

Aplicação de Modelos e Técnicas de Balanceamento na Previsão do Acidente Vascular Cerebral (AVC)

Nadiellen de Mello¹, Naira Gama¹

¹Universidade Federal da Bahia - Salvador/BA

nadiellen.melo@gmail.com, nairagama210@gmail.com

Resumo

O Acidente Vascular Cerebral (AVC) é uma condição séria e que pode atingir tanto homens quanto mulheres de diferentes idades, além de ser uma das principais causas de morte no mundo. Diversos fatores podem desencadear a doença, entre estão a hipertensão, o diabetes, o sobrepeso, além de fatores externos, ligados ao estilo de vida do paciente, como o tipo de trabalho e de residência. Com isso, o objetivo deste trabalho foi aplicar os modelos de Regressão Logística e Random Forest para prever a probabilidade de ocorrência de AVC. Durante as análises, notou-se um desbalanceamento nos dados, então aplicou-se técnicas para balancear estes dados. Os modelos foram aplicados antes e depois do tratamento e a técnica SMOTEENN se destacou no Random Forest.

Abstract

Stroke is a serious condition that can affect both men and women of different ages, and is one of the leading causes of death worldwide. Several factors can trigger the disease, including hypertension, diabetes, overweight, and external factors related to the patient's lifestyle, such as type of work and residence. Therefore, the objective of this study was to apply Logistic Regression and Random Forest models to predict the probability of stroke occurrence. During the analyses, an imbalance in the data was

noted, so techniques were applied to balance this data. The models were applied before and after treatment, and the SMOTEENN technique stood out in the Random Forest.

1 Introdução

O Acidente Vascular Cerebral (AVC) é uma das principais causas de morte no mundo, pode atingir homens e mulheres, e acontece quando os vasos sanguíneos responsáveis pelo transporte do sangue até o cérebro se rompem ou ficam obstruídos, causando confusão mental, alterações na visão, na fala e no equilíbrio, além de formigamentos e paralisia um lado do corpo. Diversos fatores como o sobrepeso, a hipertensão, o diabetes, o colesterol alto e o sedentarismo podem influenciar no desenvolvimento da condição.

Com estes fatores que podem desencadear o AVC, a previsão da doença deve ser feita o quanto antes para evitar problemas futuros, então de acordo com as condições de saúde do paciente e seu estilo de vida, é possível verificar se o mesmo está em situação de risco ou não. Pensando nisso, buscou-se uma forma de prever a doença utilizando algoritmos de predição e um conjunto de dados com as características de mais de 5.000 pacientes.

Para tanto, este trabalho foi organizado em seções, além desta Introdução. Na seção 2, encontra-se o referencial teórico, na 3 a metodologia, detalhando todo processo, desde

a escolha do ambiente de desenvolvimento até a avaliação dos modelos aplicados; e na seção 4, apresentam-se os resultados obtidos. Por fim, na seção 5 encontram-se as considerações finais sobre o projeto.

2 Referencial Teórico

Esta seção apresenta os principais conceitos teóricos relacionados ao Aprendizado de Máquina (AM) e seus modelos, os algoritmos utilizados no projeto, o desbalanceamento de classe e as técnicas de balanceamento.

2.2 Aprendizado de Máquina

O aprendizado de máquina, ou machine learning, é um campo da Inteligência Artificial (IA) e permite que um sistema aprenda com base em experiências. Oliveira *et al.* (2023) apresentam em seu trabalho as três categorias de técnicas de aprendizado de máquina, a supervisionada, em que o modelo é treinado com dados já rotulados, a não supervisionada, em que os dados não possuem rótulos e o modelo descobre padrões, e a aprendizagem por reforço, em que o aprendizado acontece por meio de tentativa e erro.

2.3 Modelos de Aprendizado de Máquina

Em seu trabalho, Oliveira *et al.* (2023) também cita alguns algoritmos típicos de aprendizado supervisionado, como as regressões, árvores de decisão, redes neurais, e máquinas de vetores.

Para este trabalho, utilizou-se dois modelos, a Regressão Logística e o Random Forest, muito utilizados em conjuntos com variáveis binárias.

2.3.1 Regressão Logística

Este algoritmo é utilizado quando uma variável categórica binária (0/1, Sim/Não) é resposta, assim temos a probabilidade de

ocorrência, neste caso, do AVC. Fernandes *et al.* (2020) incentiva o uso da regressão logística em seu trabalho e destaca que ela é a melhor ferramenta para trabalhar com variáveis dependentes dicotômicas, aquelas em que o y só possui duas categorias.

