Project Milestone -2

Debasish Panda

DSC680- Applied Data Science

Prof. Catherine Williams

Project Proposal: **Can we predict Stroke?**

Business Problem:

To design a robust system that works efficiently and will be able to predict whether a given person is likely to have a stroke or not based on the risk factors.

Background/History:

A stroke, or brain attack, happens when blood flow to our brain is stopped. It is an emergency situation. The brain needs a constant supply of oxygen and nutrients in order to work well. If blood supply is stopped even for a short time, this can cause problems. Brain cells begin to die after just a few minutes without blood or oxygen. When brain cells die, brain function is lost. We may not be able to do things that are controlled by that part of the brain. For example, a stroke may affect our ability to move, speak, eat, think and remember, controlling emotions, controlling bowel and bladder etc. In other words, a stroke can happen to anyone at any time.

- According to CDC, every year more than 795,000 people in United States have a stroke. Every 4 minutes someone dies of stroke.
- Even though stroke is a leading cause of death for Americans, but the risk of having a first stroke is nearly twice as high for blacks as for whites, and blacks have the highest rate of death due to stroke.
- Recurrent stroke is frequent; about 25 percent of people who recover from their first stroke will have another stroke within 5 years.

Since stroke is something that can happen to anyone at any age, if we can use some of the prediction modelling techniques to predict whether a given person is likely to have a stroke or not based on the risk factors, we discussed then we can save a lot more lives.

Data Source:

Data for this project is downloaded from Kaggle. This data set has more than 5000 observations.

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

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_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 notBelow is the link to the actual data set.

Methodology:

Initially, exploratory data analysis was performed to understand the data better and compute some major descriptive statistics. I produced some illustrative visualizations representing the dataset as well. Then, various machine learning algorithms were employed to see the best- resulting model for the prediction model. Mainly, SVM, Logistic Regression, Decision Tree, Random Forest, and KNN are the potential algorithms used.

The collected dataset is analyzed using Python programming language, and Jupyter Notebook is used as a scripting interface. Several python libraries are used for the various purposes of exploratory data analysis, visualization, and machine learning model creation. Some of the notable ones are:

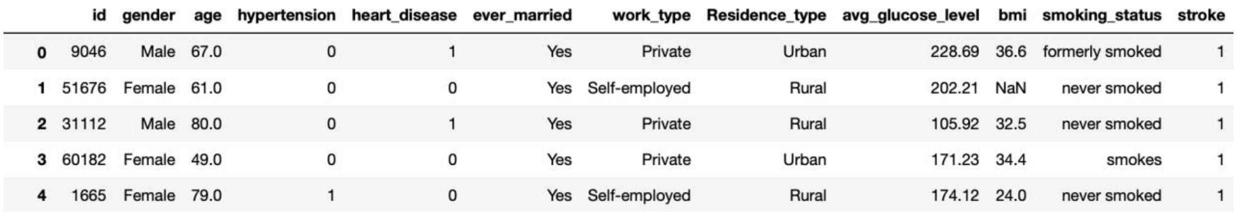
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn

After loading the dataset into the pandas data frame, different measures about the dataset were generated, such as the shape of the dataset, data types, missing values, duplicated observations, etc. Then Exploratory data analysis was performed using visualization tools. Then the critical step of the project, modeling, was performed. Finally, data scaling, numerical and categorical data, and encoding were accomplished.

Data was trained with several classification algorithms as the data processing was complete. Trained models were attempted to improve by hyper-parameter tuning. Data was first trained with Logistic Regression, SVM, GaussianNB, BernoulliNB, Decision tree, Random Forest classifier- Nearest neighbor. Fine-tuning of each of those models was performed and evaluated in each step with testing data. The model evaluation mainly used accuracy, F1 score, and whole classification report. In the case of the KNN classifier, various k values were tested, and the best one was selected to train the model.

Exploratory Data Analysis:

Dataset was imported in Python's Jupyter notebook. Initially, the dataset was imported as a panda’s data frame. Here's what the dataset looked like.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Figure 1 : Dataset preview

Dataset has 12 columns and 5110 rows of observations. As soon as the data was loaded, missing and duplicated values were checked. It was found that there were 201 rows missing values in bmi. I have imputed the missing values with the median value of the column.

There is one row with gender value as "other" (not Male and Female). I have removed that row from the data set as a cleanup.

Below is a histogram of numerical variables in the feature set.

