Stroke Prediction using Healthcare Dataset

Kanthimathinathan Ramasubranian, Ashish Kuzhiveli, Mishal Ashraf, Sagarika Patial

Binghamton University - SUNY

Abstract:

The revolutions that helped shape our lives were agricultural revolution, industrial revolution, information revolution and digital revolution. Humans have been benefitted immensely by all these revolutions however, with it also came sedentary and unhealthy lifestyles which has impacted our wellbeing and health to a great extent. These unhealthy and sedentary lifestyles have led to deaths and a few of the major causes of deaths in the world are heart diseases, cancer, and strokes. In the United States alone 795,000 deaths are caused due to strokes. Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. Stroke is a leading cause of serious long-term disability. Stroke reduces mobility in more than half of stroke survivors aged 65 and older [2]. As part of this project, we intend to help the patient, medical and healthcare community by providing them a tool that can help predict strokes in people with risk factors such as heart disease, obesity, hypertension, diabetes, and smokers. We were able to do so with the help of the healthcare dataset that was processed with machine learning algorithms that provide foresight for patients who are at a higher risk for stroke.

Keywords: Python, Jupyter Notebook, Stroke, Logistic Regression, SVM, MLP

Introduction

Stroke is one of the main causes of mortality and disability across the world. In 2020 alone, 1 in 6 deaths from cardiovascular disease was due to stroke [1]. Prior to the diagnosis of a stroke for significant periods, Doctors were unable to obtain a comprehensive overview of the disease, which resulted in major deaths and the abrupt cause of the stroke [3].

In order to predict strokes in patients with high risk factors of smoking, heart disease, obesity, hypertension and diabetes, we applied various machine learning algorithms with the help of python programming language and were able to visualize and collaborate on this project with the help of jupyter notebook.

The dataset used for this project is from Kaggle and includes comprehensive information such as gender, age, heart disease, hypertension, avg glucose level, BMI, and smoking status of about 44400 people. This dataset comprises of people who have had a stroke and of

people who haven't. It also has various data points which helped provide predictive outputs of whether a person was likely to have a stroke or not.

Purpose of the Project:

- Provide the medical and healthcare community with a tool that would help early prediction of stroke in a person with sedentary habits and heart diseases.
- Early Prediction analysis which would help healthcare providers give recommendations to people with a positive analysis on how to change their lifestyle to avoid a stroke.
- It also helps propagate the use of machine learning tools in the field of medicine.

Statistics and Labelling of Data:

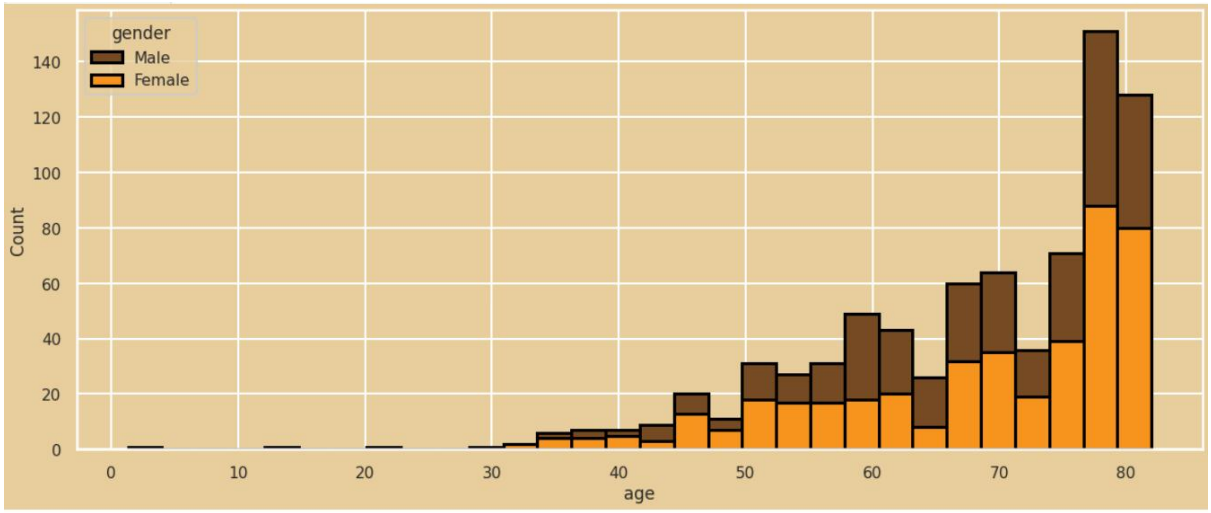
The dataset comprises of people that are split into genders of male, female, and others. It consists of age groups of children, adults and senior members. It includes the following attributes of people which include such as hypertension, heart disease, ever married, work type, Residence type, avg glucose level, BMI and smoking status and also whether the person has suffered a stroke or not. This helps the algorithm learn to predict outcomes with accuracy.

