Análise de Características Influentes na Densidade dos Vinhos Portugueses

Bruna Heloisa Feitosa Veloso[1], **Nara Raquel Dias Andrade**[1]

[1]Universidade Federal do Piauí (UFPI)
Picos – PI – Brazil

bruna.veloso@ufpi.edu.br, nara.andrade@ufpi.edu.br

Abstract. *This article seeks to delve deeper into how characteristics influence the density of Portuguese wines through the analysis of data from the Vinho Verde producer. The ultimate objective of this analysis is to be able to identify the specific attributes of the wine that influence its final density; This identification will enable the producer to have a targeted approach to improvement. Our methodology involves some steps such as data pre-processing, division of training and testing sets followed by the application of simple and multiple linear regression algorithms. Upon completion, these results are expected to provide the production team with information that can help in their production process in order to optimize production and increase product quality.*

Resumo. *Este artigo busca aprofundar em como as características influenciam na densidade dos vinhos portugueses através da análise de dados do produtor Vinho Verde. O objetivo final desta análise é conseguir identificar os atributos específicos do vinho que influenciam a sua densidade final; esta identificação fará com que o produtor consiga ter uma abordagem direcionada para melhoria. A nossa metodologia envolve algumas etapas como pré-processamento de dados, divisão de conjuntos de treinamento e teste seguido pela aplicação de algoritmos de regressão linear simples e múltipla. Após a conclusão, espera-se que estes resultados forneçam à equipe de produção informações que poderão ajudar no seu processo de produção no sentido de otimizar a produção e elevar a qualidade do produto.*

1. Introdução

Como veremos durante o estudo, para o sucesso de uma vinícola há a necessidade de compreender de maneira mais profunda como são os vinhos, com isso surge algumas possibilidades, entre elas a possibilidade de algumas características afetarem diretamente sua densidade, que pode afetar a qualidade do vinho. Para haver esse entendimento com essa pesquisa a produtora passa a ter uma noção de quais características afetam a densidade dos seus produtos, permitindo que saibam onde focar seus esforços de pesquisa ou ferramentas para tratar isso, conseguindo por fim melhores resultados do produto.

Como vimos anteriormente, a base de dados utilizada para este estudo é relevante para a produtora, suas variáveis são classificadas de forma numérica, o que facilita o processamento e análise dos dados, e por ser uma base de fácil entendimento, tornou o trabalho um pouco mais simples, pois não necessitou a realização de um tratamento

muito extenso durante o pré-processamento dos dados. Com uma abordagem direta conseguimos alcançar algumas conclusões sobre os fatores que influenciam na densidade dos vinhos.

2. Trabalhos Relacionados

Nesta seção, serão apresentados os trabalhos que se relacionam com este projeto e têm alguma relevância. Os eventuais trabalhos são apresentados na Tabela 1. Nesta tabela, apresentamos um resumo comparativo entre os trabalhos relacionados encontrados, de acordo com seus objetivos, técnicas aplicadas e características da pesquisa. Essa análise comparativa nos ajuda a trazer uma visão geral dos trabalhos relacionados, ajudando a compreensão de como eles se relacionam entre si e como se comparam com o nosso projeto em alguns pontos.

No primeiro estudo analisado [Sachet and de Avila e Silva 2024] destacam no seu estudo a aplicação de técnicas de mineração de dados para explorar os processos de fermentação de vinhos, buscando revelar informações importantes sobre diferentes características presentes durante todo o processo e ainda demonstrar a ligação entre compostos quimícos e seus efeitos. Com base nisso, o trabalho conseguiu confirmar a hipótese ao conseguir informações que auxiliam no desenvolvimento de novos dados importantes para os especialistas, e com isso sugere que a aplicação dessas técnicas de mineração de dados podem ser uma ferramenta importante para o auxílio e otimização dos processos.

De [Pires et al. 2023] o principal objetivo era investigar e avaliar entre os atributos e a qualidade dos vinhos da região de Campanha do Rio Grande do Sul, para atingir esse objetivo foram desenvolvidos modelos preditivos utilizando algumas técnicas de modelagem estatística. Com isso, os resultados obtidos mostraram que os métodos utilizados mostraram alta eficiência durante os testes realizados obtendo uma precisão de mais de 90% na qualidade sensorial com base nas características químicas dos vinhos.

Para o estudo [Yao 2023] temos uma abordagem voltada para a previsão de vendas da Walmart que utiliza três modelos de aprendizado de máquina: Decision Tree, Random Forest e K Neighbors Regressor, com essa análise foi possível perceber que o modelo que teve o efeito mais eficaz, em comparação com os outros modelos testados na previsão das vendas foi o Random Forest, isso foi constato através de três critérios: o número de correlação entre os valores previstos e os valores reais, o erro absoluto médio e o erro quadrático médio.

O presente estudo busca explorar e compreender o funcionamento de diferentes ferramentas para análise de relação entre as variáveis presentes em um conjunto de dados de vinhos. Para isso utilizamos algoritmos de regressão linear simples e múltipla, além dos testes com algoritmos da biblioteca sklearn, que incluem, o Support Vector Regressor (SVR), RandomForestRegressor (RFR), KneighborsRegressor e DecisionTreeRegressor. Com diferentes implementações e testes buscamos identificar o melhor desempenho e os atributos mais relevantes para a predição da densidade do vinho, isso trará informações importantes para produtores de vinhos pois a partir dessa análise é possível perceber quais características físico-químicas têm maior influência para a sua densidade.

Table 1. Trabalhos Relacionados

Autor	Aplicação	Objetivo	Dados	Técnicas Aplicadas
Sachet and de Avila e Silva 2024	Análise de Processos de Fermentação de Vinhos	Explorar características de leveduras em fermentação de vinhos.	Características químicas das leveduras e dados da avaliação de degustadores.	Mineração de Dados (Árvores de Decisão, Apriori)
Pires et al. 2023	Modelagem Estatística e Classificação de Vinhos	Investigar a relação dos atributos químicos com a qualidade dos vinhos	Doze atributos com onze medições de características químicas e a respectiva qualidade sensorial.	Modelagem Estatística, Classificação
Yao 2023	Predição de Vendas com Aprendizado de Máquina	Explorar métodos de previsão para dados de vendas das lojas da walmart	Dezesseis atributos coletados sobre quarenta e cinco lojas	Decision Tree, Random Forest e K Neighbors Regressor

3. Metodologia

Esta seção descreve a metodologia utilizada para resolver o problema, nela detalharemos o processo de análise de dados, a seleção das variáveis e a implementação dos algoritmos de regressão linear simples e múltipla.

3.1. Base de Dados

A base de dados escolhida é uma pesquisa realizada para saber sobre a qualidade de vinhos portugueses da produtora Vinho Verde, ela foi dividida em duas partes: a primeira contém os dados de vinhos tradicionais e a segunda sobre vinhos brancos; foi separada dessa forma, pois essa característica poderia influenciar e mostrar diferentes resultados sobre a qualidade com base nas outras características escolhidas. Essa base pode ser utilizada tanto para estudos sobre classificação como também para regressão, simples e múltipla, que serão as técnicas que abordaremos durante esse trabalho.

