset de qualidade do vinho. A avaliação do desempenho do modelo foi realizada com base em diversas métricas, e também foram geradas visualizações que ajudam a entender melhor a distribuição das classes, as correlações entre as variáveis e a performance do modelo.

3.1. Distribuição das Classes

A primeira análise realizada foi a visualização da distribuição das classes de qualidade do vinho. Através do gráfico a seguir, é possível observar a quantidade de vinhos em cada nível de qualidade, o que ajuda a entender se o dataset está balanceado ou se há alguma classe predominantemente maior que pode influenciar a performance do modelo.

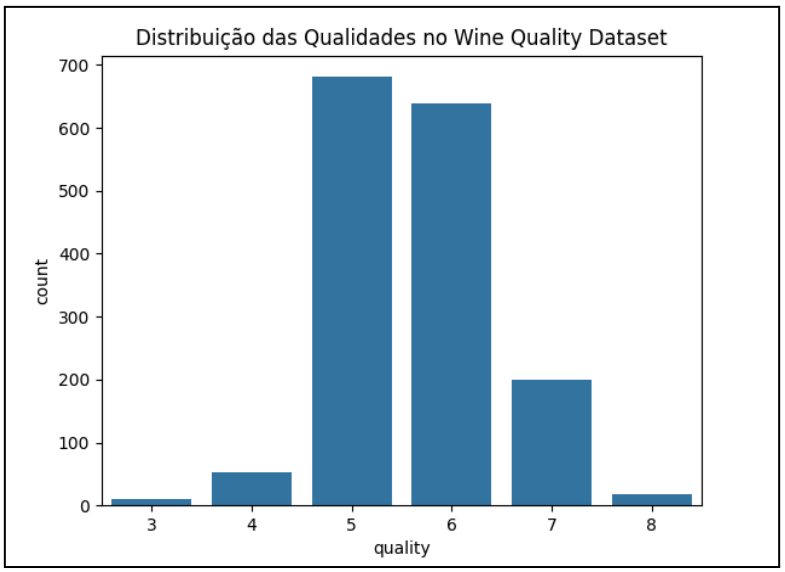

Figura 1: Gráfico da Distribuição das Qualidades (Autor)

3.2. Correlação Entre as Variáveis

A matriz de correlação entre as variáveis do dataset foi analisada para entender as relações entre as características químicas do vinho. Abaixo, segue a visualização das correlações, que revela como as variáveis estão relacionadas entre si e quais delas podem ter maior impacto na qualidade do vinho.

