Aprendizado de Máquina - Comparativo de Algoritmos de Classificação

Davi Esmeraldo da Silva Albuquerque

Brasília, Fevereiro de 2025

Sumário

1 Introdução

A produção agrícola eficiente e sustentável depende de uma série de fatores ambientais e técnicos. Um dos aspectos fundamentais para a maximização da produtividade e da qualidade das colheitas é a composição química do solo, que influencia diretamente no crescimento das plantas. A escolha do cultivo mais adequado para determinada região exige o conhecimento das propriedades do solo, incluindo a concentração de nutrientes essenciais.

Tradicionalmente, a análise do solo para fins agrícolas envolve métodos laboratoriais que, embora precisos, são custosos e demorados. Esses fatores limitam a frequência e abrangência das avaliações, dificultando a tomada de decisões para pequenos e médios produtores. Com o avanço da tecnologia e do aprendizado de máquinas, tornou-se possível desenvolver modelos preditivos capazes de auxiliar na recomendação de cultivos agrícolas com base em dados coletados de maneira mais acessível e eficiente.

O conjunto de dados *Soil Measures*, disponível em:
`https://www.kaggle.com/datasets/mohamedmostafa259/soil-measures`,
oferece informações sobre a composição química do solo e o tipo de cultivo mais apropriado para cada um, servindo como uma ferramenta valiosa para aprimorar a escolha de cultivos e a gestão da produtividade agrícola.

Dentre os diversos métodos de classificação disponíveis, este estudo compara três algoritmos amplamente utilizados em aprendizado de máquina:

- **K-Nearest Neighbors (KNN)**: Um método baseado em instâncias que classifica novos dados com base na similaridade com exemplos anteriores.
- **Random Forest**: Um algoritmo de aprendizado conjunto que combina múltiplas árvores de decisão para aumentar a robustez e precisão do modelo.
- **Support Vector Machines (SVM)**: Um modelo que busca encontrar um hiperplano ótimo para separar diferentes classes, utilizando transformações de kernel para casos não-lineares.

Desse modo, o objetivo deste trabalho é avaliar o desempenho desses três algoritmos na classificação dos cultivos agrícolas mais adequados, com base nas propriedades químicas do solo. Para isso, os dados serão pré-processados, os modelos treinados e testados, e seus desempenhos comparados. Ao final, espera-se identificar o modelo mais eficaz para este contexto específico.

2 Metodologia

Nesta seção, serão apresentados os três algoritmos de classificação utilizados neste estudo: K-Nearest Neighbors (KNN), Random Forest e Support Vector Machines (SVM). Cada um desses algoritmos possui características específicas que os tornam mais adequados para diferentes tipos de problemas.

2.1 K-Nearest Neighbors (KNN)

O K-Nearest Neighbors (KNN) é um algoritmo de aprendizado supervisionado baseado em instâncias. O referido algoritmo classifica um novo dado verificando quais são os k vizinhos mais próximos e atribui a classe mais frequente entre eles. A proximidade entre os pontos geralmente é medida por uma métrica de distância, como a distância Euclidiana, por exemplo.

O KNN é frequentemente utilizado quando há um padrão claro de agrupamento dos dados e quando o conjunto de treinamento é relativamente pequeno. É um algoritmo bastante útil para problemas onde a similaridade local entre as observações desempenha um papel fundamental na classificação.

As principais vantagens associadas à utilização do KNN são as seguintes:

- Simples de entender e implementar, pois baseia-se apenas na distância entre os pontos.
- Pode funcionar bem em conjuntos de dados menores e com classes bem separadas.
- Pode funcionar bem com dados não lineares e distribuições mais complexas, já que o modelo não assume nenhuma hipótese sobre a distribuição dos dados.

Por outro lado, as desvantagens comuns do uso do KNN são:

- Sensível à escolha do número de vizinhos (k).
- Alto custo computacional para grandes bases de dados, pois cada nova previsão requer o cálculo da distância com todas as amostras de treinamento.
- Depende da escala das variáveis, sendo necessário aplicar normalização para evitar viés em variáveis com magnitudes diferentes.

2.2 Random Forest

O Random Forest é um modelo baseado em árvores de decisão que utiliza o conceito de aprendizado de conjunto (ensemble learning). Essa técnica cria várias árvores de decisão a partir de subconjuntos aleatórios dos dados e das variáveis, combinando os resultados por meio de votação para produzir uma previsão mais robusta e precisa.

Este algoritmo é amplamente utilizado em problemas de classificação onde os dados apresentam padrões não lineares e podem conter valores atípicos. O mesmo é ideal quando a interpretabilidade não é um fator crítico e o foco está no desempenho do modelo.

As principais vantagens proporcionadas pelo uso do Random Forest são:

- Menos propenso ao overfitting em comparação com uma única árvore de decisão.
- Pode lidar com grandes volumes de dados e alto número de variáveis.
- Funciona bem mesmo sem normalização, pois não é sensível à escala das variáveis.