O *dataset* conta com 4998 instâncias de dados e 12 atributos. Esses 12 atributos, são as doze características que consideraremos durante a implementação e observação dos algoritmos, sendo elas: quantidade de ácidos envolvidos no vinho, sendo eles fixos ou não voláteis; a quantidade de ácido acético no vinho; a quantidade de ácido cítrico no vinho; a quantidade de açúcar restante após a fermentação; a quantidade de sal no vinho; a quantidade de dióxido de enxofre livre no vinho, sendo aqueles disponíveis para reagir e, portanto, exibir propriedades germicidas e antioxidantes; quantidade de formas livres e ligadas de SO2; a medição da sua densidade; a descrição de quão ácido ou básico é um vinho em uma escala de 0 (muito ácido) a 14 (muito básico); a porcentagem de teor alcoólico do vinho; e por fim, a qualidade, que está pontuada entre 3 e 8.

3.1.1. Estatística da base

Tendo em vista as características citadas anteriormente, agora faremos uma análise detalhada de algumas estatíscas desses dados, que nos trará uma melhor compreensão e visualização de distribuição dos dados, tendências entre eles e possíveis correlações. Essa técnica de análise é necessária para nos auxiliar posteriormente na apliação das técnicas de regressão. As medidas exploradas estarão na Tabela 2 e são elas: contagem dos dados de cada coluna, média, valores mínimos, valores máximos e desvio padrão.

Atributo	Quantidade	Média	Desvio Padrão	Valor Mínimo	Valor máximo
Acidez Fixa	4898	6.854788	0.843868	3.8	14.2
Acidez Volátil	4898	0.278241	0.100795	0.08	1.1
Ácido Cítrico	4898	0.334192	0.121020	0.0	1.66
Açúcar Residual	4898	6.391415	5.072058	0.6	65.8
Cloretos	4898	0.045772	0.021848	0.009	0.346
Dióxido de Enxofre Livre	4898	35.308085	17.007137	2.0	289.0
Dióxido de Enxofre Total	4898	138.360657	42.498065	9.0	440.0
Densidade	4898	0.994027	0.002991	0.98711	1.03898
pH	4898	3.188267	0.151001	2.72	3.82
Sulfatos	4898	0.489847	0.114126	0.22	1.08
Álcool	4898	10.514267	1.230621	8.0	14.2
Qualidade	4898	5.877909	0.885639	3.0	9.0

Table 2. Estatísticas Descritivas dos Atributos dos Vinhos

Como podemos observar os dados que estamos trabalhando possuem uma variedade considerável, essa variedade sugere uma diferença substancial na produção dos vinhos, ou seja, a análise desses atributos pode haver um potencial significativo nos vinhos. Para termos uma visão melhor dessas estatísticas descritivas, utilizaremos alguns gráficos para detalhar os dados de forma adequada.

Esses gráficos são: os **Histogramas**, que ajudam a visualizar a distribuição dos valores de cada atributo, auxiliando na identificação de possíveis concentração de dados, outliers ou dessimetrias; **Boxplots** que serão usados para resumir essa distribuição dos atributos, deixando claro como funciona os quartis, medianas e também possíveis outliers; e por fim a **Matriz de Correlação** que trata de um mapa de calor que mapeia as correlações entre todos os atributos de maneira abrangente, destacando quais atributos estão mais correlacionados entre si.

Ao analisarmos os histogramas dos atributos dos vinhos da produtora Vinho Verde, visto na Figura 1 é possível fazer algumas observações sobre os dados:

- **Acidez Fixa:** A distribuição da acidez fixa mostra uma concentração em torno da média de 6.85. A maioria dos vinhos possui acidez fixa entre 6 e 7, com poucos valores extremos acima de 10.
- **Acidez Volátil:** Apresenta uma distribuição levemente assimétrica à direita, com a maioria dos valores concentrados entre 0.2 e 0.4. Poucos vinhos possuem acidez volátil acima de 0.5.
- **Ácido Cítrico:** O histograma do ácido cítrico revela que muitos vinhos possuem valores baixos ou nulos, com um pico em torno de 0.3. A distribuição é assimétrica à direita, com poucos vinhos apresentando valores altos.
- **Açúcar Residual:** A distribuição do açúcar residual é altamente assimétrica à direita. A maioria dos vinhos tem baixos níveis de açúcar residual (menores que

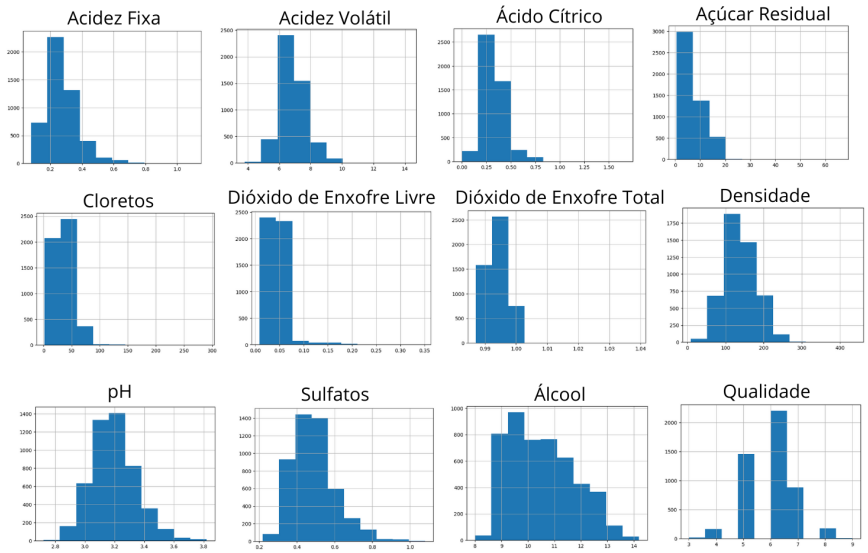

Figure 1. Histogramas Atributos

5), mas há uma longa cauda que se estende até 65.8, indicando a presença de vinhos muito doces.

- **Cloretos:** O histograma dos cloretos mostra uma distribuição assimétrica à direita, com a maioria dos valores concentrados abaixo de 0.1. Há alguns outliers com valores significativamente mais altos.
- **Dióxido de Enxofre Livre:** A distribuição do dióxido de enxofre livre é bastante ampla, com um pico em torno de 30. A cauda direita se estende consideravelmente, indicando que alguns vinhos possuem níveis muito altos deste composto.
- **Dióxido de Enxofre Total:** Similar ao dióxido de enxofre livre, a distribuição do dióxido de enxofre total tem uma cauda longa à direita. A maioria dos vinhos apresenta valores entre 100 e 150.
- **Densidade:** A densidade dos vinhos apresenta uma distribuição concentrada em torno da média de 0.994, com pouca variabilidade. A maioria dos valores está entre 0.99 e 1.0.
- **pH:** O histograma do pH mostra uma distribuição próxima da normal, centrada em torno da média de 3.19. A maioria dos vinhos tem pH entre 3.0 e 3.3.
- **Sulfatos:** A distribuição dos sulfatos é assimétrica à direita, com a maioria dos vinhos apresentando valores entre 0.3 e 0.6. Há poucos valores extremamente altos.
- **Álcool:** O histograma do álcool apresenta uma distribuição levemente assimétrica à direita, com a maioria dos valores concentrados entre 9 e 12. Poucos vinhos têm teor alcoólico acima de 13.
- **Qualidade:** A qualidade dos vinhos, avaliada numa escala de 3 a 9, mostra uma distribuição próxima da normal com um pico em torno de 6. Poucos vinhos receberam notas extremas, sejam elas muito baixas ou muito altas.

