REPORT

ML PROJECT

Harsh Kumar

Datasets

The given dataset is for stroke prediction. It has 5110 rows and 12 columns.

METHODOLOGY

OVERVIEW

There are various classification algorithms present out of which we shall implement the following

- Random Forest Classification
- KNN
- Logistic Regression
- SVM
- Decision Tree classifier

We also make use of PCA and LDA for dimensionality reduction. Also we will check the output of the model by hyperparameter tuning of some models.

Exploring the dataset and pre-processing

First I label encode all the column which need label encoding.Then I check the dataset for null values and I found 201 Nan values.

So , one solution is to drop these values and other is to fill these with ffill or using KNNemputer .I chose to drop these values , its because when i am training the data and then finding the score it giving me lesser accuracy then the accuracy i got if i drop the column. I have checked this on different models.

Then I dropped the unnecessary columns like id. Finally our dataset is now ready to use.

Implementation of Classification Algorithm

- **Model1 = Decision tree classifier** - It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
  Three type of Decision Tree classifier I made in this project
  1) Simple model
  2) Decision tree with PCA
  3) Decision tree with tuned hyperparameters
- **Model2 = Logistic Regression -** Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
  Three type of Logistic regression I made in this project -
  1)Simple logistic regression
  2)model with PCA
  3)model with tuned hyperparameters
- **Model3 = KNN -** K Nearest Neighbor or KNN algorithm falls under the Supervised Learning

category and is used for classification and regression
Three types of knn i have made in this project

1)Simple KNN
2)KNN with PCA
3) KNN with tuned hyperparameters

- **Model4 = SVC(Support Vector Classifier)-**SVC is a nonparametric clustering algorithm that does not make any assumption on the number or shape of the clusters in the data.
  Three types of SVC i have mad in this project

  1)Simple SVC
  2)SVC with PCA
  3) SVC with LDA

- **Model5 = Random Forest Classifier -** Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
  Three types of Random Forest i have made in this project

  1)Simple Random Forest
  2)Random forest with PCA
  3) Random Forest with tuned hyperparameters

EVALUATION OF MODELS

The models implemented were evaluated using techniques like - Classification report : precision , recall , f1 score , ROC plots and accuracy score.

Below are the tables that shows the results on the basis evaluation techniques

- Table 1

Accuracy with Decision tree Regression
Logistic KNN SVC Random Forest

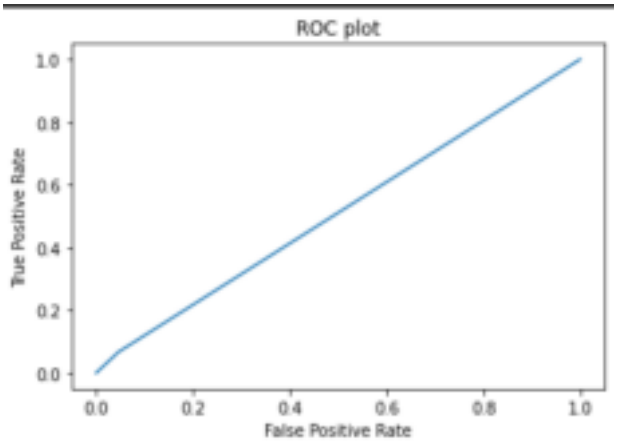
Simple model	92.05	95.31 95.21 95.41	95.21
PCA	91.14	95.4 95.3 95	95.42
Tuned hyperpara met ers	95.01	95.41 95.417	95.4

