BDSN ENDTERM PROJECT

EDA and modelling of stroke data

Stroke prediction using PySpark pipelines

Professor: Dr. Prithwis Mukerjee

Presentation by: Chinmaya Venkataraman
Roll no: A22014

SIGNIFICANCE OF PROJECT

- Stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes.
- A stroke can sometimes cause temporary or permanent disabilities, depending on how long the brain lacks blood flow and which part is affected. Stroke is the leading cause of disability worldwide and the second leading cause of death.
- Strokes cause immense health and economic burdens in India and globally. Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. This included the cost of health care services, medicines to treat stroke, and missed days of work.
- Being able to predict and prevent stroke can make a serious impact on quality of life and economic productivity lost on a daily basis.

DATA SOURCES AND SUBMISSION

Data source:

Kaggle

Kaggle link to dataset :

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Colab notebook link:

https://colab.research.google.com/drive/1FUmRHyN9TbqaB-fAKGt_He9M__4v-FZ7?usp=sharing

VARIABLES:

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_job", "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

RESEARCH OBJECTIVE AND METHODOLOGY

Objective:

- We try to determine whether the patient will have a stroke or not based on the input parameters like gender, age, various diseases, and smoking status by applying machine learning algorithms.

Methodology

- We use MongoDB server hosted on CleverCloud to manage our data remotely.
- We use MongoDB Compass to establish a connection and manage documents.
- We perform EDA on the data using Pandas, Matplotlib and Seaborn libraries in Python.
- We perform further EDA using spark.sql from PySpark.
- We treat the data as appropriate for the situation and use PySpark to fit machine learning pipelines to predict stroke.

METHODOLOGY

- After that the relevant predictor variables were assembled we divided the original dataset into an 80-20 training and test set.
- We then applied the following algorithms:
  - Decision Tree Classification
  - Logistic Regression
  - Random Forest Classification
  - Gradient Boosting Classification
  - Support Vector Machine – Linear
  - Naive Bayes.

ANALYSIS

We compared the performances of the algorithms and results are as follows:

Accuracy

```
Comparing pipeline accuracy scores:

Accuray of the Decision Tree pipeline:  94.766 %
Accuray of the Logistic Regression pipeline:  95.133 %
Accuray of the Random Forest pipeline:  95.133 %
Accuray of the Gradient Boosting pipeline:  94.674 %
Accuray of the Support Vector Machine pipeline:  95.133 %
Accuray of the Naive Bayes pipeline:  80.533 %
```

Area Under ROC Curve

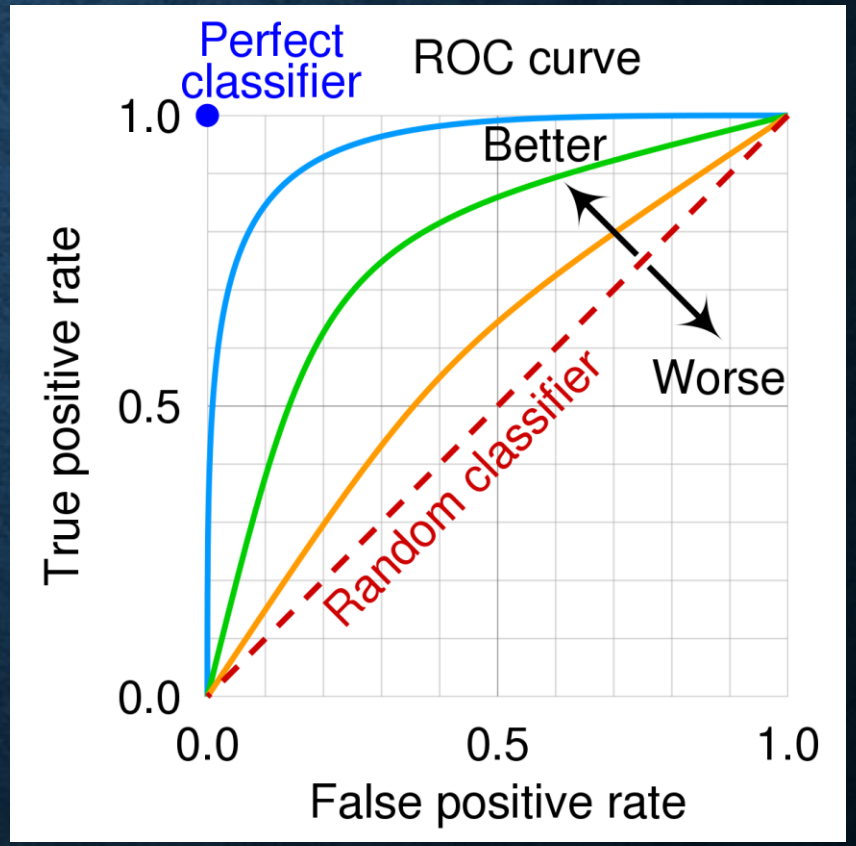
```
Comparing area under ROC curve:

Area under ROC curve of the Decision Tree pipeline:  0.57849
Area under ROC curve of the Logistic Regression pipeline:  0.83148
Area under ROC curve of the Random Forest pipeline:  0.83749
Area under ROC curve of the Gradient Boosting pipeline:  0.83206
Area under ROC curve of the Support Vector Machine pipeline:  0.68606
Area under ROC curve of the Naive Bayes pipeline:  0.18717
```