2.3.2 Random Forest

O Random Forest constrói múltiplas árvores buscando uma previsão precisa, e por isso Lorenzetti e Teloken (2016) destacam algumas características, como o fato do algoritmo ser mais poderoso se comparado somente a uma árvore de decisão, a sua boa taxa de acertos, a classificação aleatória das árvores, entre outras.

2.4 Desbalanceamento de Classe

Ao iniciar as análises do dataset, notou-se um desbalanceamento de classe, pois haviam mais dados de pessoas que nunca tiveram AVC, do que de pessoas que já tiveram a doença, por isso buscou-se uma forma de balancear os dados antes da aplicação dos algoritmos. Nos seus estudos, Oliveira *et al.* (2023) utilizou este mesmo dataset e aplicou algumas técnicas de pré-processamento para balancear os dados e evitar o enviesamento, que pode levar ao diagnóstico equivocado e a inviabilização da internação e do tratamento adequado do paciente, ou em casos mais graves, o óbito.

2.4.1 Técnicas de balanceamento

Para realização do balanceamento, são utilizadas técnicas para igualar a quantidade de registros em cada classe. Oliveira *et al.* (2023) mostra em seu trabalho duas técnicas, a *Oversampling*, em que é realizada a criação artificial de novas amostras com objetivo de deixar a classe minoritária com a mesma quantidade de registros que a majoritária, e a *Undersampling*, em que é feita a remoção da classe majoritária, também objetivando igualar a quantidade de registros no conjunto de dados.

3 Metodologia

Nesta seção, são apresentados os métodos utilizados para a realização do projeto. O trabalho caracteriza-se como um estudo experimental, ao qual dois modelos de aprendizado de máquina preditivo foram aplicados e comparados, descrevem-se o conjunto de dados utilizado, as etapas de pré-processamento, os modelos aplicados e a avaliação dos desempenhos preditivos.

3.1 Ambiente de Desenvolvimento e ferramentas utilizadas

Para a realização do projeto, foi utilizada a linguagem de programação Python, amplamente usada em aplicações de aprendizado de máquina. O ambiente de desenvolvimento adotado foi o Google Colab, devido a sua organização e facilidade de visualização dos resultados.

As principais bibliotecas utilizadas foram o *scikit-learn*, para a implementação dos algoritmos; o Pandas e o NumPy, para manipulação e análise dos dados; e por fim, o Matplotlib e o Seaborn, para a visualização gráfica.

3.2 Conjunto de Dados

Neste trabalho, utilizou-se o dataset “*Stroke Prediction Dataset*”, com dados de 5110 pacientes com diferentes idades. Na base contém o gênero do paciente, seu estado civil, o tipo de trabalho, tipo de residência, além de informações sobre seu estado de saúde, como a média de glicose, o Índice de Massa Corpórea (IMC), doenças cardíacas, e outros.

3.3 Pré-processamento dos dados

O pré-processamento dos dados foi realizado com o objetivo de garantir a qualidade, a consistência e a adequação das informações para posterior aplicação aos modelos de aprendizado de máquina. Inicialmente, na etapa de limpeza, identificou-se a presença de valores ausentes exclusivamente na variável

IMC, totalizando 201 registros, representando 4% do conjunto de dados. Uma análise da distribuição desses valores em relação à variável alvo indicou que a exclusão dessas observações poderia intensificar o desbalanceamento da classe positiva (AVC). Sendo assim, optou-se pela imputação dos valores ausentes pela mediana, realizada de forma estratificada por classe da variável AVC.

Durante a análise exploratória, foram detectadas inconsistências na variável idade, representadas por valores decimais. Como a maior parte dessas observações estava associada à classe negativa, decidiu-se pela remoção dessas linhas. Além disso, a categoria “*Other*” da variável gênero foi removida, por conter apenas um registro e não contribuir de forma significativa para a modelagem.

Após a limpeza, as variáveis categóricas foram transformadas em variáveis numéricas por meio da técnica de *One-Hot Encoding*, garantindo a compatibilidade com os algoritmos de aprendizado de máquina. A variável alvo AVC foi codificada de forma binária.

Com o conjunto de dados devidamente pré-processado, o projeto foi conduzido considerando dois cenários distintos. No primeiro, os modelos foram treinados utilizando os dados sem aplicação de técnicas de balanceamento, mantendo a distribuição original das classes. No segundo, aplicaram-se técnicas de balanceamento no conjunto de treinamento.