Table 2-

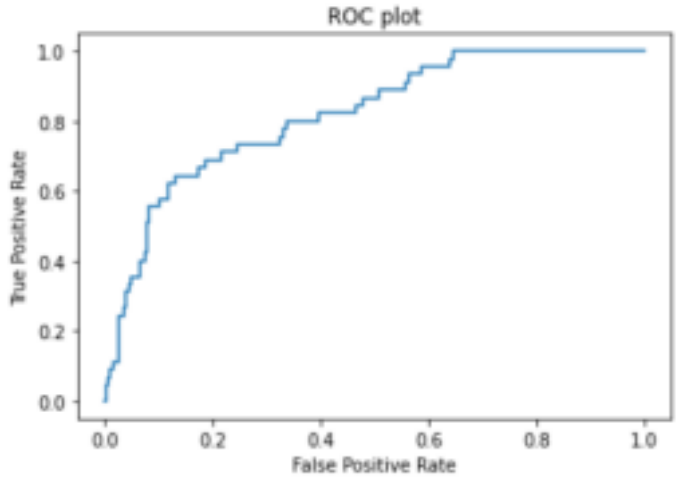
Model Name	AUC value Accuracy
Decision tree	0.51 92.05
Logistic Regression	0.81 95.31
KNN	0.65 95.21
SVC	0.72 95.41
Random Forest	0.76 95.2

ROC PLOTS

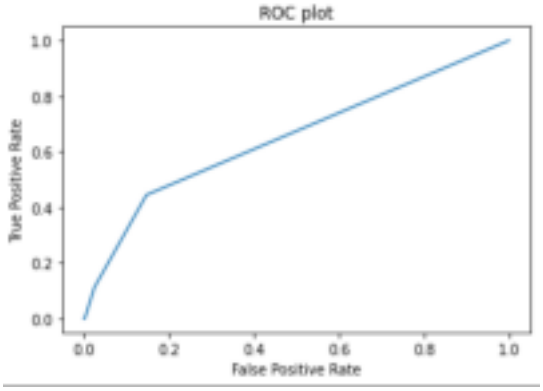
Model1 -



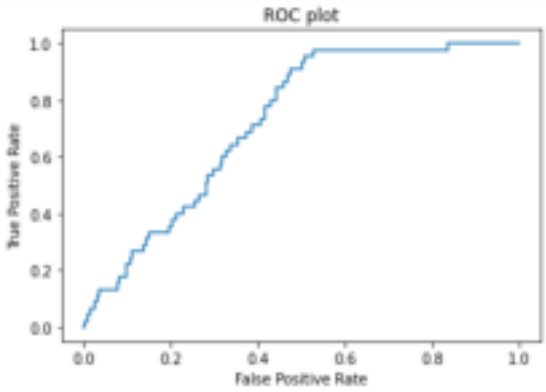
Model2-



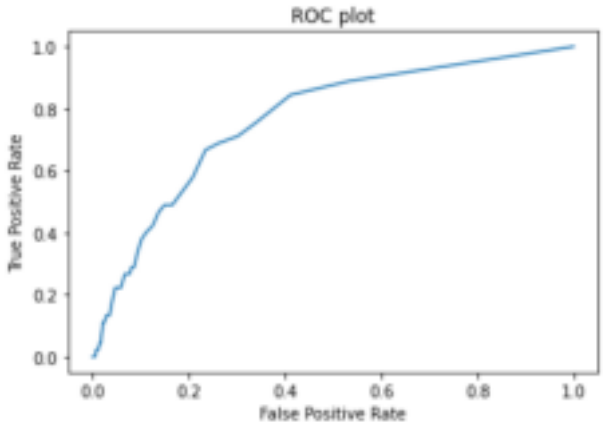
Model3-



Model4-



Model 5 -



Result and Analysis

Here we seen that the simple models logistic regression and random forest are performing very well with score of 95% and then when i used the dimensionality reduction technique like LDA and PCA the accuracy slightly decreased. But the speed

of the slow models like random forest is increased by dimensionality reduction but there was no effect seen in the case of powerful models like SVC.Also the other models like decision tree and logistic regression showed better performance after hypertunnng the parameters but that doesnt seems to happened with other models and finally I calculated the AUC values and Roc curves to visulaise the output. Finally the hypertuned decsion tree stand out best among all the models. The model name is model1__ in the project. Its accuracy score is highest 95.51 also by seeing the values precision and recall we can say that the false negative rate is lowest and false positive rate is highest among all other models, which is definitely required if we are making the perfect model to predict stroke.