Relatório de projeto da disciplina de LPAA – 2023.1

João Gustavo Cavalcanti Beltrão da Silva [1,2] orcid.org/0000-0001-8392-3115

Fernando Marques Teixeira [1,2] orcid.org/0000-0003-1381-2422

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Graduação em Engenharia de Controle e Automação, Escola Politécnica de Pernambuco, Pernambuco, Brasil,

E-mail do autor principal: João Gustavo, jgcbs@poli.br

Resumo

Este relatório avaliou o desempenho de três algoritmos de classificação (SVM, KNN e Random Forest) na detecção de câncer de mama usando um conjunto de dados do Scikit-Learn. Todos os modelos mostraram resultados promissores, com altas métricas de precisão, recall e F1 Score. O SVM se destacou na minimização de erros críticos, o KNN equilibrou bem precisão e recall, enquanto o Random Forest se destacou na distinção entre classes. A escolha do modelo depende das necessidades clínicas, e esses resultados podem ser valiosos para diagnósticos médicos. No entanto, é crucial lembrar que esses modelos são ferramentas de apoio, não substituindo o julgamento clínico. Sua implementação requer validação em ambientes clínicos reais e monitoramento contínuo. Esse estudo fornece insights sobre a aplicação de algoritmos de classificação em diagnósticos médicos e destaca a importância da escolha do modelo adequado para atender às necessidades clínicas específicas.

Palavras-Chave: classificação, SVM, KNN, Random Forest, câncer.

Abstract

This report evaluated the performance of three classification algorithms (SVM, KNN, and Random Forest) in detecting breast cancer using a Scikit-Learn dataset. All models showed promising results, with high precision, recall and F1 Score scores. SVM excelled at minimizing critical errors, KNN balanced precision and recall well, while Random Forest excelled at distinguishing between classes. The choice of model depends on clinical needs, and these results can be important for medical diagnoses. However, it is crucial to remember that these models are supportive tools and do not modify clinical judgment. Its implementation requires validation in real clinical environments and continuous monitoring. This study provides insights into the application of classification algorithms in medical diagnostics and highlights the importance of choosing the appropriate model to meet specific clinical needs.

Key-words: *classification, SVM, KNN, Random Forest, cancer.*

DOI: 10.xxxx/s11468-014-9759-3

1 Introdução

O diagnóstico precoce do câncer de mama desempenha um papel fundamental na melhoria das taxas de sobrevivência e no tratamento eficaz da doença. Com o avanço da tecnologia e o aumento da disponibilidade de dados médicos, o uso de técnicas de aprendizado de máquina tornou-se uma ferramenta valiosa no campo da medicina, especialmente na área de diagnóstico de câncer.

Neste estudo, exploramos a aplicação de algoritmos de aprendizado de máquina para o diagnóstico de câncer de mama. Utilizamos um dataset fornecido pela biblioteca Scikit-Learn [1], que contém informações sobre dados do câncer de mama e aplicamos três algoritmos de classificação: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Random Forest. Nosso objetivo é avaliar o desempenho desses algoritmos e determinar sua eficácia na identificação de casos de câncer de mama.

Neste relatório, apresentaremos uma análise detalhada dos resultados obtidos com cada algoritmo, incluindo métricas de desempenho, como acurácia, precisão, recall, F1 Score e a área sob a curva ROC (AUC). Além disso, discutiremos as implicações práticas desses resultados no contexto do diagnóstico de câncer de mama.

Este estudo representa um passo significativo em direção à utilização de ferramentas de aprendizado de máquina para aprimorar o diagnóstico médico e contribuir para a detecção precoce e tratamento eficaz do câncer de mama.

2 Métodos

2.1 Análise Teórica

Para este estudo, utilizamos um conjunto de dados do câncer de mama que contém informações detalhadas sobre as características dos tumores mamários. O conjunto de dados é composto por 569 amostras, onde cada amostra é descrita por 30 características numéricas. Essas características incluem medidas de raio, textura, perímetro, área, suavidade, compactação e outras informações relacionadas à forma e textura dos tumores.

A matriz de correlação é uma representação visual que mostra a relação entre variáveis. Cores próximas do branco indicam correlações mais fortes, enquanto valores próximos a zero sugerem correlações fracas, como podemos ver na figura 1.

