UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

Trabalho de Tópicos de Especiais de Estatística

Beatriz Da Silva Rizzoli

Guilherme De Conde Reis

Thaii Céu

Presidente Prudente

2024

1- Introdução

O vinho é uma das bebidas mais antigas e valorizadas pela humanidade, carregando um rico simbolismo cultural, social e histórico. Produzido a partir da fermentação das uvas, trazendo aromas e sabores que variam de frescos e frutados a complexos e encorpados, dependendo de seu estilo e processo de produção.

No Brasil o vinho foi introduzido pelos primeiros portugueses a chegarem no país, mas devido às condições climáticas as plantações não deram certo. Quando em 1808, com a vinda da corte real, houve um crescimento da produção para atender as necessidades da realeza.

A verdadeira mudança iniciou com a chegada dos imigrantes italianos no Rio Grande do Sul, que foi intensificada em 1871. Nesse período começaram a surgir as primeiras vinícolas centenárias. O sucesso da produção de vinho na região se deu pela utilização de uvas de origem americana e a formação de cooperativas, que promoveram o desenvolvimento da viticultura.

Atualmente, o consumo de vinho no Brasil é de, em média, 2 litros per capita por ano, segundo a Associação Brasileira de Sommeliers (ABS). O país conta com mais de 1.100 vinícolas, que empregam mais de 50 mil famílias.

Reconhecido como uma bebida refinada, o vinho é tradicionalmente servido em taças elegantes e apresenta uma ampla variação de preços, desde opções acessíveis ao público até rótulos raros e de alto valor. A qualidade do vinho é um elemento essencial, garantindo uma experiência mais satisfatória ao consumidor e sua escolha de rótulos que atendam às preferências pessoais ou às necessidades de uma ocasião específica. O conceito de qualidade do vinho vai além do sabor, sendo determinado por diversos fatores, sendo essa avaliação muito importante para valorização e controle de qualidade.

Nesse contexto, o monitoramento da qualidade do vinho se torna uma necessidade, não apenas para garantir seus padrões na produção, mas também para assegurar uma boa classificação no mercado e valorização da bebida. Trazendo os métodos atuais de machine learning junto da avaliação química e física da composição do vinho, pode-se gerar métodos eficientes de avaliação de qualidade de maneira precisa.

1.1 Objetivo

O trabalho tem como objetivo principal comparar o desempenho de quatro diferentes modelos de machine learning em tarefas de classificação e regressão, ou seja, capacidade de prever as classes e valores, respectivamente. Atrelado ao contexto visto anteriormente, o foco da análise é avaliar a eficácia dos modelos na previsão da qualidade do vinho e na estimativa do seu açúcar residual.

2- Metodologia

Este estudo tem como objetivo treinar diferentes modelos de *machine learning* (ML) para avaliar e comparar o desempenho desses modelos na tarefa de previsão e classificação da qualidade do vinho, com base em dados químicos e físicos da bebida. A metodologia adotada será dividida nas seguintes etapas:

2.1- Análise exploratória

Nessa etapa, o intuito é entender a base de dados e avaliar se é possível a realização dos modelos, averiguando a consistência dos dados em cada uma de suas variáveis. também realizar a avaliação das correlações entre as variáveis químicas e físicas e a qualidade do produto.

2.2- Tratamento dos dados

Realizar um tratamento na base com o intuito de deixar a base pronta para ser modelada dentro do software Rstudio, com tratamento para dados faltantes e tipo da variável.

A base de dados será preparada para modelagem no software RStudio. Essa etapa inclui a identificação e o tratamento de valores ausentes, ajustes no tipo de variável (como numérica ou categórica) e a normalização ou padronização dos dados, se necessário, para melhorar o desempenho dos modelos.

2.3 - Modelagem

Nessa etapa, quatro modelos de machine learning serão treinados, todos com o objetivo de classificação como também de regressão. Os modelos escolhidos para comparação foram Árvore de decisão, Florestas aleatórias, Redes Neurais e KNN (K-Nearest Neighbors). A escolha dos modelos deve-se a sua flexibilidade, permitindo tanto a classificação quanto à previsão de valores numéricos.

2.4. Avaliação dos Resultados

As medidas de avaliação têm como objetivo tanto avaliar os modelos individualmente quanto fazer comparações entre eles e entender qual pode atingir melhor desempenho.

As principais medidas para comparação entre diferentes modelos de classificação, são:

- **Accuracy:** Mede a proporção de previsões corretas (tanto verdadeiros positivos quanto verdadeiros negativos) em relação ao total de previsões. É uma métrica global que reflete a taxa de acertos do modelo, mas pode ser enganosa em casos de classes desbalanceadas.

- **Precision:** Mede a proporção de previsões positivas que são realmente verdadeiras. Indica a confiabilidade dos resultados positivos previstos pelo modelo.
- **Negative predictive value:** Mede a proporção de previsões negativas que são realmente verdadeiras. Avalia a confiabilidade dos resultados negativos previstos pelo modelo.
- **Recall:** Representa a capacidade do modelo de identificar corretamente todas as instâncias positivas. Também conhecido como "taxa de verdadeiros positivos".
- **Specificity:** Mede a capacidade do modelo de identificar corretamente as instâncias negativas. Também chamado de "taxa de verdadeiros negativos".
- **F1 score:** É a média harmônica da precisão e do recall, sendo uma métrica balanceada que considera tanto os falsos positivos quanto os falsos negativos.
- **False positive rate:** Mede a proporção de instâncias negativas que foram incorretamente classificadas como positivas.
- **False negative rate:** Mede a proporção de instâncias positivas que foram incorretamente classificadas como negativas.

Já para medir a capacidade do modelo no quesito regressão, as medidas mais comuns para comparação entre eles estão

- **MSE (Mean Squared Error - Erro Quadrático Médio)**: Mede a média dos quadrados das diferenças entre os valores reais e as previsões. Penaliza mais fortemente erros maiores devido à elevação ao quadrado, sendo útil para avaliar a precisão geral do modelo.
- **RMSE (Root Mean Squared Error - Raiz do Erro Quadrático Médio)**: É a raiz quadrada do MSE, expressando o erro médio em termos das mesmas unidades que os dados originais. Facilita a interpretação dos erros ao compará-los diretamente com os valores observados.
- **MAE (Mean Absolute Error - Erro Absoluto Médio):** Mede a média das diferenças absolutas entre os valores reais e as previsões. É menos sensível a erros extremos em comparação ao MSE, sendo útil para capturar erros médios sem amplificar outliers.
- **MAPE (Mean Absolute Percentage Error - Erro Percentual Absoluto Médio)**: Mede o erro absoluto médio como uma porcentagem dos valores reais. Indica, em média, qual é o percentual de erro entre as previsões e os valores reais. É uma métrica útil para interpretar erros em termos relativos, mas pode ser sensível a valores reais próximos de zero, o que pode distorcer os resultados

3- Análise exploratória

Foram analisadas separadamente todas as variáveis presentes na base de dados, observando suas medidas de posição, medidas de dispersão e gráficos de boxplot e histograma. As variáveis que usamos como variáveis respostas estão representadas aqui, enquanto a análise das outras variáveis se encontra no Anexo A.

