BANANA LEARNING: USO DA INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO DE QUALIDADE DE BANANAS

Eduardo dos Santos Rocha[1], Clésio Rodrigues da Silva Júnior[1], Gabriel Alves Gama[1], Leonardo Cardoso de Moura[1], Vania de Fatima Lemes de Miranda[1], Waldemar Patrique Flores Silva[1]

[1] Universidade Federal de Uberlândia, Monte Carmelo, MG (eduardo.rocha@ufu.br)

RESUMO: Atualmente, com os avanços da computação, as novas tecnologias estão cada vez mais presentes em diversas áreas, como na agricultura. O presente trabalho tem por objetivo apresentar um estudo sobre técnicas de aprendizado de máquinas e técnicas estatísticas no âmbito da identificação e classificação da qualidade de bananas por meio da análise e interpretação de vários fatores mensurados. Inicialmente, foi realizado o pré-processamento dos dados onde inconsistências (Ruídos, Outliers e Valores missing) foram devidamente tratadas, permitindo uma maior compreensão sobre informações vigentes desses dados.O modelo mais eficiente foi o K-Nearest Neighbors - KNN, obtendo acurácia de 98,17% e f1-score de 98,00%.

Palavras-chave: aprendizado de máquina, tecnologia, modelos de classificação.

INTRODUÇÃO

Conforme a revista Brasil Hortifruti (2021), bananas estão no topo do ranking (1°) de frutas mais vendidas no mundo e o desenvolvimento de técnicas capazes de detectar precocemente a qualidade é de grande valia. Para Tesla e Childress (1993), “O futuro vai mostrar os resultados e julgar a cada segundo suas realizações.”, nessa frase é possível interpretar a atuação da Inteligência Artificial - IA em muitos aspectos da vida moderna, principalmente, no auxílio direto ou indireto na área da indústria moderna. No contexto da indústria, a detecção precoce da qualidade das bananas, auxilia na produção em massa.

Utilizando técnicas de aprendizado de máquina e mineração de dados, Priyanka *et al.* (2015), desenvolveram uma investigação dos problemas relacionados a doenças causadas em bananas, utilizando máquina de vetores, rede neural convolucional e regressão.

Portanto, o presente trabalho tem por objetivo apresentar um estudo sobre técnicas de aprendizado de máquinas e técnicas estatísticas no âmbito de classificação da qualidade da banana de maneira eficiente e eficaz, empregando métodos diversificados com o intuito de alcançar a otimização dos resultados.

MATERIAL E MÉTODOS

Dentro da IA, há a linha de estudos em Aprendizado de Máquinas que emprega algoritmos capazes de auxiliar na classificação de uma base de dados. Neste trabalho, os classificadores utilizados foram KNN, Categorical Boosting - CatBoost e eXtreme Gradient Boosting - XGBoost. Para este estudo de classificação foi utilizada à base de dados "🍌 | Banana Quality" extraída da plataforma Kaggle (2024) com as variáveis tamanho, peso, doçura, suavidade, tempo de colheita, maturação, acidez, qualidade, totalizando um total de oito variáveis e oito mil observações. A avaliação desses classificadores possibilitará a verificação da eficiência dos algoritmos para identificação da qualidade das bananas.

Nessa seção, serão apresentados os resultados dos algoritmos mencionados anteriormente. A variável alvo, ou seja, aquela que queremos explicar ou prever com base em outras variáveis independentes, possuí duas classes: Ruim e Bom. As métricas de avaliação empregadas foram:

- Precisão: Mede a proporção de verdadeiros positivos em relação ao total de resultados positivos previstos pelo modelo. Em outras palavras, ela indica a acurácia das previsões positivas.
- Revocação: Mede a proporção de verdadeiros positivos em relação ao total de verdadeiros positivos existentes nos dados. Essa métrica reflete a capacidade do modelo de identificar corretamente todas as instâncias da classe positiva.
- F1-Score: É a média harmônica entre a precisão e o recall, oferecendo um equilíbrio entre essas duas métricas. O F1-score é particularmente útil quando se lida com um desbalanceamento entre as classes.
- Suporte: Número de ocorrências reais de cada classe na base de dados.
- Acurácia: Mede a proporção de previsões corretas (tanto positivas quanto negativas) em relação ao total de previsões.
- Média Macro: Calcula a média das métricas (como precisão, recall) para cada classe individualmente e, em seguida, tira a média desses valores.
- Média Ponderada: Calcula a média das métricas para cada classe, mas pondera cada classe pelo número de amostras nessa classe. Assim, classes com mais exemplos têm maior influência na métrica final.
- Validação cruzada: Técnica de avaliação que divide o conjunto de dados em múltiplos subconjuntos, treinando e testando o modelo em diferentes combinações desses subconjuntos para obter uma estimativa mais robusta de seu desempenho.