Com os boxplots presentes na Figura 2 é possível termos uma visualização clara das distribuições, pois eles destacam a presença de outliers em vários atributos que mostra

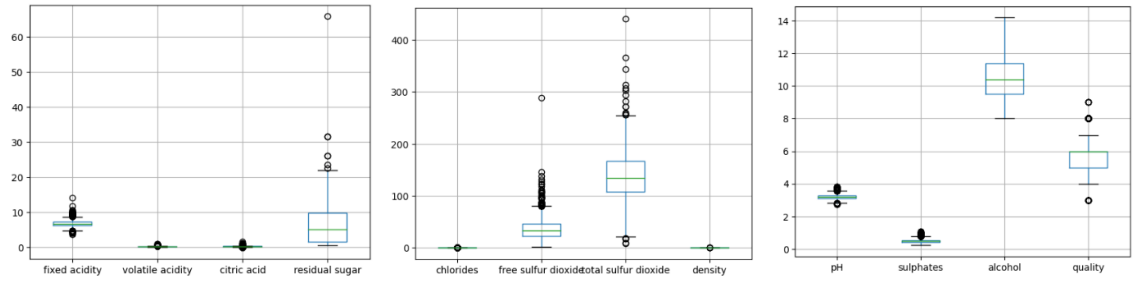

Figure 2. Boxplots Atributos

extremos nos dados, esses extremos podem influenciar as análises que faremos posteriormente.

- **Acidez Fixa:** A maioria dos vinhos possui acidez fixa entre 6.3 e 7.3, com alguns outliers acima de 10.
- **Acidez Volátil:** Valores principalmente entre 0.21 e 0.32, com alguns outliers acima de 0.5.
- **Ácido Cítrico:** Muitos valores baixos ou nulos, com outliers acima de 0.6.
- **Açúcar Residual:** Grande variabilidade, com valores entre 1.7 e 9.9, e outliers até 65.8
- **Cloretos:** Valores concentrados entre 0.036 e 0.05, com alguns outliers acima de 0.1.
- **Dióxido de Enxofre Livre:** A maioria dos valores entre 23 e 46, com muitos outliers acima de 100.
- **Dióxido de Enxofre Total:** Valores entre 108 e 167, com outliers acima de 200.
- **Densidade:** Pequena variação, com a maioria dos valores entre 0.9917 e 0.9961.
- **pH:** Principalmente entre 3.09 e 3.28, com alguns outliers abaixo de 2.8 e acima de 3.6.
- **Sulfatos:** Valores entre 0.41 e 0.55, com outliers acima de 0.8.
- **Álcool:** Principalmente entre 9.5 e 11.4, com alguns outliers acima de 12.5.
- **Qualidade:** Avaliações concentradas entre 5.0 e 6.0, com alguns outliers abaixo de 4 e acima de 7.

Como visto anteriormente, a matriz de correlação da Figura 3 é uma ferramenta que nos permite visualizar e quantificar a relação entre diferentes atributos, para isso analisaremos esse mapa de calor, que é representado por tons de azul, onde as cores mais escuras indicam correlações mais fortes e as mais claras indicam correlações fracas.

Durante a análise é possível perceber algumas relações importantes, como, a forte correlação entre dióxido de enxofre livre e dióxido de enxofre total isso indica que esses atributos estão ligados, há ainda a correlação positiva entre álcool e qualidade que indica que a característica de teor alcoólico pode ser um fator importante na avaliação da qualidade dos vinhos, e a correlação negativa entre a densidade e o álcool que mostra que vinhos mais densos tendem a ter menos álcool.

Como consequência dessa ampla análise estatística, agora possuímos uma boa base para compreendermos melhor as próximas fases do estudo. De agora em diante iremos começar a desenvolver os modelos que nos ajudarão a prever e melhorar a densidade dos vinhos, tendo como base as informações obtidas durante essa análise.

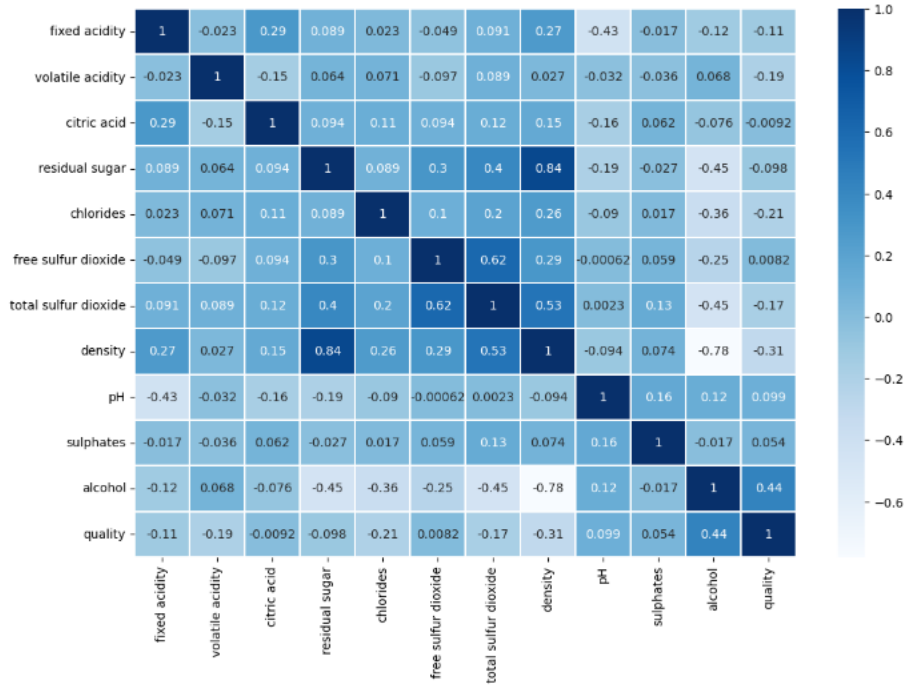

Figure 3. Matriz de Correlação

3.2. Passo-a-Passo

Esta seção detalha a metodologia adotada neste estudo. Descrevemos os passos seguidos para coletar dados, selecionar técnicas analíticas, processar informações e interpretar resultados.

3.2.1. Fluxograma

A Figura 4 apresenta o fluxograma detalhando a metodologia adotada para a resolução do problema, provendo uma visualização objetiva e detalhada dos passos e fases do projetos desde sua concepção até a conclusão.

A primeira etapa que realizamos para dar início ao projeto foi a escolha da base de dados que iríamos trabalhar; depois analisamos os dados estatisticamente para conseguirmos compreender suas características básicas; e ainda com esse objetivo tivemos a fase de geração de alguns gráficos que facilitaram o nosso entendimento sobre algumas métricas.