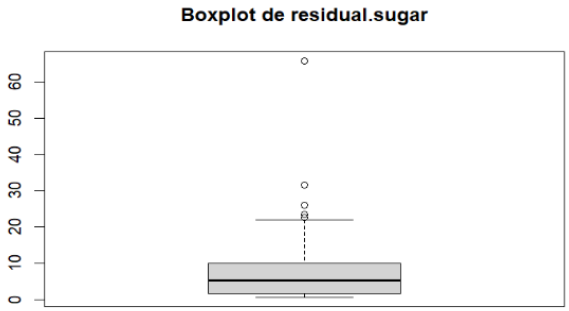
- **residual.sugar**

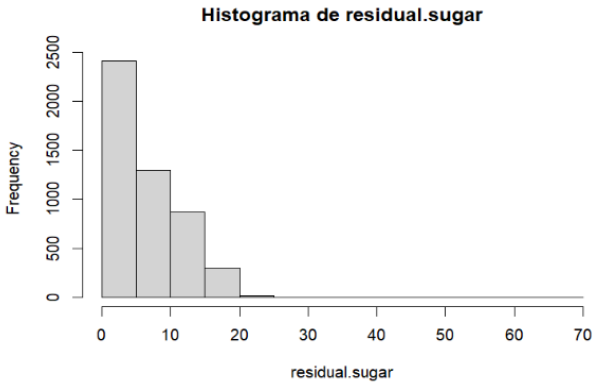
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.60	1.70	5.20	6.39	9.90	65.80

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
25.7257	5.0720	79.3573	65.2000





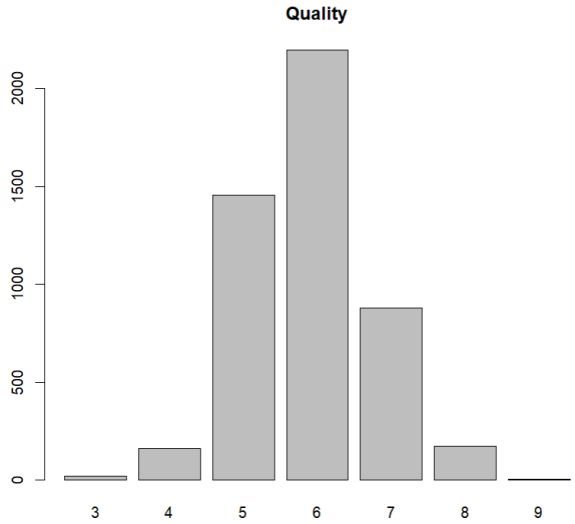
Percebe-se que **“**residual.sugar” apresenta valores próximos da média e da mediana, com um coeficiente de variação de 79%, indicando que a média não explica bem os dados. Observando os gráficos, nota-se que existe uma forte assimetria à esquerda e uma variação nos dados muito grande, apresentando um outlier discrepante.

- **quality**

Quality é a única variável categórica da base de dados, essa pode assumir valores de 0 até 10 e indica.

Classe	3	4	5	6	7	8	9

Porcentagem	0,3%	3,3%	29,7%	45%	18%	3,6%	0,1%



A qualidade mais frequente é a avaliada em “6”, representando 45% dos dados da variável.

.

4- Modelos de Classificação

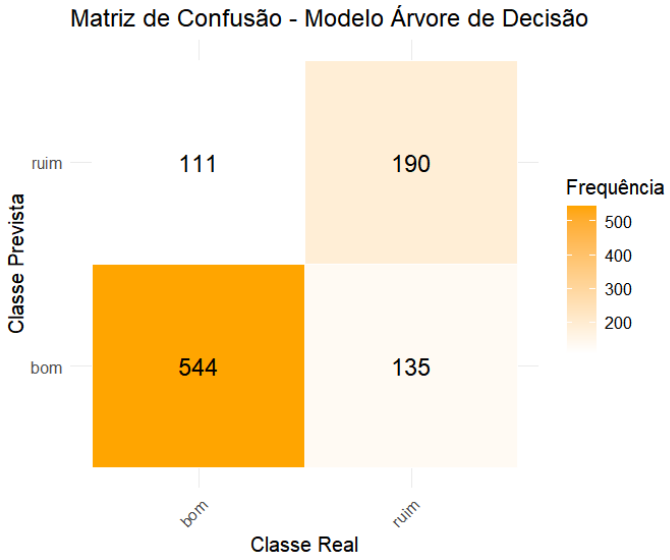
4.1 Modelo de Árvore de Decisão de Classificação

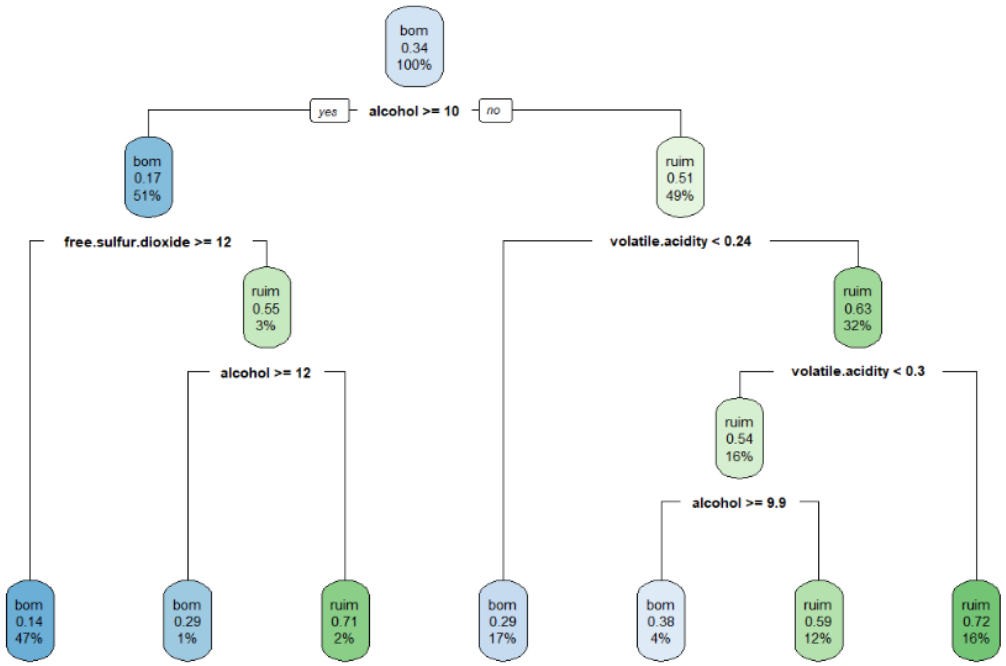
Os algoritmos de árvores de decisão são modelos de aprendizado supervisionado que podem ser usados para classificação e regressão. O modelo tem uma estrutura hierárquica e é definido por nós. Os nós são divididos em:

- Nó raiz: é o ponto de partida de uma árvore.
- Nós internos: que são pontos intermediários que continuam sendo divididos.
- Galhos: são os caminhos que conectam os nós.
- Nó folha: são os resultados finais da árvore.

Para uma árvore de classificação, o modelo seleciona uma variável e um critério que melhor separa os dados. Essa separação é realizada pelo teste do atributo escolhido, e o caminho que se segue é determinado pelo valor do atributo, levando ele para a próxima seleção do ramo correspondente. Esse processo se repete até chegar em um nó folha.

No modelo feito no trabalho foi utilizado o método class que gerou a seguinte matriz de confusão:





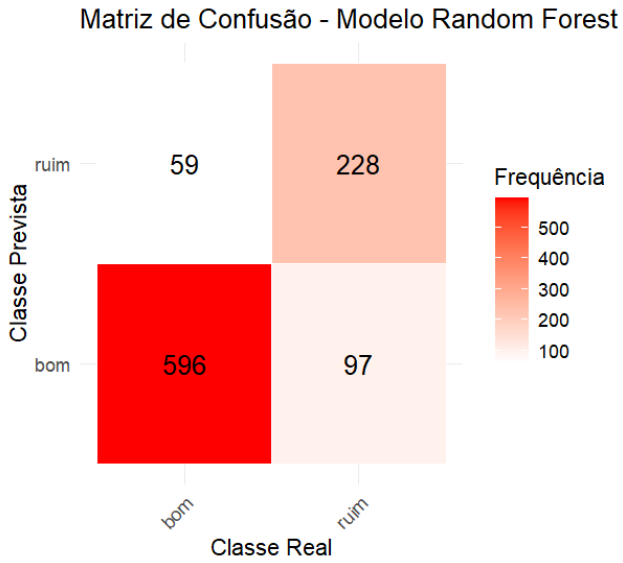
A árvore pronta nos mostra quais foram as regras criadas para chegar na classificação dos dados. Nessa imagem, cada nó carrega o termo “bom” ou “ruim”, os nós com termo “bom” são aqueles que possuem mais classes “bom” e sua base, o mesmo acontece com os que carregam o termo “ruim”. Junto a cada nó e folha existe uma probabilidade e uma porcentagem, a probabilidade representa a probabilidade do vinho ser “ruim”, e a porcentagem representa quantos por cento da base se encontram dentro daquele nó. Quando olhamos apenas para as folhas, vemos 2 cores, as azuis e as verdes, elas indicam qual classe o vinho receberá se ele seguir a regra que leva aquela folha, sendo folhas azuis o indicativo de classificação do vinho como “bom”, já as folhas verdes indicam regras que levam o vinho a ser classificado como “ruim”.

4.2 Modelo de Floresta Aleatória de Classificação

A floresta aleatória é um conjunto de árvores de decisão, independentes, que podem realizar tarefas de classificação e regressão usando o método de bootstrapp. A seleção aleatória de variáveis em cada nó de decisão ajuda na redução da correlação entre as árvores, diminuindo sua variância.

No modelo de floresta aleatória de classificação, o objetivo é prever a classe em que uma observação pertence. Todas as árvores são treinadas para prever uma classe específica, quando novas observações são introduzidas no modelo cada árvore faz uma previsão da classe mais provável para aqueles dados. A classe final é definida pelo voto majoritário de todas as árvores da floresta.

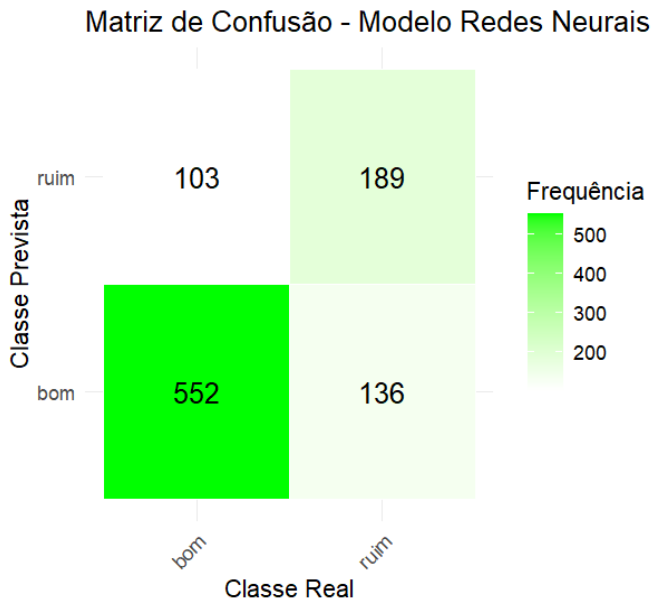
Como o modelo chega na conclusão final usando várias árvores, ao invés de uma só, o resultado final é mais consolidado. No modelo feito para o trabalho foram utilizadas 100 árvores, gerando o resultado da matriz de confusão abaixo.



4.3 Redes Neurais de Classificação

As redes neurais são compostas por camadas de nós, ou “neurônios”, que se conectam entre si. Elas têm a camada de entrada, que recebe os dados brutos, camadas ocultas, que processam as informações nos nós e enviam elas para a próxima camada, e a camada de saída que produz o resultado final.