Por outro lado, suas principais desvantagens são:

- Costuma requerer um maior tempo de treinamento devido à necessidade de construir múltiplas árvores.
- Pode ser menos interpretável do que modelos mais simples, como árvores de decisão individuais.
- Requer o ajuste de diversos hiperparâmetros para otimizar o desempenho, como o número de árvores e a profundidade máxima das árvores.

2.3 Support Vector Machines (SVM)

O Support Vector Machines (SVM) é um algoritmo de aprendizado supervisionado que busca encontrar um hiperplano ótimo para separar as classes, a fim de maximizar a margem entre os pontos mais próximos da fronteira de decisão. Caso os dados não sejam linearmente separáveis, o SVM pode utilizar funções kernel para transformar os dados em um espaço de maior dimensão, onde a separação seja possível.

O SVM é amplamente empregado em problemas de classificação onde há um baixo número de amostras e um espaço de características bem definido. Ele é especialmente útil para dados de alta dimensionalidade e problemas onde a separação entre as classes é clara.

As principais vantagens associadas ao uso do SVM são:

- Eficiência em dados de alta dimensionalidade.
- Pode lidar bem com dados não linearmente separáveis, utilizando funções kernel.
- Robusto contra overfitting, especialmente em espaços de baixa dimensão.

Por outro lado, as desvantagens do SVM incluem:

- Escolher o kernel e os hiperparâmetros corretos pode ser desafiador e exigir ajustes.
- Pode ser computacionalmente intensivo em conjuntos de dados muito grandes.
- O modelo final pode ser mais difícil de interpretar, especialmente quando são utilizados kernels não lineares.

2.4 Separação dos dados

Inicialmente, os dados foram divididos em subconjuntos para treinamento (80%) e teste (20%) de maneira estratificada, garantindo a preservação da distribuição das classes do tipo de cultivo ideal. Em seguida, a validação dos modelos foi realizada por meio de validação cruzada dentro do conjunto de treinamento, permitindo a otimização dos hiperparâmetros sem a necessidade de um conjunto de validação separado. Por fim, a avaliação final dos modelos foi conduzida no conjunto de teste, assegurando uma análise robusta do desempenho dos modelos.

2.5 Tratamento dos dados

Após a separação dos dados, foi definida uma receita de pré-processamento, que especifica que a variável alvo é o tipo de cultivo e que todas as variáveis preditoras numéricas devem ser normalizadas para aplicação dos modelos KNN e SVM. Essa normalização visa padronizar os dados, ajustando-os para uma mesma escala, o que é particularmente importante para modelos como o KNN e SVM.

2.6 Validação Cruzada

A validação cruzada foi aplicada utilizando 5 subconjuntos (folds) no conjunto de treinamento. Esse método divide os dados, permitindo que em cada iteração, um dos subconjuntos seja utilizado para teste, enquanto os demais são usados para treinar o modelo. Dessa forma, todos os dados participam tanto do treinamento quanto do teste, garantindo uma avaliação mais robusta e confiável do desempenho dos modelos.

2.7 Tunagem de Hiperparâmetros

Com a validação cruzada definida, prosseguiu-se para a tunagem dos hiperparâmetros por meio da busca em grades pré-definidas (grid search). Esse processo permite encontrar a melhor configuração entre os valores de hiperparâmetros definidos a priori para cada modelo, otimizando seu desempenho. Assim, com o intuito de se ter bom equilíbrio entre ocorrer overfitting ou underfitting foram ajustados os seguintes hiperparâmetros dos modelos:

- **KNN:**
  - O número de vizinhos (*neighbors*) foi ajustado para os valores de 3 a 25.
- **Random Forest:**
  - O número de variáveis consideradas em cada árvore (`mtry`) variou entre 1, 2 e 3.
  - Para o número de árvores (`trees`), foram testados 100, 200 e 500.
  - O número mínimo de amostras por nó (`min_n`) variou entre 1, 5 e 10.
- **SVM:**
  - O parâmetro de regularização (cost) foi definido para variar de 0.1 a 10.
  - O parâmetro de largura do kernel RBF (rbf_sigma) foi definido para variar de 0.01 a 1.

A busca pelos melhores hiperparâmetros foi realizada utilizando a acurácia como métrica principal de avaliação. Por conseguinte, após a otimização dos hiperparâmetros, os melhores valores encontrados para cada modelo foram:

- **KNN:** $neighbors = 10$

- **Random Forest:** $mtry = 1, trees = 100$, $min_n = 10$
- **SVM:** $cost = 5.96, rbf\_sigma = 1.81$

Por fim, os modelos finais, treinados utilizando esses hiperparâmetros ótimos, foram avaliados no conjunto de teste, possibilitando uma estimativa mais realista de seu desempenho em dados não vistos.

3 Análises

3.1 Banco de Dados

As variáveis presentes no conjunto de dados são:

- N: Proporção de nitrogênio no solo.
- P: Proporção de fósforo no solo.
- K: Proporção de potássio no solo.
- pH: Valor do pH do solo.
- ”Crop”: Diferentes tipos de cultivo recomendados.

