Avaliação de Modelos de Aprendizado de Máquina para Classificação de Acidente Vascular Cerebral

Andressa Caroline da Rocha Pereira[1,2], **Sheila Paloma S. Brito**[1,2]

[1]Curso de Sistemas de Informação – Universidade Federal do Piauí (UFPI),
Campus Senador Helvídio Nunes de Barros (CSHNB)
Caixa Postal 64.607-670 – Picos – PI – Brazil

[2]Coordenação de Sistemas de Informação
Universidade Federal do Piauí (UFPI) – Picos, PI – Brazil

andressacaroline@ufpi.edu.br, sheila.psb@gmail.com

Abstract. *This study aims to evaluate and compare the performance of supervised classification algorithms, including ID3, k-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP), for stroke prediction. The models were evaluated using cross-validation and the SMOTE technique for class balancing. The performance metrics analyzed included accuracy, precision, recall, F1-score and AUC. The results indicated that Random Forest stood out in all metrics, being the best performing model. KNN showed the lowest performance, with a reduction in accuracy and recall values. Thus, it is emphasized that the* ***Random Forest*** *was the most effective model for early stroke prediction in the study, with great potential for clinical applications in healthcare.*

Keywords*: Stroke prediction, Artificial Intelligence, Classification algorithms, Cross-validation, Model performance.*

Resumo. *Este estudo tem como objetivo avaliar e comparar o desempenho de algoritmos de classificação supervisionada, incluindo ID3, k-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM) e Multi-Layer Perceptron (MLP), para a previsão de Acidente Vascular Cerebral (AVC). Os modelos foram avaliados por meio de validação cruzada e utilizando a técnica de SMOTE para balanceamento das classes. As métricas de desempenho analisadas incluíram acurácia, precisão, recall, F1-score e AUC. Os resultados indicaram que o Random Forest se destacou em todas as métricas, sendo o modelo de melhor desempenho. O KNN apresentou o menor desempenho, com redução nos valores de acurácia e recall. Assim, enfatiza-se que o Random Forest foi o modelo mais eficaz para a previsão precoce de AVC no estudo, com grande potencial para aplicações clínicas na área da saúde.*

Palavras-chave*: Previsão de AVC, Inteligência Artificial, Algoritmos de classificação, Validação cruzada, Desempenho de modelos.*

1 Introdução

O Acidente Vascular Cerebral (AVC) é uma das condições médicas mais devastadoras, responsável pela morte das células cerebrais devido à interrupção do fornecimento de

sangue, o que resulta na falta de oxigênio e nutrientes essenciais para o funcionamento adequado do cérebro [de Lima Barbosa et al. 2021]. De acordo com a Organização Mundial da Saúde (2019), o AVC ocupa a segunda posição entre as principais causas de morte no mundo, sendo responsável por cerca de 11% das mortes globais. Além disso, o AVC é uma das maiores causas de incapacidade, afetando milhões de pessoas anualmente e impactando profundamente a qualidade de vida dos sobreviventes.

A detecção precoce e a previsão do AVC são cruciais para a aplicação de intervenções médicas rápidas, capazes de minimizar os danos cerebrais e aumentar as chances de recuperação. Em face desse desafio, o aprendizado de máquina emergiu como uma ferramenta poderosa e promissora na análise de grandes volumes de dados médicos, permitindo a identificação de padrões complexos e proporcionando previsões mais precisas e oportunas sobre a ocorrência de AVC [Tashkova et al. 2025, Chakraborty et al. 2024].

Diversos estudos recentes têm investigado a aplicação de algoritmos de classificação para prever o AVC a partir de dados clínicos, como idade, pressão arterial, níveis de glicose e histórico médico [Dev et al. 2022]. No entanto, um dos principais obstáculos enfrentados por esses modelos é o desbalanceamento de classes, o que pode prejudicar a precisão e a confiabilidade das previsões. Para mitigar esse problema, técnicas como o SMOTE (Synthetic Minority Over-sampling Technique) são frequentemente empregadas, gerando exemplos sintéticos da classe minoritária e, assim, melhorando o desempenho do modelo [Jing 2022].