Agora, para dar continuídade ao nosso estudo, é necessário que haja uma preparação dos dados, para que eles possam serem utilizados nos algoritmos de regressão. Depois de aplicarmos as técnicas de pré-processamento que veremos mais à frente, há a fase de implementação dos algoritmos, onde iremos falar um pouco mais sobre os algoritmos de regressão linear simples e múltipla, e como funciona a sua implementação; Após isso, faremos a implementação de testes com o auxilio da biblioteca Sklearn (biblioteca de aprendizado de máquina em Python), de algumas outras regressões disponíveis, isso vai nos permitir comparar o desempenho de diferentes modelos; Então, os resultados são verificados e discutidos para garantir a qualidade das regressões e entender a relação en-

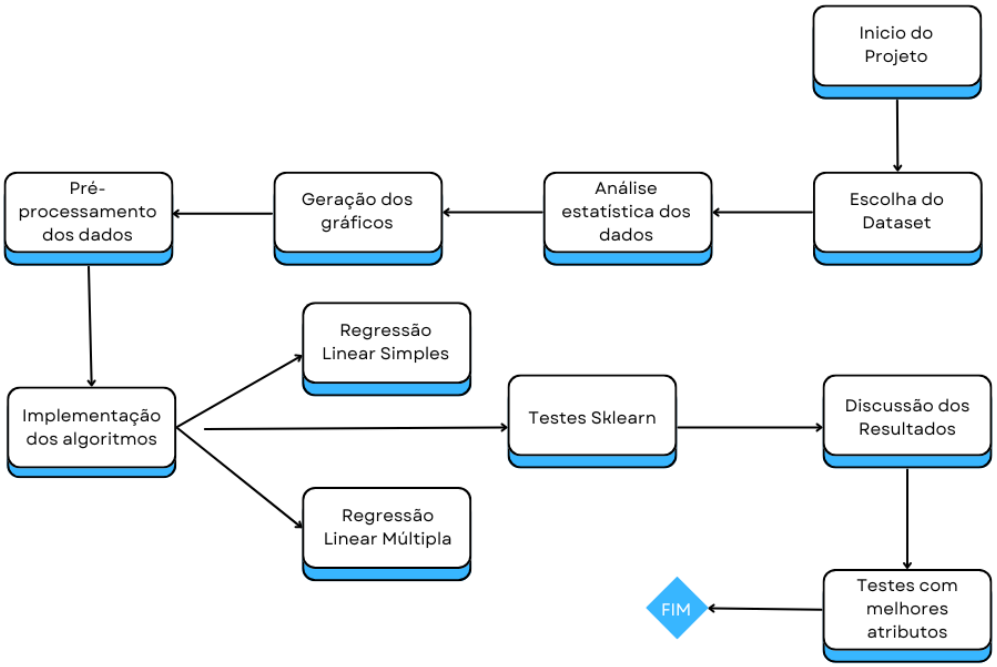

Figure 4. Fluxograma Passo-a-Passo

tre as variáveis; Por fim realizaremos testes adicionais que utilizarão apenas os melhores atributos identificados.

3.2.2. Pré-processamento dos Dados

Primeiro começamos com o pré-processamento dos dados que consisti na realização de uma análise de todas as características da base de dados e como estão distribuídas. Essa análise é realizada com o auxílio de alguns gráficos padrões como, matriz de correlação, histogramas, boxplots, entre outros, essas ferramentas ajudam a identificar relações entre as variáveis selecionadas e algumas tendências entre as mesmas.

A seguir, foi verificado se havia valores Not a Number (NaN) que poderiam comprometer a execução e o resultado do problema. Verificamos também se a existência de valores nulos em cada uma das colunas.

Após feita a análise dos dados e características, os dados foram separados em treino e teste usando a metodologia 75% treino e 25% teste. Esta metodologia foi adotada a fim de que com o treino seja possível ajustar o modelo e com o teste, seja possível avaliar seu desempenho e analisar se os resultados obtidos estão corretos ou se eles estão apenas ajustados para o nosso conjunto de treino.

3.2.3. Algoritmos implementados

A solução do problema proposto contou com a implementação dos algoritmos para o cálculo da regressão, que trata de um algoritmo que busca fazer ajustes de várias linhas nos pontos dos dados e retornar a linha que tem a menor taxa de erro. Ela pode ser dividida em: linear simples e da regressão linear múltipla dos dados. A regressão linear simples é, como o próprio nome sugere, o tipo mais simples de regressão, onde é utilizada apenas uma característica para predizer uma segunda característica. A regressão linear múltipla é o cálculo da regressão linear que utiliza N características para predizer uma única característica alvo.

- **Regressão Linear Simples -** Para a implementação do algoritmo de regressão linear simples, primeiro calculamos a correlação entre as variáveis e após a plotagem e análise do gráfico optamos por trabalhar com as variáveis selecionadas: densidade (density) e açúcar residual (residual-sugar), sendo a característica alvo a densidade. A proposta é prever o valor da variável alvo a partir da variável residual-sugar. A escolha das variáveis se deu pela alta correlação entre as duas características.
- **Regressão Linear Múltipla -** No algoritmo de regressão linear múltipla foram usadas todas as características da base, com exceção da característica quality (qualidade), sendo a característica alvo da regressão a densidade, ou seja, o objetivo é utilizar todas as características para predizer a densidade do vinho.

3.2.4. Testes Sklearn

Nessa seção do trabalho abordaremos alguns testes realizados com Sklearn, uma biblioteca do python direcionada para o aprendizado de máquina, esses testes envolveram a utilização de diferentes algoritmos de regressão disponíveis na biblioteca, para que pudessemos por fim realizarmos algumas análises comparativas com os modelos que implementamos anteriormente. Cada um dos seis algoritmos escolhidos para teste está especificado abaixo com suas vantagens e limitações, buscaremos determinar qual deles se ajusta melhor aos dados do nosso estudo.

- **Regressão Linear Simples** - Esse algoritmo segue as mesmas características apresentadas anteriormente, a diferença é que ele é derivado da própria biblioteca e o que falamos na seção 3.2.3 foram implementados do zero. Iremos utilizar as duas formas para podermos realizarmos as comparações.
- **Regressão Linear Múltipla** - No caso da regressão múltipla ele segue o mesmo sistema que o de regressão simples, pois este também será utilizado para realizar as comparações necessários com o algoritmo que implementamos.
- **Support Vector Regressor (SVR)** - Os métodos baseados em SVR são algoritmos de aprendizado de máquinas que criam um modelo baseado em uma função kernel não linear e vetores suporte escolhidos a partir de uma base de dados de treinamento [Junior and Moreno 2019]. Esse algoritmo de regressão trata-se de uma variação do SVM (Support Vector Machine), ele é adequedo para problemas com uma estrutura mais complexa e não-linear, ou seja, ele consegue encontrar uma relação funcional entre as variáveis independentes e a variável dependente mesmo que a relação entre elas seja não-linear.

- **RandomForestRegressor (RFR)** - O algoritmo RandomForestRegressor ou a regressão por floresta aleatória, é um algoritmo derivado da árvore de regressão. Ele utiliza o método de cálculo da média de várias árvores de regressão para ajustar a predição. Em Python, 'n_estimators' é usado para controlar o número de árvores de regressão selecionadas e pode melhorar o número de árvores de regressão e a precisão do ajuste no conjunto de treinamento [Yao 2023]. O algoritmo utiliza uma técnica onde há a combinação de várias árvores para assim formar um modelo mais completo e mais preciso, para isso é necessário que haja o treinamento de cada uma das árvores com uma amostra aleatória dos dados, a fim de obter previsões independentes. Para se obter as previsões finais é necessário que haja uma votação ou média das previsões individuais.
- **KneighborsRegressor** - A Regressão K Nearest Neighbor é um método não-paramétrico usado para problemas de previsão. Ele funciona com base no pressuposto de que valores de entrada semelhantes provavelmente produzem valores de saída semelhantes. No contexto de regressão, o KNN pega um número especificado (K) dos pontos de dados mais próximos (vizinhos) e faz a média de seus valores para fazer uma previsão [Kanaries 2023]. Esse é um algoritmo simples, mas que pode precisar de um alto poder computacional em conjuntos de dados grandes, já que ele faz suas previsões através de um sistema de proximidade dos exemplos de treinamento no espaço de características, ou seja, para realizar uma previsão para um novo ponto esse algoritmo encontra os "k" vizinhos que estão mais próximos do novo ponto e calcula a medida de agregação, que pode ser a média ou outra métrica utilizada.
- **DecisionTreeRegressor** - O algoritmo DecisionTreeRegressor ou Regressor de árvore de decisão constrói modelos de regressão na forma de uma estrutura de árvore, com galhos e folhas. O nó de folha representa uma decisão sobre o alvo numérico [Revathy et al. 2022]. O algoritmo é baseado em árvores de decisões, funcionando através da divisão do espaço de características em regiões distintas, tentando aumentar a uniformidade das previsões dentro de cada uma dessas regiões.