Cada uma das linhas do conjunto de dados representa uma amostra de solo de um campo específico com a indicação do cultivo mais indicado para aquele campo, considerando as medições dos outros parâmetros. Assim, o conjunto de dados contém um total de 2200 observações, com 100 amostras para cada tipo de cultivo, totalizando 22 diferentes cultivos indicados.

Os dados utilizados na análise não apresentam valores faltantes, o que elimina a necessidade de técnicas de imputação ou remoção de observações incompletas. Isso garante que todas as variáveis do conjunto de dados possam ser utilizadas em sua totalidade, preservando a integridade das informações e evitando possíveis vieses causados pela manipulação de dados ausentes.

3.2 Análise Exploratória

Em relação à distribuição dos dados, os gráficos 1 e 2 apresentados a seguir fornecem uma visualização detalhada das variáveis preditoras. O primeiro aglomerado de gráficos,1, exibe histogramas, permitindo observar a distribuição de frequência dos valores de cada variável. Já

o segundo aglomerado de gráficos,2, apresenta boxplots, os quais viabilizam a identificação de possíveis outliers e a dispersão dos dados.

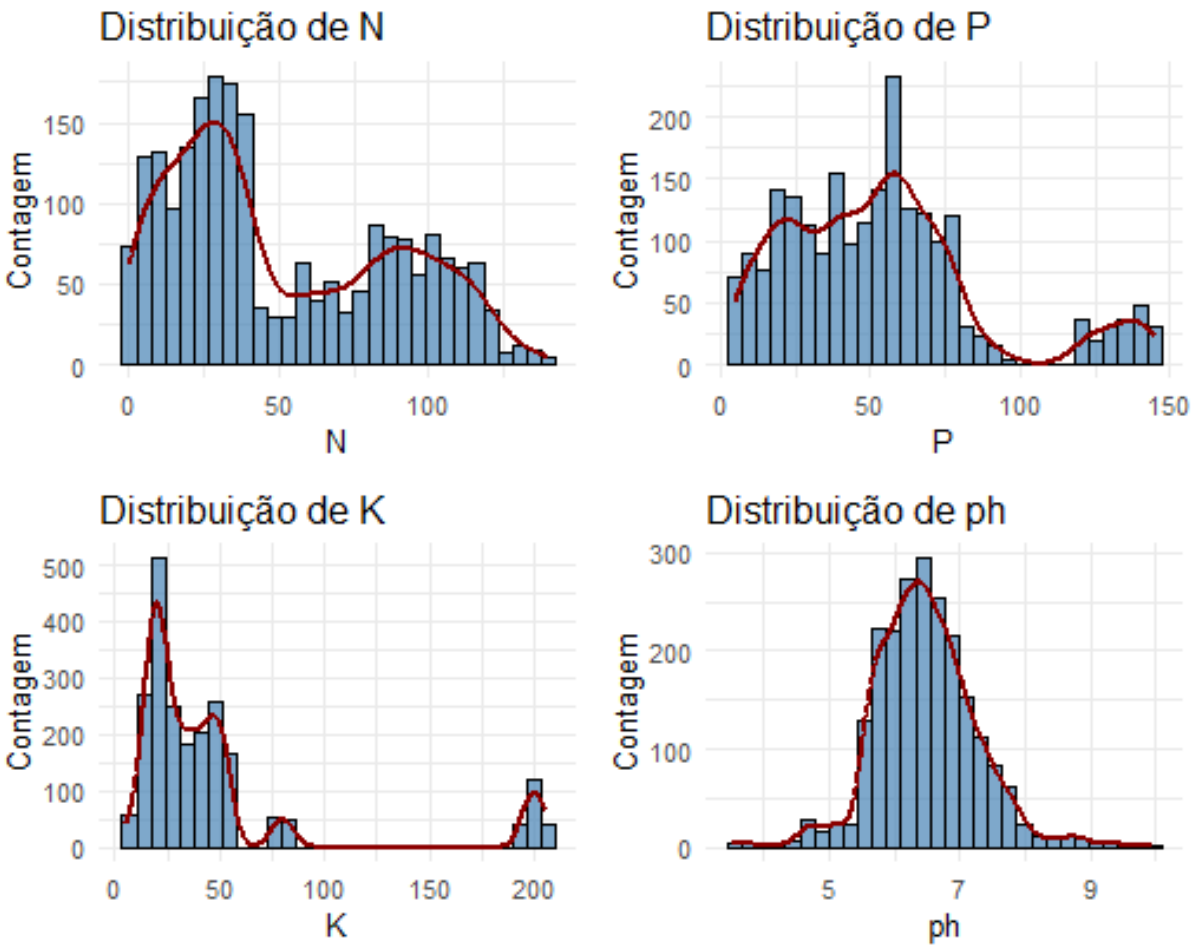

Figura 1: Histogramas das variáveis preditoras

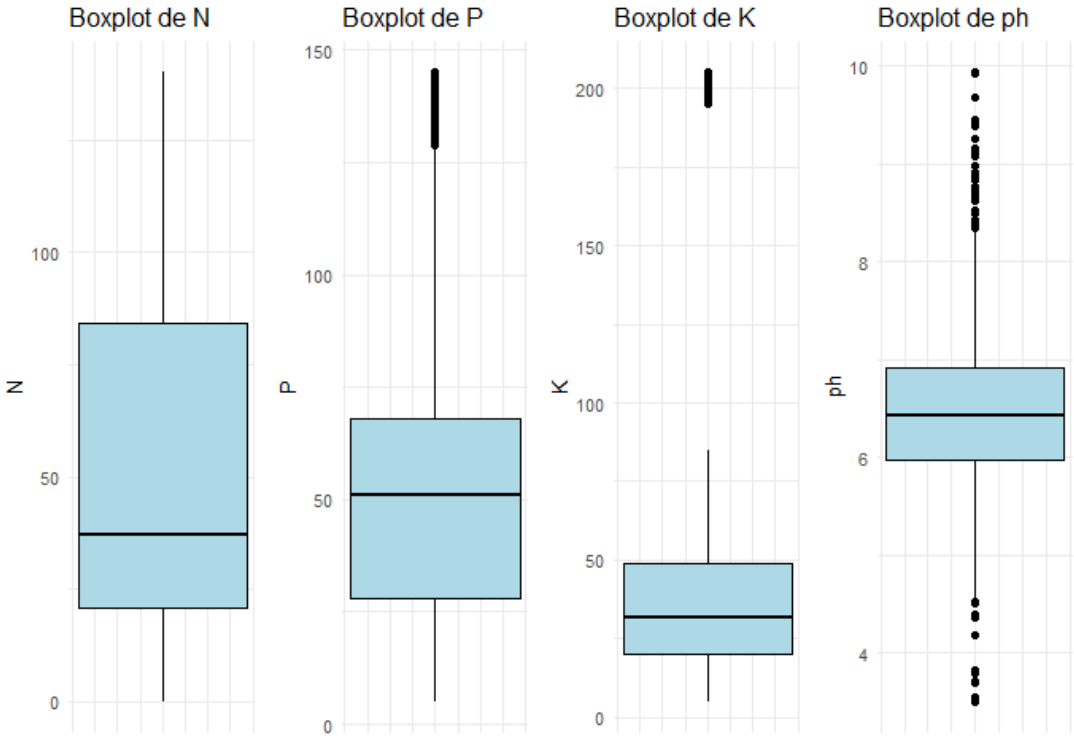

Figura 2: Boxplots das variáveis preditoras

A partir do primeiro aglomerado de gráficos,1, nota-se que as variáveis N, P e K apresentam distribuições multimodais, ou seja, com múltiplos picos, o que sugere a presença de diferentes padrões nos dados. A variável pH, por sua vez, segue uma distribuição mais próxima da normal, com uma leve assimetria à esquerda.

