

# Comparação de técnicas de regressão Qualidade do Vinho



por Leonardo Reneres dos Santos





# Introdução

Neste projeto, buscou-se realizar uma análise comparativa de modelos de regressão aplicados ao \*Wine Quality Dataset\* da UCI Machine Learning Repository, que contém dados físico-químicos de vinhos. Utilizaram-se métricas de desempenho, como erro quadrático médio (MSE), erro absoluto médio (MAE) e coeficiente de determinação ( $R^2$ ), para avaliar a precisão dos modelos e identificar o mais adequado para prever de forma precisa a qualidade dos vinhos.



# Introdução

O trabalho se estrutura nos seguintes objetivos:

1. Avaliar a precisão de modelos de regressão, incluindo Regressão Linear, SVR, Random Forest e técnicas de Boosting, para prever a qualidade do vinho.
2. Comparar e interpretar as métricas de desempenho dos modelos.
3. Identificar vantagens e limitações de cada modelo em termos de complexidade, interpretabilidade e sensibilidade a ruídos e \*outliers\*.
4. Dada as características da base de dados utilizada, explorar os impactos provenientes do tamanho da base de dados utilizada em cada um dos algoritmos.



# Trabalhos relacionados

## **Trabalho principal (original e base outras situações) :**

Cortez, P., Antonio Luíz Cerdeira, Fernando Almeida, Telmo Matos and José Reis. “Modeling wine preferences by data mining from physicochemical properties.” Decis. Support Syst. 47 (2009): 547-553.

[https://www.semanticscholar.org/paper/Modeling-wine-preferences-by-data-mining-from-Cortez-Cerdeira/bf15a0ccc14ac1deb5cea570c870389c16be019c?utm\\_source=direct\\_link](https://www.semanticscholar.org/paper/Modeling-wine-preferences-by-data-mining-from-Cortez-Cerdeira/bf15a0ccc14ac1deb5cea570c870389c16be019c?utm_source=direct_link)

## **Outros trabalhos (usam o trabalho original como citação):**

- Angus, D. C.. “Modeling Wine Quality from Physicochemical Properties.” (2019).
  - [https://www.semanticscholar.org/paper/Modeling-Wine-Quality-from-Physicochemical-Angus/f9c457828e4e26ab2ae6f0f9a4cea66c98767df6?utm\\_source=direct\\_link](https://www.semanticscholar.org/paper/Modeling-Wine-Quality-from-Physicochemical-Angus/f9c457828e4e26ab2ae6f0f9a4cea66c98767df6?utm_source=direct_link)
- Agyemang, Perpetual O.. “Modeling the Preference of Wine Quality Using Logistic Regression Techniques Based on Physicochemical Properties.” (2010).
  - [https://www.semanticscholar.org/paper/Modeling-the-Preference-of-Wine-Quality-Using-Based-Agyemang/a5c6f899b1ac4b57805102252be02ddc8ec2b5c2?utm\\_source=direct\\_link](https://www.semanticscholar.org/paper/Modeling-the-Preference-of-Wine-Quality-Using-Based-Agyemang/a5c6f899b1ac4b57805102252be02ddc8ec2b5c2?utm_source=direct_link)
- Nebot, Àngela, Francisco Mugica and Antoni Escobet. “Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques.” International Conference on Simulation and Modeling Methodologies, Technologies and Applications (2015).
  - [https://www.semanticscholar.org/paper/Modeling-Wine-Preferences-from-Physicochemical-Nebot-Mugica/e7c34d5b766df595105a9732355bb7dfb0f1dada?utm\\_source=direct\\_link](https://www.semanticscholar.org/paper/Modeling-Wine-Preferences-from-Physicochemical-Nebot-Mugica/e7c34d5b766df595105a9732355bb7dfb0f1dada?utm_source=direct_link)



# Fundamentos

O projeto utiliza o *\*Wine Quality Dataset\** do repositório da UCI Machine Learning, que contém 6497 amostras, sendo 1599 de vinhos tintos e 4898 de vinhos brancos. Cada amostra é caracterizada por 11 variáveis físico-químicas (como acidez, teor alcoólico e pH) e uma variável de saída que representa a qualidade do vinho em uma escala de 0 a 10.

A implementação foi realizada em Python, utilizando bibliotecas como scikit-learn, pandas, numpy e matplotlib para visualização dos resultados.



# Apresentação da Base de Dados

Base de dados Wine Quality - contém informações sobre a qualidade do vinho.

Variável	Descrição
fixed acidity	Acidez fixa do vinho.
volatile acidity	Acidez volátil do vinho.
citric acid	Ácido cítrico do vinho.
residual sugar	Açúcar residual no vinho.
chlorides	Cloreto no vinho.
free sulfur dioxide	Dióxido de enxofre livre no vinho.
total sulfur dioxide	Dióxido de enxofre total no vinho.
density	Densidade do vinho.
pH	pH do vinho.
sulphates	Sulfato no vinho.
alcohol	Teor alcoólico do vinho.
quality	Nota de qualidade do vinho.



# Fundamentos

Para avaliar o desempenho dos modelos de regressão aplicados, são usadas métricas como:

- \*\*Erro Quadrático Médio (MSE)\*\*: mede o erro médio ao quadrado entre as previsões e valores reais.
- \*\*Erro Absoluto Médio (MAE)\*\*: indica a magnitude média do erro entre previsões e valores observados.
- \*\*Coeficiente de Determinação ( $R^2$ )\*\*: indica a proporção da variabilidade explicada pelo modelo.
- Time: medida de tempo usada para comparação entre os algoritmos





# Fundamentos

Em busca de melhores comparações além das implementações dos algoritmos em sua forma pura, foram realizados testes com:

Tunning dos melhores modelos em busca da melhor eficacia de resulatdos

Validação cruzada \*k-fold\* com  $k=5$ , dividindo os dados em cinco partes para uma avaliação mais estável e menos sujeita a \*overfitting\*.

Implementação com Modelos híbridos regressão:

Os modelos são:

Aqui estão algumas combinações que usaremos:

1. Random Forest + Gradient Boosting (meta: Linear Regression)
1. Random Forest + Gradient Boosting (meta: Ridge Regression)
1. Random Forest + AdaBoost (meta: Linear Regression)
1. Gradient Boosting + AdaBoost (meta: Ridge Regression)



# Metodologia

## Definição do objetivo/Seleção dos dados/Limpeza dos dados.

Objetivo : Avaliação e análise comparativa e detalhada quanto ao desempenho de diversos modelos de regressão aplicados à predição da qualidade de vinhos.

Seleção da base de dados :  
\*Wine Quality Dataset\* devida a sua característica de subdivisão : 6497 amostras, sendo 1599 de vinhos tintos e 4898 de vinhos brancos.

## Aplicação das técnicas de mineração.

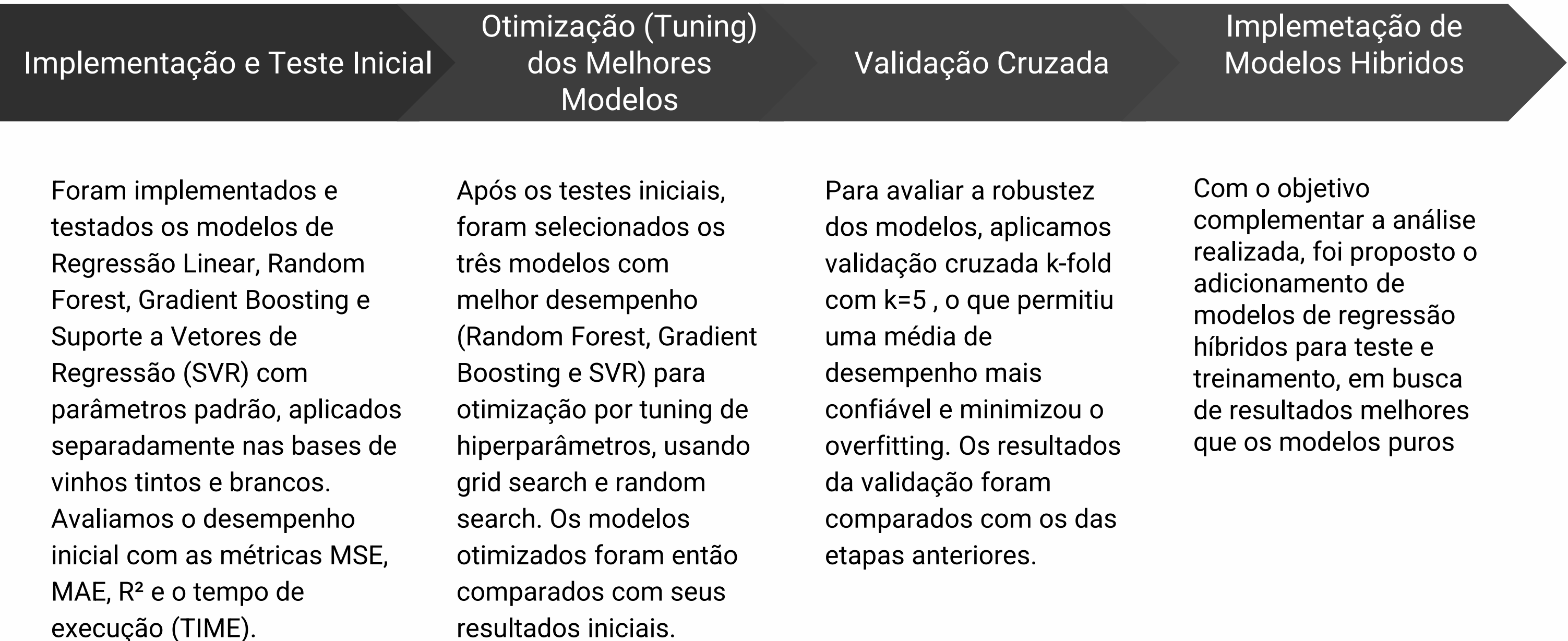
Foram implementadas 4 formas de aplicação de técnicas :

- Algoritmos puros.
- Tuning dos melhores modelos.
- Croos Validation
- Modelos Híbridos.

## Avaliação/Comparação dos resultados obtidos.

Os resultados de cada fase – teste inicial, tuning, validação cruzada e Modelos Híbridos – foram comparados usando MSE, MAE,  $R^2$  e TIME, identificando o modelo com melhor desempenho geral e o impacto das etapas de otimização e validação.

# Metodologia

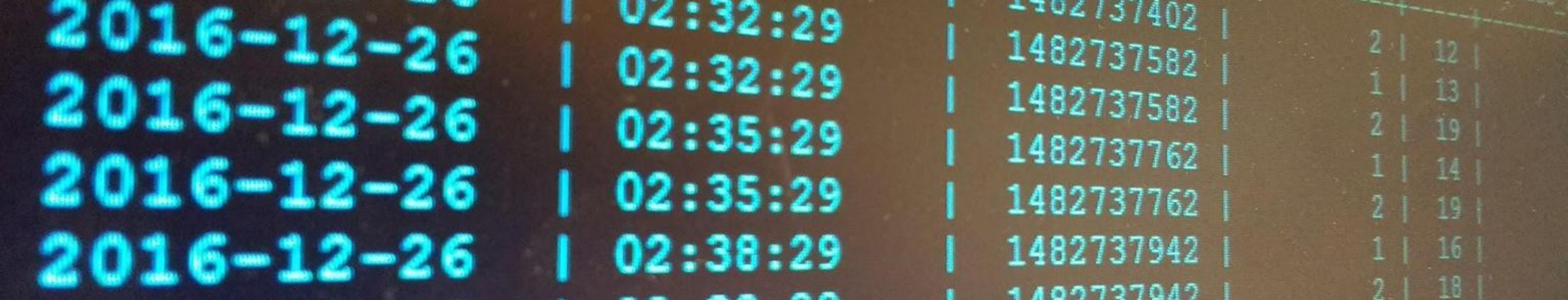






# Implementação e Resultados





2016-12-26	02:32:29	1482737402	2	12
2016-12-26	02:32:29	1482737582	1	13
2016-12-26	02:35:29	1482737582	2	19
2016-12-26	02:35:29	1482737762	1	14
2016-12-26	02:35:29	1482737762	2	19
2016-12-26	02:38:29	1482737942	1	16
		1482737942	2	18

# Pré-Processamento e Limpeza dos Dados

Preparar os dados para a análise de regressão.

1

Alteração do tipo dos dados

Com a importação, foi necessária a alteração do tipo dos dados de objeto para real (float)