3.2.5. Métricas Utilizadas para Avaliação

Para a avaliação dos modelos propostos e utilizados, foram usadas as seguintes métricas de avaliação de desempenho:

- **R^2**: Também conhecida como R-dois ou coeficiente de determinação, representa o percentual da variância dos dados que é explicado pelo modelo, os valores para esse coeficiente variam de 0 (zero) a 1 (um). Quanto maior é o valor de R^2,ou seja, quanto mais próximo de 1 for o valor, mais explicativo é o modelo em relação aos dados previstos.
- **Erro Médio Absoluto (MAE)** : A métrica MAE mede a média da diferença entre o valor real com o predito. Mas por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. Além disso, esta métrica não é afetada por valores discrepantes — os denominados outliers.
- **Erro Quadrático Médio (MSE)** : O MSE é uma métrica que calcula a média de diferença entre o valor predito com o real, como a métrica MAE. Entretanto, ao

invés de usar o módulo do resultado entre o valor dos rótulos reais e o valor dos rótulos preditos, nesta métrica a diferença é elevada ao quadrado. Desta maneira penalizando valores que sejam muito diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação as previsões.

- **Raiz do Erro Quadrático Médio (RSME)**: A raiz do erro quadrático médio (RMSE — do inglês, Root Mean Squared Error) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Para lidar com o problema da diferença entre unidades, esta métrica aplica aplica a raiz quadrática como demonstrado na equação.

4. Resultados

Apresentaremos nessa seção os resultados que foram alcançados com a aplicação dos algoritmos de regressão explicados anteriormente. A seguir, discutiremos os desempenhos desses modelos, como eles se ajustaram aos dados, e qual as interpretações sobre a influência dessas características na densidade dos vinhos. E por fim, realizaremos alguns testes adicionais com os melhores atributos vistos durante o trabalho.

4.1. Resultados com 1 característica

Nessa parte apresentamos os gráficos gerado do resultado do algoritmo que implementamos de regressão simples com apenas uma características, presente na Figura 5 e o resultado do gráfico do modelo de regressão simples da biblioteca Skelearn, Figura 6 depois criamos a Tabela 3 para tornar mais simples a visualização das métricas usadas para medir o funcionamento e o desempenho. Utilizaremos esses dados e gráficos para realizarmos uma análise mais aprofundada sobre os seus significados, essas observações serão feitas na seção de Discussão de Resultados.

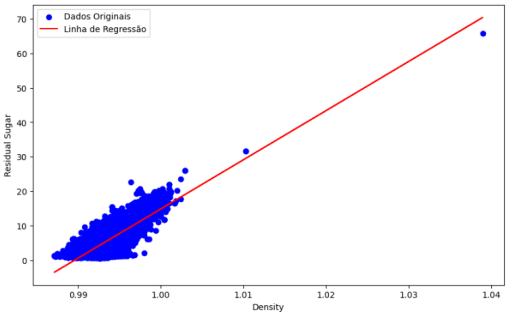

Figure 5. Regressão Linear Simples

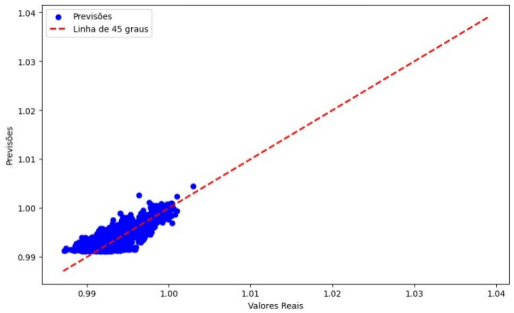

Figure 6. Regressão Linear Simples - Sklearn

	R^2	MAE	MSE	RMSE
Linear Simples	0.7104	2.2270	7.6476	2.7654
Linear Simples Sklearn	0.6753	0.0013	2.6886	0.0016
Linear Múltipla	0.9645	0.0003	3.1678	0.00
Linear Múltipla Sklearn	0.9665	0.0004	2.7713	0.0005

Table 3. Métricas para os modelos de regressão implementados.

4.2. Resultados com todas as característica

Iremos analisar agora os gráficos e métricas da regressão linear múltipla, que como apresentado anteriormente trata-se do estudo baseado em todas as variáveis para prever uma variável alvo. Na Figura 7 apresenta o resulto do algoritmo que implementamos para esse estudo, já na Figura 8 é apresentado o resultado do algoritmo da biblioteca sklearn. E podemos observar as métricas dos mesmos na Tabela 3, que foi apontada anteriormente.

