Stroke Prediction Project - CMSE 492

Sanjay Subramanian

Project Overview

Repository Link: https://github.com/sanjay7303/cmse492_project/

The following project is to identify what are the main features that affect whether a patient has a stroke or not. The following features are hypertension, bmi, work_type, smoking_status and many more. Moreover, the idea is to use machine learning to predict which model for what high quality features can produce an accurate prediction.

Project Setup

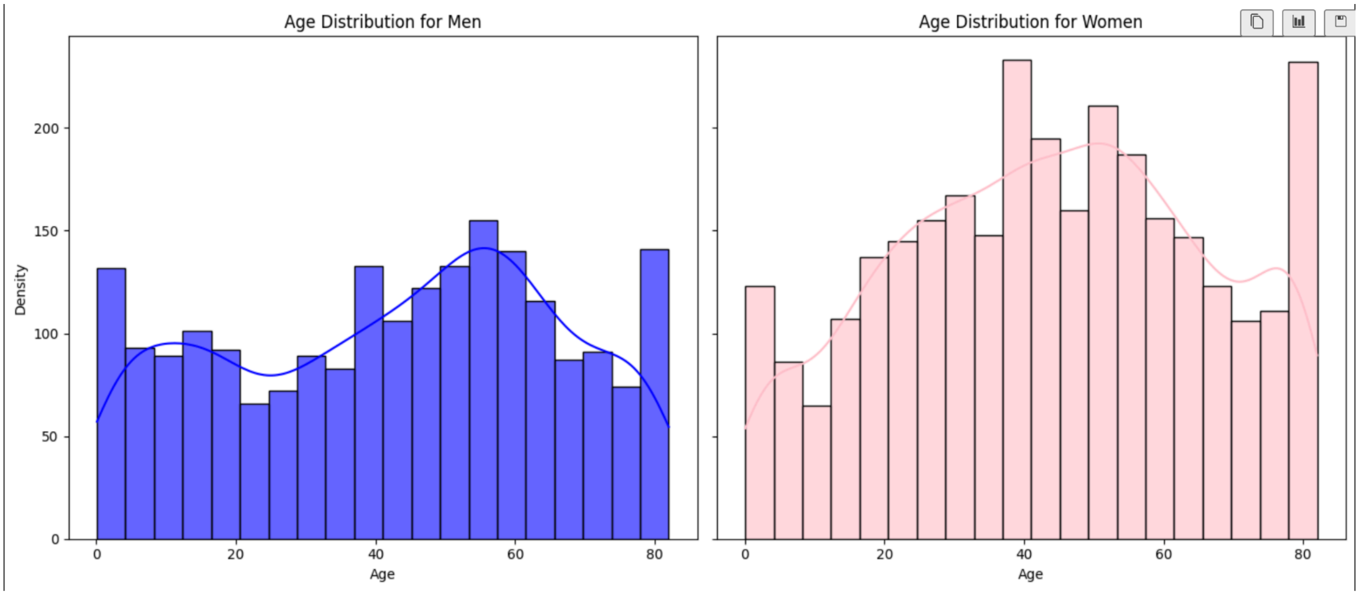
The following project has a repository to track my work and progress as well as my changes. I have a notebooks folder to keep track of my EDA notebook as well as my drafts along the way. There is also a datasets folder that contains the dataset for the project as well as any cleaned versions used for the machine learning model. Reports folder is also present to document the findings along the way and to keep track of what has been done every step of the way. Lastly, there is a results folder in the repository that will contain the results of the machine learning models used and the figures to back up those results.

Complete Process and Machine Learning Approach

The following step by step process for this project is to initially preprocess the data, which will include imputing the missing values in a very reliable way and then remove any columns that are highly not correlated with the target column. Doing this will enhance the accuracy of the machine learning model in general. Furthermore, the next step would be to do an EDA and visualize the data and see the different patterns amongst the various columns. The visualizations would include frequency plots, bar plots, box plots and scatter plots. Lastly, is when we would start the preparation of running the machine learning models. Since there are many categorical variables present in the dataset, it would be wise to perform an one-hot encoding to convert them into numeric and assign a value. The reason for this is to ensure the model is more accurate and most if not all machine learning models only take numeric variables. With regards to the machine learning model it would be wise to start off with a logistic regression since it is binary classification and then move to a more complex model like neural network and so on.

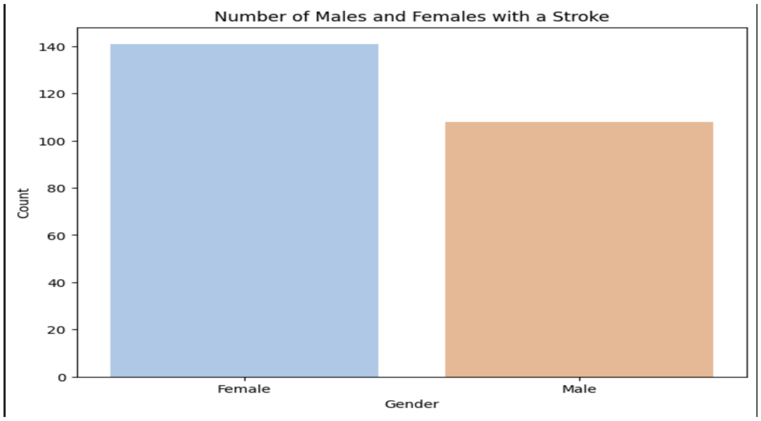
Initial Analysis and Key Findings

For this project the initial analysis for me is to see the difference between the two categories which is male and female regarding the possibility of stroke. My initial thought was to see how many males and how many females are present in the data and what is the age distribution amongst those two genders. The initial plot I created was a distribution bar plot which can be seen below.



From the above image you can see that there is a more even spread of all age groups for men compared to women. There is also a higher number of women of the age greater than 75 than men as well. So from this we can presume that more females will have a stroke compared to males.

The next figure includes the number of males and females in the data that have a stroke



From the above figure we can see that more females have been identified with a stroke than males.

Challenges and Solutions

One of the main challenges is to identify which machine learning model would be ideal and the computational power of each might be robust to handle. So, identifying the trade off between the accuracy and bias will determine what model will be the best choice to go with. Since this is a binary classification, logistic regression would be the baseline model to go with but using the hyperparameter tuning and many more we can explore the possibility of using neural networks. Next, would be imputing the missing values. Using the simple imputer function in python would be the baseline choice, but the accuracy of the final model as a whole would not be that good. So, predicting the missing values would be the best option to replace them. Using a linear model would be ideal for that sort of prediction because of the numeric values.

Next Steps

The following next steps of the project is to predict the missing values in the data preprocessing step and add more visualizations in the EDA part of the project. Visualizing correlations between the columns using scatter plot is something I will be looking to do next. Also, getting started on formulating the machine learning models and using plots to see the accuracy is also one of the next steps. As for a timeline, I would be looking to finish these tasks or at least get a working draft of all of these by the second week of November.

Conclusion

To conclude, I was able to gather the data to get ready for the project and did my initial preprocessing to clean the data. Then I was able to perform a preliminary EDA and was intrigued with the results I got. Reflecting on the progress I have made till now, I am able to see this project has a lot of nuances with regards to the different features and utilizing them to its full potential is one of the lessons I have learned as of now. I will be looking to visualize and explain my findings thoroughly to best convey my results.

References

[Stroke Prediction Dataset | Kaggle]