XGBOOST: O modelo XGBOOST é um algoritmo empregado na classificação e regressão, originado da evolução do Random Forest, Boosting e Gradient Boosting. Onde implementa técnicas que oferecem um alto desempenho e escalabilidade, criando um algoritmo otimizado, que visa trabalhar com dados avulsos com um aprendizado ponderado, lidando com pesos ao utilizar árvores de decisão. Ademais, implementa técnicas de podas de árvores, regularização L1 (lasso) penaliza a somas dos valores obsoletos dos pesos e L2 (Ridge) em que penaliza a soma dos quadrados dos pesos, etc.

O objetivo principal do XGBoost é minimizar a função de perda enquanto adiciona regularização para evitar *overfitting* (quando o modelo se ajusta excessivamente aos dados de treino, mas tem baixo desempenho em novos dados), melhorando a precisão das previsões. Ele constrói árvores de decisão sequenciais para corrigir os erros das anteriores, em um processo conhecido como *boosting* (uma técnica em que cada novo modelo corrige os erros dos anteriores, aprimorando a desempenho).

CATBOOST: CatBoost é adequado para dados categóricos heterogêneos, muito utilizado em aprendizagem supervisionada em tarefas de classificação e regressão. Seu destaque é a alta velocidade de processamento e transformação das variáveis categóricas e numéricas automaticamente.

O principal objetivo do CatBoost é reduzir a função de perda ao mesmo tempo que incorpora regularização para prevenir o *overfitting*, aumentando a precisão das previsões. Ele gera árvores de decisão em sequência, ajustando os erros anteriores por meio do boosting e é especialmente eficiente no tratamento de variáveis categóricas.

KNN: O modelo KNN é um algoritmo de classificação em aprendizado supervisionado, que utiliza conceitos de aprendizagem baseada em instâncias em seu modelo matemáticos, ou seja, encontra os K vizinhos mais próximos, considerando medidas de distância Manhattan ou Euclidianas. Portanto, graficamente um novo ponto possuirá uma classe definida a partir do cálculo de distância entre esse ponto e seus vizinhos mais próximos.

O principal objetivo do KNN é classificar ou prever com base na proximidade dos dados. Ele identifica os **k** vizinhos mais próximos de um ponto e toma decisões considerando a maioria para classificação ou a média para regressão, sem treinamento explícito. O KNN evita *overfitting* ao escolher um valor adequado de **k**, equilibrando a sensibilidade do modelo a ruídos ou padrões.

RESULTADOS E DISCUSSÃO

XGBOOST:

Tabela 1. Métricas de desempenho do modelo XGBOOST

Classe	Precisão	Revocação	F1-Score	Suporte
Ruim	0,97	0,98	0,98	1211
Bom	0,97	0,98	0,98	1189
Acurácia	-	-	0,98	2400
Média Macro	0,98	0,98	0,98	2400
Média Ponderada	0,98	0,98	0,98	2400

O modelo XGBoost apresentou um desempenho excelente em todas as métricas. A acurácia geral de 97,62% indica que o modelo consegue classificar corretamente a maioria das instâncias. A validação cruzada, que fornece uma estimativa mais robusta do desempenho do modelo, reforça essa conclusão com uma acurácia de 97,45%. O equilíbrio nas classes "Ruim" e "Bom" é um ponto positivo, evidenciado por métricas como a precisão e o F1-score. Isso sugere que o modelo tem uma capacidade elevada de generalizar e não está enviesado para nenhuma das classes, o que é essencial para aplicações práticas.