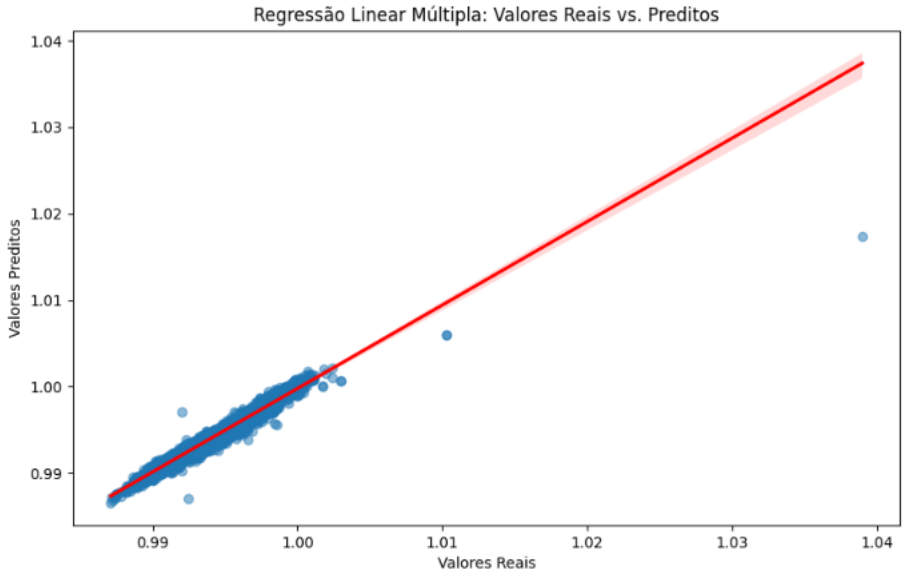

Figure 7. Regressão Linear Múltipla

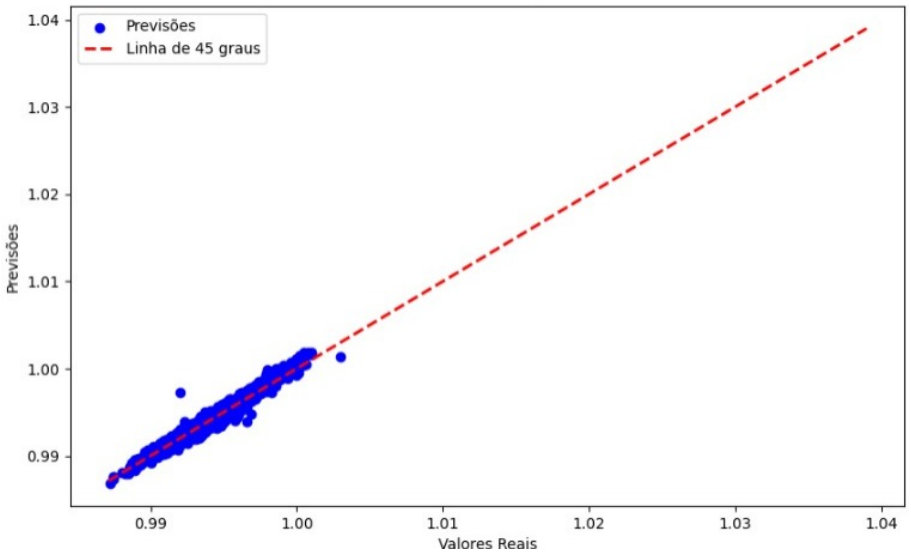

Figure 8. Regressão Linear Múltipla - Sklearn

E para complementar esse estudo, testamos alguns outros algoritmos apresentados na subseção 3.2.5, que foram utilizados da mesma maneira que a regressão múltipla (com várias variáveis para predizer apenas uma) e suas respectivas métricas foram implementados utilizando as funções prontas disponíveis na biblioteca *Scikit-Learn*, do python.

	R^2	MAE	MSE	RMSE
Support Vector Regression	-44.20	0.019	0.0003	0.019
Random Forest Regressor	0.975	0.0003	2.033	0.0004
Kneighbors Regressor	0.93	0.0005	5.65	0.0007
Decision Tree Regressor	0.94	0.0003	4.83	0.0006

Table 4. Métricas para os modelos de regressão testados.

A Tabela 4 apresenta os valores das métricas para todos os 4 modelos de regressão abordados neste projeto. Além dos valores da métricas de avaliação, para cada algoritmo linear foi implementado o gráfico de linha de regressão dos modelos, para que fosse verificado o quão bem eles se ajustaram aos dados utilizados. Para os algoritmos baseados em árvores (Decision Tree e RnadomForest) foram implementados o gráfico com a impotrância das caracteristicas para as predições dos algoritmos.

4.2.1. Support Vector Regressor

O algoritmo SVR foi implementado com o hiperparâmetro 'kernel' definido como linear, para que assim o algoritmo se comportasse como um algoritmo de regressão linear.

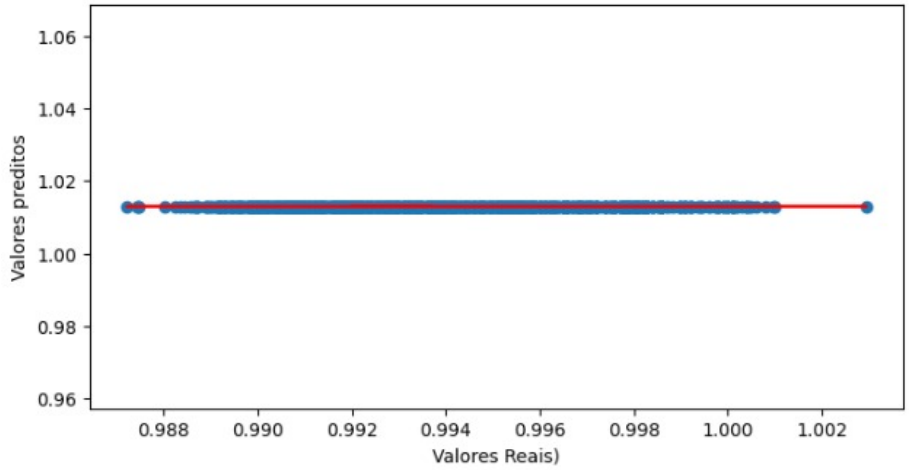

Figure 9. Linha de Regressão algoritmo SVR

A Figura 9 apresenta as predições do modelo SVR. Os pontos azuis representam as previsões do modelo, enquanto a linha vermelha representa a linha de 45 graus, onde os valores previstos seriam iguais aos valores reais. Os pontos localizados sobre a linha, que idealmente seria onde o predito seria igual ao real são os pontos onde o modelo acertou a previsão.

A análise do gráfico apresentado na Figura 9 nos mostra que o modelo, apesar de erroneamente aparentar se ajustar bem aos dados, as predições foram as mesmas para todas as amostras, fazendo com que a dispersão se apresentasse como uma linha horizontal, como mostra a Figura.

4.2.2. Random Forest Regressor

O algoritmo Random Forest Regressor é um algoritmo baseado em árvores aleatórias, ou seja, o Random Forest Regressor não assume uma relação linear entre as variáveis de entrada e a variável de saída, em vez disso, ele opera de forma não linear.

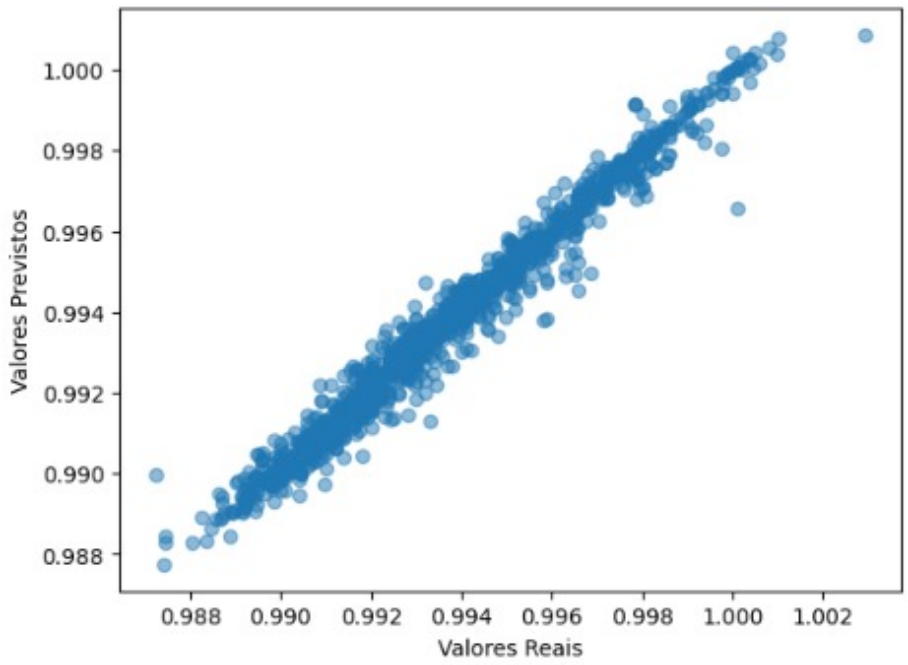

Figure 10. Dispersão do Algoritmo Random Forest Regressor

A Figura 10 demonstra o desempenho do modelo, apresentando no eixo X os valores reais da base, enquanto no eixo y são apresentados os valores preditos pelo modelo. Os algoritmos baseados em árvores permitem a visualização da impotância das caracteristicas, de modo que pode ser observado qual das caracteristicas teve maior importância para a previsão do modelo. A Figura 11 apresenta a importância que cada uma das caracteristicas teve durante as predições do modelo RFR.

