United International University

Project Proposal

Course Name: Pattern Recognition Laboratory

Course Code: CSI 416 (B)

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

State-of-the-art

From the code section of this dataset, we have seen models being used. We've found that 91.10% accuracy, 94% precision and recall is the maximum score till now.

Proposed Approach

We'll use all features except id, and stroke as class label. At first we will preprocess the dataset, i.e. remove missing values, convert categorical variables to numerical, handle dataset if it is imbalanced, etc. Then using visualization libraries, we will do some data visualizations by plotting various plots in order to understand the dataset better, and to find out the correlation between the variables. After that, we'll use some machine learning classification algorithms on this dataset and will observe their performances. Finally, we'll compare these results based on some score.

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

Result/Comparison Metrics

At first, we will find out the number of True positive, True negative, False positive, and False negative by plotting a confusion matrix. Then we will use accuracy, precision, recall and F1 score. F1 score is needed for comparison as we know that there is a tradeoff between precision and recall.

Therefore,

1. Confusion Matrix
2. Accuracy = (True positive + True negative) / (True positive + True negative + False positive + False negative)
3. Precision = True Positive / True positive + False positive
4. Recall = True Positive / Total positive + False negative
5. F1 score = (2 * Precision * Recall) / (Precision + Recall)