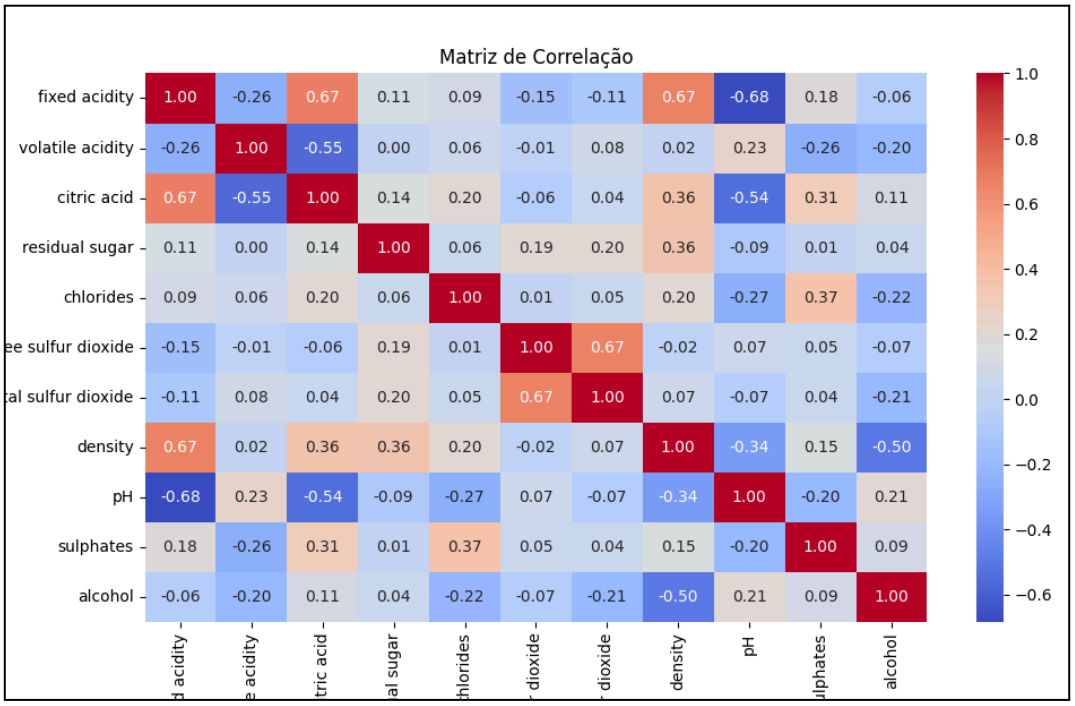

Figura 2: Gráfico da Distribuição da Matriz de Correlação(Autor)

Ao analisar a matriz de correlação, o objetivo é entender como as variáveis do dataset estão interligadas. Cada célula da matriz mostra o grau de correlação entre duas variáveis, que pode variar de -1 a 1.

- **Correlação positiva (próxima de 1)**: Isso significa que, à medida que uma variável aumenta, a outra também tende a aumentar. Por exemplo, uma

correlação forte entre "álcool" e "qualidade" pode sugerir que vinhos com mais álcool tendem a ter uma qualidade superior.

- **Correlação negativa (próxima de -1)**: Indica que, à medida que uma variável aumenta, a outra tende a diminuir. Por exemplo, uma correlação negativa entre "ácido volátil" e "qualidade" pode sugerir que vinhos com maior acidez volátil têm qualidade inferior.
- **Correlação próxima de 0**: Significa que não há uma relação clara entre as duas variáveis, ou seja, elas não se influenciam diretamente.

Na análise, procuramos por variáveis com correlações fortes (positivas ou negativas) com a qualidade do vinho, pois essas variáveis podem ser mais relevantes para prever o sabor ou a avaliação geral do vinho. Variáveis com correlações fracas com a qualidade, por outro lado, provavelmente têm menos impacto e podem ser menos importantes para a modelagem preditiva.

A visualização da matriz ajuda a identificar rapidamente esses padrões, facilitando a escolha das variáveis mais relevantes para um modelo de previsão ou análise mais aprofundada.

3.3. Avaliação da Performance do Modelo

Após o treinamento do modelo SVM, a acurácia, o relatório de classificação e a matriz de confusão foram utilizados para avaliar sua performance.

A acurácia do modelo foi calculada para verificar a porcentagem de previsões corretas em relação ao total de previsões feitas.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
0	7.4	0.70	0.00	1.9	0.076
1	7.8	0.88	0.00	2.6	0.098
2	7.8	0.76	0.04	2.3	0.092
3	11.2	0.28	0.56	1.9	0.075
4	7.4	0.70	0.00	1.9	0.076

Figura 3 e 4: Relatório da Acurácia(Autor)

free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
11.0	34.0	0.9978	3.51	0.56	9.4	5
25.0	67.0	0.9968	3.20	0.68	9.8	5
15.0	54.0	0.9970	3.26	0.65	9.8	5
17.0	60.0	0.9980	3.16	0.58	9.8	6
11.0	34.0	0.9978	3.51	0.56	9.4	5

O relatório de classificação detalha as métricas de precisão, recall e F1-score para cada classe de qualidade do vinho.

Precisão é a proporção de verdadeiros positivos (TP) em relação ao total de exemplos classificados como positivos pelo modelo. Em outras palavras, é a acurácia das previsões positivas feitas pelo modelo.

Recall (ou Sensibilidade) é a proporção de verdadeiros positivos (TP) em relação ao total de exemplos que realmente são positivos. Ele mede a capacidade do modelo de identificar todas as amostras positivas.

O F1-score é a média harmônica entre precisão e recall. Ele foi criado para ser uma métrica equilibrada, que leva em consideração tanto os falsos positivos quanto os falsos negativos. O F1-score é particularmente útil quando há um desbalanceamento entre as classes, já que ele tenta balancear a prioridade entre a precisão e o recall.