Já por meio da visualização do segundo conjunto de gráficos,2, tem-se nítida a presença de outliers, especialmente nas variáveis P, K e pH, que apresentam pontos extremos acima ou abaixo dos limites do boxplot. Isso sugere que há valores significativamente diferentes da maioria dos dados, o que pode impactar nas análises estatísticas e modelos preditivos.

Desse modo, considerando as diferentes escalas das variáveis e a presença de outliers, torna-se essencial normalizar ou padronizar os dados. Ao alinhar suas escalas, espera-se uma melhor robustez dos modelos SVM e KNN aplicados neste trabalho, uma vez que esses métodos são sensíveis a distorções nas distâncias e na variação dos dados.

Já os gráficos apresentados a seguir 3 e 4, fornecem uma visão detalhada das relações entre as variáveis preditoras. O primeiro agrupamento de gráficos 3, exibe diagramas de dispersão entre cada par de variáveis, permitindo identificar padrões de associação e tendências lineares nos dados. As linhas vermelhas representam tendências ajustadas, destacando possíveis relações entre as variáveis.

Já o segundo grupo de gráficos, 4, apresenta um mapa de calor das correlações, ilustrando a intensidade e a direção das relações entre as variáveis. Nesse gráfico de correlação, os valores variam entre -1 e 1, onde valores positivos indicam correlação positiva (quando uma variável aumenta, a outra também tende a aumentar), e valores negativos indicam correlação negativa (quando uma variável aumenta, a outra tende a diminuir).