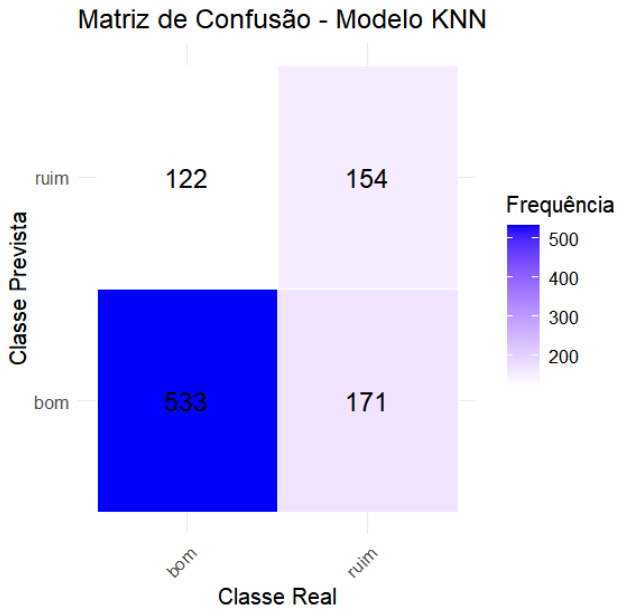
Para o trabalho foram utilizados 10 neurônios em uma única camada com 10, com um número máximo de 1000 interações.



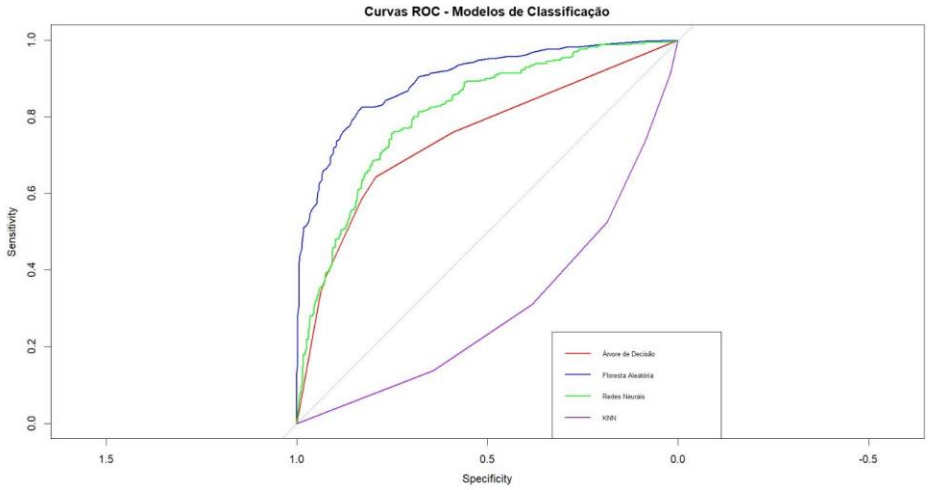
4.4 K-nn de Classificação

O K-NN de classificação é um método que utiliza a distância entre os indivíduos da base para classificar novas observações, um ponto de observação sobre esse método é que ele sempre precisa trazer com si a base de dados de treinamento para poder prever valores futuros, pois esse se baseia nos “k” pontos mais próximos a ele para classificar o novo ponto, a classe mais frequente dentre os pontos próximos é determinada como classe da nova observação.

Para o modelo do trabalho foi utilizado a distância euclidiana com k = 5, ou seja, o modelo identifica as características dos cinco pontos mais próximos para prever a classe e cada ponto que será avaliado.



4.5 Conclusões da classificação



Métricas	Árvore de Decisão	Floresta Aleatória	Redes Neurais	K-nn
Acurácia	0.7489796	0.8408163	0.772449	0.7010204
IC 95% da acurácia	[0.7218324,0.7761268]	[0.8179111, 0.8637216]	[0.7462002, 0.7986978]	[0.6723574, 0.7296834]
Precisão	0.8011782	0.8600289	0.801676	0.7571023
Valor preditivo negativo	0.6312292	0.7944251	0.6931818	0.557971
Sensibilidade	0.8305344	0.9099237	0.8763359	0.8137405

Especificidade	0.5846154	0.7015385	0.5630769	0.4738462
F1 - Score	0.8155922	0.884273	0.837345	0.7844003
Taxa de Falso Positivo	0.4153846	0.2984615	0.4369231	0.5261538
Taxa de Falso Negativo	0.1694656	0.2984615	0.1236641	0.1862595
Kappa	0.4230189	0.09007634	0.462199	0.2989404
IC 95% do Kappa	[0.3605097, 0.4855281]	[0.5766958, 0.68334414]	[0.3989575, 0.5234822]	[0.2315988, 0.3662819]

A análise das métricas dos modelos de classificação motras que a Floresta Aleatória é o modelo com melhor desempenho. Este modelo apresenta a maior acurácia (0,8408), maior precisão (0,8600), sensibilidade (0,9099), especificidade (0,7015) e F1-Score (0,8843). Além disso, possui a menor taxa de falsos positivos (0,2985), o que reforça sua capacidade de classificar corretamente tanto as classes positivas quanto as negativas. No entanto, o valor de Kappa para a Floresta Aleatória (0,0908) é o menor entre os modelos, indicando baixa concordância entre as previsões e os valores reais.

As Redes Neurais, por outro lado, apresentam bom equilíbrio entre as métricas, com destaque para o maior valor de Kappa (0,4622), indicando boa concordância global. Também apresentam alta sensibilidade (0,8763), precisão (0,8017) e um F1-Score elevado (0,8373). Além disso, possuem a menor taxa de falsos negativos (0,1237), o que é um ponto positivo na identificação de verdadeiros positivos. A acurácia (0,7724) é inferior à da Floresta Aleatória, mas ainda é consistente.

A Árvore de Decisão apresenta desempenho mediano. Sua acurácia é de 0,7489, e o F1-Score é de 0,8156. Este modelo tem boa sensibilidade (0,8305), mas apresenta baixa especificidade (0,5846) e uma taxa relativamente alta de falsos positivos (0,4154). O valor de Kappa (0,4230) é menor que o das Redes Neurais, mas ainda indica concordância razoável.

O K-NN apresenta o menor desempenho geral, com acurácia de 0,7010, a menor precisão (0,7571), menor valor preditivo negativo (0,5580) e baixa especificidade (0,4738). Além disso, possui a maior taxa de falsos positivos (0,5262), indicando uma maior tendência de classificar negativamente instâncias positivas. Apesar disso, o modelo tem uma sensibilidade razoável (0,8137) e um valor de Kappa de 0,2989, que é superior ao da Floresta Aleatória, mas inferior ao dos demais.

Sendo assim, a Floresta Aleatória se destaca como o melhor modelo, especialmente para tarefas que requerem alta precisão e sensibilidade. As Redes Neurais também são uma excelente opção, com bom equilíbrio entre as métricas. A Árvore de Decisão apresenta resultados satisfatórios em algumas métricas, mas é inferior às duas primeiras. Por fim, o K-NN, apesar de algumas métricas razoáveis, é o modelo com pior desempenho global. As matrizes e a curva ROC complementam tudo o que foi comentado aqui.

