PRML Project Report

Team Members:-

- Ashish Kumar (B20EE011)
- Krishna Mohan (B20EE031)
- Shailesh Yadav (B20EE062)

Topic:- Stroke Prediction

Abstract- We have been given a dataset for stroke prediction and we had to use three to four classifiers to predict stroke and report the accuracy. So we have performed random forest, decision tree, SVM, and logistic regression on the given dataset.

INTRODUCTION

Stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status.

Dataset:

The file healthcare-dataset-stroke-data.csv is used in training and testing in ratio 70:30 respectively.

The dataset contains 5110 rows and 11 columns.

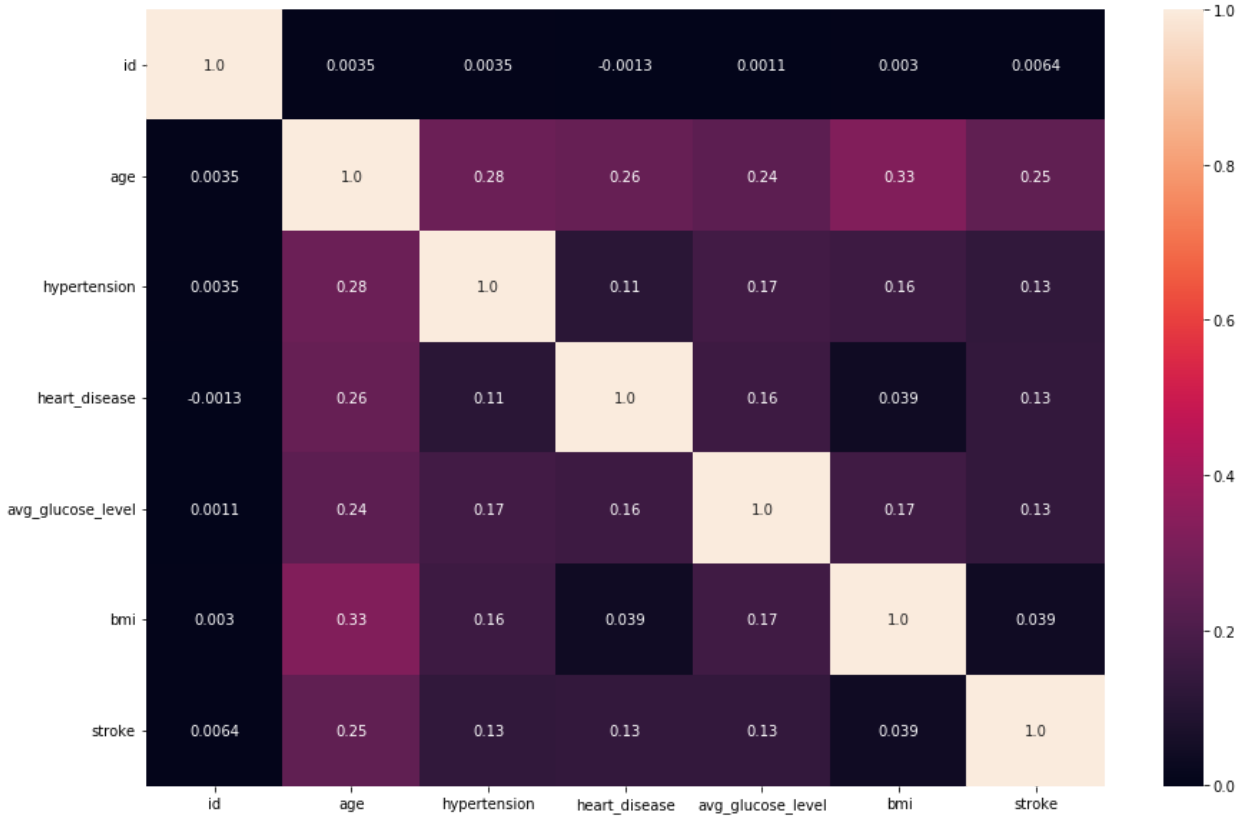
1 Id columns
2 gender-Male, female
3.age-continuous datatype
4.hypertension-binary datatype (0,1)
5.heart-disease-binary datatype (0,1)
6.ever_married- binary datatype (yes,No)
7.residence_type-Urban,ruler
8.avg_glucose_level-continuous datatype
9.BMI-continuous datatype
10.smoking_status-nominal
11.work_type-nominal

12.smoking_status-nominal
13. stroke-binary datatype (0,1)

PREPROCESSING THE DATASET

On counting the null value it is found that BMI feature has 201 null values. Since BMI contains some NAN values so replacing the nan values with the mean of BMI feature.

CORRELATION MATRIX



FROM THE ABOVE CORRELATION MATRIX WE CAN SEE THAT BMI HAS VERY LESS ASSOCIATED WITH THE STROKE. ID SEEMS NOT USEFUL COLUMN. LIGHTER IS COLOR MORE IT IS CORRELATED

Removing the irrelevant column Id from the dataset.

METHODOLOGY:

Below is the algorithm that we have implemented:

1. Decision tree classifier
2. Random forest classifier
3. Logistic regression
4. SVM

Decision tree: A decision tree is a supervised learning technique that is commonly used to solve classification problems.

Random forest classifier: Random Forest is a classifier that consists of a number of decision trees on various subsets of a given dataset and takes the average to enhance the dataset's prediction accuracy.

Logistic regression: Logistic regression is a common Machine Learning method that is part of the Supervised Learning approach. It is used to forecast the categorical dependent variable based on a set of independent factors.

SVM: The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may simply place fresh data points in the proper category in the future. This optimal choice boundary is referred to as a hyperplane.

EVALUATION OF MODEL:

MODEL	ACCURACY
Decision tree	0.9419
Random forest	0.9406
Logistic Regression	0.9452
SVM	0.9420

DECISION TREE	RANDOM FOREST
TP=1 FP=6 TN=1443 FN=83	TP=3 FP=10 TN=1439 FN=81
LOGISTIC REGRESSION	SVM
TP=0 FP=0 TN=1449 FN=84	TP=0 FP=0 TN=1449 FN=84

Individual Contribution

`Ashish Kumar (B20EE011)` -Preprocessing, LogisticRegression, Report.

`Krishna Mohan (B20EE031)` -Decision Tree, Random Forest, Report

`Shailesh Yadav (B20EE062)` -SVM, Exploratory analysis, Report

REFERENCES:

[1].https://www.geeksforgeeks.org/implementation-of-logistic-regression-from-scratch-using-python/

[2].https://python-bloggers.com/2021/04/master-machine-learning-decision-trees-from-scratch-with-python/

[3].https://www.python-engineer.com/courses/mlfromscratch/07_svm/

[4].https://www.python-engineer.com/courses/mlfromscratch/10_randomforest/

[5]https://www.hindawi.com/journals/jhe/2021/7633381/