Relatório de Classificação:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	1
4	1.00	0.00	0.00	17
5	0.61	0.75	0.67	195
6	0.51	0.61	0.56	200
7	1.00	0.00	0.00	61
8	1.00	0.00	0.00	6
accuracy			0.56	480
macro avg	0.85	0.23	0.21	480
weighted avg	0.64	0.56	0.51	480

Figura 5: Relatório de Classificação(Autor)

A matriz de confusão compara os valores reais com as previsões feitas pelo modelo. A diagonal principal mostra os acertos, enquanto as fora da diagonal indicam os erros. Cada célula fora da diagonal revela onde o modelo confundiu uma classe com outra. Esse gráfico ajuda a identificar quais classes são mais difíceis de prever corretamente e onde o modelo pode ser melhorado.

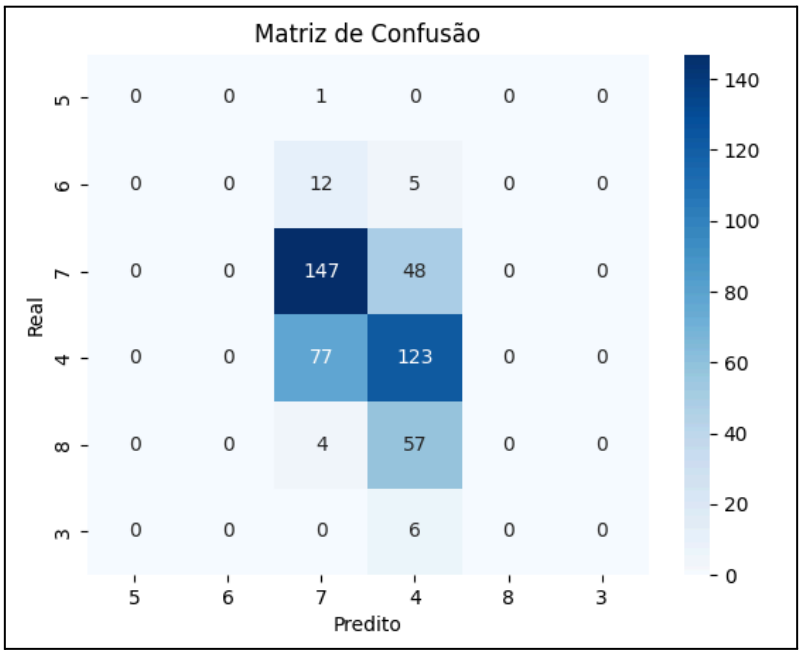

Figura 6: Gráfico da Distribuição da Matriz de Confusão(Autor)

Esses resultados oferecem uma visão completa da eficácia do modelo SVM na previsão da qualidade do vinho para posterior análise, bem como, destaca as áreas em que ele pode ser melhorado.

CONCLUSÃO

Este trabalho teve como objetivo aplicar técnicas de análise de dados e aprendizado de máquina para prever a qualidade do vinho a partir de suas características químicas. Utilizando o modelo Support Vector Machine (SVM), foi possível identificar as correlações entre as variáveis e entender como elas influenciam a classificação da qualidade do vinho. A avaliação do modelo, por meio de métricas como acurácia, relatório de classificação e matriz de confusão, revelou que, embora o modelo tenha apresentado dificuldades em prever algumas classes, ele conseguiu realizar previsões razoáveis para a maioria dos casos.

Embora o desempenho do modelo tenha sido satisfatório em termos gerais, algumas classes, especialmente as menos representadas, apresentaram resultados mais fracos. Isso indica que o modelo pode se beneficiar de ajustes, como a alteração dos parâmetros ou a utilização de técnicas que melhorem o balanceamento das classes. A análise das correlações entre as variáveis também foi valiosa, pois permitiu entender quais características são mais relevantes para determinar a qualidade do vinho.

Em resumo, este trabalho demonstrou que o aprendizado de máquina pode ser uma ferramenta eficaz para prever a qualidade do vinho, mas também mostrou que há espaço para aprimorar o modelo e obter previsões mais precisas.