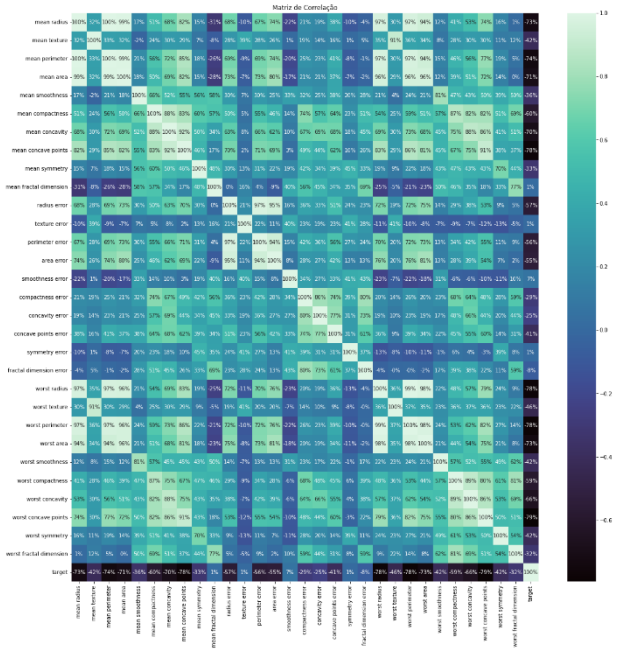

Figura 1: Matriz de correlação

2.1.1 Algarismos de classificação

Exploramos três algoritmos de classificação amplamente utilizados:

- **Support Vector Machine (SVM):** Implementamos o SVM com um kernel linear, uma escolha comum para problemas de classificação, e com parâmetros padrão. O SVM é conhecido por sua eficácia na criação de fronteiras de decisão que separam eficientemente classes em espaços multidimensionais.

- **K-Nearest Neighbors (KNN):** Aplicamos o KNN com um valor de vizinhos igual a 5. Este método de aprendizado baseia-se na proximidade de pontos de dados e é adequado para problemas nos quais a estrutura local dos dados é importante.

- **Random Forest:** Utilizamos o algoritmo Random Forest com 100

estimadores e um estado aleatório fixo para garantir resultados reprodutíveis. O Random Forest é uma técnica de conjunto que combina múltiplas árvores de decisão para melhorar a precisão e a robustez do modelo.

2.2 Análise Prática

Para avaliar o desempenho dos modelos, dividimos o conjunto de dados em dois subconjuntos: treinamento e teste. O conjunto de teste representou 50% dos dados originais e foi usado para avaliar o desempenho final dos modelos. A divisão foi estratificada para garantir que a proporção de tumores malignos e benignos fosse preservada em ambos os conjuntos.

Foi utilizada a validação cruzada para garantir a robustez de nossa análise, realizamos a validação cruzada com validação k-fold, usando k=5. Essa abordagem nos permitiu obter estimativas mais confiáveis do desempenho dos modelos e evitar problemas de overfitting. A validação cruzada foi aplicada ao conjunto de treinamento em cada iteração, enquanto o conjunto de teste permaneceu separado para a avaliação final. Além disso foi utilizado parâmetros para avaliação do desempenho dos métodos utilizados, como:

- **Acurácia:** Essa métrica mede a proporção de previsões corretas feitas pelo modelo. Em outras palavras, representa a precisão global do modelo em classificar corretamente os tumores como malignos ou benignos.
- **Precisão:** A precisão avalia a proporção de previsões positivas corretas (tumores malignos) em relação ao total de previsões positivas feitas pelo modelo. Essa métrica é valiosa quando desejamos garantir que as previsões positivas sejam altamente confiáveis.
- **Recall (Sensibilidade)**: O recall mede a proporção de verdadeiros positivos (tumores malignos corretamente identificados) em relação ao total de tumores malignos reais. É particularmente importante quando a detecção de tumores malignos é uma prioridade, pois avalia a capacidade do modelo de identificar eficazmente casos positivos.
- **F1 Score**: O F1 Score é uma métrica que combina a precisão e o recall para fornecer uma medida geral do desempenho do modelo. Ele é útil quando desejamos equilibrar a precisão e o recall em nossas previsões.
- **Área sob a curva ROC (AUC)**: A AUC avalia a capacidade do modelo de distinguir entre as classes positivas (maligno) e negativas (benigno) calculando a área sob a curva da característica de operação do receptor (ROC). Quanto maior a AUC, melhor o modelo é em distinguir entre as classes.
- **Matrizes de Confusão**: Para uma análise detalhada das previsões feitas pelos modelos, geramos matrizes de confusão. Essas matrizes fornecem informações sobre os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, permitindo uma compreensão mais profunda do desempenho do modelo em diferentes cenários.