2

Verificação de valores faltantes

Não foi necessária a exclusão de valores faltantes

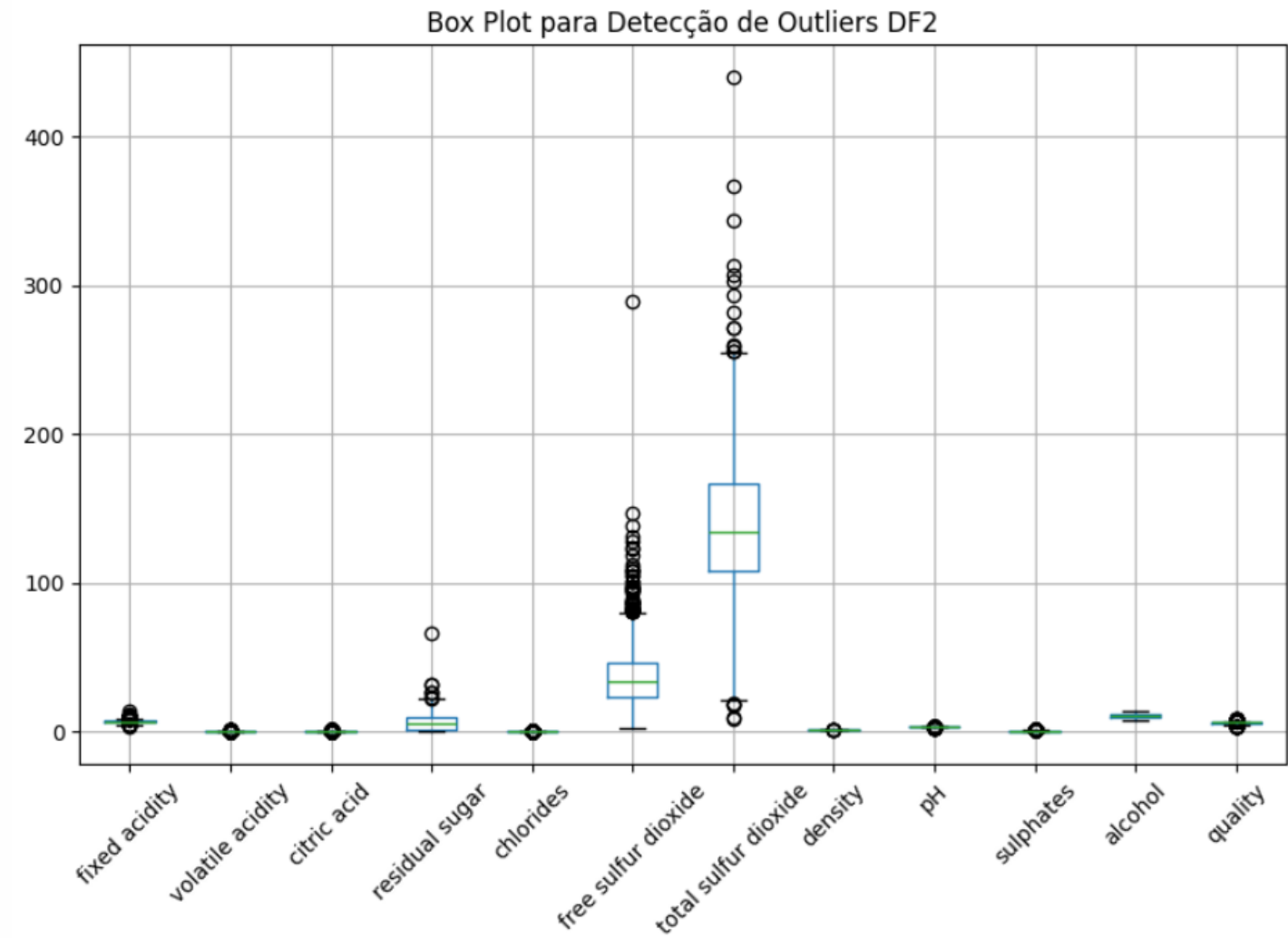
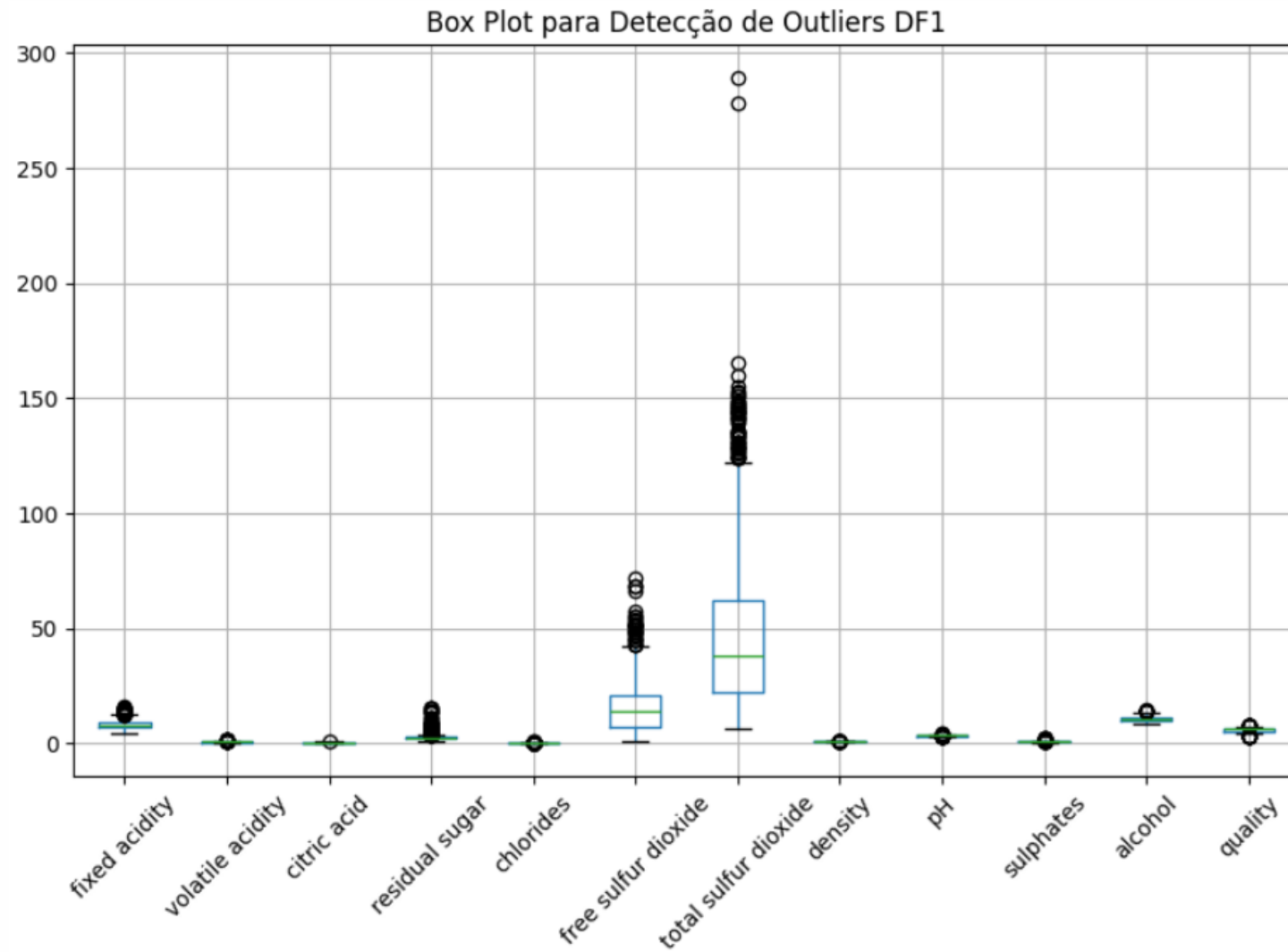
3

Remoção de outliers

Converter dados para o formato adequado para a análise

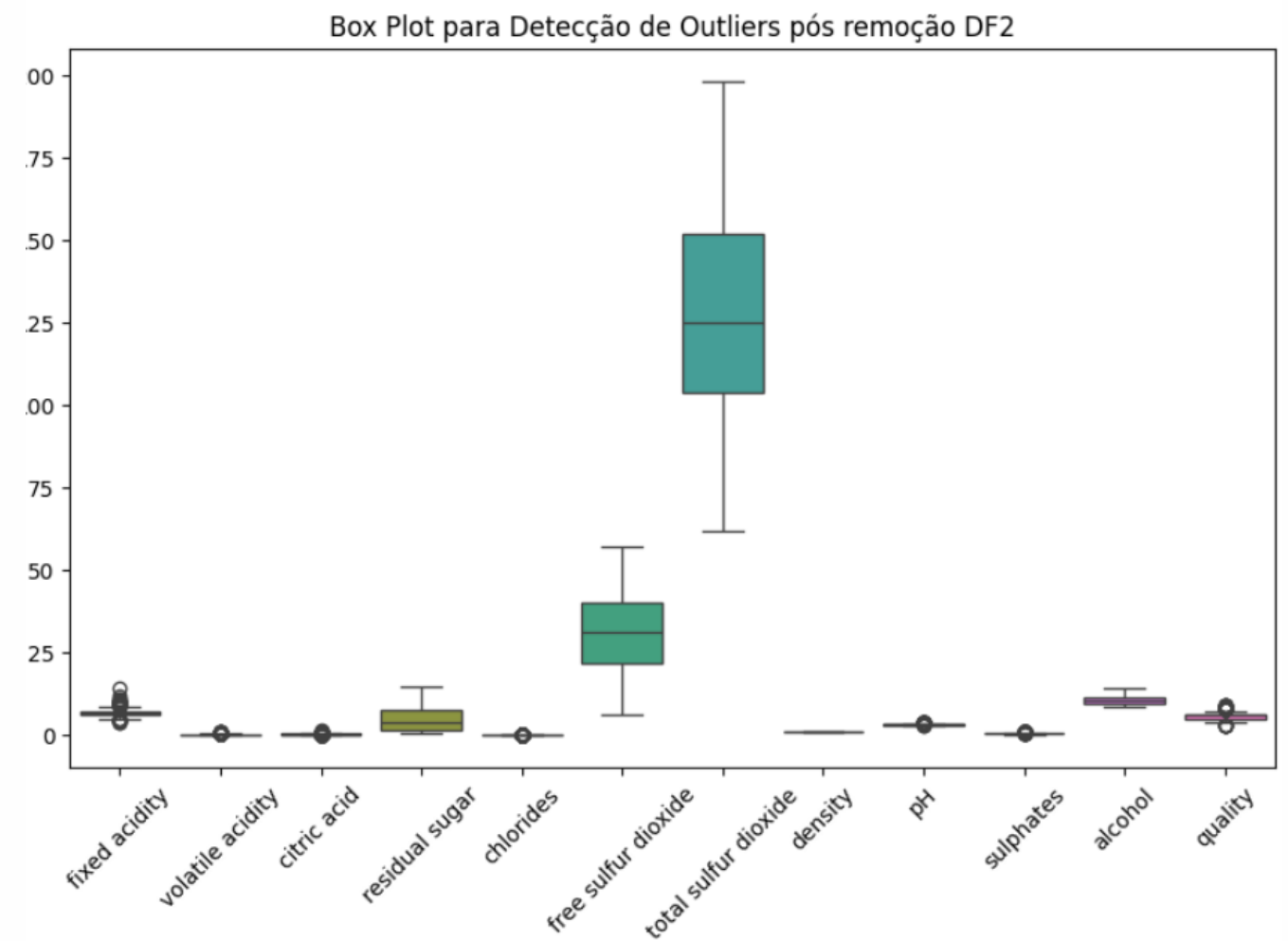
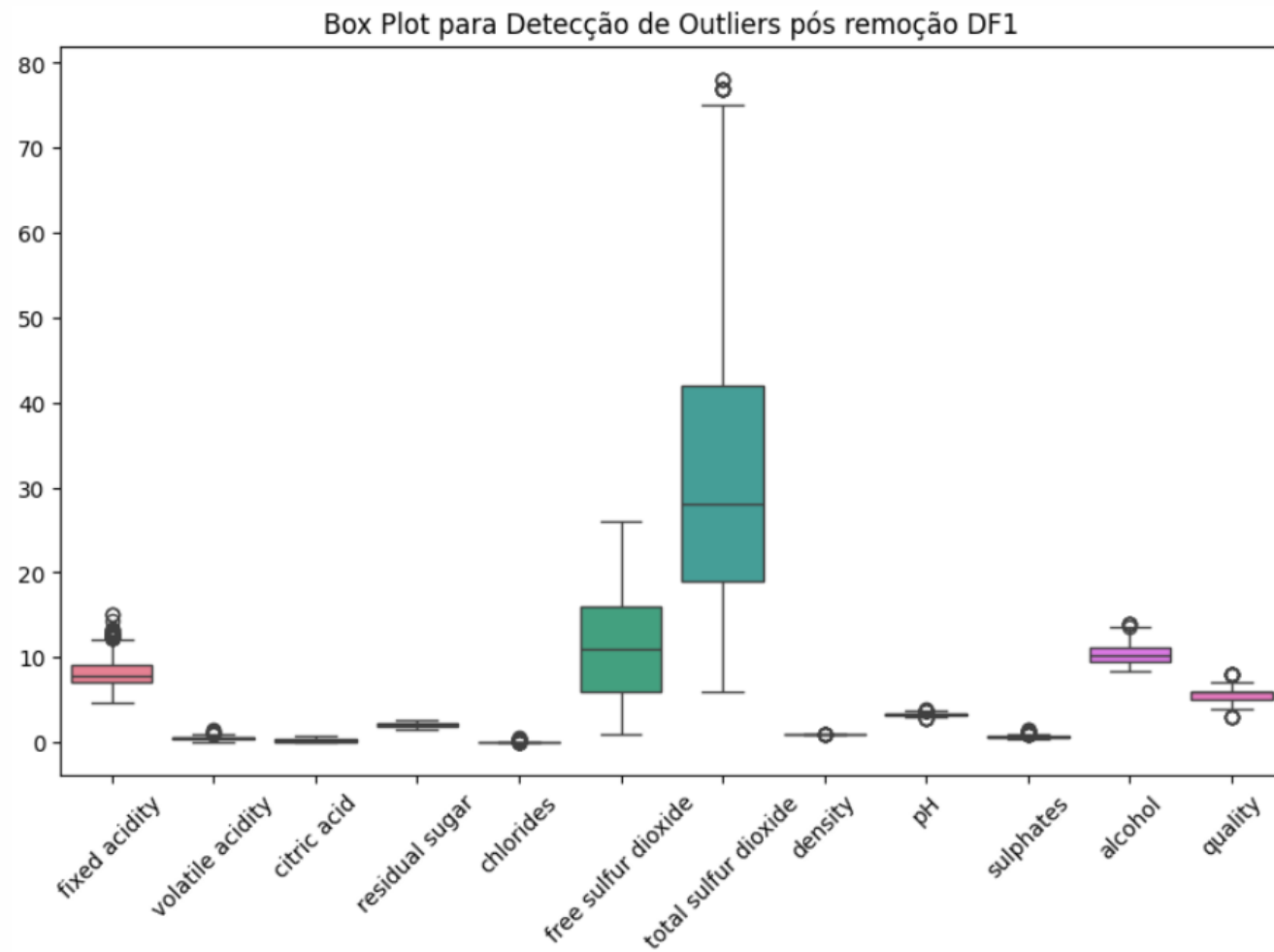
# Remoção de outliers

Antes



# Remoção de outliers

Depois





# Normalização/Escalonamento de Variáveis numéricas

↔

Dados normalizados DF1:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
1	0.269231	0.397260	0.000000	0.272727	0.066318	
2	0.307692	0.520548	0.000000	0.909091	0.104712	
3	0.307692	0.438356	0.052632	0.636364	0.094241	
4	0.634615	0.109589	0.736842	0.272727	0.064572	
5	0.269231	0.397260	0.000000	0.272727	0.066318	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
1	0.40	0.388889	0.685039	0.625000	0.182540	
2	0.96	0.847222	0.586614	0.326923	0.277778	
3	0.56	0.666667	0.606299	0.384615	0.253968	
4	0.64	0.750000	0.704724	0.288462	0.198413	
5	0.40	0.388889	0.685039	0.625000	0.182540	

	alcohol	quality
1	0.178571	0.4
2	0.250000	0.4
3	0.250000	0.4
4	0.250000	0.6
5	0.178571	0.4

