Stroke Prediction Using Traditional Machine Learning Algorithms

9 April 2025

Sarthak Pandit (C24CS1001)
Siddharth Ranka (C24CS1002)
Mansi Muneshwar (B21CS047)

Abstract

This report presents a comparative analysis of various classical machine learning techniques for the binary classification problem of predicting strokes based on health and demographic data. Using SMOTE to address class imbalance, we evaluate performance using metrics such as accuracy, precision, recall, and F1-score on the Kaggle Stroke Prediction dataset.

1 Introduction

Stroke is one of the leading causes of death and long-term disability. Accurate prediction of stroke using health data can help in timely diagnosis and prevention. The goal of this project is to analyze the effectiveness of different machine learning models in predicting the risk of stroke based on clinical features.

2 Dataset and Preprocessing

We used the publicly available dataset from Kaggle: `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`. The dataset contains 5110 samples after SMOTE balancing. Key preprocessing steps included:

- Handling missing values
- Label encoding for categorical features
- Feature scaling using StandardScaler
- Balancing data using SMOTE (Synthetic Minority Over-sampling Technique)

3 Methodology

We evaluated the following models:

3.1 1. Random Forest

An ensemble method using 5 trees was used for classification. The model achieved the best performance among all methods.

- Accuracy: 92.54%
- Precision: 89.95%
- Recall: 95.78%
- F1 Score: 92.78%

3.2 2. Multilayer Perceptron (MLP)

Implemented using a single hidden layer (100 neurons, 300 iterations). The model underperformed significantly:

- Accuracy: 65.86%
- Precision: 59.52%
- Recall: 98.97%
- F1 Score: 74.34%

3.3 3. Support Vector Machine (SVM)

Trained using Stratified 5-Fold Cross-Validation with RBF kernel.

- Accuracy: 79.16%
- Precision: 92.62%
- Recall: 79.16%
- F1 Score: 84.65%

3.4 4. Naive Bayes

GaussianNB with 5-Fold Cross-Validation and final evaluation on a hold-out test set.

- Average CV Accuracy: 72.97%
- Final Test Accuracy: 71.82%
- Final F1 Score: 79.81%

3.5 5. Gaussian Mixture Model (GMM)

Unsupervised clustering approach with two components (stroke and no stroke).

- Accuracy: 62.82%
- F1 Score: 73.40% (weighted)
- Very poor recall for stroke class (6.02%)

4 Results Summary

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	92.54%	89.95%	95.78%	92.78%
MLP	65.86%	59.52%	98.97%	74.34%
SVM	79.16%	92.62%	79.16%	84.65%
Naive Bayes	71.82%	93.39%	71.82%	79.81%
GMM	62.82%	88.68%	62.82%	73.40%

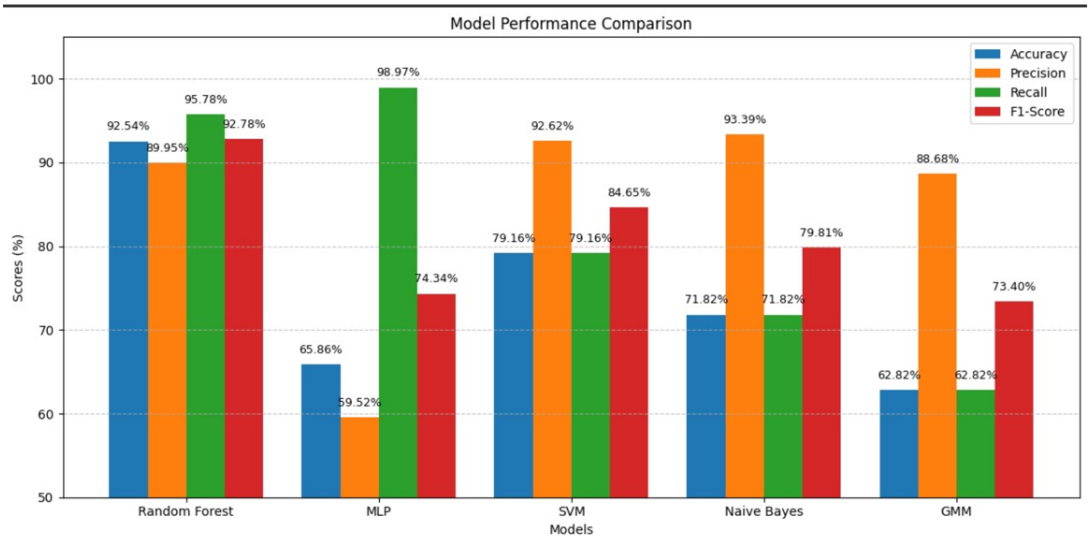

Figure 1: Comparison of Model Performance Metrics

5 Qualitative Analysis

In this section, we analyze the qualitative performance of each model:

1. Random Forest: The Random Forest model demonstrated great stability and handled outliers well. It was able to generalize better on unseen data. Visual inspection of predicted vs actual cases showed that the model performed well across the different classes.

2. MLP: MLP performed poorly in terms of recall, specifically for the stroke class. We analyzed the misclassified cases and found that the model struggled with very imbalanced features. The network's architecture and training strategy might need adjustment.

3. SVM: SVM performed relatively well, with the highest precision of 92.62%. However, it struggled with recall for the stroke class, especially when identifying stroke cases. The recall value (79.16%) indicates that the model still missed many stroke cases, which suggests that fine-tuning the SVM hyperparameters and exploring alternative kernels may improve performance.

4. Naive Bayes: Naive Bayes struggled with handling the class imbalance, which led to high precision but low recall for the stroke class.

5. GMM: GMM performed poorly in classification, especially in recall for the stroke class. This is expected, as GMM is not designed for classification tasks.

6 Failure Analysis

The failure analysis highlights several key issues:

1. Class Imbalance: The major failure for many models, especially MLP and Naive Bayes, was the inability to correctly predict the minority class (stroke class). Despite using SMOTE, the imbalance affected the performance, particularly recall.

2. Overfitting: The Random Forest model, while successful, did show some signs of overfitting. A cross-validation approach with more diverse data could help mitigate this.

3. Model Complexity: For models like MLP, more sophisticated architectures (e.g., adding more layers or neurons) and tuning of hyperparameters may be needed to improve recall.

7 Conclusion

Among the evaluated models, the Random Forest classifier demonstrated the best overall performance, particularly in identifying stroke cases with high recall. While SVM showed the highest precision, it still missed a significant portion of stroke cases, as reflected in its recall. Future work may involve experimenting with ensemble techniques or deep learning architectures for better generalization.