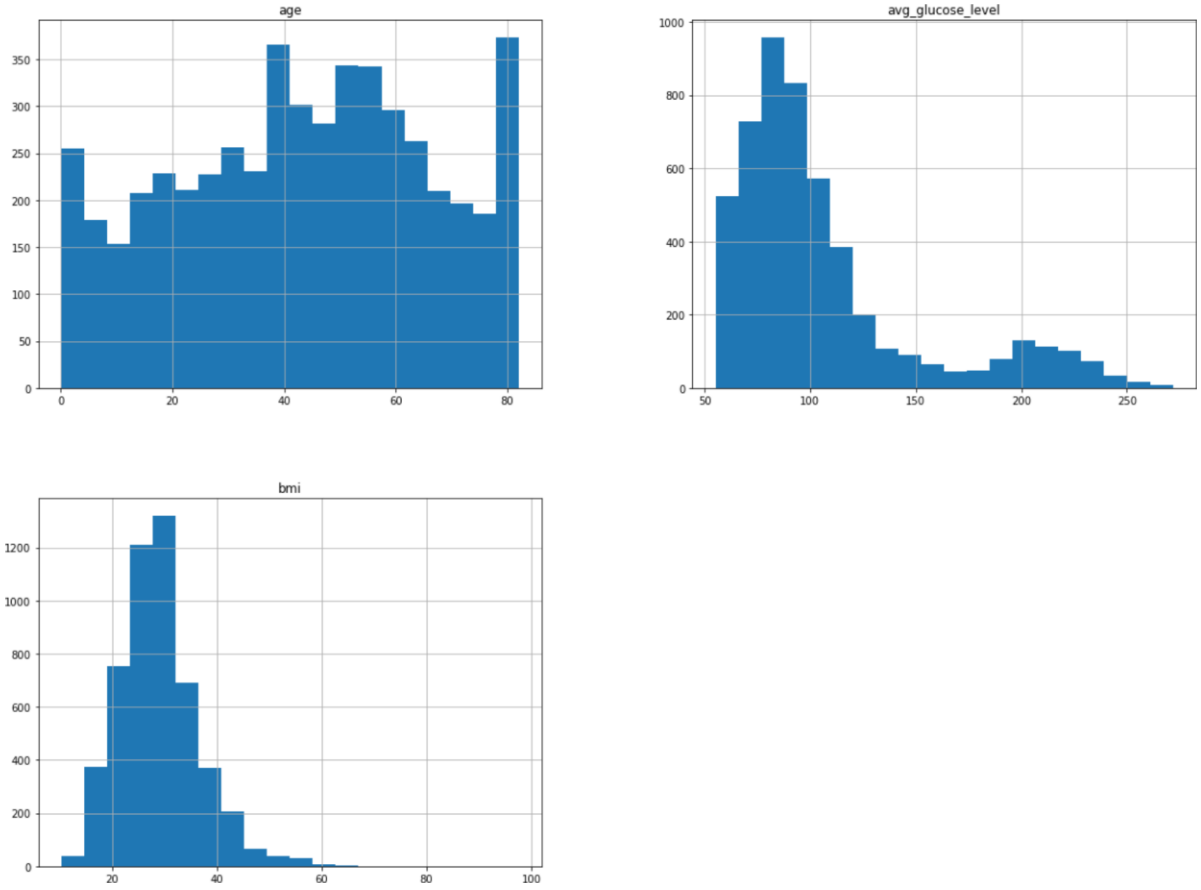


Figure 2: Histograms of Numerical variables

Figure 2 shows that we have a reasonable variation in the age in the input data set. Also, none of the numerical variables have outliers to clean. For example, glucose value and BMI are centered around 90 and 30 weights, respectively. This data behavior is expected as most incoming patients tend to be centered around this mode value as slightly borderline glucose (100 max) and Overweight/obese bmi range of 30.

Conclusion:

Based on the results this predictive model was found out to be a KNN classifier with hyperparameter tuning, which has the highest classification accuracy of 90%. I must admit that predicting a stroke is a very sensitive undertaking for anyone. I am confident that this model is the best model for the given limited features as in this study. But it is not the best for predicting a stroke in general because the model doesn't consider other important risk factors like exercise, alcohol consumption, nutrition, etc.

Limitations:

Logistic regression models are not perfect and have their disadvantages, "Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So, on high dimensional datasets, Regularization techniques should be considered to avoid over-fitting.

Future uses/Additional Applications:

If we can gather more amount of data and train our model, then the accuracy as well as the predictability of the model will improve a lot. This model can be deployed in the healthcare to take better decisions and save many lives.

Recommendation:

This model can predict stroke disease based on 11 features included in this study with 90 percent accuracy. Still, it can be improved further if more data and features have all other important risk factors of Stroke and a more balanced dataset with higher volume for train data. Therefore, I would recommend any stakeholder use this model with caution.

Ethical Assessment:

As this project is related to the health of individuals and as it contains sensitive information, I made sure to not use any PII information such as individual names. Also, the data needs to be presented in accurate form without any modifications or misrepresentations. Also, steps must be taken to avoid any bias towards any gender. The users of this model also need to be aware of the limitations of the model and not use it for self-medications and should seek doctor incase of any discomfort even though their result may suggest otherwise using the model.

References:

1. NHLBI. (March24, 2022). What is a Stroke? National Heart, Lung and Bone Institute. Retrieved December 12, 2022, from https://www.nhlbi.nih.gov/health/stroke
2. Chris Albon. (2018). *Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning*.
3. Stroke Prediction Dataset | Kaggle. *https://www.kaggle.com/fedesoriano/stroke- prediction-dataset*
4. Stroke background information by CDC from *https://www.cdc.gov/stroke/about.htm*

Questions:

1.Why do we need to predict the Risk of Stroke?

2. How does this project helps in predicting Stroke?

3. Does this disease affect more people of color? If yes, how do you know?

4. What parameters are used in the model?

5. What are the different models that are built as part of this project?

6.Are there any new features built from the existing features?

7. Can adding any other features affect the model accuracy?

8. is there any additional data to supplement the dataset used?

9. Why is accuracy a poor model for imbalanced data sets?

10. Can this methodology be used in other areas of medical sciences?