Este estudo tem como objetivo avaliar e comparar o desempenho de diversos modelos de classificação supervisionada, incluíndo o *Inductive Decision Tree* (ID3), *k-Nearest Neighbors* (KNN), *Random Forest*, *Support Vector Machine* (SVM) e *MultiLayer Perceptron* (MLP) na previsão e classificação da ocorrência de AVC.

2 Metodologia

As etapas a seguir foram aplicadas para avaliação de diferentes algoritmos de classificação supervisionada, mais especificamente: ID3, KNN, Random Forest, SVM e MLP. Foram aplicadas técnicas de pré-processamento, como a normalização dos dados e a seleção de características, além de utilizar validação cruzada (k-fold = 5) como forma de garantir a robustez e generalização dos modelos.

A análise será realizada por meio de validação cruzada, e as métricas de desempenho, como Acurácia, Precisão, Recall, F1-Score AUC e curva ROC, serão cuidadosamente comparadas para identificar o modelo mais eficaz para a previsão do AVC.

2.1 Base de Dados

Este estudo utiliza a base de dados ”Stroke Prediction Dataset”, disponível na plataforma Kaggle [Soriano 2021]. Com 5.110 registros e 12 atributos, cada entrada representa um paciente com dados sobre características demográficas e condições de saúde relacionadas ao risco de Acidente Vascular Cerebral (AVC).

2.1.1. Descrição dos Atributos

A base de dados utilizada contém 12 atributos clínicos associados ao risco de AVC, incluindo variáveis como id, gênero, idade, hipertensão, estado civil, ocupação, tipo de

residência, presença de doença cardíaca, nível médio de glicose, IMC, tabagismo e AVC. Os dados incluem variáveis numéricas e categóricas e são essenciais para a construção de modelos preditivos eficazes, que visam identificar padrões relacionados à ocorrência do AVC

2.2 Pré-Processamento dos Dados

O pré-processamento iniciou-se com a imputação de valores ausentes nas variáveis numéricas, preenchendo-os com a mediana, e nas variáveis categóricas, com a moda. As variáveis categóricas foram transformadas em valores numéricos utilizando o método de *Label Encoding*.

A variável alvo foi definida com base na coluna "stroke", convertendo-a em classificação binária. Os dados foram então divididos utilizando validação cruzada estratificada com *StratifiedKFold*, garantindo que as classes fossem distribuídas de forma equilibrada nos conjuntos de treino e teste.

Após essas etapas, o conjunto de dados estava livre de valores nulos e pronto para a modelagem, com o desbalanceamento de classes tratado por meio da técnica *SMOTE*, assegurando a robustez e precisão das previsões.

2.3 Modelos de Classificação

Este estudo utilizou cinco modelos de aprendizado de máquina para prever o *Acidente Vascular Cerebral (AVC)*, com os seguintes detalhes:

- **K-Nearest Neighbors (KNN)**: Classifica com base nos *K* vizinhos mais próximos. Utilizou-se $k = 5$, com validação cruzada estratificada para otimizar o desempenho e evitar overfitting.
- **Random Forest**: é um conjunto de árvores de decisão onde cada árvore realiza uma previsão e o resultado final é determinado pela votação majoritária. Foi configurado com 100 árvores (*n_estimators = 100*), sendo robusto para dados desbalanceados.
- **ID3 (Decision Tree)**: Árvore de decisão que usa ganho de informação para dividir os dados. A profundidade máxima da árvore foi controlada para evitar overfitting, utilizando validação cruzada para otimização.
- **Support Vector Machine (SVM)**: Utiliza um hiperplano para separar as classes. Foi aplicado com o kernel *RBF* e ajuste do parâmetro de penalização *C* para balancear o overfitting.
- **Multi-Layer Perceptron (MLP)**: Rede neural com 3 camadas ocultas, com o número de neurônios na camada oculta calculado dinamicamente como a média entre o número de entradas (*n_inputs*) e o número de saídas (*n_outputs*), ou seja, $n\_hidden = \frac{n\_inputs + n\_outputs}{2}$. O número de iterações foi definido como 300 (*max_iter = 300*), garantindo a convergência do modelo.