Note: Values obtained when file run at - 2023-02-05 20:53:43.394779+05:30

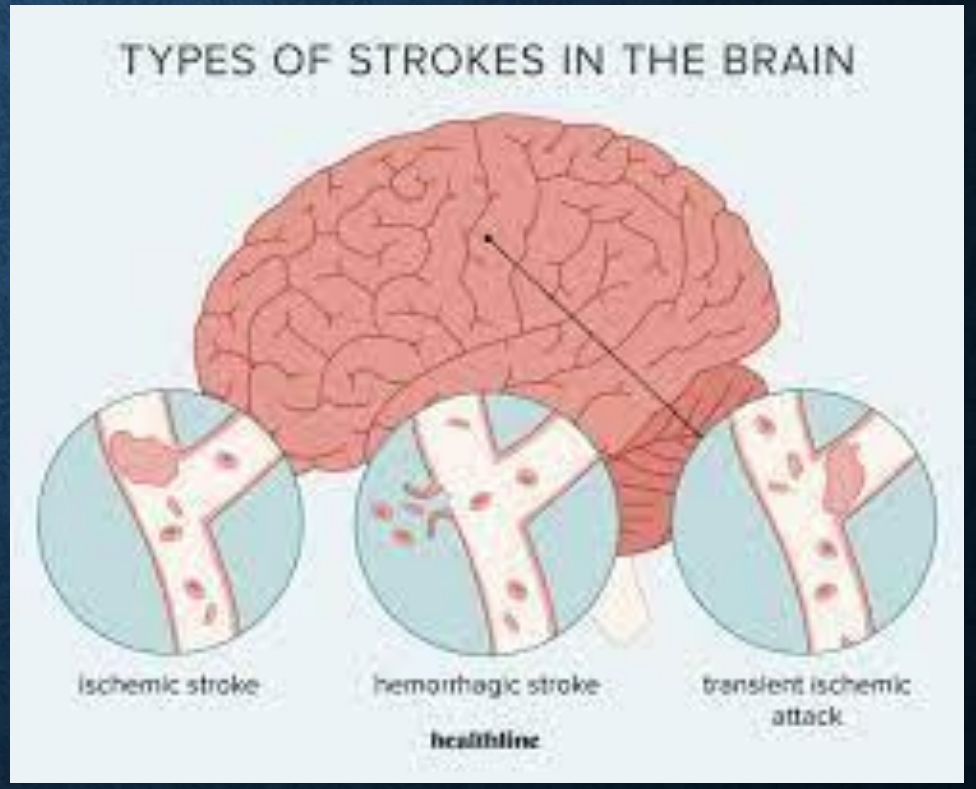
INTERPRETATION

- Random Forest pipeline along with logistic regression and gradient boosting gave the best result for our data.
- Most models gave an accuracy score between 94 and 96% except for Naïve Bayes which gave an accuracy score of ~80%.
- Depending on the train-test split, the best model varied.



CONCLUSION

- Strokes can be effectively predicted with machine learning models.
- Random forest along with logistic regression and gradient boosting were the best for this purpose based on data.
- The factors that we chose to predict stroke are likely not independent, despite giving us an accuracy of 80% since other models performed better.



THANK YOU