



Previsão de AVC com Aprendizado de Máquina

Cleverson Pereira da Silva¹

Leandro Zerbinatti

¹Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

10391119@mackenzista.com.br

Resumo. O Acidente Vascular Cerebral (AVC) é uma das principais causas de morte e incapacitação no mundo. A detecção precoce de fatores de risco é fundamental para prevenção. Este projeto aplica técnicas de aprendizado de máquina para prever a ocorrência de AVC a partir de dados clínicos e demográficos. O dataset utilizado, Stroke Prediction Dataset (Kaggle), contém informações como idade, sexo, hipertensão, doença cardíaca, nível de glicose, índice de massa corporal (IMC), tabagismo e tipo de trabalho. O objetivo é desenvolver modelos preditivos que auxiliem profissionais da saúde na identificação precoce de pacientes em risco. Para o desenvolvimento, serão aplicados algoritmos de classificação como Random Forest e XGBoost, com técnicas de balanceamento de classes (SMOTE) e métricas de avaliação como precisão, revocação e F1-score. Os resultados esperados incluem a obtenção de um modelo capaz de detectar pacientes de risco com maior sensibilidade, apoando a tomada de decisão clínica.

1. Introdução

a) Contextualização

O Acidente Vascular Cerebral (AVC) é um problema de saúde pública que causa elevada morbidade e mortalidade em todo o mundo. A identificação de fatores de risco é essencial para implementar ações preventivas e reduzir seus impactos sociais e econômicos.

b) Justificativa

O uso da Inteligência Artificial em saúde permite identificar padrões ocultos em bases de dados médicos. Este projeto busca contribuir para o apoio à decisão clínica, fornecendo uma ferramenta que auxilie na detecção precoce de risco de AVC.

c) Objetivo

Desenvolver e avaliar modelos de aprendizado de máquina capazes de prever a ocorrência de AVC em pacientes com base em informações clínicas e demográficas.

d) Opção do Projeto

Este projeto segue a opção 'Framework', utilizando bibliotecas como scikit-learn e XGBoost para modelagem e análise.

2. Descrição do Problema

O desafio consiste em prever a ocorrência de AVC com base em variáveis clínicas e demográficas. O problema é formulado como uma tarefa de classificação binária, onde a variável alvo indica se o paciente sofreu ou não AVC. A dificuldade principal reside no desbalanceamento do dataset, com poucos casos positivos, exigindo técnicas de reamostragem e modelos robustos.

3. Aspectos Éticos e Responsabilidade

O uso de Inteligência Artificial em saúde exige cautela quanto à privacidade e ao uso ético dos dados. O dataset empregado está anonimizado, garantindo confidencialidade. É importante ressaltar que a solução proposta não substitui o diagnóstico médico, mas atua como ferramenta de apoio à decisão. O modelo deve ser avaliado quanto a vieses e seu uso deve respeitar os princípios de responsabilidade social e ética em saúde.

4. Dataset

O dataset utilizado é o 'Stroke Prediction Dataset' disponível no Kaggle. Ele contém variáveis como idade, sexo, hipertensão, doença cardíaca, nível de glicose, IMC, tipo de trabalho e hábitos de tabagismo. O dataset possui cerca de 5 mil registros, com a variável alvo 'stroke' (0 = não teve AVC, 1 = teve AVC).

Durante a preparação dos dados foram realizadas etapas de limpeza (remoção de valores nulos em IMC), codificação de variáveis categóricas (LabelEncoder),

normalização (StandardScaler) e balanceamento de classes (SMOTE). A análise exploratória incluiu estatísticas descritivas e visualizações para entender a distribuição das variáveis.

4.1 Escolha dos Modelos

Random Forest Classifier: Modelo robusto baseado em múltiplas árvores de decisão.

XGBoost Classifier: Algoritmo de boosting gradiente com alto desempenho em dados estruturados.

4.2 Pré-processamento

- Remoção da coluna 'id' Padronização com
- Imputação de valores StandardScaler
- ausentes em 'bmi'
- Codificação com • Divisão treino/teste com
- LabelEncoder estratificação

Aplicação de SMOTE: Técnica usada para平衡ar os dados sinteticamente. Ela gera novas amostras da classe minoritária (AVC), ajudando os modelos a aprenderem melhor seus padrões.

4.3 Implementação dos Modelos

Random Forest:

```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
```

XGBoost:

```
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss',
n_estimators=100, random_state=42)
```

4.4 Métricas Utilizadas

- | | |
|------------------------|------------------|
| Matriz de Confusão | - F1-Score |
| - Precisão (Precision) | - Acurácia Geral |
| - Revocação (Recall) | |

