

ANÁLISE COMPARATIVA ENTRE MODELOS DE MACHINE LEARNING EM CONJUNTOS DE DADOS NA ÁREA MÉDICA

Davi dos Reis Soares (UNIFEI) – soaresreisdavi@gmail.com
Rafael Spinola Cambraia (Hospital das Clínicas de Itajubá) -
rafaelcambraia@yahoo.com.br
João Luiz Junho Pereira (UNIFEI) – joaoluizjp@unifei.edu.br

Palavras-chave: *Machine Learning*, Diagnóstico Médico, *Recall*, Precisão, Acurácia.

1. INTRODUÇÃO

A Inteligência Artificial (IA), especialmente com o *Machine Learning* (ML), está cada vez mais presente no cotidiano, expandindo-se para a área médica.

O uso dessas novas tecnologias em diagnóstico tem sido amplamente estudado. Pesquisadores como Li et al. (2024, p. 328) demonstram a superioridade dos métodos de ML em relação a escores de risco e marcadores laboratoriais, destacando modelos com melhor desempenho.

Entretanto, a performance de um algoritmo varia conforme a complexidade e as características do *dataset*, sendo que um modelo eficaz em um contexto pode falhar gravemente em outro.

Com o intuito de realizar uma análise comparativa, este estudo avalia o desempenho de quatro modelos de ML - *K-Nearest Neighbors* (k-NN), Regressão Logística, *Gaussian Naive Bayes* (GaussianNB) e *Random Forest* - em quatro *datasets* clínicos: Risco de Ataque Cardíaco, Câncer de Pulmão, COVID-19 e Risco de Acidente Vascular Cerebral.

O k-NN é um algoritmo não-paramétrico que classifica um novo dado pela maioria dos votos de seus *k* vizinhos mais próximos, baseando-se na distância.

O *GaussianNB* tem raízes no Teorema de Bayes; conforme Rish (2001, p. 41-46.), ele se baseia em probabilidade, permitindo um cálculo de classificação rápido e robusto.

A Regressão Logística é um modelo de classificação linear que utiliza a função logística para mapear a combinação linear das *features* para uma probabilidade entre 0 e 1.

Por contraste, o *Random Forest*, proposto por Breiman (2021), é um algoritmo *ensemble* que combina centenas de Árvores de Decisão, reduzindo o *overfitting* e oferecendo, geralmente, boa performance preditiva.

Para garantir a confiabilidade, a avaliação não se baseia apenas na acurácia, mas em métricas clínicas essenciais: acurácia, *F1 Score*, *recall* e precisão. Essa abordagem visa identificar o algoritmo mais adequado para a previsão das doenças em cada problema.

2. DESENVOLVIMENTO

2.1 Metodologia

Com o objetivo de avaliar a eficácia de diferentes algoritmos de ML na área da saúde, selecionamos quatro conjuntos de dados clínicos, todos provenientes da plataforma Kaggle:

- *Lung Cancer*: Visa estabelecer o risco de um paciente desenvolver câncer de pulmão. A predição é baseada na análise de múltiplos fatores, como idade, histórico de tabagismo e peso corporal.
- Previsão de risco de ataques cardíacos: Este *dataset* prediz o risco de ataques cardíacos, combinando dados binários, quantitativos e qualitativos.
- Dados dos sintomas de COVID-19: O foco deste conjunto de dados é a previsão da presença do vírus COVID-19. A classificação é realizada com base em um espectro de sintomas.
- Dados de AVC: Visa prever o risco de Acidentes Vasculares Cerebrais. Sua relevância reside no fato de ser um conjunto de dados desbalanceado.

Inicialmente, o pré-processamento foi crucial para equilibrar a ordem de grandeza e o tipo de variáveis. As preditoras foram separadas em colunas numéricas e categóricas. Em seguida, os *datasets* foram divididos: 80% para treinamento e 20% para teste.

Montamos os pipelines para o pré-processamento, tratando dois pontos. A ausência de dados foi resolvida preenchendo valores faltantes: com a média nas colunas numéricas e com o valor mais frequente nas categóricas. Para a transformação de variáveis, as numéricas foram padronizadas (média zero e desvio-padrão um) e as categóricas convertidas para sistema binário.

Os hiperparâmetros foram definidos usando valores fixos. Essa escolha teve como objetivo estabelecer uma linha de base de desempenho, utilizando valores *default* ou aqueles comumente empregados como pontos de partida eficazes na literatura de *ML*.

Tabela 1 – Hiperparâmetros

Algoritmo	Hiperparâmetro Definido	Valor Utilizado
KNN	n_neighbors	5
GaussianNB	(nenhum)	<i>Default</i>
Regressão Logística	max_iter	1000
	penalty	l2 (Ridge)
Random Forest	n_estimators	100
	criterion	<i>gini</i>
	max_depth	<i>None</i>
	min_samples_split	2
	min_samples_leaf	1
	max_features	<i>"sqrt"</i>

Fonte: Autoria própria (2025)

Com isso nós podemos treinar os modelos e testar cada um deles e assim comparar com os valores obtidos com os valores reais e ver como as métricas se comportaram.

3. RESULTADOS E DISCUSSÃO

Uma vez executados os códigos podemos comparar os resultados de cada método em cada conjunto de dados

3.1 RISCO DE ATÁQUE CARDIACO

Tabela 2 – Dados *Heart Attack Risk Prediction Dataset*

	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 Score</i>
--	-----------------	------------------	---------------	-----------------

KNN	0,607	0,35	0,12	0,17
GaussianNB	0,4535	0,35	0,63	0,45
Regressão Logística	0,6229	0,34	0,06	0,10
Random Forest	0,6412	0,0	0,0	0,0

Fonte: Autoria própria (2025)

Embora o *GaussianNB* tenha a menor Acurácia (0,4535), ele alcançou o maior *Recall* (0,63). Em um contexto médico, o *Recall* é a métrica crucial, pois mede a capacidade de o modelo identificar corretamente os pacientes em risco e, consequentemente, minimizar falsos negativos (deixar a doença passar despercebida).