Description of the datatypes of the attributes of columns from the dataset:

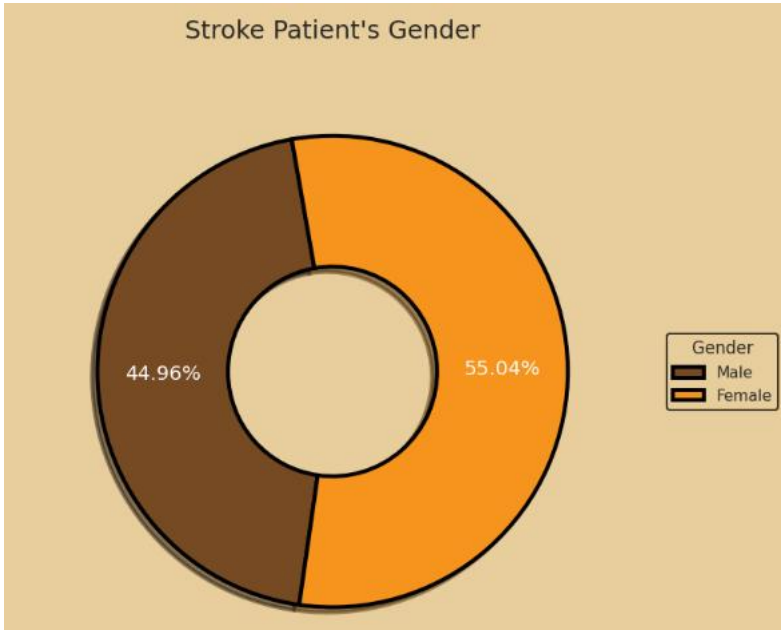
```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43389 entries, 0 to 43388
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 43389 non-null  int64
 1   gender             43389 non-null  object
 2   age                43389 non-null  float64
 3   hypertension       43389 non-null  int64
 4   heart_disease      43389 non-null  int64
 5   ever_married       43389 non-null  object
 6   work_type          43389 non-null  object
 7   Residence_type     43389 non-null  object
 8   avg_glucose_level  43389 non-null  float64
 9   bmi                41931 non-null  float64
 10  smoking_status     30099 non-null  object
 11  stroke             43389 non-null  int64
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

Data Visualization:

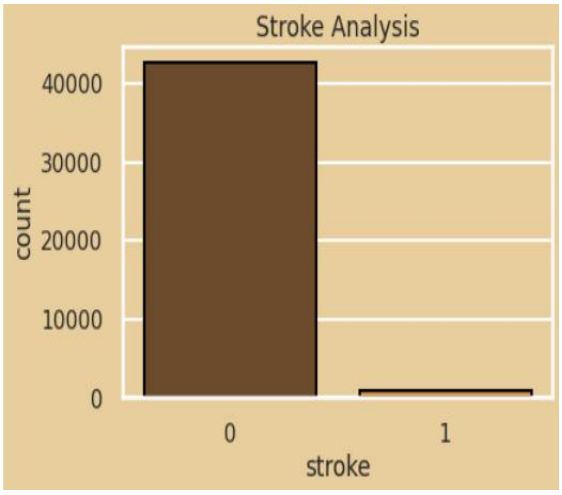
We can visualize the provided dataset with the help of importing the seaborn library and its various plots. For instance, below is a histogram of men and women of different age groups who have had strokes.

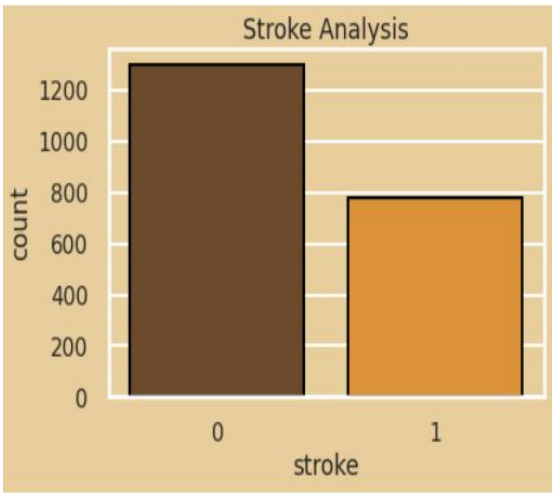


Below is a pie chart of the percentage of men and women who have had a stroke:

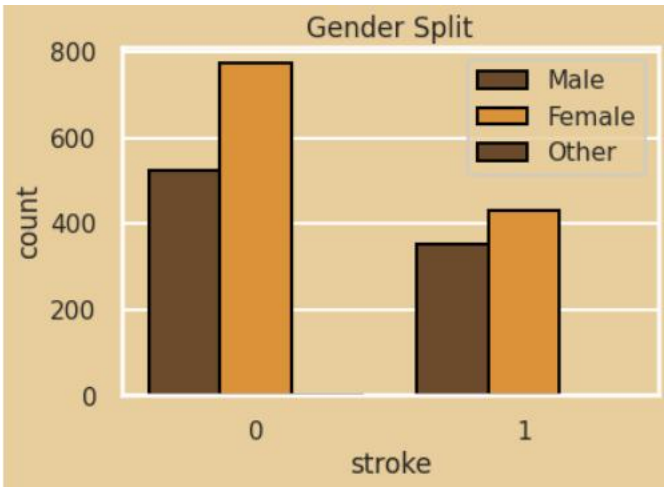


Since the data of people who have had strokes were significantly lower than those that didn't have strokes in the data set of 44k, it would not have been possible to create a model with it as that data would heave been considered as noise and hence, we have reduced it to develop a model with the following sample set where the number of people who had had strokes is almost above 50% of people who have not had strokes. A comparison of their plots has been provided below:

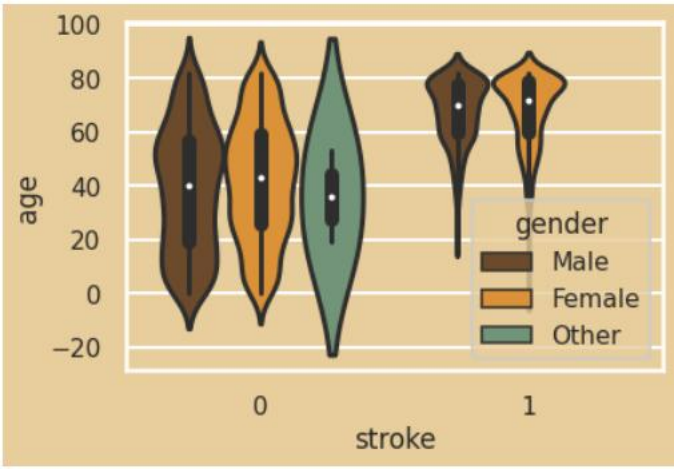




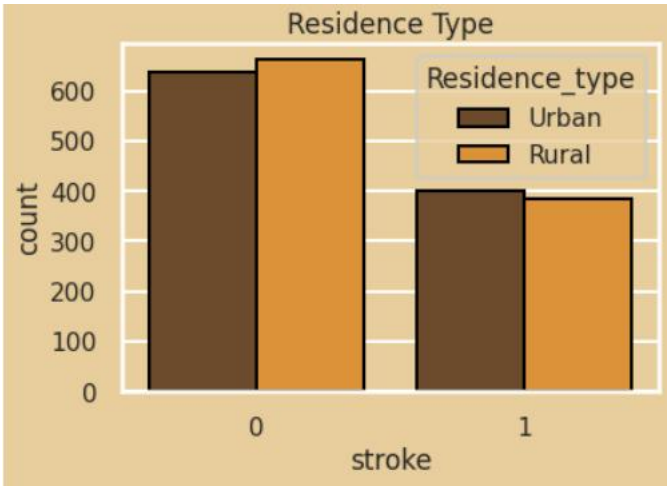
A bar plot of men and women who have had a stroke against those who haven't had a stroke.



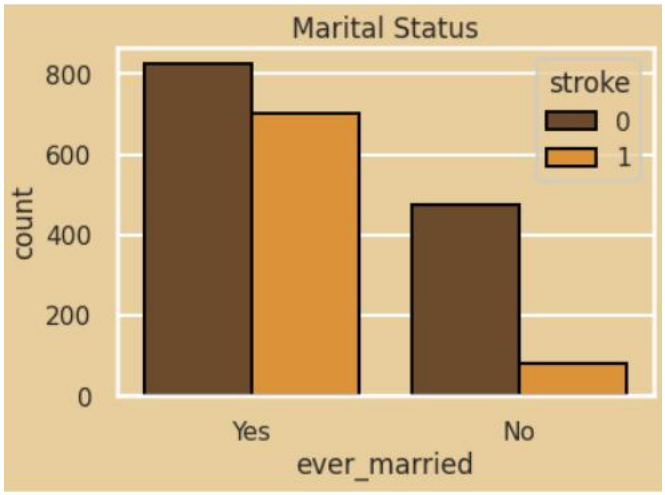
A violin plot depicting age groups of genders who have had a stroke against those who haven't had a stroke.



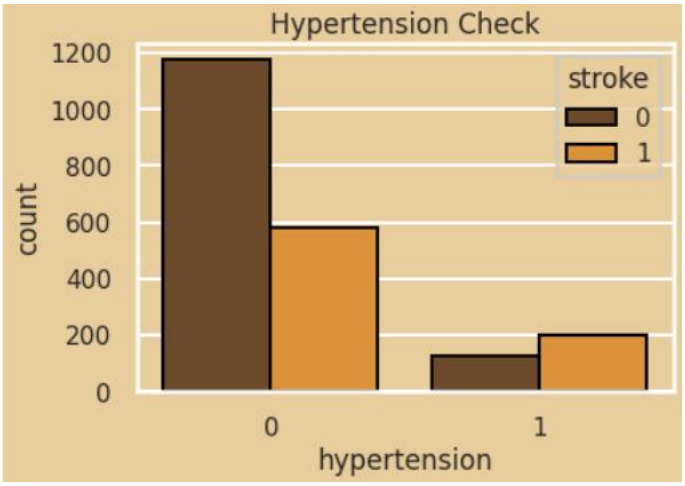
A bar plot which represents the number of people residing in rural and urban areas who have had a stroke against people who haven't had a stroke.



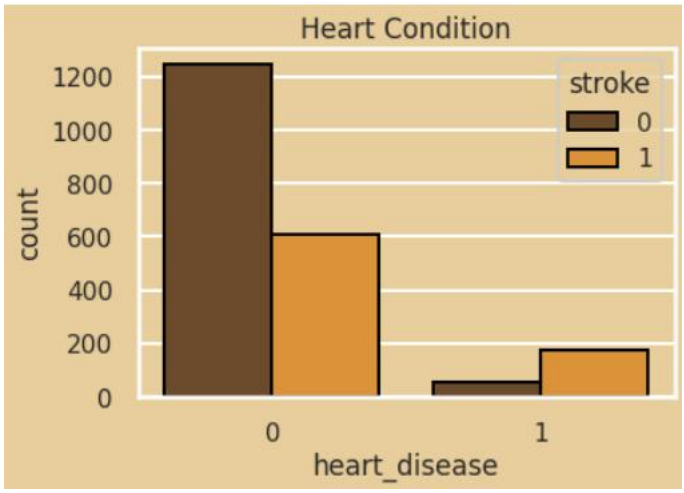
Bar plot illustrating the difference in strokes between the number of people who answered in the affirmative and negative to marriage.



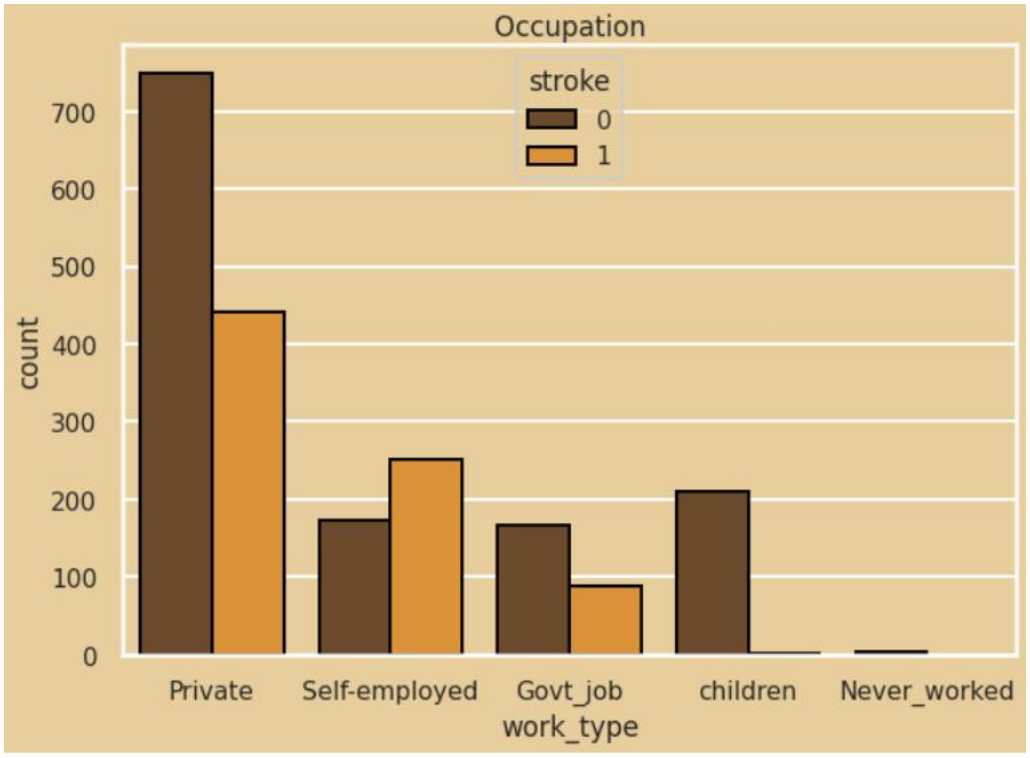