Todos os modelos foram avaliados utilizando *SMOTE* para balanceamento das classes e validação cruzada estratificada. As métricas de *Acurácia, Precisão, Recall, F1-Score* e *ROC AUC* foram usadas para comparar o desempenho de cada modelo, assegurando uma análise robusta e confiável.

2.4 Avaliação dos Resultados

A avaliação dos resultados foi realizada utilizando validação cruzada estratificada, garantindo uma análise robusta e eficaz. A técnica de *SMOTE* foi aplicada para balancear as classes e melhorar a performance dos modelos.

As métricas de desempenho utilizadas para avaliar os modelos de algoritmos foram:

- **Acurácia**: Proporção de previsões corretas.
- **Precisão**: Taxa de verdadeiros positivos entre as previsões positivas.
- **Recall**: Capacidade de identificar corretamente os casos positivos.
- **F1-Score**: Média harmônica entre precisão e recall.
- **ROC AUC**: Medida da capacidade de distinguir entre as classes.

Os modelos foram comparados com base nessas métricas, buscando identificar qual deles obteve o melhor equilíbrio entre sensibilidade e especificidade, para uma visão abrangente do desempenho de cada modelo, permitindo a escolha do melhor modelo para a previsão de AVC com base na análise dos múltiplos critérios de avaliação.

2.5 Trabalhos Relacionados

A previsão de AVC tem sido amplamente abordada utilizando algoritmos de aprendizado de máquina, visando melhorar a detecção precoce e a precisão no diagnóstico. No estudo de [Tashkova et al. 2025] realizaram uma análise comparativa entre modelos como *Random Forest* e *SVM*, destacando a eficácia desses modelos, mas também apontando o desafio do *desbalanceamento de classes*. A necessidade do uso de técnicas como o *SMOTE* para lidar com esse problema foi ressaltada.

O estudo de [Chakraborty et al. 2024] propuseram uma abordagem empilhada de aprendizado de máquina, combinando múltiplos modelos preditivos e destacando a importância da *seleção de características*. No entanto, o estudo também identificou a limitação dos métodos tradicionais ao lidar com dados desbalanceados, reforçando o uso do *SMOTE*.

Já o estudo de [Dev et al. 2022] utilizaram redes neurais para a previsão de AVC, obtendo bons resultados, mas identificando que a imputação de dados ausentes pode comprometer a qualidade das previsões, destacando a importância de técnicas de *regularização* para evitar o *overfitting*.

No estudo de [Jing 2022] também investigou o impacto do desbalanceamento de classes e a eficácia do *SMOTE* para melhorar a sensibilidade dos modelos preditivos. Embora o *SMOTE* tenha sido eficaz, o estudo evidenciou a necessidade de novas abordagens para lidar com o viés de classe.

Assim, estes estudos contribuem para a área ao comparar diversos modelos de aprendizado de máquina, incluindo *KNN*, *Random Forest* e *MLP*, aplicando o *SMOTE* e validação cruzada estratificada para avaliar a precisão dos modelos. A análise das métricas de *Acurácia, Precisão, Recall, F1-Score e ROC AUC* oferece novos insights sobre a aplicação de técnicas de *pré-processamento* e *balanceamento de dados*, ampliando a compreensão sobre o impacto dessas abordagens na previsão do AVC.