A partir da análise da Figuras 10 é perceptivel que o modelo se ajustou bem aos dados, fazendo predições com valores bem próximos aos valores reais. Isso pode ser observado pelos pontos estarem bem próximos à linha de regressão. Isso indica um bom ajuste do modelo aos dados. Já a Figura 11 demonstra que as variaveis que posseum maior peso, ou seja, são mais representativas do dados e consequentemente possuem maior importância para a predição são as caracteristicas *Alcohol e Residual Sugar*.

4.2.3. Kneighbors Regressor

O algoritmo KNeighbors Regressor é uma variação do algoritmo KNN para classificação, usado para regressão. Diferentemente do algoritmo SVR e semelhante aos algoritmos baseados em árvores implementados para este trabalho, o algoritmo KNeighbors não é um algoritmo linear, mas sim pertencente à classe de algoritmos K Vizinhos mais próximos. A Figura 12 apresenta o gráfico de dispersão dos dados.

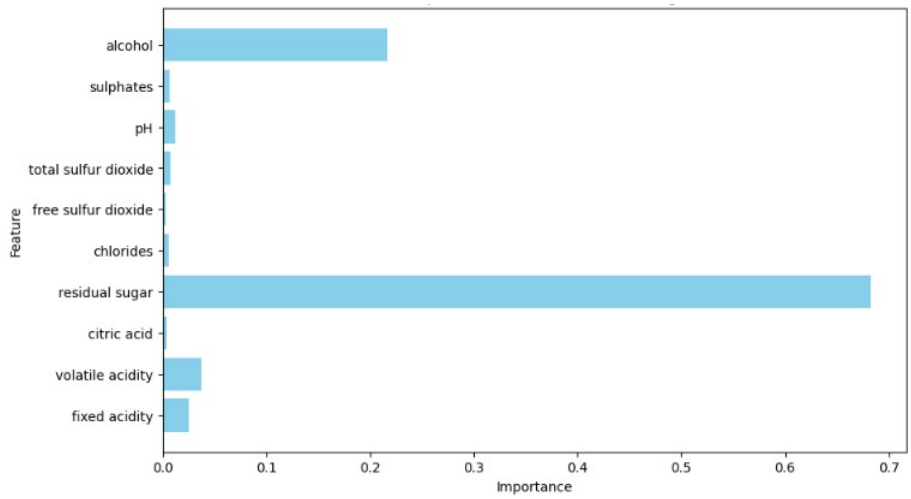

Figure 11. Importância das caracteristicas do Algoritmo Random Forest Regressor

A análise do gráfico deixa evidente que o algoritmo se ajustou bem aos dados, considerando a dispersão dos pontos no plano, em que a grande maioria das amostras apresentam valor real próximo ao predito. Pequenas dispersões e alguns outliers podem ser visualizados, mas a maioria dos pontos está concentrada ao longo da linha diagonal, indicando um bom desempenho geral do modelo.

4.2.4. Decision Tree Regressor

Assim como o algoritmo Random Forest Regressor, o algoritmo Decision Tree Regressor também é um algoritmo baseado em árvore e também é um algoritmo não linear, ou seja, não assume uma relação linear entre as variáveis de entrada e a variável. A Figura 13 apresenta a disperção dos valores previstos x valores reais.

A partir da analise dos resultados do grafico é perceptivel que o modelo se ajustou bem aos dados utilizados, apesar de que muitos dados ainda estão dispersos, é observável que a grande maioria se dispõe como uma diagonal, evidenciando que os valores preditos são próximos dos valores reais. A Figura 14 apresenta a importância das caracteristicas para a regressão.

A partir da análise da importância das caracteristicas, é possível observar que dentre as 11 caracteristicas, todas possuem o mesmo grau de importância para o cálculo da regressão pelo algoritmo Decision Tree Regressor.

4.3. Discussão dos Resultados

Na análise do nosso algoritmo de **regressão simples**, é possível perceber a inclinação na linha de regressão que representa a relação linear entre as duas variáveis de estudo (residual sugar e density), essa inclinação é positiva o que representa que à medida que o açúcar residual tende a aumentar, a densidade do vinho também aumenta. Podemos ver também como a maioria dos dados está concentrada em valores mais baixos, com alguns poucos pontos dispersos, o que pode indicar um padrão comum na produção ou uma preferência por determinada característica nos vinhos. No caso da regressão simples com

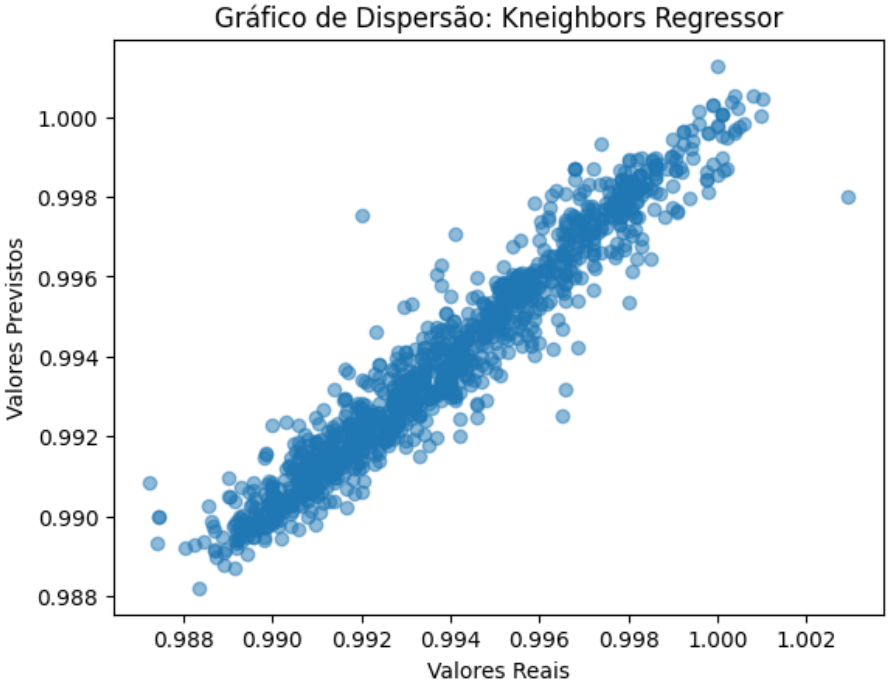

Figure 12. Dispersão do algoritmo KNeighbors Regressor

sklearn é possível perceber a semelhança com o anterior, pois ambos os gráficos mostram que os modelos de regressão linear simples tiveram um ajuste razoável, com os pontos estando relativamente próximos das linhas (linha de regressão e linha de 45 graus), mas como há muitos pontos ao redor da linha isso nos mostra que há espaço para melhoria no ajuste dos modelos.

Na observação das métricas obtidas para esses algoritmos, vemos que para o R^2 do algoritmo implementado obtivemos 71%, que indica a variância dos dados que é explicada pelo modelo e cerca de 68% pelo modelo Sklearn. E as métricas de erro mostraram que o modelo do Sklearn tem erros menores, principalmente MAE e RMSE, por isso as previsões individuais desse modelo são mais próximas dos valores reais.