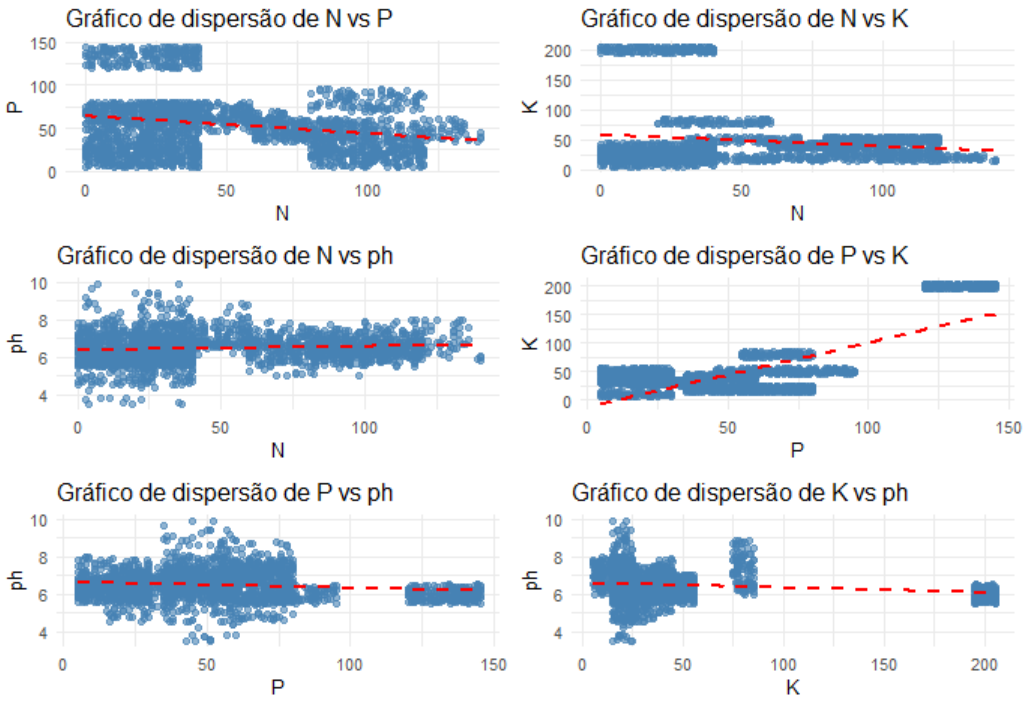

Figura 3: Gráfico de dispersão das variáveis preditoras

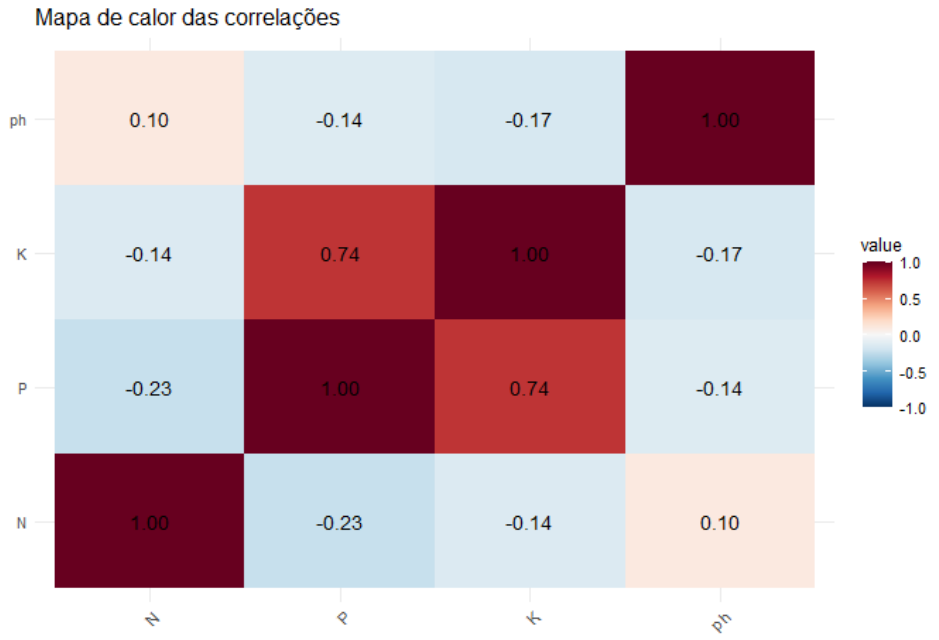

Figura 4: Gráfico de correlação das variáveis preditoras

Por meio da visualização conjunta dos gráficos 3 e 4, desprende-se que algumas variáveis possuem relações bem definidas, como P e K, que apresentam uma correlação forte e positiva (0.74), sugerindo que um aumento no valor de P está frequentemente associado a um aumento nos valores de K. Em contrapartida, a variável N possui correlações fracas com as demais variáveis, indicando uma relação menos linear. O pH também apresenta correlações baixas com os outros atributos, sugerindo que essa variável pode ser relativamente independente no conjunto de dados.

Por fim, os dois seguintes gráficos 5 e 6 objetivam viabilizar uma melhor visualização dos dados. No gráfico 5 foi aplicada a técnica de redução de dimensionalidade, análise de componentes principais (PCA). A mesma transforma as variáveis originais em componentes ortogonais, permitindo observar a distribuição das culturas agrícolas em um espaço de menor dimensionalidade. Já no gráfico 6 são representados boxplots para cada tipo de cultivo ideal.