3 Resultados

Os resultados médios indicam que o modelo Random Forest se destacou com uma acurácia média de 92,45% e um recall médio de 96,50%, demonstrando excelente desempenho tanto na precisão quanto na capacidade de identificar os casos positivos. O modelo KNN apresentou uma acurácia média de 80,74%, com um recall de 82,75%, evidenciando um desempenho sólido, mas com alguma dificuldade em identificar todos os casos positivos. Por sua vez, o MLP, com uma acurácia média de 74,34%, obteve uma precisão média de 98,08%, destacando-se pela alta capacidade de discriminação, embora tenha apresentado um recall de 74,49%, sugerindo desafios na identificação de todos os casos positivos (Tabela 1).

Table 1. Resultados de Desempenho dos Algoritmos de Classificação (média ± desvio padrão)

Modelo	Acurácia	Precisão	Recall	F1-Score	ROC AUC
KNN	0.81 ± 0.02	0.97 ± 0.00	0.83 ± 0.02	0.89 ± 0.01	0.62 ± 0.04
Random Forest	0.92 ± 0.01	0.96 ± 0.00	0.97 ± 0.01	0.96 ± 0.00	0.55 ± 0.02
ID3	0.89 ± 0.01	0.96 ± 0.00	0.93 ± 0.01	0.94 ± 0.01	0.55 ± 0.04
SVM	0.79 ± 0.02	0.97 ± 0.00	0.80 ± 0.02	0.88 ± 0.01	0.64 ± 0.02
MLP	0.74 ± 0.01	0.98 ± 0.01	0.74 ± 0.02	0.85 ± 0.01	0.73 ± 0.05

A Área Sob a Curva (AUC), é uma métrica utilizada para avaliar a capacidade do modelo em distinguir entre as classes positivas e negativas, com valores próximos de 1 indicando excelente desempenho. Em seus resultados de aplicação, demosntrou que modelo Random Forest obteve o melhor desempenho, com uma AUC de 0.7984, seguido pelo MLP, com AUC de 0.8172. O KNN, com AUC de 0.6288, apresentou o desempenho mais fraco, sugerindo uma menor capacidade de discriminação entre as classes. Os modelos SVM e ID3 apresentaram AUCs de 0.7456 e 0.5595, respectivamente, sendo que o ID3 teve o menor desempenho geral (Figura 1).

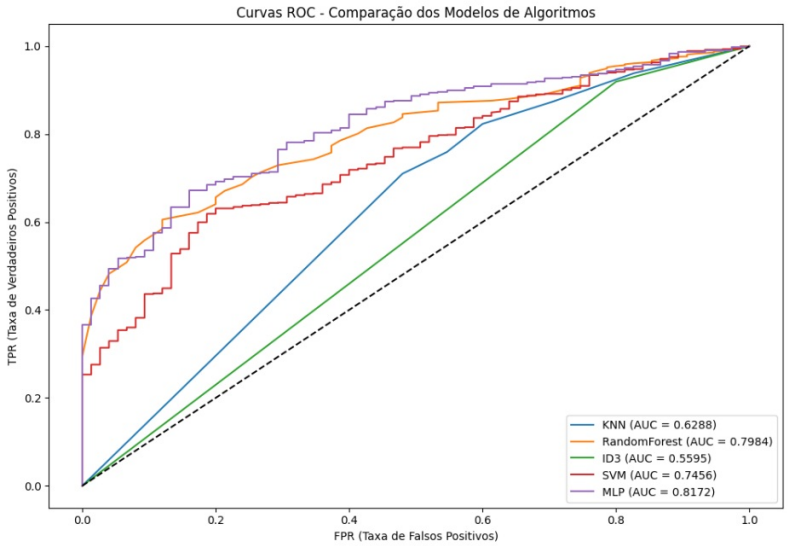

Figure 1. Curva ROC: Avaliação Comparativa entre os Modelos de Algoritmos

O gráfico a seguir, representa o comparativo das métricas de desempenho dos algoritmos, demonstrando que o Random Forest é o modelo que mais se destaca entre todas as métricas, com valores elevados de acurácia, precisão e recall, demonstrando um bom equilíbrio geral. O MLP apresenta um bom desempenho em precisão e AUC, mas com recall mais baixo. O KNN apresenta os menores valores de acurácia e recall, enquanto o ID3 tem o desempenho mais fraco em várias métricas, especialmente em F1-score e AUC (Figura 2).