↔

Dados normalizados DF2:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
1	0.269231	0.397260	0.000000	0.272727	0.066318	
2	0.307692	0.520548	0.000000	0.909091	0.104712	
3	0.307692	0.438356	0.052632	0.636364	0.094241	
4	0.634615	0.109589	0.736842	0.272727	0.064572	
5	0.269231	0.397260	0.000000	0.272727	0.066318	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
1	0.40	0.388889	0.685039	0.625000	0.182540	
2	0.96	0.847222	0.586614	0.326923	0.277778	
3	0.56	0.666667	0.606299	0.384615	0.253968	
4	0.64	0.750000	0.704724	0.288462	0.198413	
5	0.40	0.388889	0.685039	0.625000	0.182540	

	alcohol	quality
1	0.178571	0.4
2	0.250000	0.4
3	0.250000	0.4
4	0.250000	0.6
5	0.178571	0.4

# Aplicação de Técnicas de Regressão Iniciais

Resultados para df1:

	MSE	MAE	R2	Time
Linear Regression	0.015127	0.099742	0.471607	0.027801
SVR	0.015227	0.100440	0.468109	0.071398
Random Forest	0.013461	0.084473	0.529821	0.542246
Gradient Boosting	0.015179	0.095235	0.469785	0.226127

Resultados para df2:

	MSE	MAE	R2	Time
Linear Regression	0.016421	0.101326	0.303542	0.005392
SVR	0.014714	0.096663	0.375961	0.376163
Random Forest	0.011536	0.076035	0.510738	2.041423
Gradient Boosting	0.014849	0.094340	0.370217	0.629812

## Análise dos resultados iniciais

### Base de dados do vinho tinto = df1

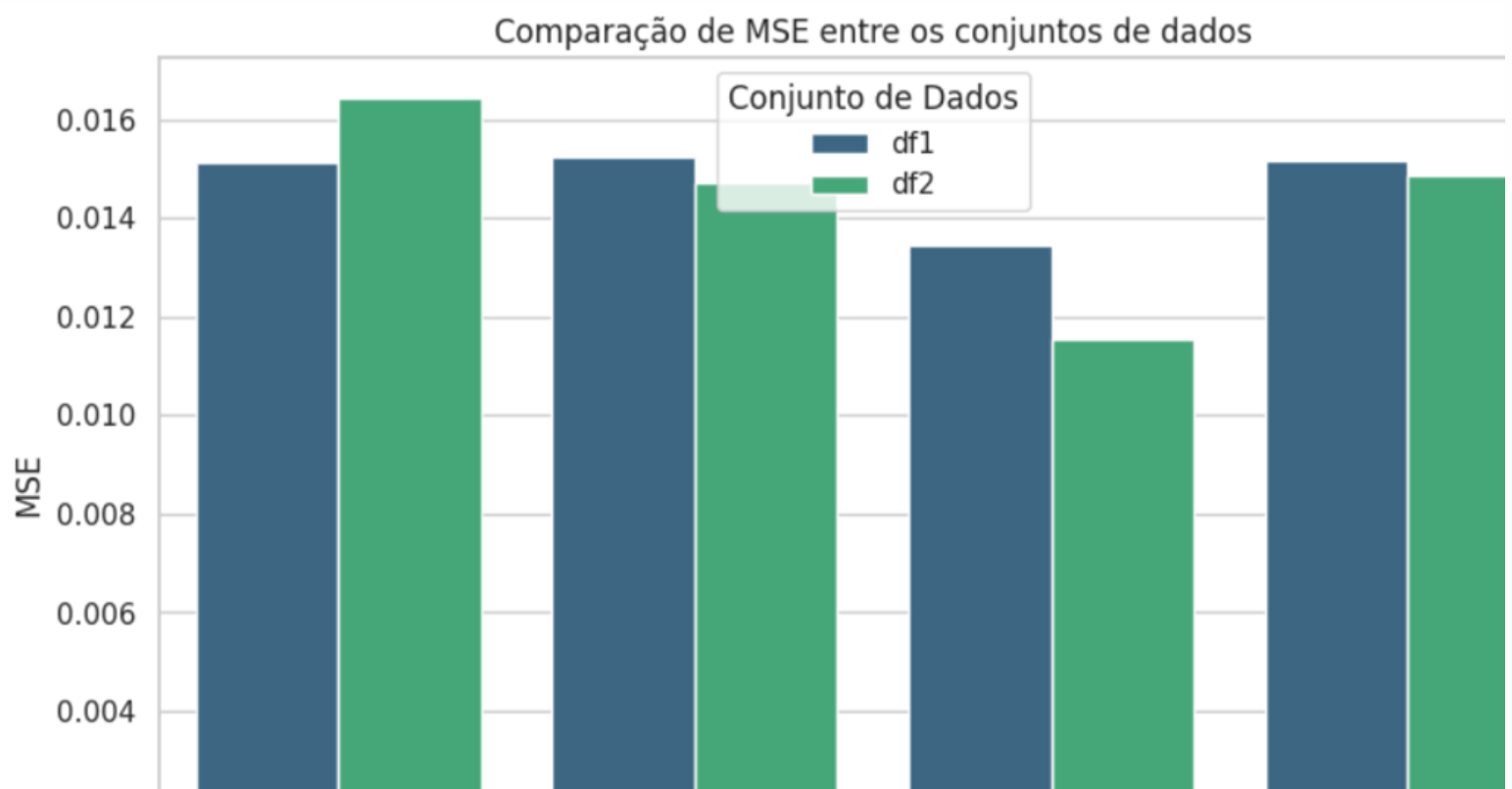
Para primeira base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, uma vez que esta teve os **menores valores** de MSE (Erro Quadrático Médio) e MAE (erro médio absoluto) e um maior valor de R2 (Coeficiente de Determinação)

### Base de dados do vinho branco = df2

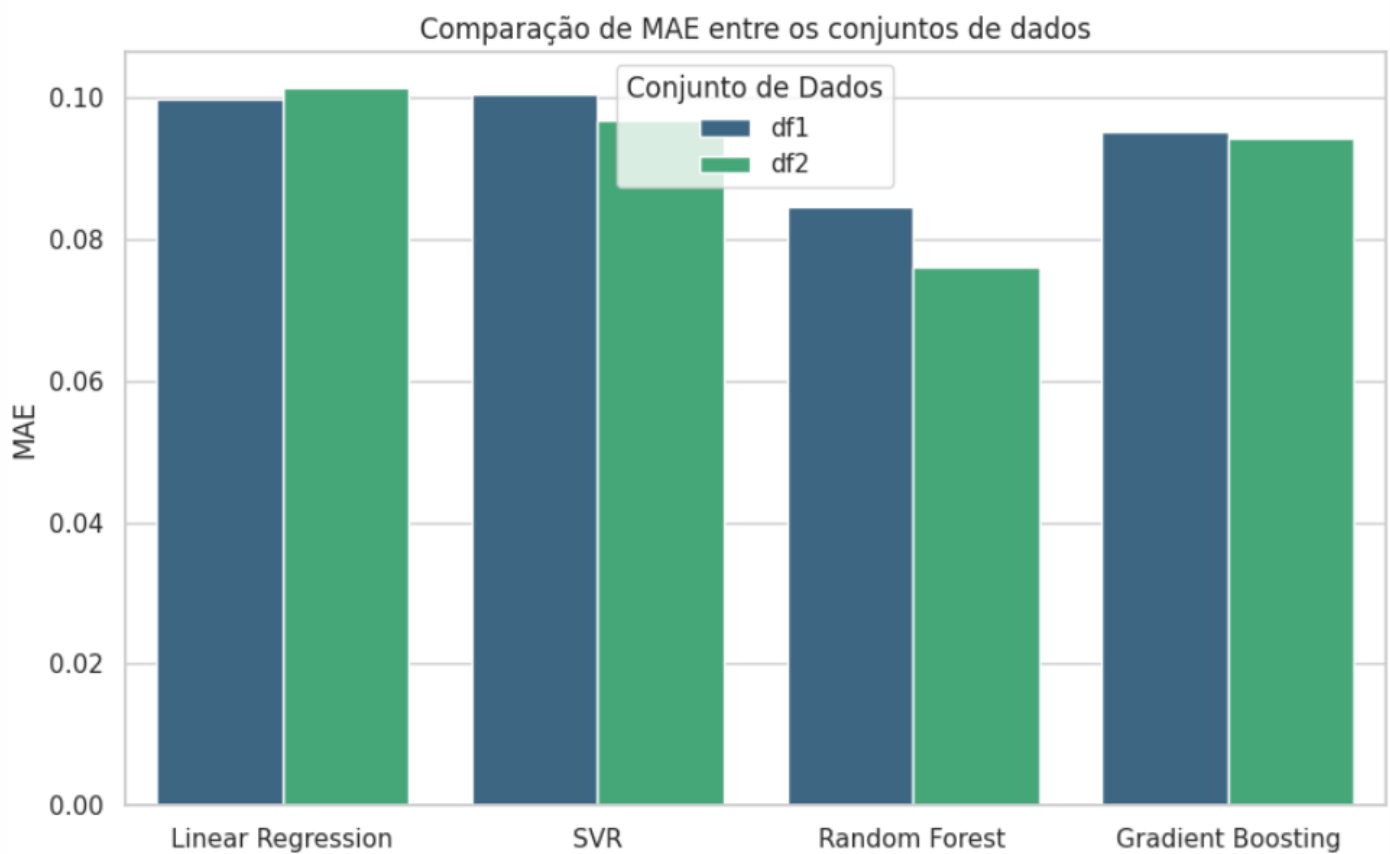
Também para segunda base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, uma vez que esta teve os **menores valores** de MSE (Erro Quadrático Médio) e MAE (erro médio absoluto) e um **maior valor** de R2 (Coeficiente de Determinação)

Portanto em situações normais o Modelo Random forest se mostra mais eficiente e ambas as situações

# Aplicação de Técnicas de Regressão Iniciais

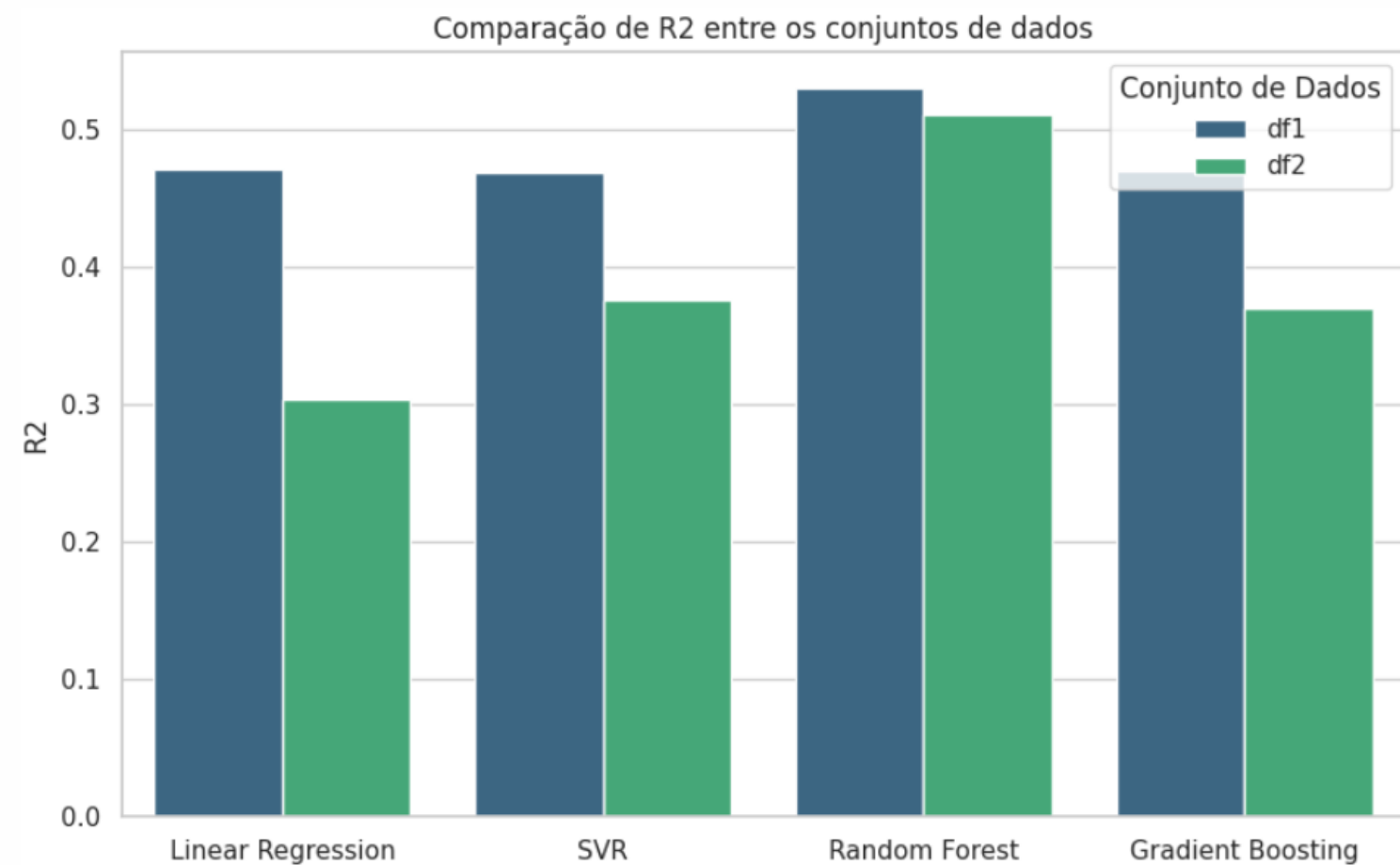
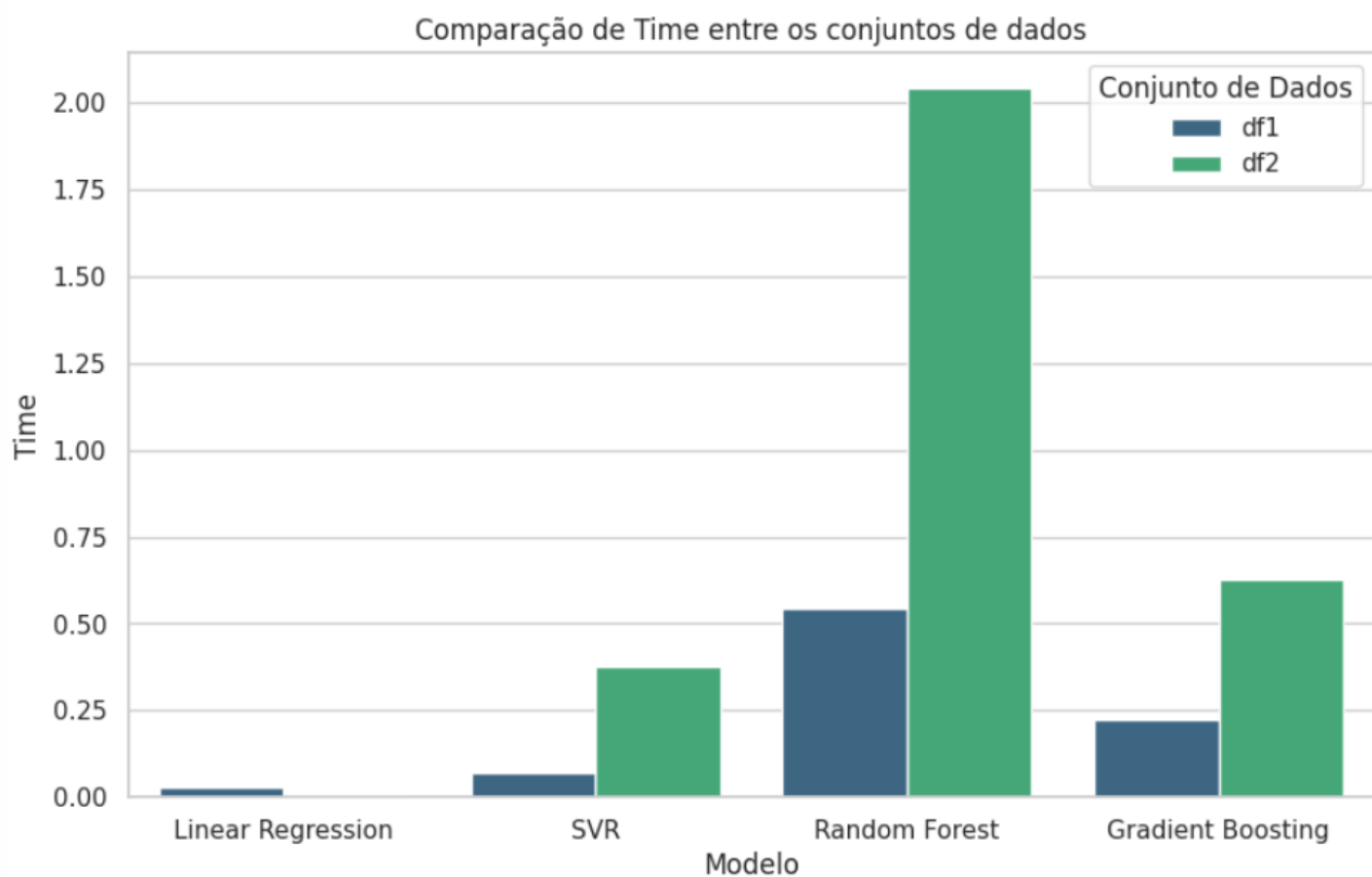