CATBOOST:

Tabela 2. Métricas de desempenho do modelo CATBOOST

Classe	Precisão	Revocação	F1-Score	Suporte
Ruim	0,98	0,98	0,98	1211
Bom	0,98	0,98	0,98	1189
Acurácia	-	-	0,98	2400
Média Macro	0,98	0,98	0,98	2400
Média Ponderada	0,98	0,98	0,98	2400

Assim como o XGBoost, o CatBoost também demonstrou um desempenho notável, com uma acurácia de 98,00% e uma validação cruzada indicando 97,94%. A similaridade entre as métricas de precisão, revocação e F1-score (todas em 0,98) reflete a robustez do modelo. A capacidade de lidar bem com variáveis categóricas, característica nativa do CatBoost, pode ter contribuído para esse excelente desempenho. Este resultado reforça a eficácia do CatBoost em cenários onde as variáveis categóricas são transformadas em numéricas, mantendo o equilíbrio entre as classes e garantindo um alto grau de acerto nas previsões.

KNN:

Tabela 3. Métricas de desempenho do modelo KNN

Classe	Precisão	Revocação	F1-Score	Suporte
Ruim	0,98	0,98	0,98	1211
Bom	0,98	0,98	0,98	1189
Acurácia	-	-	0,98	2400
Média Macro	0,98	0,98	0,98	2400
Média Ponderada	0,98	0,98	0,98	2400

O KNN, por sua vez, também apresentou um desempenho impressionante, com uma acurácia ligeiramente superior aos outros modelos, alcançando 98,17%. A validação cruzada confirmou esse alto desempenho, indicando 98,07%. O equilíbrio entre as classes "Ruim" e "Bom" é bem representado, com ambas as classes atingindo 0,98 nas métricas de precisão, revocação e F1-score. Vale destacar que, embora o KNN seja um algoritmo relativamente simples, ele conseguiu competir de igual para igual com algoritmos mais sofisticados como o XGBoost e o CatBoost. Isso pode ser explicado pela natureza dos dados e pela escolha adequada das métricas e técnicas de pré-processamento.

Comparação e Análise dos Modelos:

A análise dos três modelos XGBoost, CatBoost e KNN evidencia um excelente desempenho geral, com acurácia próxima ou superior a 98% para todos eles. A precisão e o F1-score consistentes indicam que os modelos conseguem equilibrar bem a classificação entre as classes “Ruim” e “Bom”, sem enviesamento. Embora o XGBoost e o CatBoost sejam conhecidos por sua robustez em problemas complexos, é interessante observar que o KNN, um algoritmo mais simples, obteve resultados similares. Isso sugere que o pré-processamento e a escolha dos hiperparâmetros foram eficazes, e que, para o conjunto de dados utilizado, não houve uma diferença significativa entre algoritmos mais sofisticados e um mais básico como o KNN. A validação cruzada reafirma a estabilidade dos modelos, tornando-os adequados para aplicação prática.

CONCLUSÕES

O KNN se destacou como o superior, ainda que as diferenças entre CatBoost e XGBOOST fossem mínimas em termos de métricas de desempenho, como precisão, revocação, F1-Score e suporte. Portanto, na prática, a superioridade de um algoritmo em relação a outro ocorre em função da obtenção de resultados superiores em métricas de avaliação relevantes ao problema.

REFERÊNCIAS

BRASIL HORTIFRUTI. O que mudou no consumo brasileiro de frutas e hortaliças nos últimos anos? Uma publicação do CEPEA - ESALQ/USP, Ano 20, n. 209, mar. 2021. ISSN 1981-1837.

KAGGLE. 🍌 | Banana Quality. Disponível em: https://www.kaggle.com/datasets/l3llff/banana. Acesso em: 11 jul. 2024.

SAHU, Priyanka; SINGH, Amit Prakash; CHUG, Anuradha; SINGH, Dinesh. A systematic literature review of machine learning techniques deployed in agriculture: A case study of banana crop. *IEEE Access*, v. 10, p. 87333-87360, 2022.

TESLA, Nikola; CHILDRESS, David Hatcher. The fantastic inventions of Nikola Tesla. Adventures Unlimited, 1993.