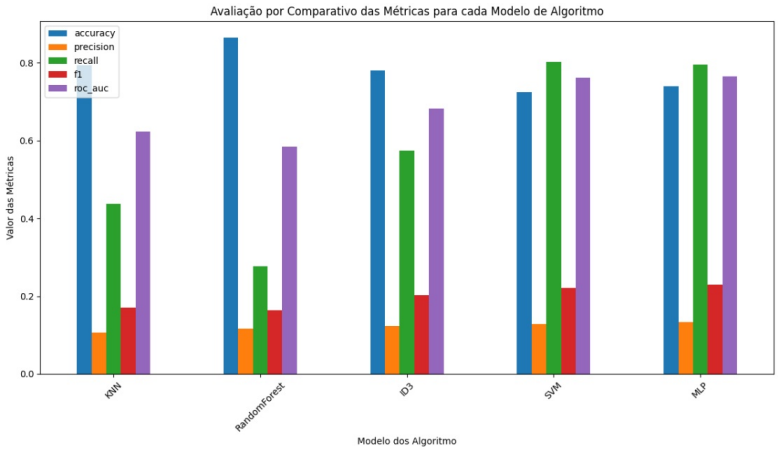

Figure 2. Avaliação Comparativa via Métricas dos Modelos de Algoritmos

4 Conclusão

Este estudo demonstrou a aplicabilidade das técnicas de aprendizado de máquina na previsão de Acidente Vascular Cerebral (AVC), destacando a eficácia do modelo Random Forest, que apresentou o melhor desempenho geral em todas as métricas avaliadas. Os resultados sugerem que modelos baseados em árvores de decisão, como Random Forest, são altamente eficazes em cenários com possibilidades de dados desbalanceados, oferecendo uma excelente combinação de desempenho em identificação de casos positivos e negativos. O MLP, embora tenha mostrado bom desempenho em precisão e AUC, evidenciou uma redução em recall, indicando a limitação de registros para os casos positivos de AVC na base de dados.

A pesquisa contribui significativamente para a compreensão das capacidades desses modelos e abre caminho para futuras investigações, com o objetivo de melhorar a generalização e a aplicabilidade em cenários no contexto clínico e em outras áreas de grande complexidade. Futuros estudos podem explorar ajustes mais finos nos hiperparâmetros e a utilização de técnicas avançadas de seleção de características, como RFE, para melhorar ainda mais a performance dos modelos em bases de dados com maior complexidade ou maior desbalanceamento.

Assim, este trabalho não apenas contribui para o campo da previsão de AVC com aprendizado de máquina, mas também oferece insights valiosos sobre como esses modelos podem ser aplicados com sucesso em contextos clínicos e em outras áreas desafiadoras.

References

Chakraborty, S. et al. (2024). Predicting stroke occurrences: A stacked machine learning approach with feature selection and data preprocessing. *BMC Bioinformatics*.

de Lima Barbosa, A. M., Pereira, C. C. M., Miranda, J. P. R., de Lima Rodrigues, J. H., de Carvalho, J. R. O., and Rodrigues, A. C. E. (2021). Perfil epidemiológico dos pacientes internados por acidente vascular cerebral no nordeste do brasil. *Revista Eletrônica Acervo Saúde*, 13(1):e5155–e5155.

Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., and John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *arXiv*.

Jing, Y. (2022). Machine learning performance analysis to predict stroke based on imbalanced medical dataset. *arXiv*.

Soriano, F. (2021). Stroke prediction dataset.

Tashkova, A., Eftimov, S., Ristov, B., and Kalajdziski, S. (2025). Comparative analysis of stroke prediction models using machine learning. *arXiv*.