Bar plot portraying the relationship between hypertension and stroke.



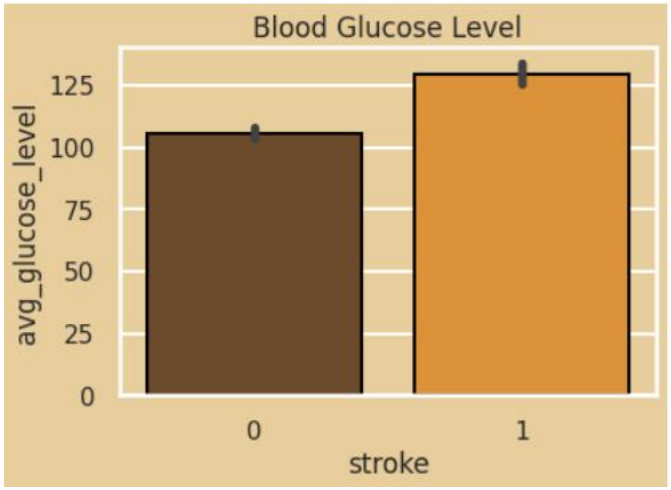
Bar plot displaying the relationship between heart disease and stroke.



Bar plot illustrating the relationship between occupation and stroke.



Bar plot showing the relationship between strokes and people who have diabetes or higher level of sugar content in their blood against people who don't.



Preprocessing of Data:

As part of preprocessing of data, BMI and Smoking status consisted of 179 and 571 records of null values. This data needed to be replaced to avoid inaccuracies on the prediction of whether a person would suffer from a stroke or not. The data was replaced with average values of the dataset to provide a more accurate picture of stroke prediction.

Genders were categorized as well to help with the normalizing the data. Residence type was also classified as 1 and 0 for Urban and Rural.

Occupation or work types which consisted of Government Job, never worked, Private, Self-employed, and groups of children were also categorized along with the different groups to represent their status in binary format.

Smoking Status was also normalized with the help of get dummies function to include categories of people who formerly smoked, smokes and never smoked.