4.5 Comparação dos Resultados (com SMOTE)

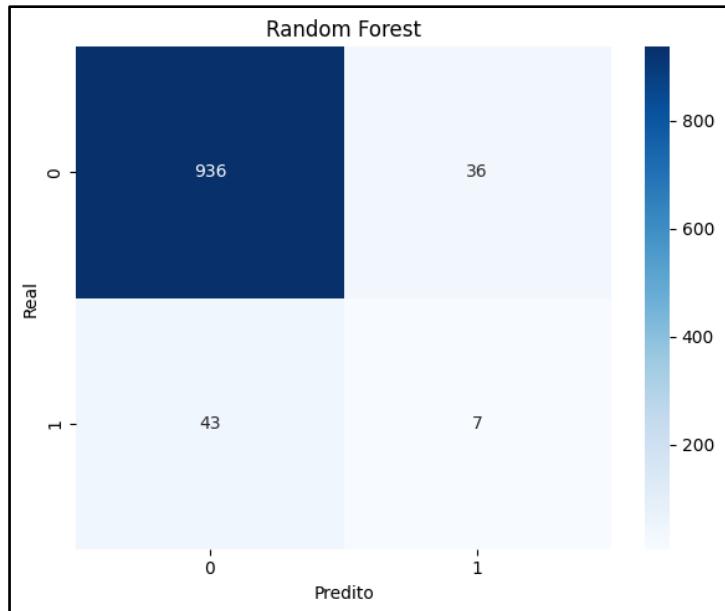
Resultados obtidos após aplicação do SMOTE no conjunto de treino:

Modelo	Precisão (classe 1)	Revocação (classe 1)	F1-score (classe 1)
Random Forest (SMOTE)	0.16	0.14	0.15
XGBoost (SMOTE)	0.19	0.14	0.16

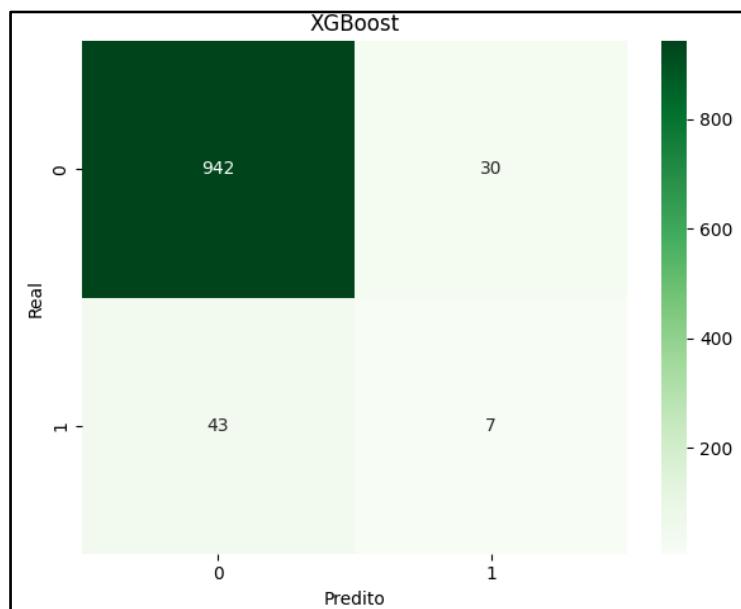
Ambos os modelos conseguiram detectar alguns casos positivos de AVC. O XGBoost apresentou desempenho ligeiramente superior, com melhor precisão e F1-score. A técnica de balanceamento SMOTE foi essencial para esse ganho.

4.6 Matrizes de Confusão (com SMOTE)

◆ Matriz de Confusão - Random Forest



■ Matriz de Confusão - XGBoost



A aplicação do SMOTE permitiu que os modelos aprendessem melhor sobre a classe minoritária (pacientes com AVC), que inicialmente era ignorada. Embora os valores de revocação e F1-score ainda sejam baixos, o modelo XGBoost se destacou ligeiramente. A acurácia geral manteve-se alta (~93%). A inclusão de técnicas de balanceamento é essencial nesse tipo de problema. Futuras melhorias podem incluir seleção de atributos, ajuste de limiares e ensembles.

5. Metodologia e Resultados Esperados

A metodologia envolve a aplicação de algoritmos de aprendizado supervisionado, em particular Random Forest e XGBoost, por serem adequados para dados tabulares e desbalanceados. As etapas incluem: divisão treino/teste com estratificação, aplicação de SMOTE para balanceamento, treino dos modelos e avaliação por métricas de classificação. Os resultados esperados incluem a melhora da sensibilidade (recall) na detecção de pacientes de risco e a obtenção de métricas equilibradas de precisão e F1-score. Espera-se que o modelo contribua como uma ferramenta de suporte à decisão em saúde, destacando pacientes com maior risco de AVC.

6. Referências bibliográficas

KAGGLE. Stroke Prediction Dataset. Disponível em:
<<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>>. Acesso em:
10 set. 2025.

NATURE. Scientific Reports – Predicting Stroke Using Machine Learning. Disponível
em: <<https://www.nature.com/articles/s41598-024-61665-4>>. Acesso em: 12 set. 2025.

IBM. Random forest. Disponível em: <<https://www.ibm.com/br-pt/think/topics/random-forest>>. Acesso em: 12 set. 2025.

IBM. XGBoost. Disponível em: <<https://www.ibm.com/br-pt/think/topics/xgboost>>. Acesso em: 12 set. 2025.