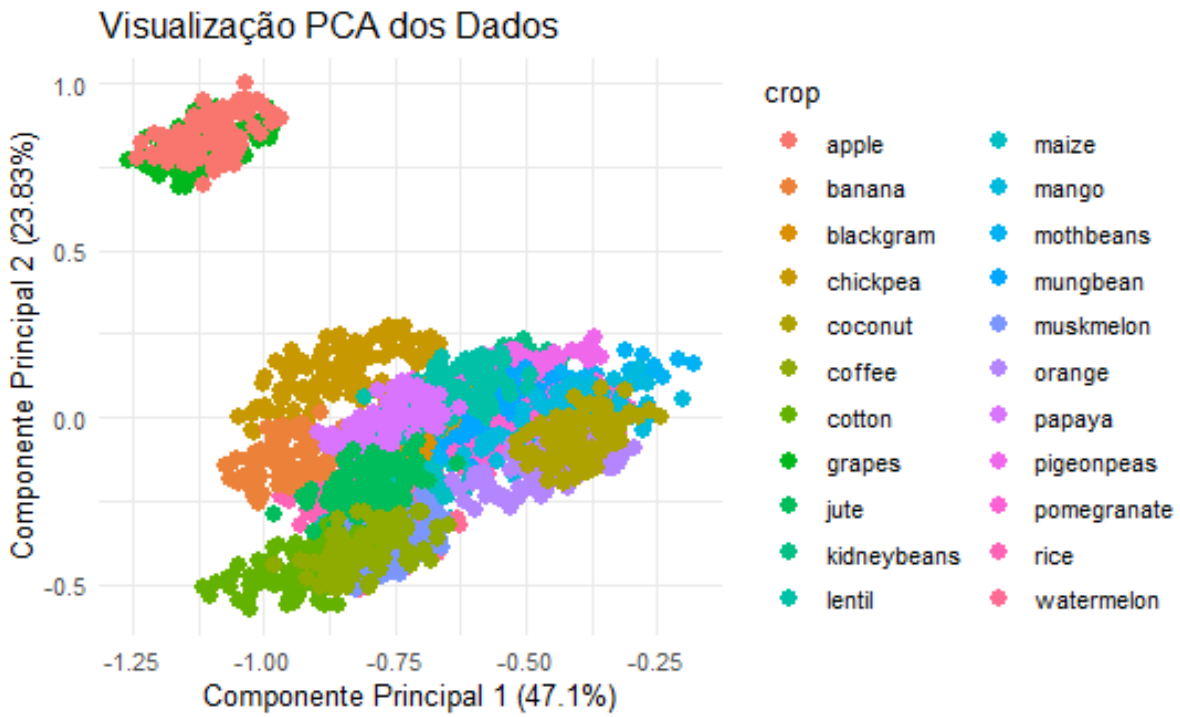

Figura 5: PCA das variáveis preditoras

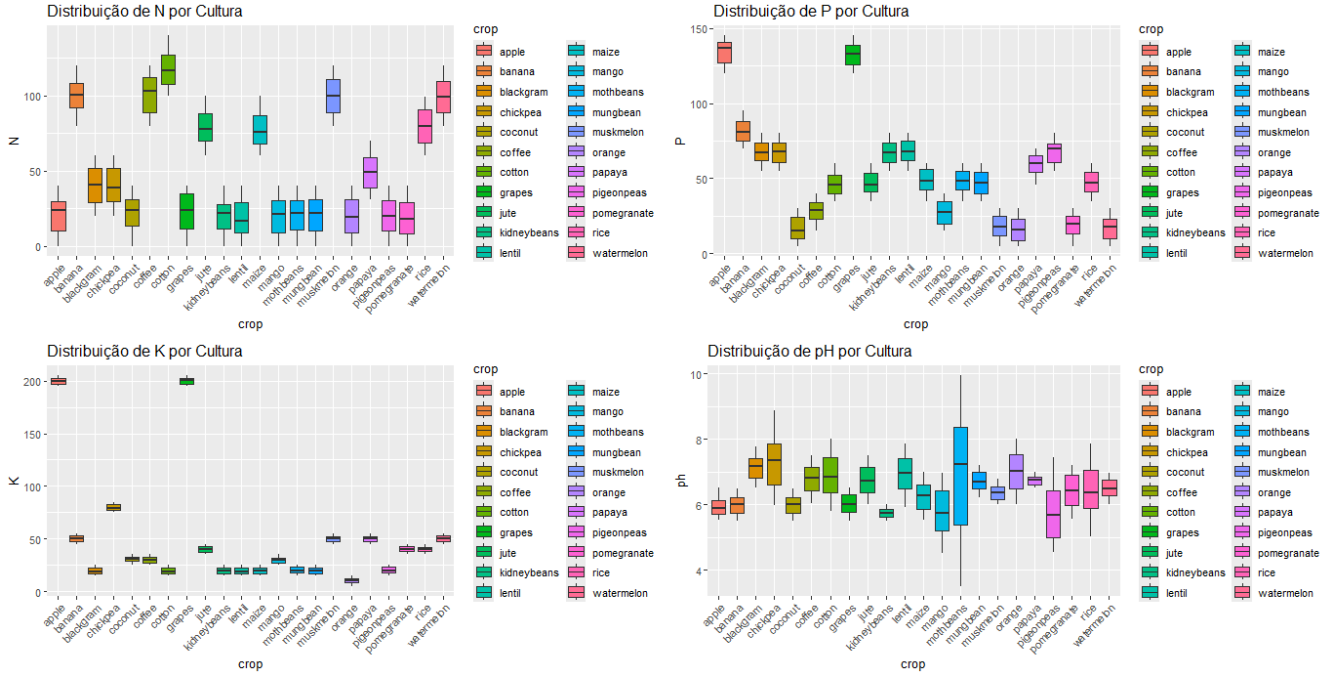

Figura 6: Boxplot das variáveis preditoras para cada cultivo recomendado

No gráfico 5, observa-se que o primeiro componente principal explica, aproximadamente, 47% da variabilidade dos dados, enquanto o segundo componente principal explica, aproximadamente, 24%, totalizando 71% da variabilidade dos dados, indicando que a maior parte da informação presente nas variáveis originais está bem representada nesses dois componentes principais.

Além disso, percebe-se que os diferentes tipos de cultivo recomendados apresentam agrupamentos bem definidos, sugerindo que as combinações das características analisadas do solo influenciam diretamente na recomendação do tipo de cultivo ideal. No entanto, o fato de os dados estarem, em geral, condensados pode dificultar a diferenciação dos cultivos e afetar a performance de modelos de aprendizado de máquina na classificação.

Adicionalmente, um aspecto notável é a presença de um grupo isolado no canto superior esquerdo do gráfico 5 e boxplots com características distintas dos demais em 6. Esse grupo, que inclui os cultivos de ”apple” e ”grape”, destaca-se por suas propriedades que diferem substancialmente das demais culturas. Isso reforça a ideia de que, embora existam agrupamentos generalizados, cultivos específicos podem exigir condições diferenciadas.

4 Avaliação dos modelos

Primeiramente, ao analisar o tempo de processamento como uma medida relacionada ao custo computacional, observou-se o ajuste dos modelos e seus respectivos hiperparâmetros

ótimos. O KNN foi o mais rápido, com um tempo de execução de 4,67 segundos. O Random Forest, embora ainda relativamente rápido, levou 87,64 segundos, um tempo 18,76 vezes maior que o do KNN. Já o SVM apresentou o maior tempo de processamento, atingindo 483,03 segundos, o que evidencia sua maior complexidade e demanda computacional, devido principalmente à estrutura do algoritmo.

A tabela 1 a seguir apresenta a comparação entre os modelos Random Forest, SVM e KNN em diferentes métricas de desempenho. Esses indicadores auxiliam a avaliar a precisão das previsões, a capacidade de identificar corretamente os casos positivos e a confiabilidade geral de cada modelo.