Ao analisarmos a **regressão múltipla** é possível perceber que em ambos os gráficos, os pontos azuis estão muito próximos das linhas isso demonstra que os modelos possuem um bom ajuste e que os valores preditos estão bem alinhados com os valores reais. Durante a análise comparativa das métricas temos valores de R^2 muito altos, indicando que ambos os algoritmos foram excelentes em explicar a variância nos dados, e erros (MAE, MSE, RMSE) extremamente baixos, mas o modelo Sklearn mostra uma pequena superioridade no desempenho.

E com os testes utilizando os algoritmos da biblioteca sklearn, percebemos que através dos indicadores de desempenho os resultados mostram que o **Random Forest Regressor** teve o melhor desempenho, seguido do **Decision Tree Regressor**, e por fim o **KNeighbors Regressor**. Já o **Support Vector Regression** obteve um desempenho pior do que uma simples média, isso pode ter ocorrido pois os dados podem não possuir um comportamento linear, ou o algoritmo não se ajustou bem a eles, bem como alguma configuração do modelo.

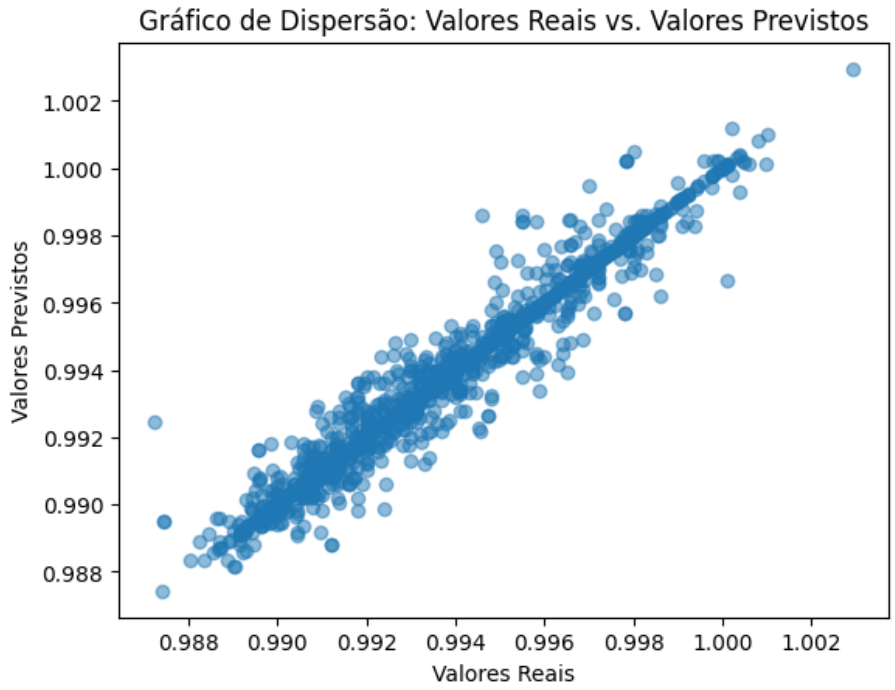

Figure 13. Dispersão do algoritmo Decision Tree Reegressor

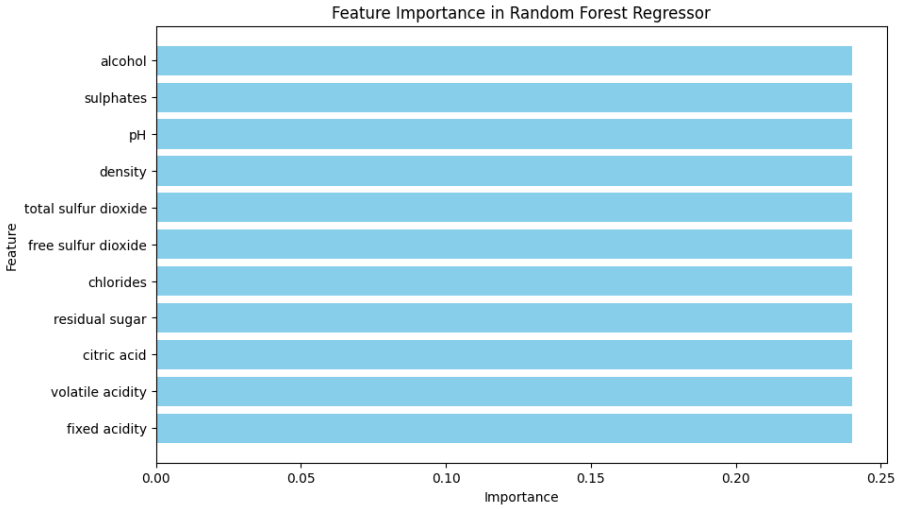

Figure 14. Importância das caracteristicas do Decision Tree Regressor

Em resumo, na análise de regressão simples, tanto o algoritmo implementado manualmente como o do Sklearn demonstraram um bom ajuste, mas o modelo da biblioteca possui erros reduzidos. Para a regressão múltipla, ambas as técnicas tiveram um excelente ajuste, mas o Sklearn também possui vantagem, com algumas métricas superiores. E para os algoritmos de testes, podemos perceber que o mais eficaz foi o Random Forest Regressor, enquanto o Support Vector Regression não encontrou compatibilidade com o conjunto de dados.

5. Conclusão

Este foi um estudo de caso para determinar os melhores algoritmos de regressão para modelar as variáveis da relação de produção de vinho, para isso, consideramos a regressão linear simples e a regressão múltipla, exploramos a implementação manual e também a biblioteca Sklearn para os dois modelos; também nos aprofundamos em algoritmos mais avançados da biblioteca, incluindo Random Forest, Decision Tree, KNeighbors e Support Vector Regression, para testes com várias variáveis. E resumindo, a nossa análise sustenta que a regressão múltipla é superior à regressão simples. Verificamos a importância de selecionar um modelo apropriado, pois trata-se de um passo fudamental para melhorar a precisão das previsões. E destacamos a superioridade de métodos de regressão múltiplos e algoritmos de conjunto como o Random Forest, que podem lidar com problemas de previsão complexos.

References

Junior, D. H. and Moreno, U. F. (2019). Estimaçao de pressao de fundo de poço utilizando svr e ukf. In *Congresso Brasileiro de Automática-CBA*, volume 1.

Kanaries (2023). k-nn (k nearest neighbor) regressão em python.

Pires, A. S., Trentin, G., Moraes, C. C., and da Silva Camargo, S. (2023). Modelos computacionais para predição da qualidade sensorial de vinhos a partir de características químicas. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Computational Wine Analysis*. Universidade Federal do Pampa (UNIPAMPA), Caixa Postal 242 – 96.413-170 – Bagé – RS – Brasil.

Revathy, G., Rajendran, V., Rashmika, B., Kumar, P. S., Parkavi, P., and Shynisha, J. (2022). Random forest regressor based superconductivity materials investigation for critical temperature prediction. *Materials Today: Proceedings*, 66:648–652.

Sachet, M. and de Avila e Silva, S. (2024). Aplicação de técnicas de mineração de dados na análise de processos de fermentação de vinhos. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Anais do Congresso de Mineração de Dados*. Universidade de Caxias do Sul (UCS), Alameda João Dal Sasso, 800 – 95705-266 – Bento Gonçalves – RS – Brasil.

Yao, B. (2023). Walmart sales prediction based on decision tree, random forest, and k neighbors regressor. *Highlights in Business, Economics and Management*, 5:330–335.