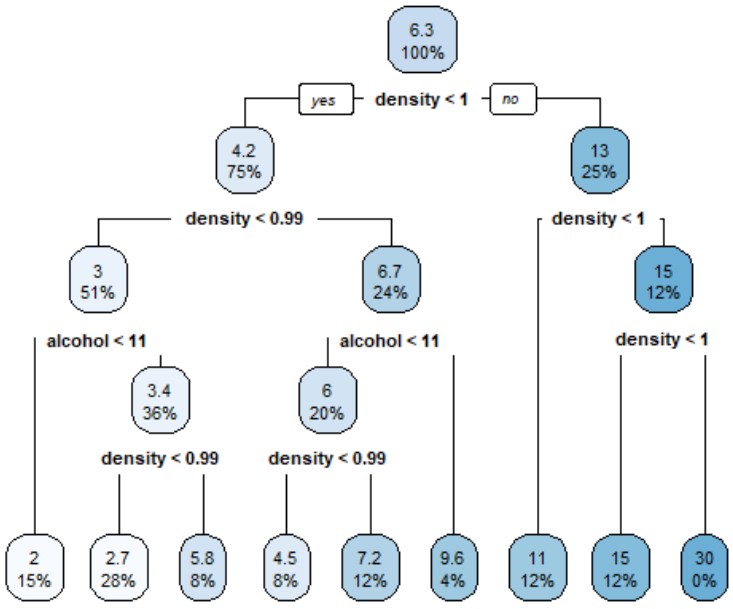
5- Modelos de Regressão

5.1 Modelo de Árvore para regressão

Para a árvore de regressão o modelo utiliza critérios de poda para otimizar as previsões que ele realiza. Existem dois tipos de poda, a Pré-Poda, que estabelece um limite antes da árvore crescer muito, e a Pós-Poda, onde a árvore se desenvolve totalmente, mas seus galhos são revisados para a remoção dos que não estão sendo úteis, geralmente os transformando em folhas.

As árvores de regressão também realizam a extração de forma autônoma, usando variáveis que são necessárias para a tomada de decisões. As regras de extração são expressões lógicas que descrevem a relação entre atributos e resultados em um conjunto de dados.

No modelo feito no trabalho foi utilizado a anova para gerar as métricas que serão avaliadas.



A árvore pronta é representada por um conjunto de de regras que nos leva até suas folhas com o objetivo de reduzir a variação nos dados dentro dessa, em que é definido qual o valor estimado do açúcar residual, os valores previstos podem variar de 2 até 30.

5.2 Modelo de Floresta Aleatória de Regressão

O modelo de floresta aleatória de regressão, diferente do modelo de classificação, prevê valores numéricos. Cada árvore resulta em um valor e a floresta calcula a média desses valores para obter a previsão final, tendo como objetivo minimizar o erro quadrático médio.
A metodologia usada na Floresta para prever um valor numérico parte do conceito de ampliar a metodologia de árvore aleatória, unindo vários modelos de árvores aleatórias fracos, então é feita média do resultado de cada árvore individual. No modelo feito para o trabalho foram utilizadas 100 árvores.

5.3 Redes Neurais de Regressão

O modelo de Redes Neurais para Regressão segue o mesmo padrão do modelo para classificação. Porém, a saída é uma estimativa numérica, e não uma classe.

Para o trabalho foram utilizados 10 neurônios em uma única camada com 10, com um número máximo de 1000 interações, além disso, foi utilizado a função linout para que a saída do modelo use uma função de ativação linear para gerar os resultados.

5.4 K-nn de Regressão

O K-NN para previsão de valores numéricos se assemelha muito ao de classificação, onde o agora o objetivo é estimar o valor de novas observações, assim como na classificação, agora com objetivo de regressão é utilizado a distância para ordenar os vizinhos mais próximos do novo ponto, assim que determinado os “k” vizinhos de menor distância, para estimar o valor do novo ponto existem diferentes estratégias, a mais comum é fazer a média dos “k” vizinhos mais próximos, assim como também poderia ser feita a média ponderada pela distância.

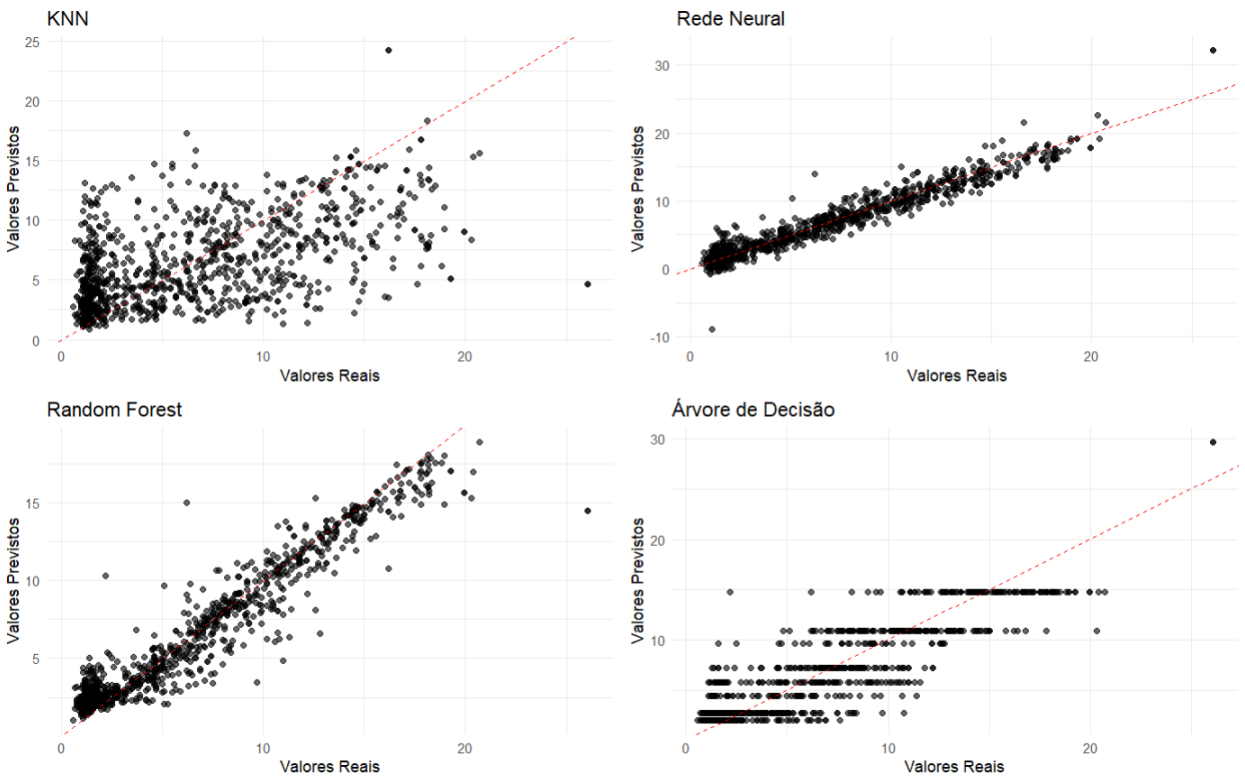
Para o modelo do trabalho foi utilizado as mesmas configurações do modelo de classificação. O modelo identifica as características dos cinco pontos mais próximos para prever o valor de cada ponto que será avaliado.