**DF1 = Base de dados Vinho Tinto**  
**DF2 = Base de dados Vinho Branco**





# Aplicação de Técnicas de Regressão Iniciais



## Análise dos resultados após a tunagem

### Base de dados do vinho tinto = df1

Para primeira base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, uma vez que esta teve os **menores valores** de MSE(Erro Quadrático Médio) e MAE (erro médio absoluto) e um **maior valor de R2** (Coeficiente de Determinação)

Todavia devido a seu **alto tempo** de treinamento e teste ser longo, em aplicações em que o tempo de execução seja um fator importante indicado seria o **SRV com Tunagem**

### Base de dados do vinho branco = df2

Já para segunda base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, com uma leve vantagem sobre o Gradient Boostion uma vez que esta teve os **menores valores** de MSE(Erro Quadrático Médio) e MAE (erro médio absoluto) e um **maior valor de R2** (Coeficiente de Determinação)

Todavia este sofre o mesmo problema do Random Forest quando ao seu tempo de execução, sendo extremamente lento.

Portanto é necessária interpretação do utilizador da ferramenta para determinar qual o melhor algoritmo que pode ser aplicado para seu objetivo.

# Tunagem dos Parâmetros

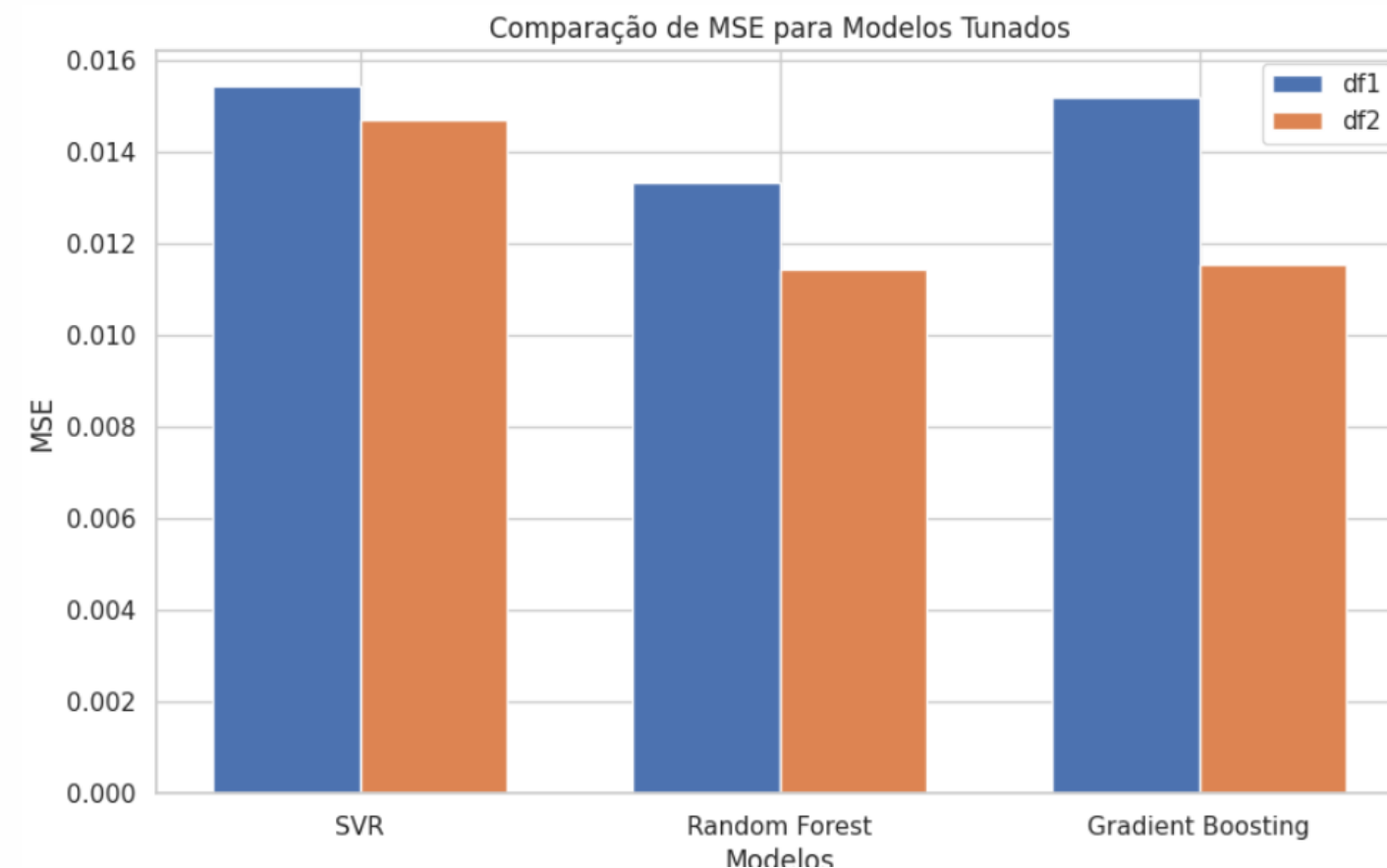
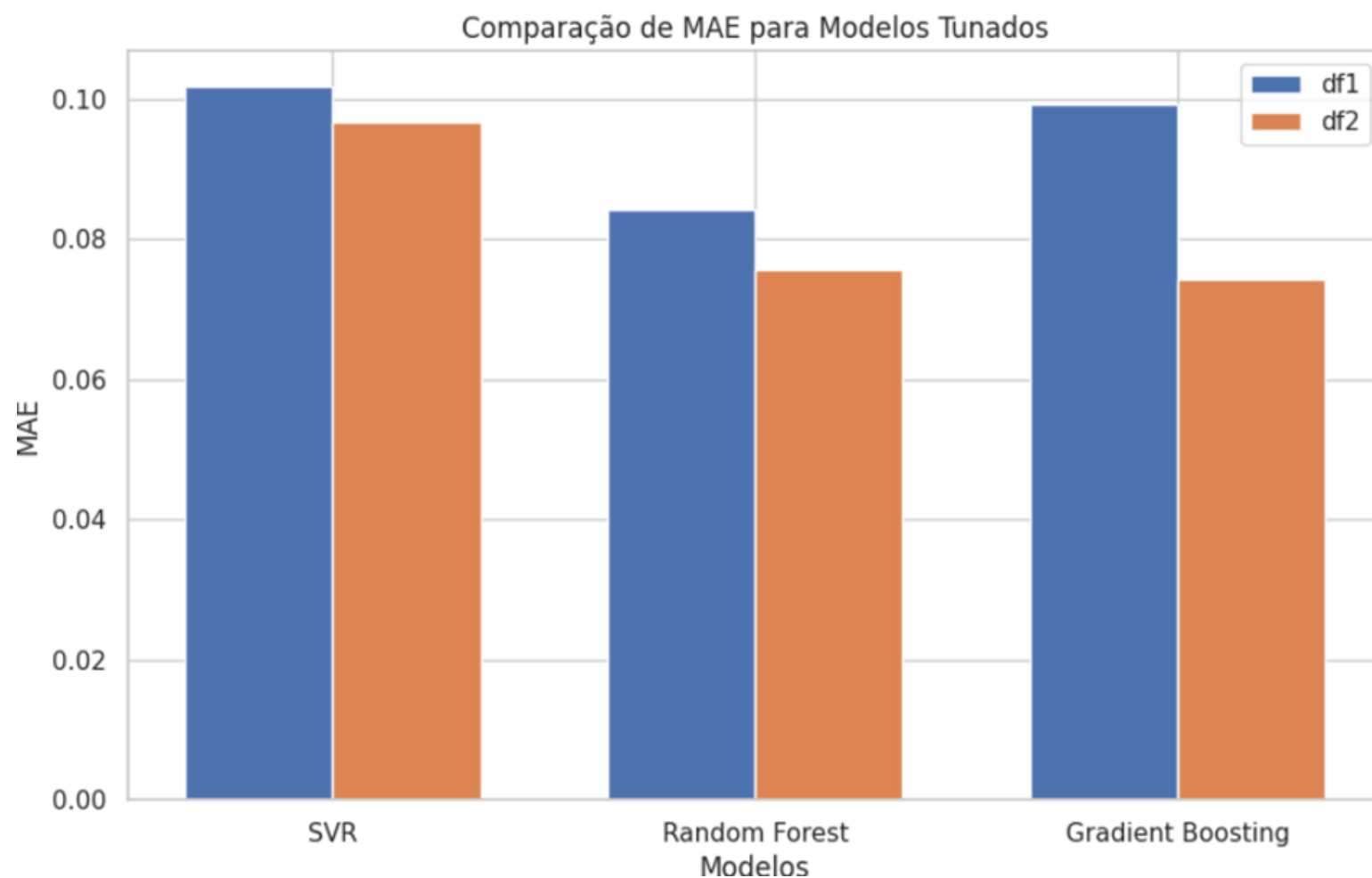
Assim como especificado, foram escolhidos os modelos de SVR, Random Forest e Gradient Boosting para a avaliação da tunagem de parâmetros

Resultados após tunagem para df1:					
Modelo	Melhor Modelo	MSE	MAE	R²	Tempo (s)
SVR	SVR(C=10, kernel='linear')	0.015443	0.101737	0.460562	3.032590
Random Forest	DecisionTreeRegressor(max_features=1.0, random_state=42)	0.013358	0.084333	0.533409	12.062052
Gradient Boosting	DecisionTreeRegressor(criterion='friedman_mse')	0.015194	0.099294	0.469265	22.042699

Resultados após tunagem para df2:					
Modelo	Melhor Modelo	MSE	MAE	R²	Tempo (s)
SVR	SVR(C=1)	0.014714	0.096663	0.375961	10.095348
Random Forest	DecisionTreeRegressor(max_features=1.0, random_state=42)	0.011468	0.075741	0.513625	41.716616
Gradient Boosting	DecisionTreeRegressor(criterion='friedman_mse')	0.011551	0.074339	0.510099	64.207976