Tabela 1: Resultados das métricas para diferentes técnicas de classificação

Métrica	**SVM**	**KNN**	**Random Forest**
Acurácia	75.91%	73.18%	77.95%
Precisão	76.13%	74.11%	78.42%
Recall	76.38%	73.70%	78.83%
F1-Score	75.41%	73.21%	77.72%

Ao comparar a acurácia dos três modelos, o Random Forest apresentou o melhor desempenho, alcançando 78%, seguido pelo SVM, com 75,9%. O KNN, por sua vez, obteve a menor acurácia, com 73,2%. Esses resultados sugerem que, especificamente para essa tarefa e para os conjuntos de dados analisados, o Random Forest, ajustado com os hiperparâmetros ótimos obtidos, foi o modelo mais preciso, seguido pelo SVM e, por último, pelo KNN.

Ainda referente às métricas de desempenho observadas na tabela 1, observa-se que, além da acurácia, o Random Forest também se destacou em outras métricas importantes. A precisão, que mede a proporção de predições positivas corretas, foi mais alta no Random Forest, com 78,42%, seguida pelo SVM com 76,13%, e por último o KNN, com 74,11%. Esse resultado reforça a ideia de que o Random Forest, além de ter a maior acurácia, também conseguiu identificar corretamente a maior proporção de positivos, o que é crucial em cenários onde a precisão das predições positivas é importante.

Em relação ao Recall, que indica a capacidade do modelo de identificar corretamente as instâncias positivas, o Random Forest novamente se destacou, alcançando 78,83%. O SVM e o KNN apresentaram valores mais baixos, com 76,38% e 73,70%, respectivamente. Isso sugere que o Random Forest teve um desempenho superior ao identificar as instâncias positivas, o que pode ser particularmente relevante em contextos em que minimizar os falsos negativos é essencial.

O F1-Score, que é a média harmônica entre a precisão e o recall, foi igualmente mais alto no Random Forest, com 77,72%, comparado ao SVM (75,41%) e ao KNN (73,21%). O F1-Score equilibra os trade-offs entre precisão e recall, e o valor superior do Random Forest sugere um melhor compromisso entre essas duas métricas, o que é ideal em problemas onde tanto a identificação correta de positivos quanto a minimização de falsos positivos são importantes.