5.5 Conclusões da regressão

O modelo com melhor desempenho é o de Redes Neurais, além de possuir as menores medidas de erro ele apresenta o maior R^2, demonstrando melhor capacidade de prever valores e demonstrando que conseguiu capturar a melhor variabilidade dos dados.

Já o K-nn apresentou o pior desempenho, pois tem os maiores erros e o menor R^2 = 0.2607, demonstrando não ser eficaz em capturar a variabilidade dos dados, além disso, foi o modelo que menos se ajustou aos dados.

Métricas	Árvore de Decisão	Floresta Aleatória	Redes Neurais	K-nn
R^2	0.7897903	0.9185048	0.9381434	0.2607403
MSE	5.611364	2.175443	1.651209	19.7339
RMSE	2.368832	1.474938	1.284994	4.442285
MAE	1.801151	1.010525	0.9445241	3.330969
MAPE	50.90812	33.00964	28.33492	98.17581



Referências

https://vinhobrasileiro.org/o-vinho-no-brasil

https://oglobo.globo.com/patrocinado/dino/noticia/2023/04/brasileiro-consome-em-media-dois-litros-de-vinho-por-ano.ghtml

https://www.divinho.com.br/blog/o-que-e-vinho/?srsltid=AfmBOoqg2Lr4I4cZkgeSykV0feTWavb2OGDl3oajeG8SNe2CWgwPZ7eA

Anexo A - Análise exploratória

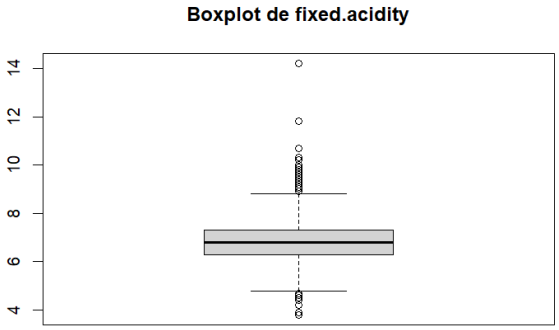
- **fixed.acidity**

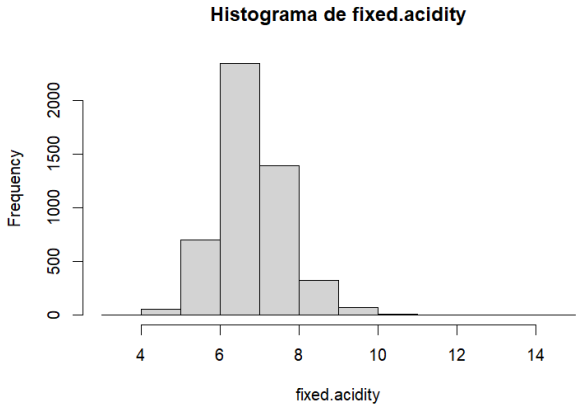
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
3.8	6.3	6.8	6.8	7.3	14.2

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.7121	0.8438	12.3106	10.4000





Apesar do boxplot acusar diversos outliers, os dados parecem não serem tão dispersos, principalmente ao olhar seu desvio padrão baixo e coeficiente de variação de aproximadamente 12%.

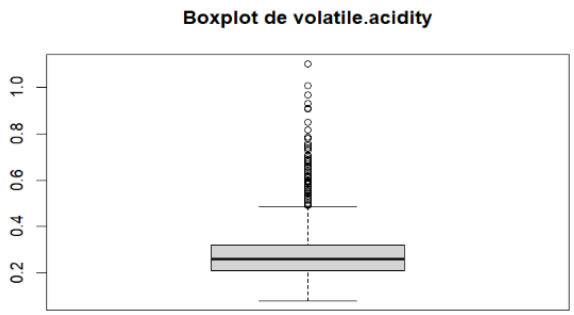
- **volatile.acidity**

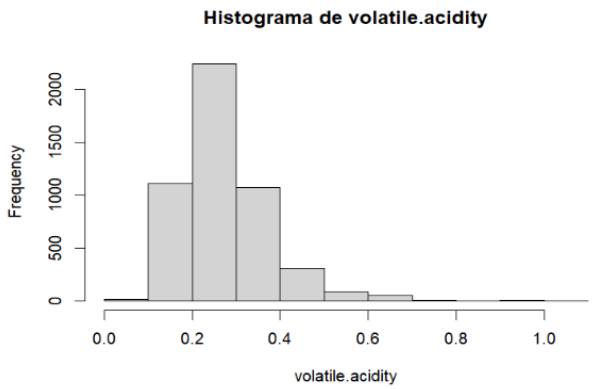
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.08	0.21	0.26	0.27	0.32	11.10

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.0101	0.1007	36.2256	1.0200





“volatile.acidity” apresenta uma assimetria leve à esquerda apesar da proximidade da média com a mediana, apresenta também um coef. de variação relativamente alto.

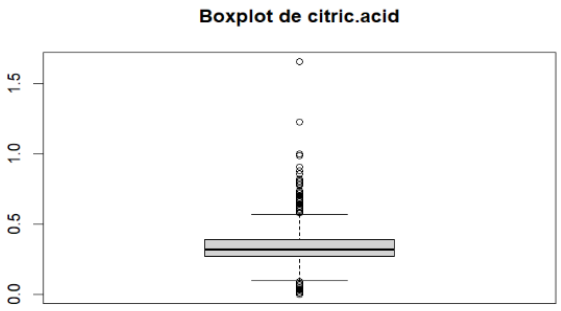
- **citric.acid**

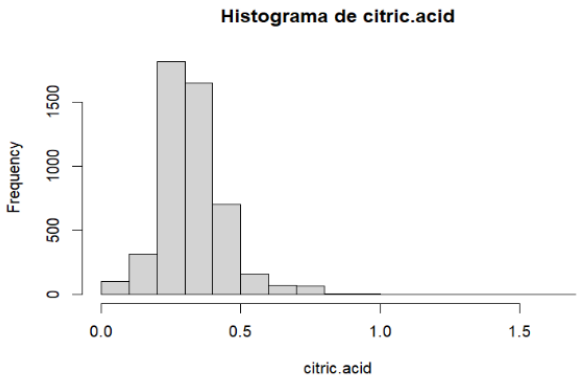
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.00	0.27	0.32	0.33	0.39	1.66

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.0146	0.1210	36.2127	1.6600





“citric.acid” apresenta uma assimetria leve à esquerda apesar da proximidade da média com a mediana, apresenta também um coef. de variação relativamente alto.