Os modelos com maior Acurácia são clinicamente inviáveis para este *dataset*, pois falharam na detecção de risco.

3.2 RISCO DE CÂNCER DE PULMÃO

Tabela 3 – conjunto de dados *Lung Cancer*.

	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 Score</i>
KNN	0,9032	0,98	0,91	0,94
GaussianNB	0,8548	0,92	0,91	0,92
Regressão Logística	0,9032	0,94	0,94	0,94
Random Forest	0,8871	0,94	0,93	0,93

Fonte: Autoria própria (2025)

O modelo mais indicado para o diagnóstico de Câncer de Pulmão é a Regressão Logística.

Este modelo obteve a maior Acurácia e alcançou o melhor equilíbrio entre as métricas de segurança. Com um *Recall* de 0,94, ele garante uma alta sensibilidade para evitar falsos negativos, e sua precisão de 0,94 mantém uma baixa taxa de falsos positivos. Esta combinação de alta sensibilidade e confiabilidade o torna a escolha mais segura e eficaz na prática clínica.

3.3 RISCO DE COVID-19

Tabela 4 –conjunto de dados COVID-19 Symptoms Checker.

	Accuracy	Precision	Recall	F1 Score
KNN	0,9715	0,98	0,98	0,98
GaussianNB	0,7682	1,0	0,71	0,83
Regressão Logística	0,9678	0,98	0,99	0,98
Random Forest	0,9788	0,99	0,98	0,99

Fonte: Autoria própria (2025)

O modelo *Random Forest* apresentou o melhor desempenho geral para o *dataset* de COVID-19, obtendo a maior Acurácia (0,9788) e o melhor *F1 Score* (0,99). Sua performance é a mais segura clinicamente, pois alcança um *Recall* de 0,98 – demonstrando alta capacidade de detecção da doença. Simultaneamente, sua Precisão de 0,99 garante uma taxa mínima de alarmes falsos.

3.1 RISCO DE ACIDENTE VASCULAR CEREBRAL

Tabela 5 – Conjunto de dados Cerebral Stroke Prediction - Imbalanced Dataset.

	Accuracy	Precision	Recall	F1 Score
KNN	0,9691	0,06	0,04	0,05
GaussianNB	0,3115	0,03	0,99	0,05
Regressão Logística	0,9819	0,00	0,0	0,0
Random Forest	0,982	1,0	0,01	0,01

Fonte: Autoria própria (2025)

O *dataset* de Risco de AVC se mostrou extremamente desbalanceado, o que resultou em modelos com alta Acurácia enganosa.

A Regressão Logística e o *Random Forest* formam os mais afetados pelo desbalanceamento: embora com alta Acurácia, o *Recall* de 0,0 a 0,01 significa que eles deixaram de detectar 99% a 100% dos casos reais de AVC.

O *GaussianNB* obteve o maior *Recall* (0,99), identificando quase todos os casos de AVC. No entanto, sua Precisão extremamente baixo (0,03), evidenciado pela baixa Acurácia (0,3115), mostra que ele não é uma boa opção para este *dataset*.

Portanto, nenhum dos quatro modelos com estes hiperparâmetros é recomendado para ser utilizado no estado atual dos dados.

4. CONSIDERAÇÕES FINAIS

Por meio deste estudo, demonstramos que o desempenho de cada modelo tende a variar de acordo com o conjunto de dados específico que está sendo trabalhado.

Os resultados obtidos foram:

- Risco de Ataque Cardíaco: O *Gaussian Naive Bayes* foi o modelo mais indicado, devido ao seu maior *Recall* (0,63).
- Câncer de Pulmão: A Regressão Logística obteve o melhor equilíbrio de métricas, com alta Acurácia, *Recall* e Precisão, sendo a escolha mais segura.
- Sintomas de COVID-19: O *Random Forest* apresentou o melhor desempenho geral, com alta capacidade de detecção da doença.
- Risco de AVC: Nenhum dos quatro modelos é recomendado no estado atual do *dataset*.

Tendo isso em vista, torna-se essencial a avaliação e comparação de diversos modelos de ML para garantir a seleção do algoritmo mais adequado para o problema em questão.

É igualmente importante evidenciar o desafio de trabalhar com conjuntos de dados desbalanceados. Conforme observado nos testes, o desbalanceamento pode comprometer a confiabilidade dos resultados. Por essa razão, é fundamental considerar a aplicação de técnicas de balanceamento de dados.

AGRADECIMENTOS

Gostaríamos de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por investir e promover a realização deste trabalho, nada disso seria possível sem esta instituição.

REFERÊNCIAS

LI, Xiaojing et al. Interpretable prediction of 30-day mortality in patients with acute pancreatitis based on machine learning and SHAP. **BMC Medical Informatics and Decision Making**, v. 24, n. 1, p. 328, 2024.

RISH, Irina et al. An empirical study of the naive Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. 2001. p. 41-46.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

Lung cancer. Kaggle, 2021. Disponível em:
<https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer> Acesso em: 15 set. 2025.

Cerebral Stroke Prediction-Imbalanced Dataset. Kaggle, 2021. Disponível em:
<https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalanced-dataset> Acesso em: 15 set. 2025.

Heart Attack Risk Prediction Dataset. Kaggle 2024. Disponível em:
<https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset>
Acesso em: 15 set. 2025

COVID-19 Symptoms Checker. Kaggle, 2019. Disponível em:
<https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker> Acesso em: 15 set. 2025