3 Resultados

Support Vector Machine (SVM):

O modelo Support Vector Machine (SVM), demonstrou um desempenho impressionante com uma Acurácia (CV) de 0.96 e uma Acurácia de 0.96, indicando uma taxa de previsões corretas muito alta. Além disso, a Precisão de 0.97 e o Recall de 0.97 revelam uma capacidade sólida do modelo em identificar tanto tumores malignos quanto benignos. O F1 Score, que combina precisão e recall, atingiu 0.97, indicando um equilíbrio entre essas métricas. A área sob a curva ROC (AUC) foi notavelmente alta, registrando 0.99, o que aponta para a habilidade do modelo em distinguir eficazmente entre tumores malignos e benignos como podemos ver na figura 2.

DOI: 10.xxxx/s11468-014-9759-3

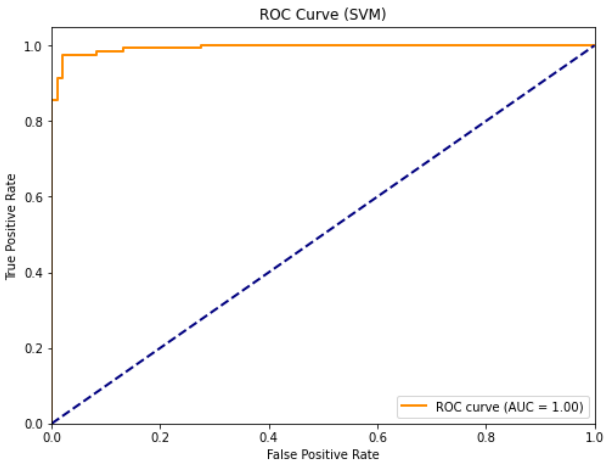

Figura 2: Curva ROC (SVM)

A Matriz de Confusão mostrou que o SVM teve apenas 5 falsos negativos e 5 falsos positivos, o que é uma representação visual das previsões do modelo, como mostra a figura 3.

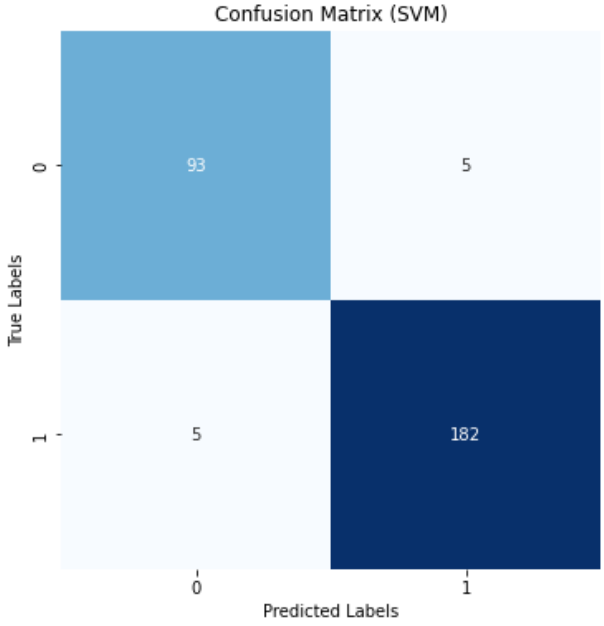

Figura 3: Matrix de confusão (SVM)

K-Nearest Neighbors (KNN):

O algoritmo K-Nearest Neighbors (KNN), apresentou uma Acurácia (CV) de 0.90 e uma Acurácia de 0.96, demonstrando uma alta taxa de previsões corretas. A Precisão de 0.96 e o Recall de 0.97 destacam a capacidade do modelo em classificar eficazmente tanto tumores malignos quanto benignos. O F1 Score alcançou 0.97, indicando um equilíbrio sólido entre precisão e recall. A área sob a curva ROC (AUC) registrou 0.98, sugerindo uma boa capacidade de discriminação entre as classes, como podemos ver na figura 4.