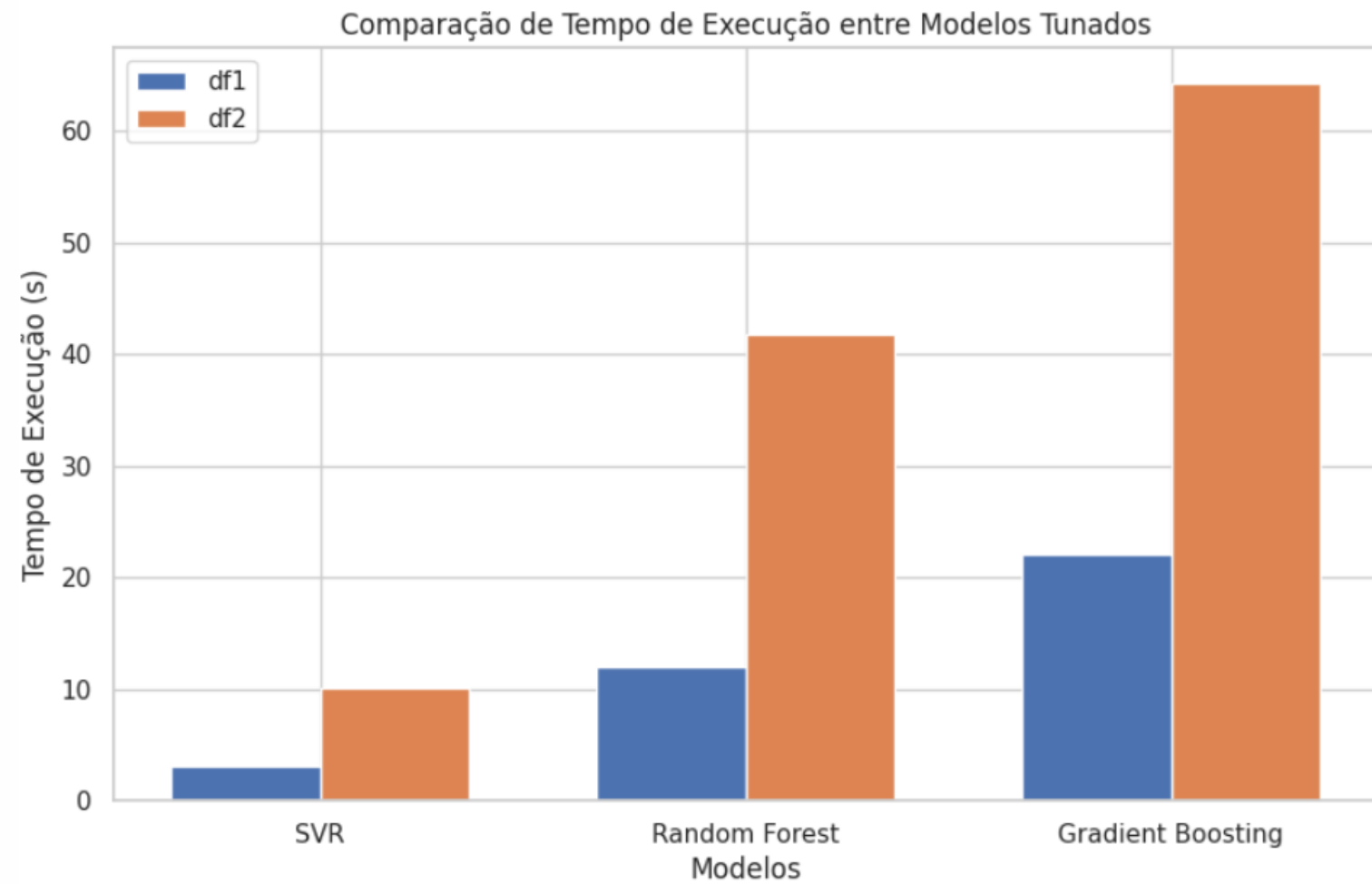
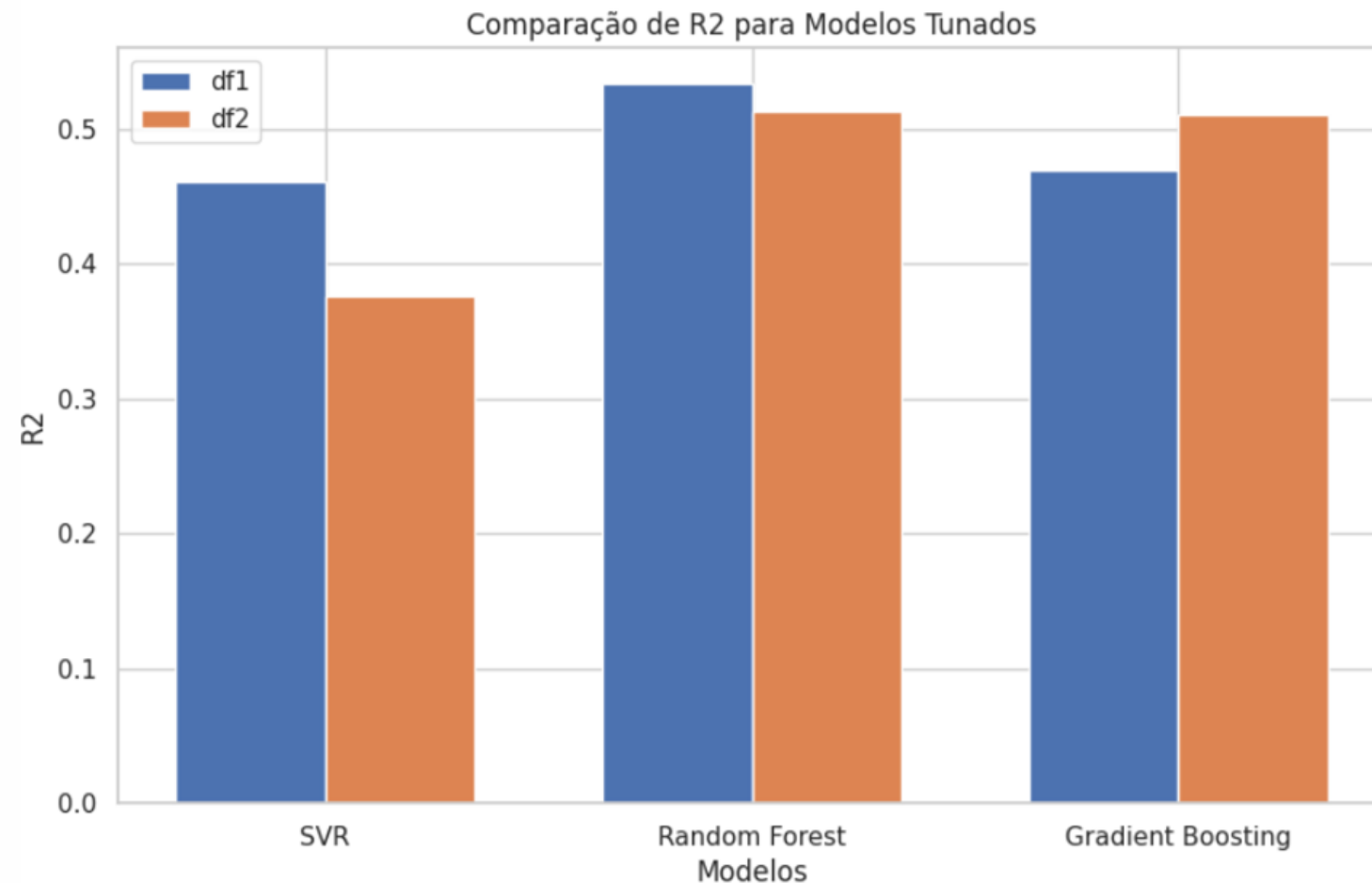
# Aplicação de Técnicas de Regressão: Tunning

**DF1** = Base de dados Vinho Tinto  
**DF2** = Base de dados Vinho Branco





# Aplicação de Técnicas de Regressão : Tunning



# Modelo de Técnicas de Regressão: Cross Validation

Resultados para Vinhos Tintos:				
Modelo	MSE	MAE	R²	Tempo (s)
Linear Regression	0.015127	0.099742	0.471607	0.003212
SVR	0.014842	0.095313	0.481579	0.072047
Random Forest	0.013498	0.084495	0.528524	0.855595
Gradient Boosting	0.015037	0.094842	0.474761	0.430642

Resultados para Vinhos Brancos:				
Modelo	MSE	MAE	R²	Tempo (s)
Linear Regression	0.016421	0.101326	0.303542	0.002931
SVR	0.014287	0.093991	0.394056	0.879670
Random Forest	0.011528	0.076149	0.511091	3.132886
Gradient Boosting	0.014840	0.094284	0.370607	0.634997

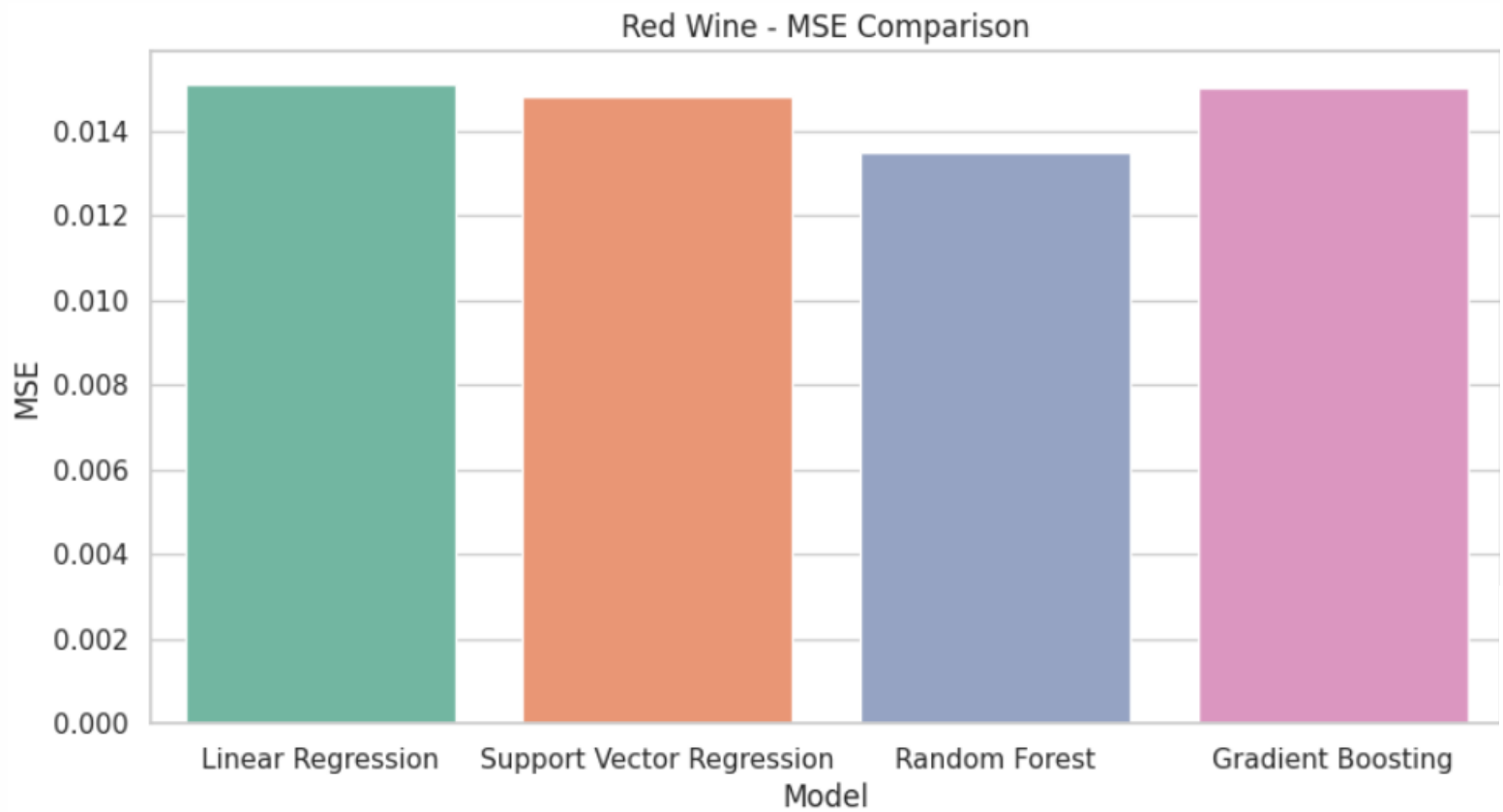
## Base de dados do vinho tinto = df1

Para primeira base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, uma vez que esta teve os menores valores de MSE(Erro Quadrático Médio) e MAE (erro médio absoluto) e um maior valor de R2 (Coeficiente de Determinação)

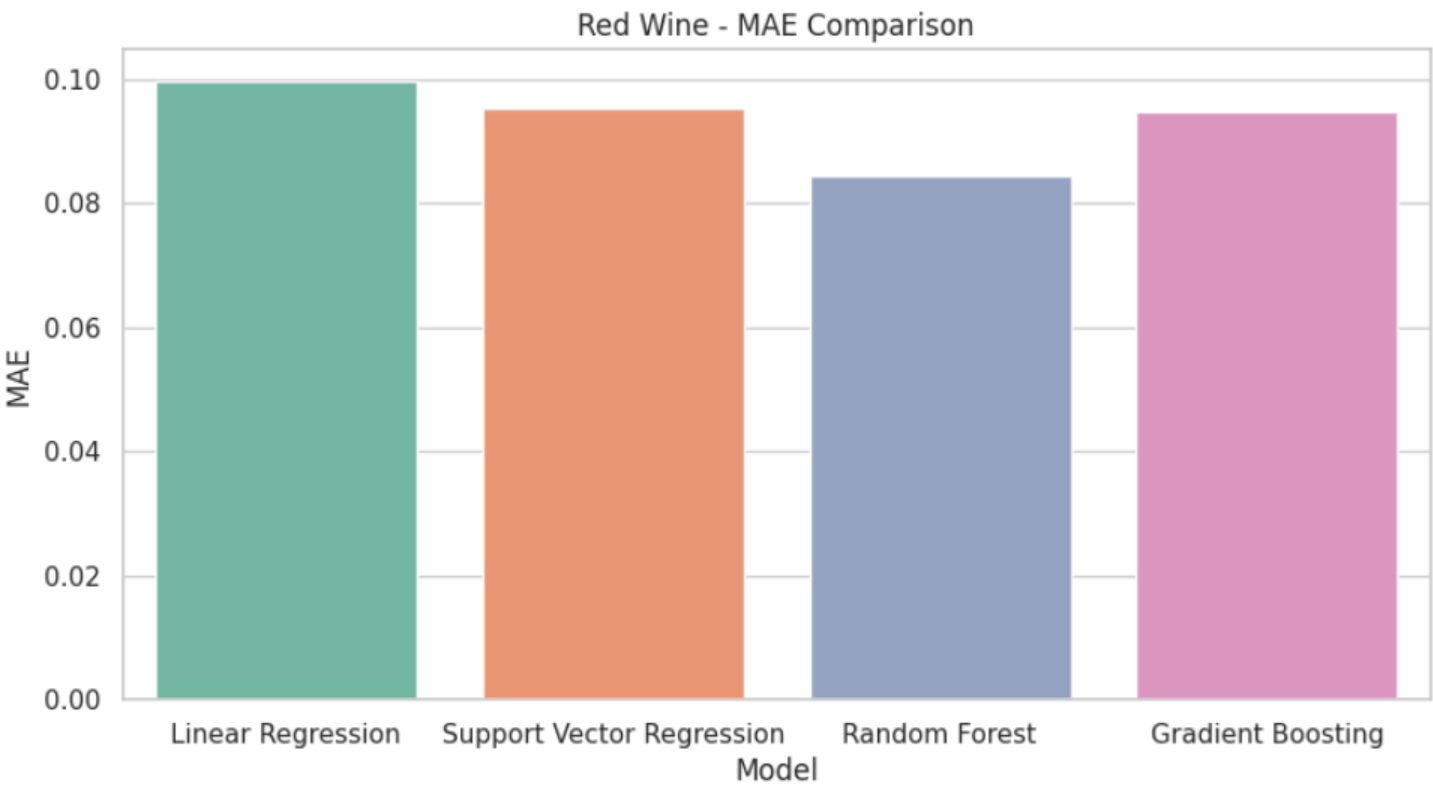
## Base de dados do vinho branco = df2

Do mesmo modo para segunda base de dados o modelo que apresentou melhor desempenho foi a **Random Forest**, uma vez que esta teve os menores valores de MSE(Erro Quadrático Médio) e MAE (erro médio absoluto) e um maior valor de R2 (Coeficiente de Determinação)