Logistic regression Model for Prediction of Stroke:

Regression analysis is a commonly used correlation analysis method. By choosing the appropriate regression model, we can get more accurate quantitative correlation among the parameters[4]. In this case we are using a binary logistic regression classifier to predict whether the person would have a stroke or not depending on various risk factors.

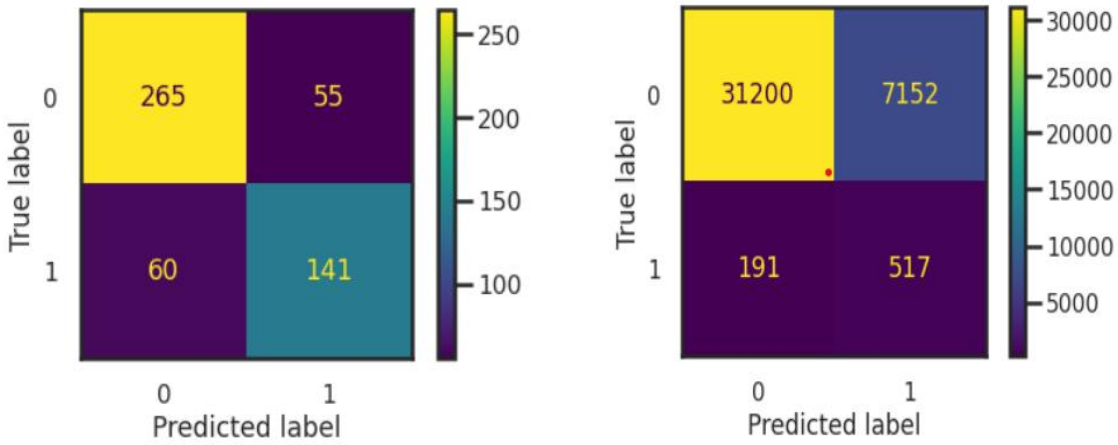
$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X)$$

'hΘ(x)' *is output of logistic function , where 0 ≤ hΘ(x) ≥ 1*

*'**β1**' is the slope*
*'**βo**' is the y-intercept*
***'X'** is the independent variable*

*(βo + β1*x) - derived from equation of a line Y(predicted) = (βo + β1*x) + Error value*

The confusion matrix that was created with the logistic regression model that had a smaller subset against the entire dataset is provided below



From the above we can comprehend that for the smaller subset used to train the model, it's True Negatives were at 265, False positives at 55, False Negatives at 60 and True Positive at 141.

When the model was run with the entire dataset it's True Negatives were at 31200, False positives at 7152, False Negatives at 191 and True Positive at 517.