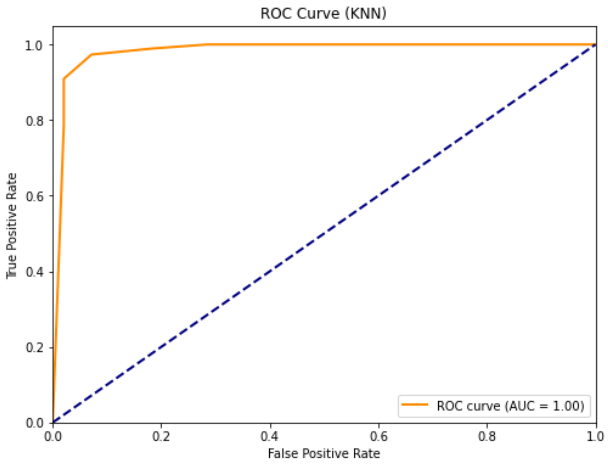

Figura 4: Curva ROC (KNN)

A Matriz de Confusão revelou que o KNN teve 7 falsos positivos e 5 falsos negativos, como observado na figura 5.

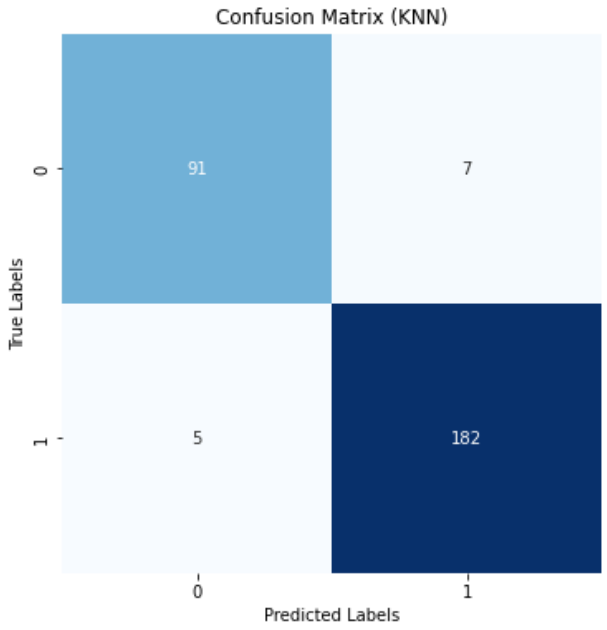

Figura 5: Matrix de confusão (KNN)

Algoritmo: Random Forest:

O Random Forest demonstrou um desempenho notável com uma Acurácia (CV) de 0.94 e uma Acurácia de 0.96, indicando uma alta taxa de previsões corretas. A Precisão de 0.97 e o Recall de 0.97 realçam a capacidade do modelo em classificar com sucesso tanto tumores malignos quanto benignos. O F1 Score, uma métrica que equilibra precisão e recall, alcançou 0.97. O destaque foi a área sob a curva ROC (AUC), que atingiu 1.00, indicando uma capacidade excepcional do modelo em distinguir entre tumores malignos e benignos, como podemos observar na figura 6.

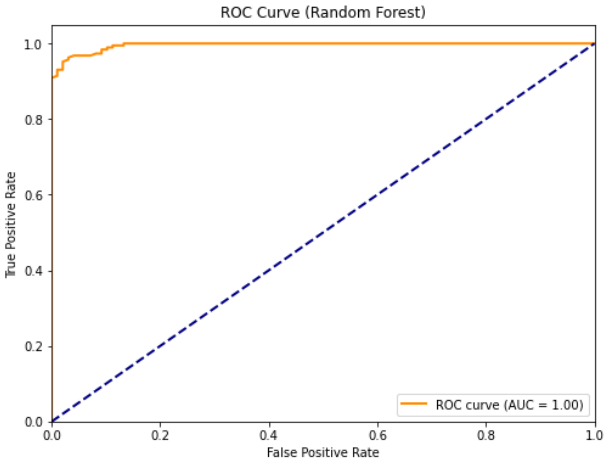

Figura 6: Curva ROC (Floresta Aleatória)

A Matriz de Confusão para o Random Forest revelou 6 falsos positivos e 6 falsos negativos, como mostra a figura 7.