# Aplicação de Técnicas de Regressão: Cross

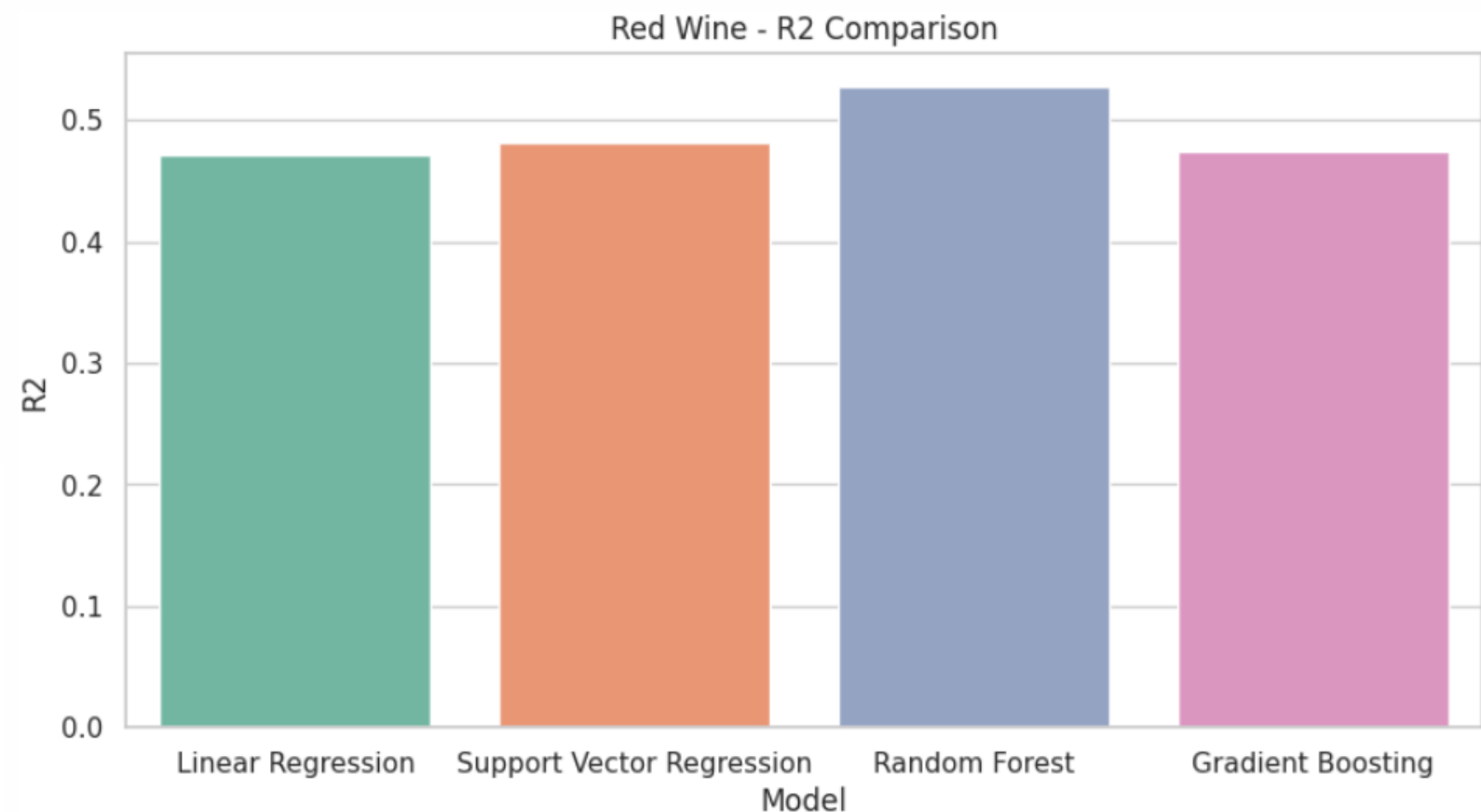


DF1 = Base de dados Vinho Tinto





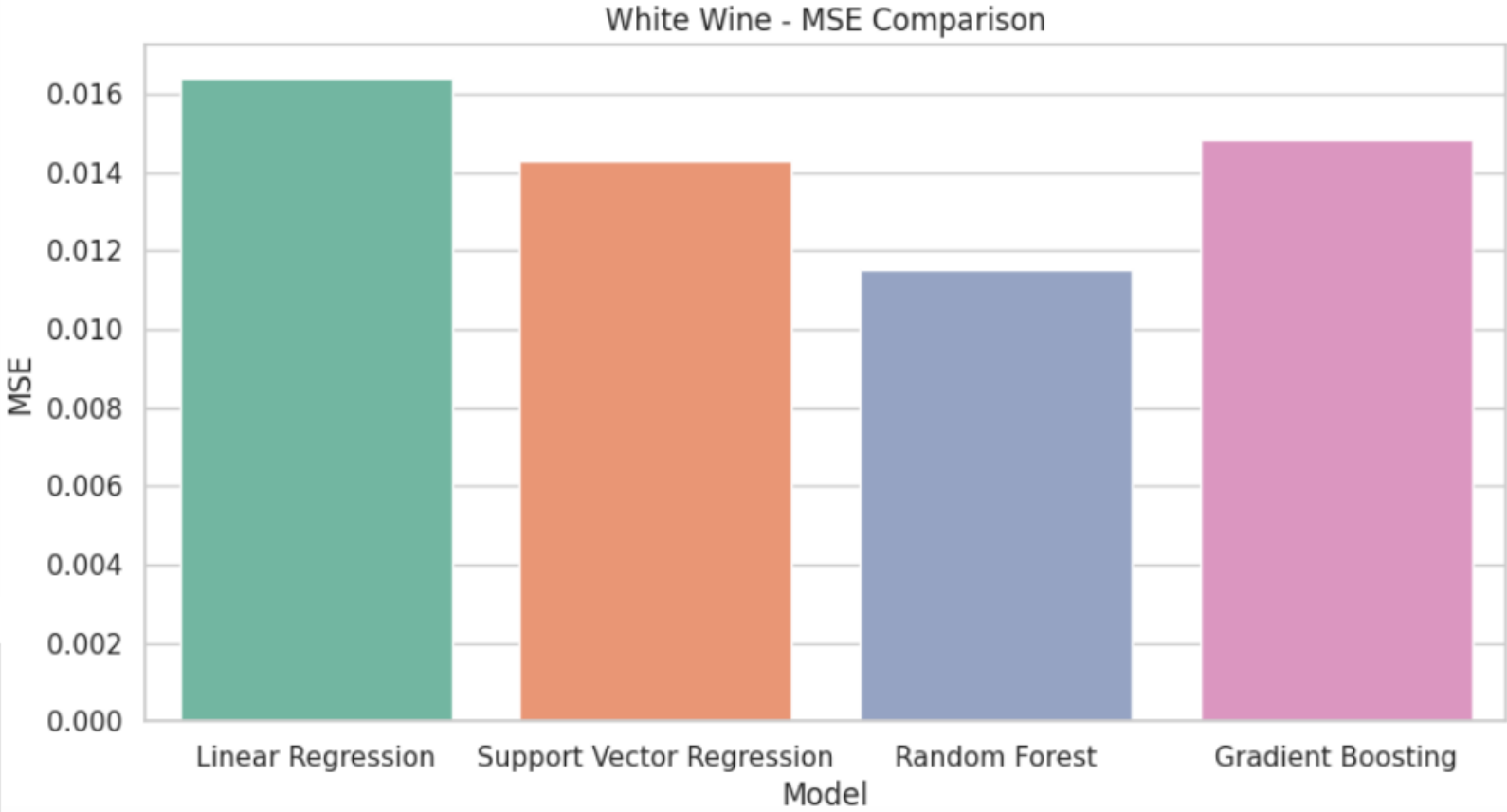
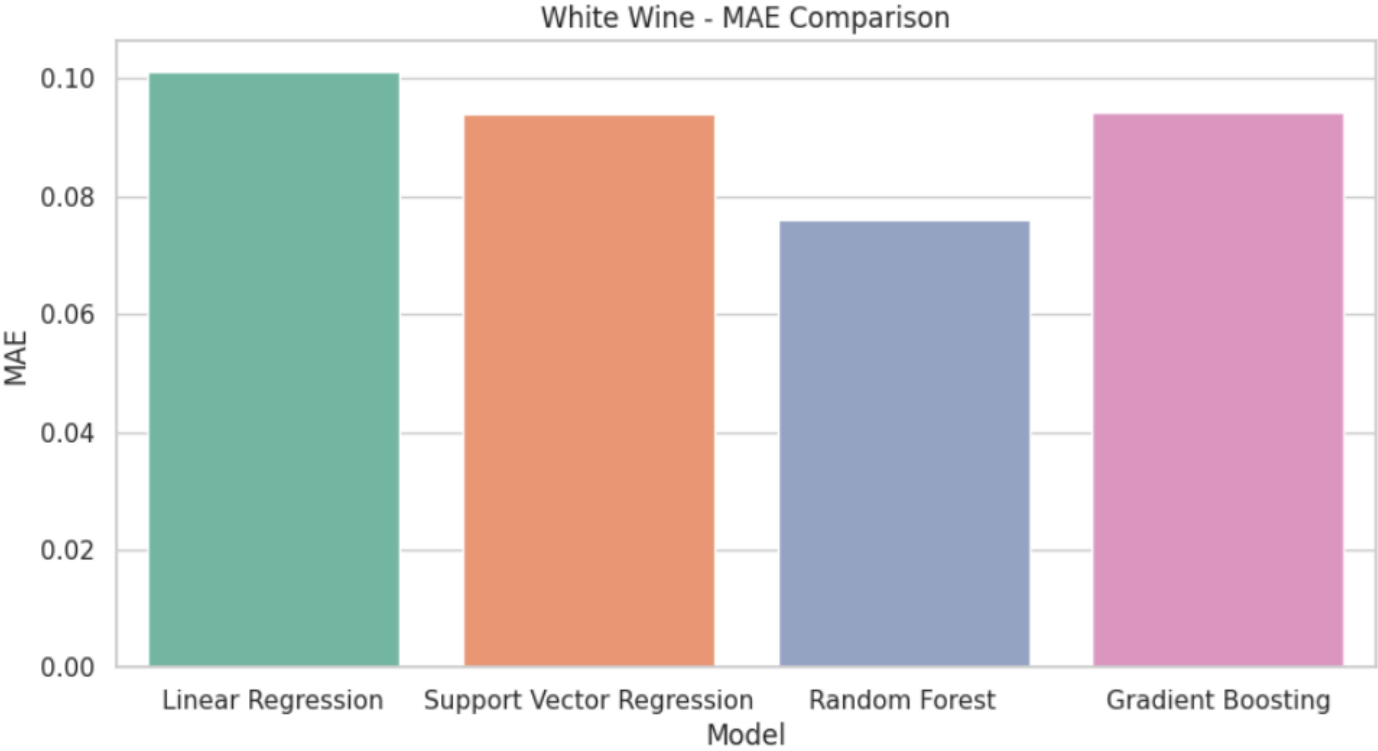
# Aplicação de de Regressão Logística: Cross Validation



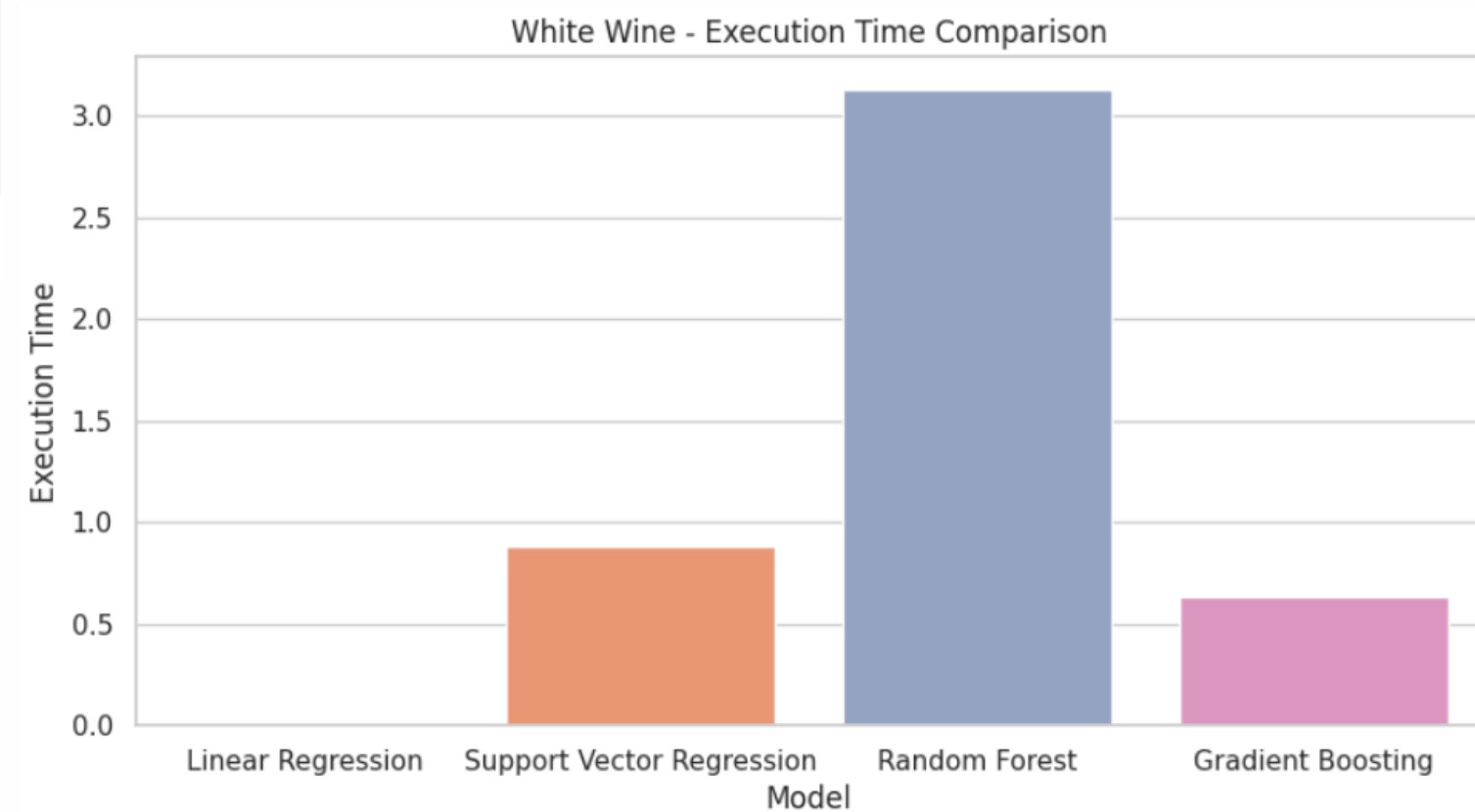
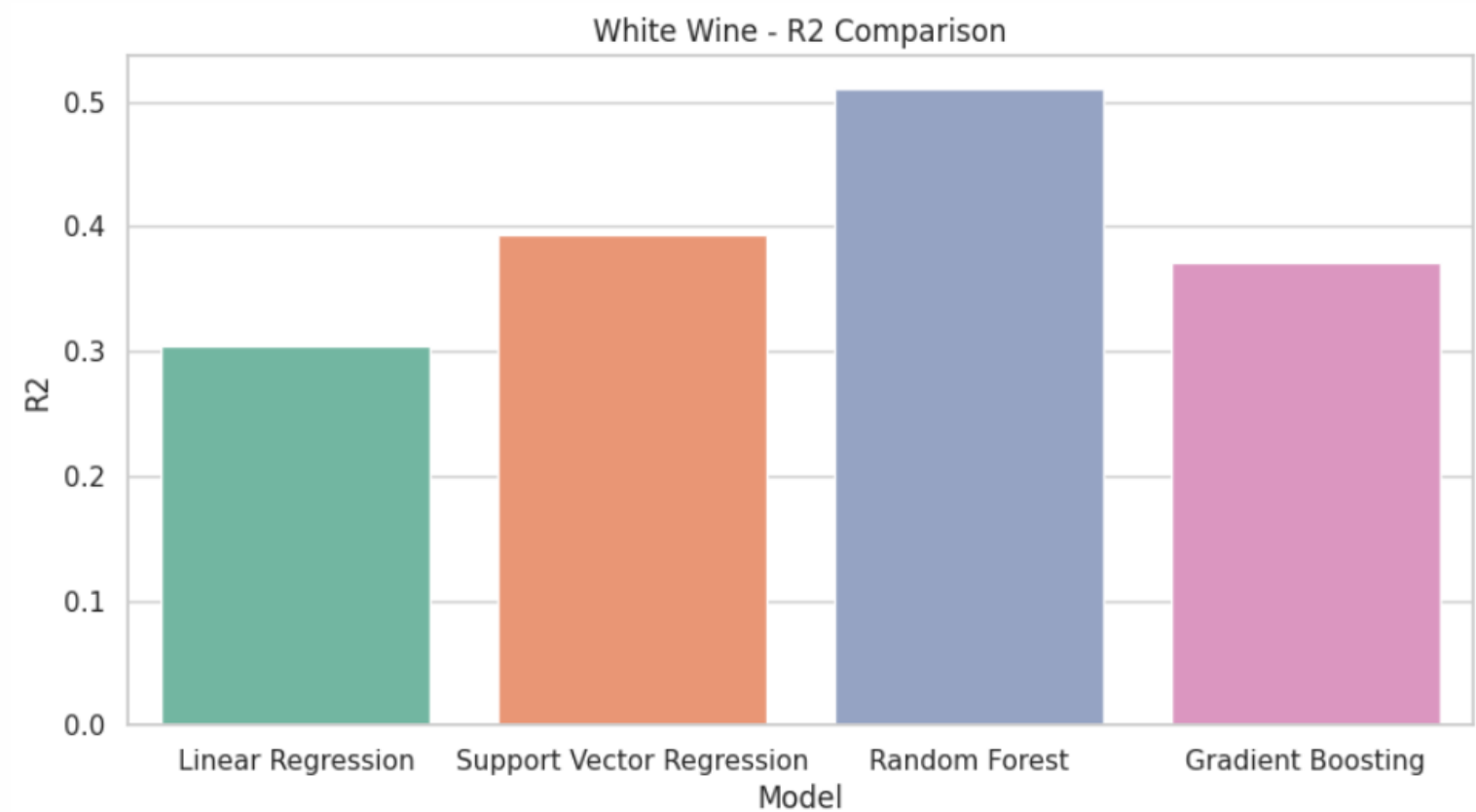
# Aplicação de de Regressão Logística: Cross Validation