Finalmente, a análise das matrizes de confusão a seguir permitiu compreender o desempenho de cada modelo em relação aos diferentes tipos de cultivo recomendados, identificando padrões de acerto e erro. Os resultados detalhados dessa avaliação estão apresentados na tabela 2.

Tabela 2: Acurácia dos modelos KNN, Random Forest e SVM para cada classe

KNN

Classe	Acurácia
Muskmelon	0.4615
Watermelon	0.4667
Grapes	0.4762
Lentil	0.4762
Apple	0.5789
Coconut	0.5833
Pigeonpeas	0.5909
Rice	0.6190
Blackgram	0.6364
Mungbean	0.6667
Jute	0.7059
Coffee	0.7692
Kidneybeans	0.7931
Pomegranate	0.8333
Cotton	0.9048
Mango	0.9091
Mothbeans	0.9333
Maize	0.9375
Papaya	0.9629
Banana	1.0000
Chickpea	1.0000
Orange	1.0000

Random Forest

Classe	Acurácia
Watermelon	0.4286
Grapes	0.4444
Muskmelon	0.4615
Apple	0.5455
Rice	0.6000
Mungbean	0.6923
Pigeonpeas	0.7059
Blackgram	0.7273
Coconut	0.7273
Jute	0.7692
Lentil	0.7778
Kidneybeans	0.8065
Mothbeans	0.8333
Mango	0.8462
Coffee	0.9333
Cotton	0.9524
Banana	1.0000
Chickpea	1.0000
Maize	1.0000
Orange	1.0000
Papaya	1.0000
Pomegranate	1.0000

SVM

Classe	Acurácia
Grapes	0.2000
Apple	0.4000
Watermelon	0.4286
Muskmelon	0.4615
Blackgram	0.6667
Lentil	0.6667
Mungbean	0.6800
Coconut	0.6957
Rice	0.7647
Pigeonpeas	0.7857
Coffee	0.8000
Jute	0.8000
Kidneybeans	0.8065
Mothbeans	0.8824
Cotton	0.9091
Mango	0.9167
Maize	0.9375
Pomegranate	0.9474
Banana	1.0000
Chickpea	1.0000
Orange	1.0000
Papaya	1.0000

Ao comparar a acurácia dos modelos para cada tipo de cultivo, observou-se que o Random Forest teve um desempenho superior em diversas culturas, especialmente nas seguintes:

- Jute e Kidneybeans: O Random Forest apresentou acurácias de 76,9% e 80,6%, respectivamente, superando o KNN, mas tendo um desempenho semelhante ao SVM em Kidneybeans e ligeiramente inferior em Jute.

- Lentil: O Random Forest alcançou 77,8%, superando o KNN (47,6%) e o SVM (66,7%).
- Coffee e Cotton: O Random Forest demonstrou uma acurácia superior, com 93,3% para Coffee e 95,2% para Cotton, sendo mais eficaz do que os demais modelos.
- Maize e Pomegranate: Essas foram as culturas que atingiram 100% de acurácia apenas com o Random Forest, evidenciando sua alta eficácia para essas classes.
- Por outro lado, para algumas culturas como Mothbeans e Mango, o Random Forest não teve o melhor desempenho, apresentando acurácias de 83,3% e 84,6%, enquanto o KNN (93,3% e 90,9%) e o SVM (88,2% e 91,7%) obtiveram resultados superiores.
- Em Grapes, Apple e Watermelon, o Random Forest teve um desempenho fraco, com acurácias de 44,4%, 54,5% e 42,9%, respectivamente. O SVM apresentou um desempenho especialmente baixo para Grapes (20%) e Apple (40%), enquanto o KNN superou ambos nesses casos.

5 Conclusão

O presente trabalho procurou trabalhar três algoritmos de classificação (KNN, Random Forest e SVM) para a recomendação de cultivos agrícolas com base nas propriedades químicas do solo. A partir da análise exploratória e dos testes realizados, notou-se que, embora o KNN apresente como vantagem um tempo de processamento mais rápido, sua acurácia ficou mais abaixo em relação aos demais modelos, evidenciando certas limitações na generalização. O SVM, por sua vez, demonstrou uma performance relativamente boa em termos de acurácia, mas seu elevado custo computacional pode inviabilizá-lo para cenários com grandes volumes de dados.

Destacou-se, por fim, que o modelo Random Forest se mostrou ligeiramente superior, alcançando uma acurácia de 78%, além de apresentar um custo computacional significativamente menor do que o algoritmo SVM. Ademais, a análise das matrizes de confusão reforçou a robustez do Random Forest, que se adaptou melhor às variações entre as classes e obteve taxas de acerto mais elevadas na maior parte dos cultivos. Dessa forma, define-se, para o estudo em questão, o Random Forest como a alternativa mais promissora, proporcionando um equilíbrio vantajoso entre precisão e eficiência.

Finalmente, faz-se válido destacar que futuras investigações podem explorar a integração de outros métodos, o aprimoramento dos hiperparâmetros ou a inclusão de mais variáveis preditoras a fim de otimizar ainda mais os resultados obtidos e ampliar o potencial de aplicação dessas técnicas no contexto agrícola.