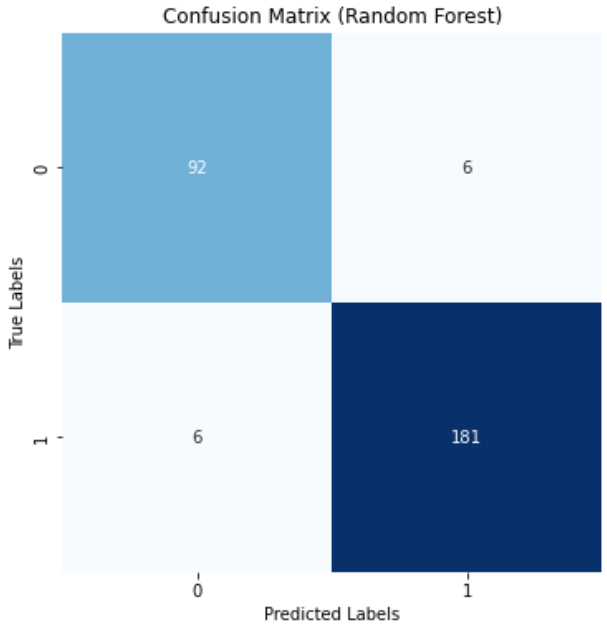

Figura 7: Matrix de confusão (Floresta Aleatória)

4 Discussão

O SVM demonstrou um desempenho notável com alta acurácia, precisão e recall. Esse modelo é particularmente eficaz quando a prioridade está na minimização de falsos positivos e falsos negativos, o que é crucial em diagnósticos de câncer. Além disso, a alta AUC indica que o SVM é capaz de distinguir com sucesso entre as classes, tornando-o uma opção sólida quando a precisão é crítica.

Por outro lado, o KNN mostrou ser robusto, alcançando altas taxas de acurácia e recall. No entanto, sua acurácia de validação cruzada ligeiramente menor pode ser um ponto de consideração. O KNN pode ser uma escolha adequada quando o foco está na obtenção de um equilíbrio entre precisão e recall, e quando a variação dos dados é um aspecto crítico.

O Random Forest, por sua vez, se destacou com uma AUC excepcionalmente alta. Este modelo é uma excelente escolha quando a distinção entre classes é fundamental, como em casos de diagnóstico de câncer. Sua capacidade de lidar com dados complexos e a natureza robusta do ensemble de árvores o tornam uma opção atraente. No entanto, é importante

DOI: 10.xxxx/s11468-014-9759-3

notar que o Random Forest pode ser mais computacionalmente intensivo e requer um ajuste cuidadoso de hiperparâmetros.

5. Conclusões

Em conclusão, a implementação prática de modelos de aprendizado de máquina, como Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Random Forest, em um cenário clínico exige uma avaliação cuidadosa das necessidades específicas. Cada um desses modelos demonstrou seu valor e destaque em diferentes contextos. O SVM, com sua ênfase na precisão de diagnóstico e na capacidade de equilibrar falsos positivos e falsos negativos, é uma escolha sólida quando a prioridade é minimizar erros críticos em diagnósticos médicos. Por outro lado, o KNN se destaca em situações em que a variabilidade dos dados é uma preocupação central e uma pequena margem de erro na precisão é tolerável, tornando-o adequado para contextos clínicos específicos. Por fim, o Random Forest brilha em cenários onde a distinção nítida entre classes é fundamental e recursos computacionais estão disponíveis para explorar sua capacidade excepcional de discriminação.

Em última análise, a escolha do modelo a ser implementado deve ser guiada pelas necessidades clínicas específicas e pelas métricas de desempenho relevantes, considerando o equilíbrio entre precisão, recall e capacidade de discriminação. Com uma abordagem informada e estratégica, esses modelos de aprendizado de máquina têm o potencial de aprimorar significativamente o diagnóstico e o tratamento de câncer, contribuindo para um melhor atendimento aos pacientes no campo da saúde.

Referências

[1] BIBLIOTECA SCIKIT-LEARN. Disponível em: <https://scikit-learn.org/stable/user_guide.html>. Acesso em: 22 set. 2023.