DF2 = Base de dados Vinho Branco



# Aplicação de Técnicas de Regressão: Cross



# Modelo de Técnicas de Regressão: Modelos Híbridos

Model	Wine Type	MSE	MAE	R2	Time
Stacked Model 1	Red	0.014646	0.090945	0.461022	3.980490
Stacked Model 1	White	0.012578	0.079236	0.460960	13.361454
Stacked Model 2	Red	0.014936	0.094623	0.450322	3.430535
Stacked Model 2	White	0.012657	0.081220	0.457584	13.326817
Stacked Model 3	Red	0.014711	0.090439	0.458634	3.658929
Stacked Model 3	White	0.012577	0.079201	0.461015	13.982182
Stacked Model 4	Red	0.015932	0.100660	0.413699	3.068662
Stacked Model 4	White	0.014805	0.094994	0.365542	6.474205

Stacked Model 1: Random Forest + Gradient Boosting (meta: Linear Regression)

Stacked Model 2: Random Forest + Gradient Boosting (meta: Ridge Regression)

Stacked Model 3: Random Forest + AdaBoost (meta: Linear Regression)

Stacked Model 4 : Gradient Boosting + AdaBoost (meta: Ridge Regression)

## 1. Eficácia dos Modelos (MSE, MAE, R²)

**Stacked Model 1** e **Stacked Model 3** são os mais **precisos para ambos os tipos de vinho**.

Para **vinhos tintos**, **Stacked Model 1** apresentou ligeiramente menor MSE (0.0146 vs 0.0147) e R² levemente maior (0.461 vs 0.458) do que Stacked Model 3.

Para **vinhos brancos**, Stacked Model 1 e Stacked Model 3 tiveram desempenho praticamente idêntico em MSE (0.0126) e R² (~0.461). Stacked Model 4 teve o **pior desempenho** para ambos os tipos de vinho, com MSE e MAE mais altos e R² mais baixos, indicando baixa capacidade de previsão.

**Conclusão de Eficácia: Stacked Model 1 e Stacked Model 3 são os melhores em precisão para ambos os vinhos.**

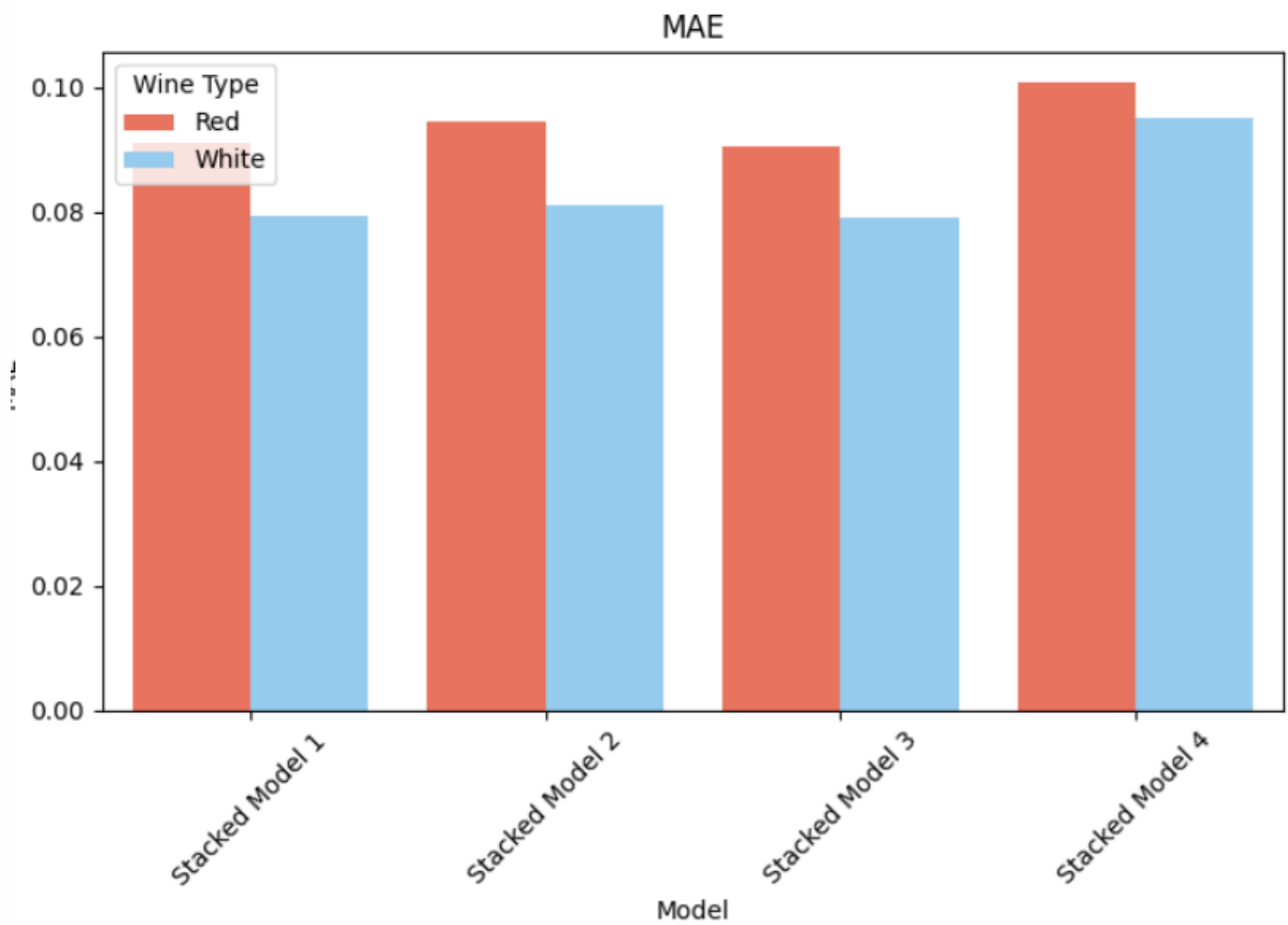
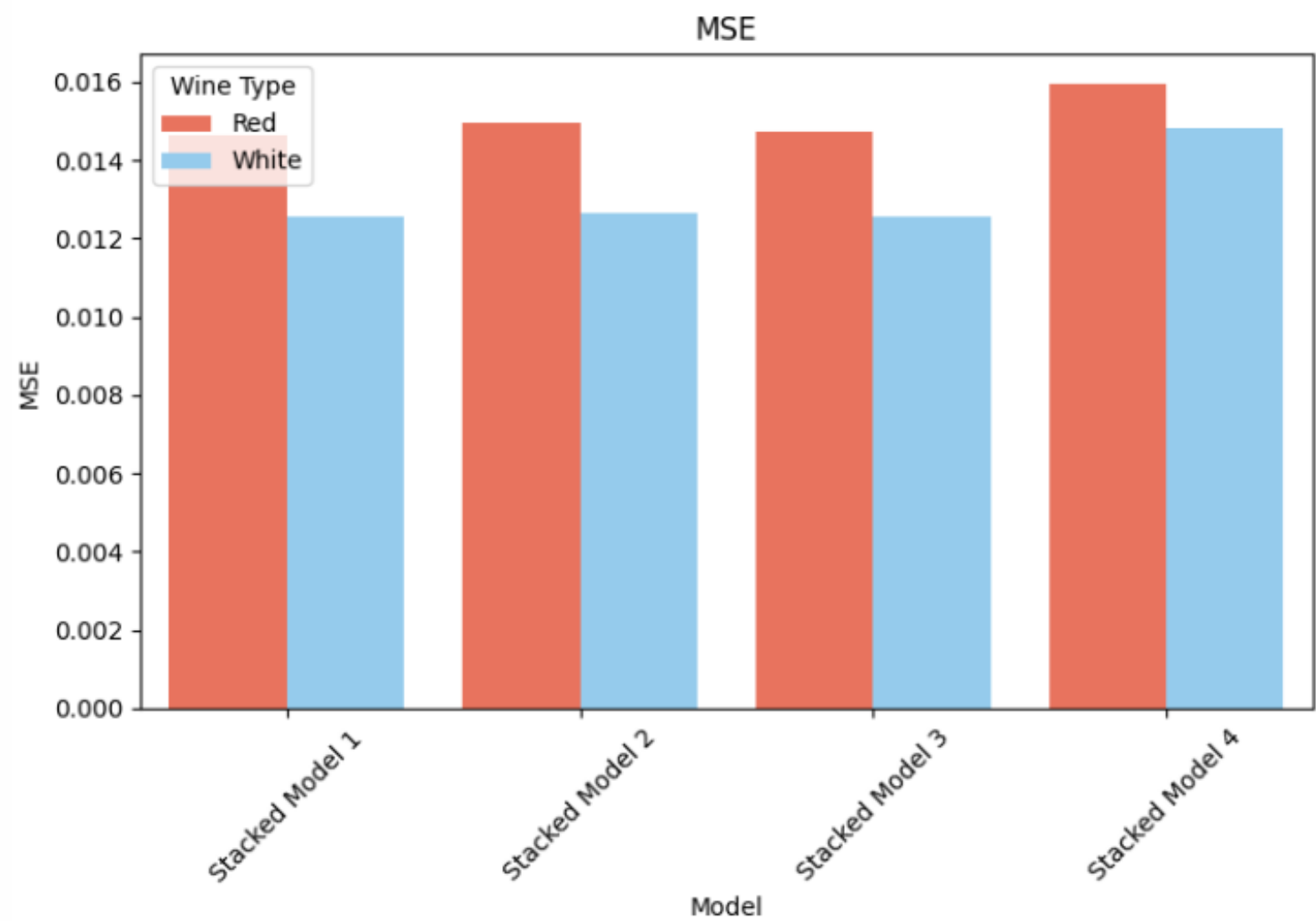
## 2. Tempo de Execução

**Stacked Model 4 foi o mais rápido** para ambos os tipos de vinho (3.07 s para tintos e 6.47 s para brancos) mas com precisão inferior. Stacked Model 1 e Stacked Model 3 **foram mais lentos para vinhos brancos** (~13 segundos) e moderados para vinhos tintos (3.4–4 segundos).