Foram utilizados três técnicas de balanceamento: SMOTE, método *Oversampling*, responsável pela geração de amostras sintéticas da classe minoritária; ENN, método *Undersampling*, que remove instâncias da classe majoritária; e SMOTEENN, método híbrido que combina a sobreamostragem SMOTE e a subamostragem ENN. Essa estratégia permitiu analisar e comparar o

impacto do balanceamento no desempenho dos modelos aplicados

3.4 Aplicação dos modelos

O processo de aplicação dos modelos se deu em dois momentos: com os dados ainda desbalanceados e após a aplicação das técnicas de balanceamento. Assim foram utilizadas as técnicas de *Oversampling* e *Undersampling*.

3.4.1 Regressão Logística

Durante as análises, verificou-se o registro de 4.746 pacientes sem AVC (classe 0) e 248 com AVC (classe 1), então aplicou-se a Regressão Logística com estes dados não balanceados para averiguar se isso impactaria no desempenho do algoritmo.

Então, com esta aplicação, o objetivo foi verificar entre todos os positivos reais, quantos o modelo utilizado encontrou, e entre todos aqueles que ele julgou como positivo, quantos realmente eram. Após a análise dos resultados e do balanceamento do dataset, o modelo foi aplicado novamente.

3.4.2 Random Forest

Assim como a Regressão Logística, o Random Forest também é um algoritmo de Aprendizado de Máquina e foi utilizado com o mesmo objetivo. A sua aplicação também foi feita antes e depois do balanceamento, observando a porcentagem que ele sinalizou como positivo, e quantos realmente eram. Esta análise é importante para que a previsão seja precisa e não erre nos resultados.

3.5 Avaliação

A avaliação do desempenho dos modelos foi realizada por meio das métricas Acurácia, Precisão, Recall, F1-score e Área sob a Curva ROC (AUC). Essas métricas foram escolhidas por serem amplamente utilizadas em problemas

de classificação e, especialmente, por fornecerem uma análise mais adequada em cenários com classes desbalanceadas, como o presente projeto. Sendo assim, considerando o contexto do problema e o desbalanceamento entre as classes, as métricas Recall, Precisão e F1-score da classe positiva (AVC) foram adotadas como critérios principais para a comparação e seleção do melhor modelo.

A acurácia mede a proporção total de predições corretas realizadas pelo modelo. Porém, por si só, essa métrica pode ser insuficiente em bases desbalanceadas, pois pode mascarar um baixo desempenho na identificação da classe minoritária.

A precisão avalia a proporção de instâncias corretamente classificadas como positivas em relação ao total de predições positivas, enquanto o recall indica a capacidade do modelo em identificar corretamente os casos positivos. Por outro lado, o F1-score corresponde a média harmônica entre a precisão e o recall, útil para avaliar o equilíbrio entre essas duas métricas. Por fim, a métrica AUC foi utilizada para mensurar a capacidade discriminativa dos modelos, indicando o quão bem eles distinguem entre as classes positiva e negativa.

4 Resultados e discussões

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação dos diferentes modelos de aprendizado de máquina, comparando os resultados para os dois cenários, base de dados desbalanceada e balanceada. As análises baseiam-se nas métricas descritas na seção 3.5, com ênfase no recall, precisão e F1-score.

4.1 Desempenho da Regressão Logística

Na Tabela 1, observa-se que, no cenário sem balanceamento, a Regressão Logística apresentou acurácia moderada, porém desempenho limitado na identificação da classe positiva, apresentado no valor baixo da precisão. Apesar disso, o recall da classe positiva obteve um valor significativamente elevado, indicando que o modelo conseguiu

identificar uma parcela significativa dos casos de AVC, ainda que com muitos falsos positivos.

Com a aplicação da técnica SMOTE, verificou-se um aumento na acurácia comparado ao cenário desbalanceado. Porém, a precisão e o F1-score permaneceram baixos. Além disso, observou-se uma redução do recall da classe positiva, indicando que o método de desbalanceamento utilizado não foi o suficiente para melhorar o desempenho do modelo na identificação da classe minoritária.

Ao aplicar a técnica ENN, observou-se um aumento significativo no recall da classe positiva, indicando maior sensibilidade do modelo. Por outro lado, observou-se uma redução da acurácia e da precisão, o que evidencia um aumento no número de classificações incorretas.

