United International University

Project Report

Course Name: Pattern Recognition Laboratory

Course Code: CSI 416 (B)

Submitted By: Group 2 – Team Kingsmen

Project Name:

Performance Comparison using Machine Learning Classification Algorithms on a Stroke Prediction dataset.

Team Members:

Name	Student ID	Email Address
Mohammed Jawwadul Islam	011 181 182	mislam181182@bscse.uiu.ac.bd
MD Fahad Al Rafi	011 181 201	mrafi181201@bscse.uiu.ac.bd
Pranto Podder	011 181 202	ppodder181202@bscse.uiu.ac.bd

Dataset in Kaggle

Stroke Prediction Dataset | Kaggle

Problem Definition

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

In our project we want to predict stroke using machine learning classification algorithms, evaluate and compare their results.

Dataset Description

Number of instances = 5111

Number of attributes = 12

Attribute Information:

1) **id:** unique identifier
2) **gender:** "Male", "Female" or "Other"
3) **age:** age of the patient
4) **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) **ever_married:** "No" or "Yes"
7) **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) **Residence_type:** "Rural" or "Urban"
9) **avg_glucose_level:** average glucose level in blood
10) **bmi:** body mass index
11) **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) **stroke:** 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Acknowledgement:

Dataset was provided by fedesoriano

Our Approach

Data Analysis:

At first, using visualization libraries, we did some data visualizations by plotting various plots like pie chart, count plot, curves, etc. in order to understand the dataset better, and to find out the correlation between the attributes.

Data Preprocessing:

We've used all features except id, and stroke as class label. We dropped the id column from the dataframe. The full dataset was then split by 80-20, where 80% is used for training and 20% for testing.

We then found unique values present in each column. The output showed 201 feature vectors were missing from BMI feature, i.e. N/A (null values). Since BMI type is continuous, we replaced the null values by median using Sklearn Simple Imputer.

As sklearn libraries can work only with numeric valued data, so we've converted the text features into numeric value using LabelEncoder from Sklearn.

The dataset was imbalanced, where the no. of negative samples in the Label class far outweigh the positive samples by a large margin. So we oversampled the dataset, i.e. increase the number of positive samples, by using RandomOverSampler from imblearn.

Finally, we’ve scaled the dataset using StandardScaler from Sklearn.

Model Building:

After our dataset was finally ready, we have used some machine learning classification algorithms on this dataset and observed their performances.

Some of the algorithms that we’ve chosen to apply on this dataset are:

1. Logistic Regression
2. Naive Bayes
3. k Nearest Neighbors
4. Support Vector Machine – Gaussian SVM
5. Random Forest Classifier

Model Evaluation:

We then compared these results based on various classification metrics.

The metrics are: accuracy, precision, recall, f1 score and mcc score. The results are displayed at the end of the report.

Hyperparameter Tuning:

Finally, we performed hyperparameter tuning on 2 algorithms, SVM and Random Forest, to find the best model and parameters.

SVM:

Best Accuracy: 97.56%

Best parameters: kernel': 'rbf', **'gamma'**: 0.8, **'C'**: 1

Random Forest:

Best Accuracy: 97.56%

Best parameters: n_estimators': 80, **'max_features'**: 'sqrt', **'criterion'**: 'gini

Challenges

The dataset was imbalanced. Number of instances of Class label 0 were much higher than class label 1. So we had to oversample the dataset (by 70%) in order to get a good performance on our models.

Result/Comparison Metrics

We've used accuracy score and F1 score to compare between the algorithms. As we know, there's a tradeoff between precision and recall. To get the result from this we need to calculate F1 score.

1. **Accuracy** = (True positive + True negative) / (True positive + True Negative + False positive + False negative)
2. **Precision** = True Positive / True positive + False positive
3. **Recall** = True Positive / Total positive + False negative
4. **F1** score = (2 * Precision * Recall) / (Precision + Recall)
5. **MCC score** (Matthews correlation coefficient): Range of MCC is from -1 to +1, where +1 indicates a good model, and -1 indicates a bad model.

Formula Model	Accuracy	Precision	Recall	F1 score	MCC score
SVM	82.29%	0.14	0.46	0.22	0.18
KNN	76.32%	0.13	0.59	0.21	0.19
Random Forest	94.81%	1.00	0.02	0.04	0.13
Logisitic Regression	81.02%	0.17	0.65	0.27	0.26
Naive Bayes	79.45%	0.16	0.7	0.27	0.27

Conclusions

From this dataset, we learned a couple of important things, based on correlations between the variables:

1. The older you get, the more you are at risk of getting a stroke.

2. Overweight people have a higher chance of suffering a stroke

3. People with glucose level less than 100 suffers stroke more.

4. Imbalanced dataset can have poor performance on your levels.

5. If dataset is not very large, it is better to impute the missing values rather than dropping them.