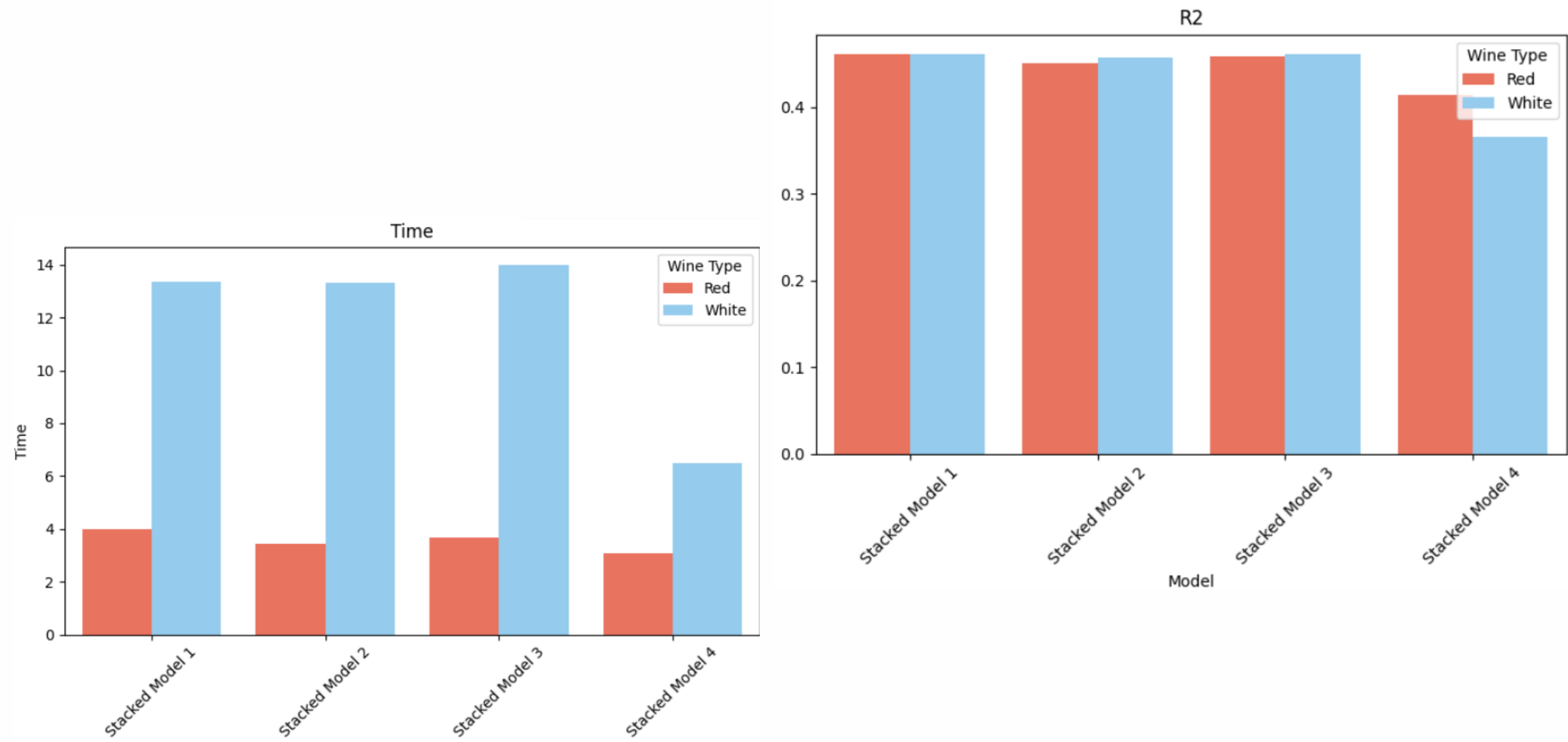
**Conclusão de Tempo: Stacked Model 4 é mais rápido, mas sacrifica precisão; Stacked Model 1 e Stacked Model 3 são moderados para vinhos tintos, mas consideravelmente mais lentos para brancos.**



# Modelo de Técnicas de Regressão: Modelos Híbridos



# Modelo de Técnicas de Regressão: Modelos Híbridos



# Comparação dos Resultados Obtidos

## Aplicação dos algoritmos puros:

### Resultados para df1:

Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
Linear Regression	0.015127	0.099742	0.471607	0.036982
SVR	0.015227	0.100440	0.468109	0.071032
Random Forest	0.013461	0.084473	0.529821	0.518728
Gradient Boosting	0.015179	0.095235	0.469785	0.222424

### Resultados para df2:

Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
Linear Regression	0.016421	0.101326	0.303542	0.006713
SVR	0.014714	0.096663	0.375961	0.392594
Random Forest	0.011536	0.076035	0.510738	2.027817
Gradient Boosting	0.014849	0.094340	0.370217	0.626956

## COM TUNING

### Resultados após tunagem para df1:

Modelo	Melhor Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
SVR	SVR(C=10, kernel='linear')	0.015443	0.101737	0.460562	3.032590
Random Forest	DecisionTreeRegressor(max_features=1.0, random_state=42)	0.013358	0.084333	0.533409	12.062052
Gradient Boosting	DecisionTreeRegressor(criterion='friedman_mse')	0.015194	0.099294	0.469265	22.042699

### Resultados após tunagem para df2:

Modelo	Melhor Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
SVR	SVR(C=1)	0.014714	0.096663	0.375961	10.095348
Random Forest	DecisionTreeRegressor(max_features=1.0, random_state=42)	0.011468	0.075741	0.513625	41.716616
Gradient Boosting	DecisionTreeRegressor(criterion='friedman_mse')	0.011551	0.074339	0.510099	64.207976

## Com Cross-Validation

### Resultados para Vinhos Tintos:

Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
Linear Regression	0.015127	0.099742	0.471607	0.003212
SVR	0.014842	0.095313	0.481579	0.072047
Random Forest	0.013498	0.084495	0.528524	0.855595
Gradient Boosting	0.015037	0.094842	0.474761	0.430642

### Resultados para Vinhos Brancos:

Modelo	MSE	MAE	R <sup>2</sup>	Tempo (s)
Linear Regression	0.016421	0.101326	0.303542	0.002931
SVR	0.014287	0.093991	0.394056	0.879670
Random Forest	0.011528	0.076149	0.511091	3.132886
Gradient Boosting	0.014840	0.094284	0.370607	0.634997

Model	Wine Type	MSE	MAE	R2	Time
Stacked Model 1	Red	0.014646	0.090945	0.461022	3.980490
Stacked Model 1	White	0.012578	0.079236	0.460960	13.361454
Stacked Model 2	Red	0.014936	0.094623	0.450322	3.430535
Stacked Model 2	White	0.012657	0.081220	0.457584	13.326817
Stacked Model 3	Red	0.014711	0.090439	0.458634	3.658929
Stacked Model 3	White	0.012577	0.079201	0.461015	13.982182
Stacked Model 4	Red	0.015932	0.100660	0.413699	3.068662
Stacked Model 4	White	0.014805	0.094994	0.365542	6.474205

# Comparação dos Resultados Obtidos

Na análise dos modelos **Regressão Linear**, **SVR**, **Random Forest**, **Gradient Boosting** e **modelos híbridos (stacking)**, com base nas métricas **MSE**, **MAE**, **R<sup>2</sup>** e **Tempo de Execução**, observamos o desempenho em duas bases de dados de tamanhos distintos:

- **df1**: conjunto menor, com **1600 instâncias**;
- **df2**: conjunto maior, com **4899 instâncias**.

## Comparação dos Resultados:

### Modelos Padrões sem Tunagem

- No **df1**, **Random Forest** e **Gradient Boosting** alcançaram melhores resultados em MSE e R<sup>2</sup> em relação à **Regressão Linear** e ao **SVR**, mas com maior custo computacional.
- No **df2**, **Random Forest** novamente se destacou em precisão, com um tempo de execução bem maior, seguido pelo **Gradient Boosting**.

### Modelos com Tunagem

A tunagem permitiu melhorias significativas para todos os modelos, com **Random Forest** e **Gradient Boosting** apresentando melhor desempenho em precisão em ambos os conjuntos de dados, mas com aumentos expressivos nos tempos de execução, especialmente para o **Random Forest** no df2.



# Comparação dos Resultados Obtidos

## Resultados com Cross-Validation

Com **Cross-Validation**, **Random Forest** e **Gradient Boosting** continuaram com valores superiores de precisão. Embora o custo computacional tenha sido otimizado, ainda foi consideravelmente maior que o da **Regressão Linear**.

## Modelos Híbridos (Stacking)

A introdução de modelos híbridos de empilhamento (*stacked models*) buscou melhorar a robustez e a precisão combinando modelos base:

- **Stacked Model 1** e **Stacked Model 3** apresentaram os melhores valores de MSE e  $R^2$  para ambos os tipos de vinho, mantendo-se entre os modelos com melhores resultados de precisão, especialmente em relação ao **MSE** e **MAE**.
- Comparado aos modelos individuais, os modelos híbridos mostraram-se competitivos e com menor custo computacional que o **Random Forest** no conjunto maior (**df2**), com tempos de execução entre **3 a 13 segundos**, tornando-os opções de precisão e desempenho relativamente equilibrados.
- **Stacked Model 4** obteve o menor desempenho, especialmente em vinhos brancos, com menor  $R^2$  e um MSE mais alto em comparação aos outros modelos híbridos.

# Conclusão

## Modelo de Melhor Desempenho:

Entre os modelos testados, **Random Forest** permaneceu como o modelo com maior precisão em termos de MSE e  $R^2$ , tanto para df1 quanto df2. No entanto, os modelos híbridos **Stacked Model 1 e 3** destacaram-se por oferecer um desempenho competitivo com menor custo computacional.

## Impacto do Tamanho da Base:

O tempo de execução aumentou substancialmente para **Random Forest e Gradient Boosting** com o aumento do tamanho do conjunto de dados. Os modelos híbridos mantiveram uma relação de custo-benefício mais vantajosa no conjunto maior.

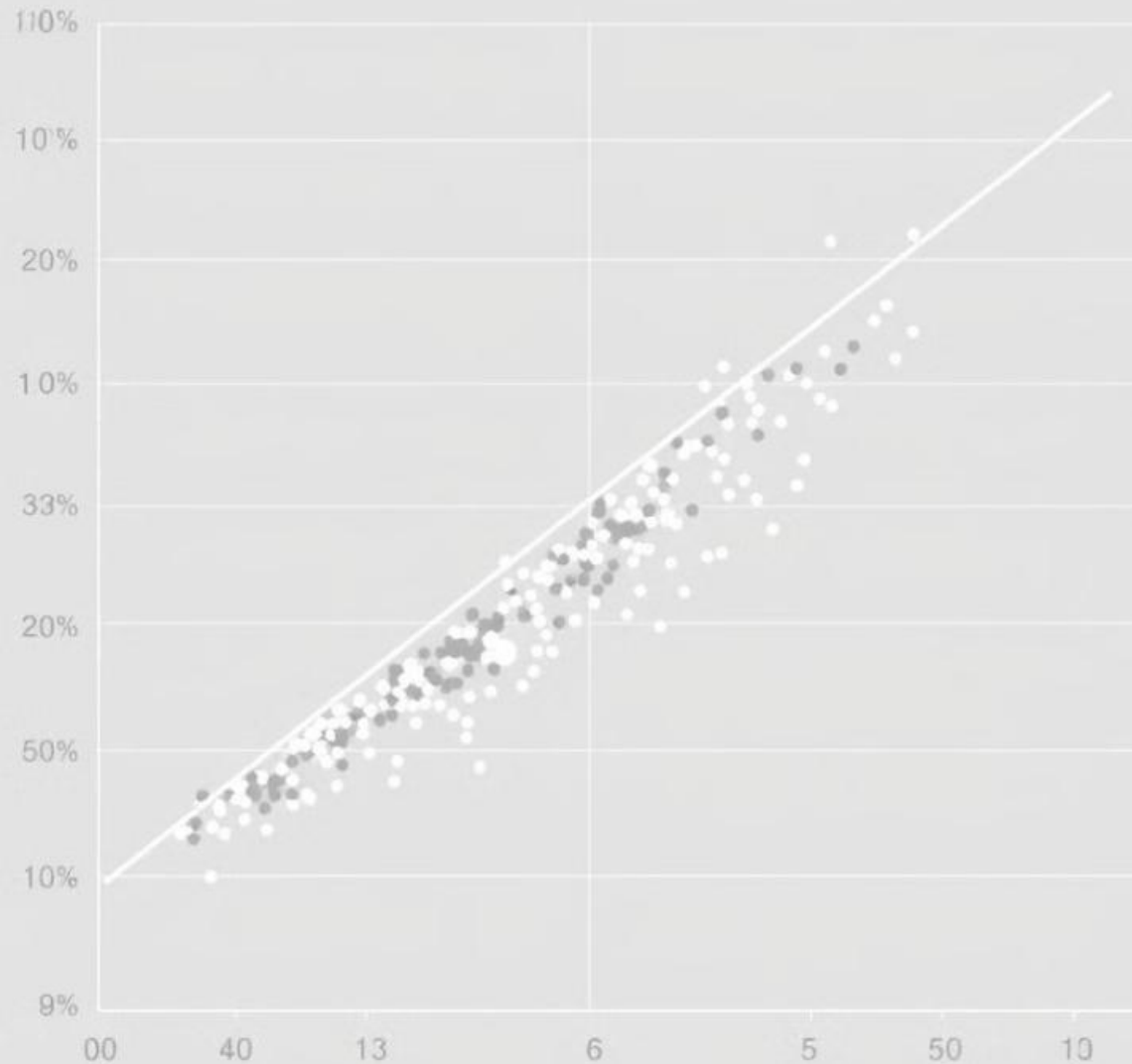
## Custo Computacional:

**Regressão Linear contínua como a opção mais rápida**, enquanto Random Forest apresenta o maior custo computacional. Os modelos híbridos proporcionaram um equilíbrio entre desempenho e custo, adequados para cenários que demandam boa precisão sem sacrificar tanto o tempo de execução.

## Recomendação Final:

Para **precisão máxima** em ambas as bases, **Random Forest com tunagem é o modelo recomendado**. No entanto, para um **compromisso entre precisão e tempo de execução**, **Stacked Model 1 e Stacked Model 3** são opções muito viáveis, especialmente em contextos com grandes volumes de dados.

# Referências:



ALMEIDA, H. M. de. Análise de regressão linear múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma indústria de calçados. 2014. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, Campina Grande, 2014. Disponível em: <https://dspace.bc.uepb.edu.br/jspui/bitstream/123456789/5167/1/PDF%20-%20Humberto%20Moreira%20de%20Almeida.pdf>. Acesso em: 25 outubro. 2024. DOI: <https://doi.org/10.24432/C56S3T>.

CORTEZ, Paulo; CERDEIRA, A.; ALMEIDA, F.; MATOS, T.; REIS, J. Wine Quality. 2009. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C56S3T>.

DOSUALDO, D. G. Investigação de regressão no processo de mineração de dados. 2003. Dissertação (Mestrado em Ciências Matemáticas e de Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-12112014-101732/pt-br.php>. Acesso em: 25 out. 2024.

LOVATO, M.; WAGNER, R. Avaliação da qualidade dos vinhos de mesa suave por análises físico-químicas. Cadernos da Escola de Saúde, Curitiba, v. 2, n. 8, 2017. Disponível em: <https://portaldeperiodicos.unibrasil.com.br/index.php/cadernossaude/article/view/2365/1937>. Acesso em: 26 out. 2024.