Por fim, com a aplicação da técnica SMOTEENN observou-se um desempenho excelente para todas as métricas avaliadas. Esse resultado indica que a combinação de oversampling e undersampling permitiu à Regressão Logística aprender padrões mais representativos da classe minoritária.

Balanceamento	ACC(%)	P-(%)	P+(%)	F1-(%)	F1+(%)	R-(%)	R+(%)	AUC (%)
Não	73,38	98,12	12,47	83,98	21,30	73,40	72,97	73,18
SMOTE	82,65	96,56	12,5	90,28	19,25	84,77	41,89	63,33
ENN	65,77	98,61	10,87	78,29	19,21	64,91	82,43	73,67
SMOTEENN	88,74	89,82	87,97	87,01	90,07	84,38	92,27	88,33

Tabela 1: Desempenho da Regressão Logística considerando diferentes técnicas de balanceamento

4.1 Desempenho do Random Forest

Na Tabela 2, observa-se que, no cenário sem balanceamento, o modelo Random Forest apresentou acurácia alta, porém desempenho limitado na identificação da classe positiva, evidenciado pelos baixos valores de recall e F1-score.

Com a utilização da técnica SMOTE, não verificou-se um aumento no desempenho do modelo, mantendo valores baixos de precisão, recall e F1-score da classe positiva.

Ao aplicar a técnica ENN, observou-se um aumento significativo no recall da classe positiva, indicando maior capacidade do modelo em identificar casos de AVC. Por outro lado, observou-se uma redução da acurácia.

Por fim, com a aplicação da técnica SMOTEENN observou-se um aumento significativo do desempenho do modelo, com valores elevados em todas as métricas avaliadas. Indicando que essa técnica permitiu o Random Forest a identificar e aprender padrões mais representativos da classe minoritária.

Balanceamento	ACC(%)	P-(%)	P+(%)	F1-(%)	F1+(%)	R-(%)	R+(%)	AUC (%)
Não	88,79	96,17	15,94	93,96	20,75	91,85	29,72	60,79
SMOTE	90,66	95,59	13,33	95,05	14,63	94,52	16,21	55,37
ENN	78,91	97,51	13,55	87,80	22,16	79,85	60,81	70,33
SMOTEENN	96,35	96,97	95,87	95,87	96,73	94,79	97,61	96,20

Tabela 2: Desempenho do Random Forest considerando diferentes técnicas de balanceamento

5 Considerações Finais

Observando os resultados obtidos, percebe-se que com dados desbalanceados o modelo de Regressão Logística conseguiu identificar uma boa parcela de casos de AVC, já o Random Forest obteve valores menores. Após a aplicação das técnicas, nota-se que o SMOTEENN se destacou, principalmente com o algoritmo Random Forest.

Com isso, conclui-se que em casos como o deste dataset, o balanceamento é a melhor solução. Modelos como a Regressão Logística e o Random Forest podem ser utilizados para uma melhor predição, principalmente em conjuntos do domínio saúde, pois o objetivo é evitar que pacientes doentes sejam classificados como saudáveis. Além disso, deve-se considerar todas as possibilidades, e

ao final de cada teste verificar qual técnica e modelo apresentaram o melhor desempenho.

Referências

Cleiane Gonçalves Oliveira et al. ABORDAGENS PARA TRATAMENTO DE DADOS DESBALANCEADOS UTILIZADAS NO APRENDIZADO DE MÁQUINA NA ÁREA DA SAÚDE: UM ESTUDO DE CASO NA PREDIÇÃO DE ACIDENTE VASCULAR CEREBRAL (AVC). In: ANAIS DO LV SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2023, São José dos Campos. Anais eletrônicos..., Galoá, 2023. Disponível em: <<https://proceedings.science/sbpo/sbpo-2023/trabalhos/abordagens-para-tratamento-de-dados-desbalanceados-utilizadas-no-aprendizado-de?lang=pt-br>>. Acesso em: 08 Nov. 2025.

FEDESORIANO (2023). Stroke Prediction Dataset [conjunto de dados]. Kaggle. Disponível em: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data>. Acesso em: 5 Out. 2025.

FERNANDES, Antônio Alves Tôrres et al. Leia este artigo se você quiser aprender regressão logística. **Revista de Sociologia e Política**, v. 28, p. 006, 2020.

TELOKEN, Alex et al. Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão. **Simpósio de Pesquisa e Desenvolvimento em Computação**, v. 2, n. 1, 2016.