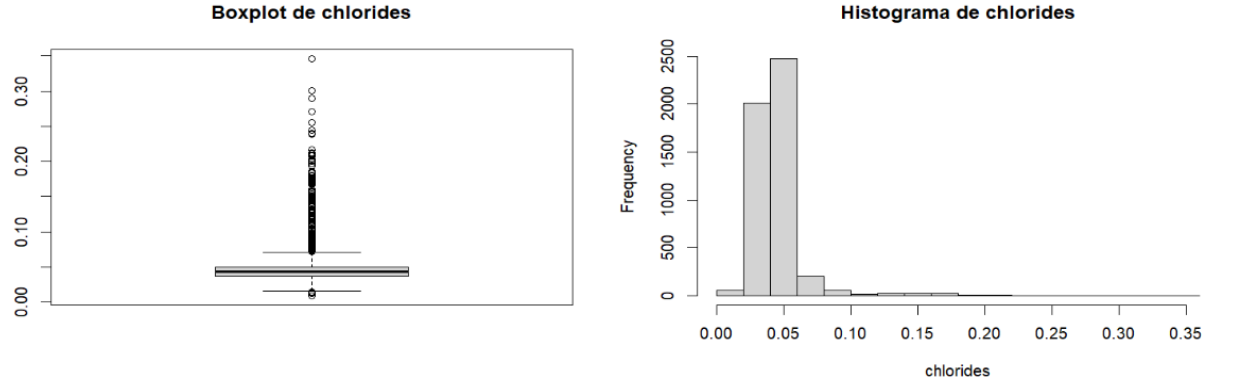
- **chlorides**

Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.009	0.036	0.043	0.045	0.050	0.346

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.0005	0.0218	47.7318	0.3370



“chlorides” apresenta uma forte assimetria à esquerda e uma forte variação nos dados, com CV de 47.7%.

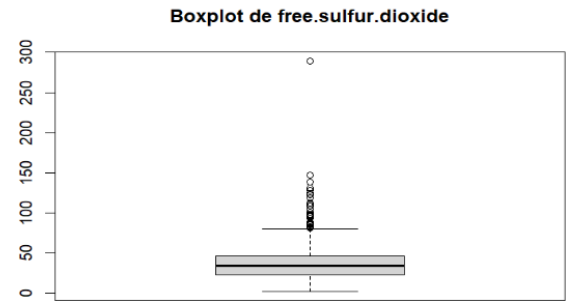
- **free.sulfur.dioxide**

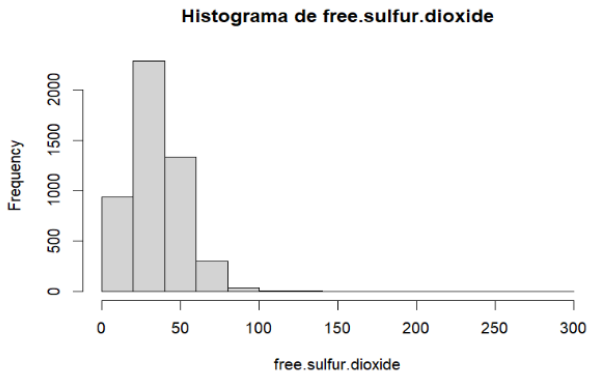
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
2.00	23.00	34.00	35.31	46.00	289.00

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
289.2427	17.0071	48.1678	287,0000





“free.sulfur.dioxide” possui uma forte assimetria à esquerda e um coef. de variação de 48% indicando uma forte variação nos dados

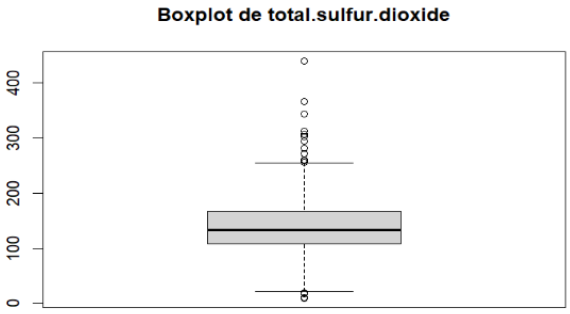
- **total.sulfur.dioxide**

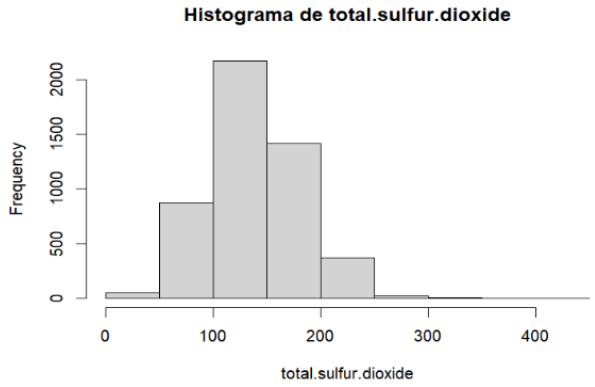
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
9.0	108.0	134.0	138.4	167.0	440.0

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
1806.0850	42.4980	30.7154	431,0000





“total.sulfur.dioxide” aparenta apresentar uma assimetria à esquerda muito leve devido a diferença da média e a mediana, possui um coef. de variação de 30.7% sendo seus dados concentrados ao redor da média.

