Stroke Prediction Using Machine Learning Techniques

CSE422 Lab Project Report

***Abstract*—This report focuses on predicting stroke by utilizing machine learning models with the help of healthcare dataset. The data set was processed accordingly, trained and analyzed to get results. Logistic regression, Decision tree and Neural Network were chosen as models to train and evaluate. The goal is to identify the best suitable model with the help of prediction accuracy, F1 score, and AUC-ROC curve. Decision tree and neural network models were deemed suitable models as they captured the nonlinear relationships of the dataset. However, class imbalance affected the F1 scores.**

I. Introduction

Nowadays, stroke is a major health issue affecting millions of people worldwide which may lead to lasting brain damage, long-term disabilities, or even death. That is why it is crucial for us to detect it as soon as possible and prevent it. That is the motivation behind this project. The plan is to utilize an AI model such that it can predict strokes in light of clinical and segment information. A dataset of more than 5000 people with their related information like age, gender, residence, BMI, marital status, heart disease, work type, glucose level, smoking status and hypertension. An analysis will be done based on different AI calculations to get their precision. The ultimate goal is to set the model as a web application that is able to take inputs from clients and provide stroke expectations.

II. Dataset Description

The dataset contains demographic information of 5,110 people and 12 features. These features are namely, id, age, gender, residence, BMI, marital status, heart disease, work type, glucose level, smoking status, hypertension and stroke. The chosen target variable is stroke. There is a mixture of categorical, binary and continuous types in these features.

III. Data Analysis

The dataset contains 5,110 records with 12 features. Age moderately correlates with hypertension and heart disease, and bmi correlates with age. The target variable, stroke, shows weak correlations, and potential outliers and class imbalance suggest further pre-processing is needed.

IV. Data Preprocessing

- Handling null and duplicate values: Missing BMI values are filled using the median. Duplicates were not found here.
- Encoding categorical values: Label encoding was used for binary categorical features like residence, marital status and gender. One-hot encoding was used on multi-class features like work type, smoky status.
- Removing irrelevant features: Features like ID do not contribute to determining eligibility and thus are discarded in data processing.
- Feature scaling: StandardScaler was utilized for normalizing continuous features.

V. Model Training and Evaluation

We trained the following models:

- Logistic Regression
- Decision Tree
- Neural Network (MLP)

A. Evaluation Metrics

Model	Accuracy	F1 Score	AUC
Logistic Regression	0.952	0.039	0.842
Decision Tree	0.912	0.198	0.584
Neural Network	0.922	0.070	0.725

TABLE I
Model Performance Comparison

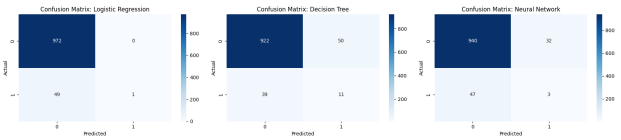

Fig. 1. Confusion Matrices of All Models

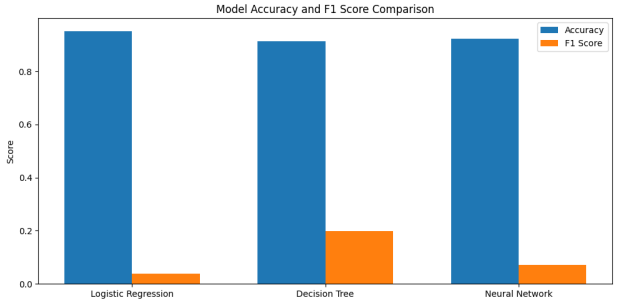

Fig. 2. Model Accuracy and F1 Score

VI. Conclusion

Logistic Regression achieved the highest accuracy and AUC but underperformed in F1 Score due to class imbalance. Decision Tree offered better F1 performance. The choice of best model depends on the application's focus—either on precision or sensitivity.

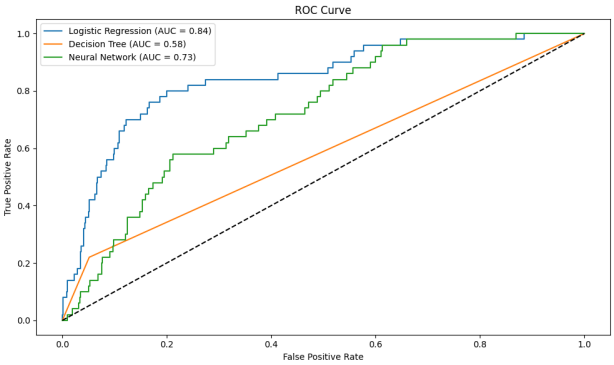

Fig. 3. ROC Curve Comparison