Predicting Stroke Risk from Health Indicators Using Machine Learning

CMSE 492 Project Proposal

Jared Frazier
frazi153@msu.edu

Date: 11/2/2025

GitHub: https://github.com/jcf454/cmse492_project/tree/main

Abstract

Stroke remains one of the most common and catastrophic medical emergencies, with a leading rank among causes of death and disability in the world. Timely identification of high-risk individuals can enable prevention and thus significantly reduce the long-term burden of both patients and healthcare systems. The aim of this project is to develop and evaluate machine learning models capable of predicting stroke incidents from a wide variety of health indicators that include age, hypertension, heart disease, glucose levels, and BMI.

I will perform data preprocessing, handle missing values, encode categorical features, and address issues related to class imbalance before modeling on a publicly available anonymous patient health records dataset. The study will compare the performance of a number of supervised learning algorithms, namely logistic regression, random forests, and a shallow neural network, in terms of the best balance between predictive accuracy and clinical interpretability. Model performance will be evaluated in terms of recall, precision, F1-score, and ROC-AUC, focusing on correct identification of the positive stroke cases. Preliminary analysis shows that all baseline models reach high accuracy with low recall, which highlights the challenge of severe class imbalance. This will lead to the development of a robust yet interpretable model that will reveal how careful data preparation and model selection can be done to increase prediction quality and reliability for medical risk.

1 Background and Motivation

When the blood supply to the brain is cut off, a stroke happens, which frequently results in fatalities or serious neurological impairment in a matter of minutes. Since many stroke risk factors are known, including high blood sugar, heart disease, and hypertension, machine learning presents the possibility of modeling how these factors interact and identifying high-risk individuals before a stroke happens. The incidence and severity of stroke cases may be decreased by preventative measures brought about by early detection.

Conventional stroke risk assessment methods may miss nonlinear dependencies or combined effects between several factors because they use linear scoring systems and fixed thresholds. On the other hand, these intricate relationships can be directly learned from data by machine learning models. By doing this, they might find risk trends that are missed by conventional approaches.

Despite the promise, stroke prediction is a challenging problem: data are inherently imbalanced, as few individuals actually have a stroke, and missing or noisy clinical variables make modeling difficult. This study will take on these challenges explicitly through careful preprocessing, balanced model evaluation, and model comparison across several model families. More broadly, this will help to establish how data-driven methods can be used to complement medical decision-making without sacrificing interpretability or clinical transparency.

2 Data Description

The ”Stroke Prediction Dataset,” which is accessible on Kaggle, was initially gathered from medical records for studies on stroke risk factor identification. A patient’s demographic, behavioral, and medical characteristics, including age, gender, marital status, type of residence, smoking status, and type of employment, are included in each entry. While the binary target variable `stroke` indicates whether the person has had a stroke, physiological measurements such as average glucose level and BMI offer quantitative indicators of general health.

There are more than 5,000 records and about a dozen features in the dataset. Data cleaning entails using the column mean to impute missing BMI values and eliminating the `id` column, which has no predictive value. One-hot encoding is used to transform categorical variables, like gender and work type, into numerical indicators. Because stroke cases make up less than 5 percent of all records, the dataset is highly unbalanced, which may cause models to predict ”no stroke.” In order to counteract this, I will make sure that minority class performance is prioritized during model training and evaluation by using stratified sampling and class weighting. To ensure reproducibility and version control, the processed dataset will be stored in the project’s `data/processed/` directory.

3 Proposed Methodology

To determine whether a patient is at risk of having a stroke, the analysis will be conducted using a supervised learning framework. I’ll divide the dataset into training, validation, and testing subsets after cleaning and encoding it. A logistic regression classifier and a majority-class predictor will be part of the baseline models, which will act as benchmarks for more intricate models. The simplicity and interpretability of logistic regression make it an ideal foundation for clinical modeling.

A shallow neural network and a random forest classifier will be used in later models. By using feature importance metrics, random forests can preserve interpretability while capturing nonlinear relationships and interactions between features. By learning more intricate decision boundaries, the neural network—which has one or two hidden layers—may enhance recall even more. Cross-validation will be used to lessen variance in results, and validation performance will be used to adjust model hyperparameters.

Using libraries like `scikit-learn` and `TensorFlow/Keras`, all models will be implemented in Python. When necessary, feature scaling, normalization, and class rebalancing will be used. The trade-off between interpretability and predictive power at progressively higher levels of model complexity will be evaluated with the aid of the comparison of model results.

4 Evaluation Framework

Given the class imbalance inherent to this dataset, evaluation metrics must go beyond overall accuracy. The analysis will focus on recall, precision, F1-score, and ROC-AUC. High recall is particularly critical, as the primary goal in a clinical setting is to identify as many true stroke cases as possible, even if some false positives occur. Precision will also be monitored to ensure that the model does not produce an excessive number of false alarms, which could reduce its clinical usefulness. The F1-score will serve as a balanced measure between these two objectives, and the ROC-AUC will provide an aggregate assessment of model discrimination capability across thresholds.

Training and testing sets will be created using stratified sampling to preserve class proportions. I will compare baseline and advanced models under consistent evaluation conditions. Early results from the majority-class baseline yield roughly 95% accuracy but negligible recall for stroke cases, underscoring the difficulty of predicting rare events. The project's success will be measured by achieving a meaningful improvement in recall and F1-score for the stroke class while maintaining reasonable precision. Additionally, feature importance analysis and model interpretability tools such as SHAP values may be used to highlight which health indicators contribute most strongly to predictions, linking quantitative results to established medical insights.