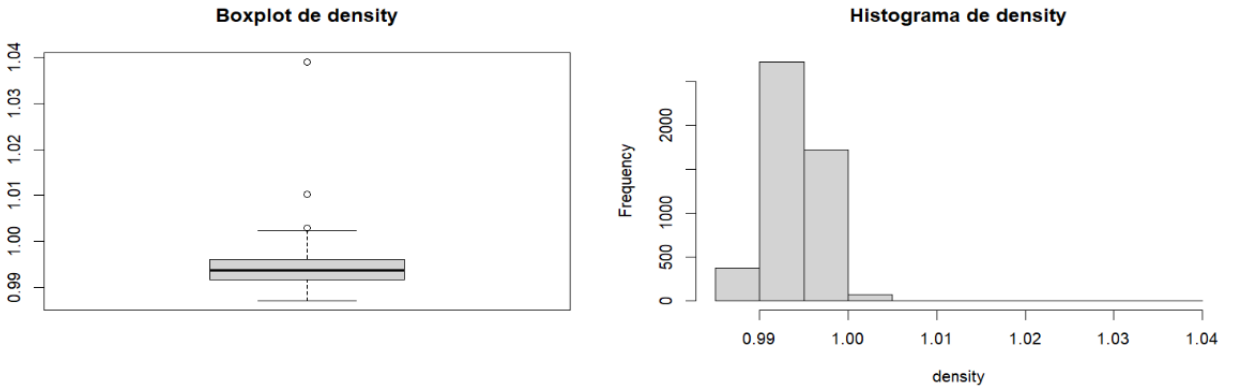
- **density**

Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.9871	0.9917	0.9937	0.9940	0.9961	1.0390

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.000009	0.002990	0.300887	0.051870



"density" aparenta ter um outlier muito grande, apesar de possuir uma variação muito pequena.

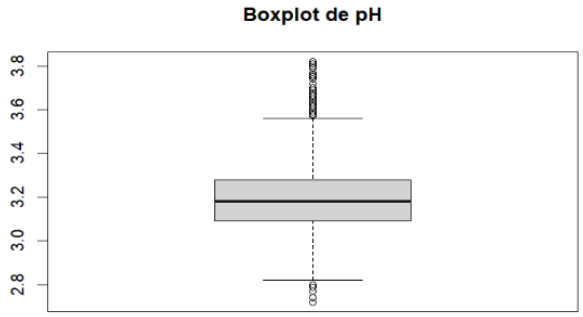
- **pH**

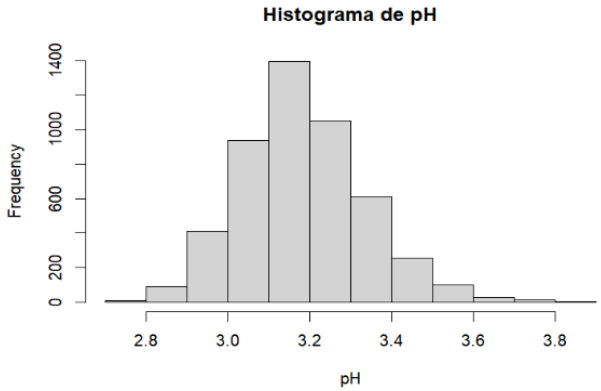
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
2.72	3.09	3.18	3.19	3.28	3.82

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.0228	0.1510	4.7361	1.1000





“pH” Apresenta distribuição quase simétrica com baixa variação, apesar do boxplot acusar muitos outliers (isso pode ocorrer pois a diferença de Q1 e Q3 é muito pequena).

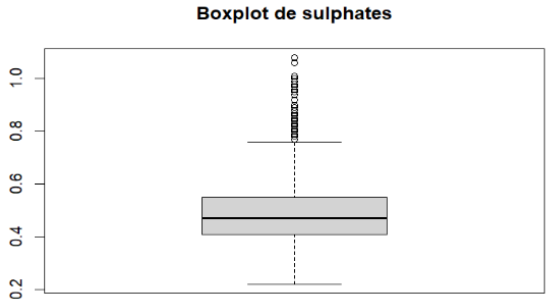
- **sulphates**

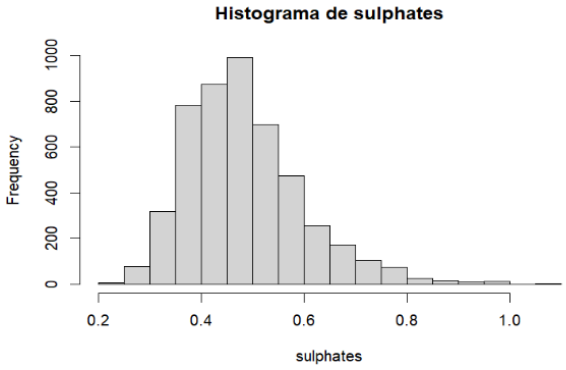
Medidas de posição

Min.	1° quartil	Mediana	Média	3° Quartil	Max.
0.22	0.41	0.47	0.49	0.55	1.08

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
0.0130	0.1141	23.2982	0.8600





“sulphates” apresenta dados quase simétricos ao ter a média e mediana muito próximos e um coeficiente de variação relativamente baixo;

- **alcohol**

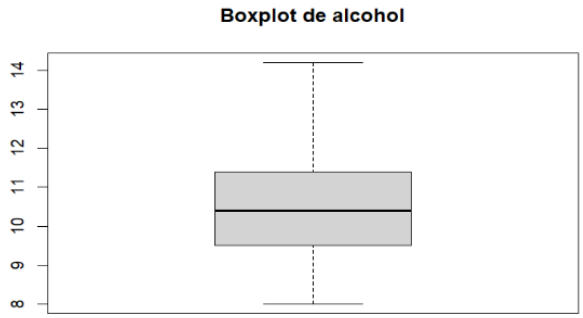
Medidas de posição

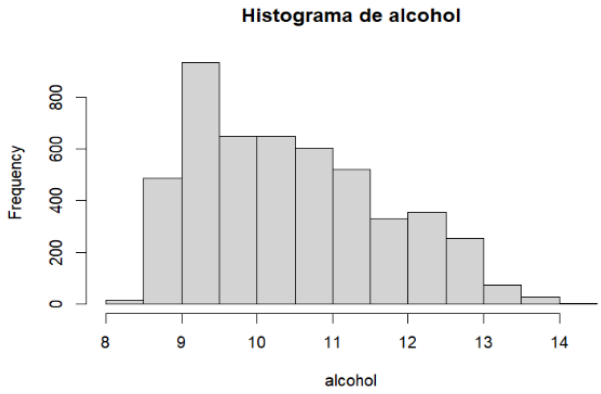
Min.	1° quartil	Mediana	Média	3° Quartil	Max.

8.0	9.5	10.4	10.5	11.4	14.2

Medidas de dispersão

Variância	Desvio Padrão	Coef. Variação(%)	Amplitude
1.5144	1.2306	11.7042	6.200





“alcohol” apresenta uma forte assimetria à esquerda, não apresenta outliers e uma variação baixa.

Anexo B - Código em R


KNN.R


Floresta Aleatória.R


Redes Neurais.R


Árvore de Decisão.R


Curva ROC.R


Observado x Previstos.R