SVM Model for Prediction of Stroke:

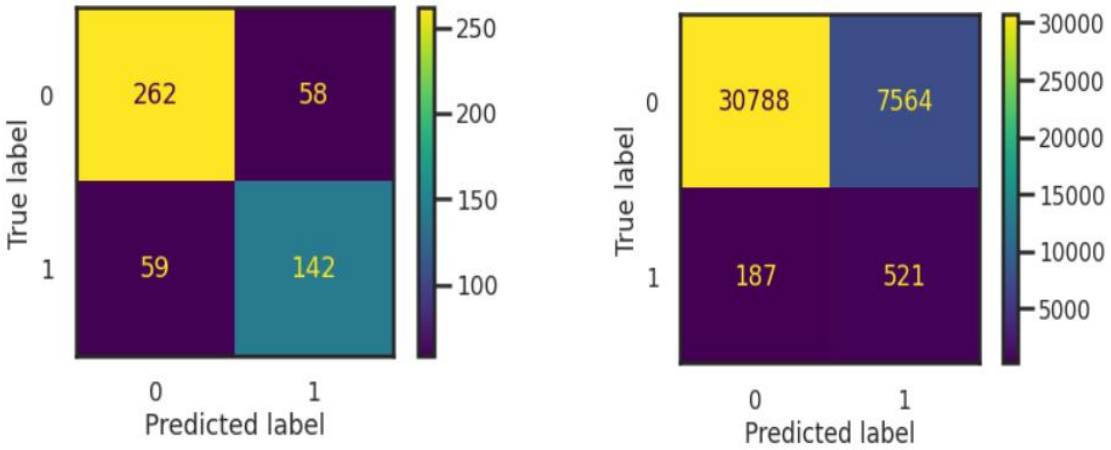
A **Support Vector Machine** or **SVM** is a machine learning algorithm that looks at data and sorts it into one of two categories. The goal of the algorithm behind SVM is to maximize the minimum distance. The product of a predicted and actual label would be greater than 0 (zero) on correct prediction, otherwise less than zero.

$$y_n[w^T\phi(x) + b] = \begin{cases} \geq 0 \; if \; correct \\ < 0 \; if \; incorrect \end{cases}$$

For perfectly separable datasets, the optimal hyperplane classifies all the points correctly, further substituting the optimal values in the weight equation.

$$w^* = arg_w max \left[min_n \frac{|w^T(\phi(x_n)) + b|}{||w||_2} \right] = arg_w max \left[min_n \frac{y_n|w^T(\phi(x_n)) + b|}{||w||_2} \right] \quad \because \; perfect \; seperation$$

The confusion matrix that was created with the SVM model that had a smaller subset against the entire dataset is provided below



From the above we can comprehend that for the smaller subset used to train the model, it's True Negatives were at 262, False positives at 58, False Negatives at 59 and True Positive at 142.

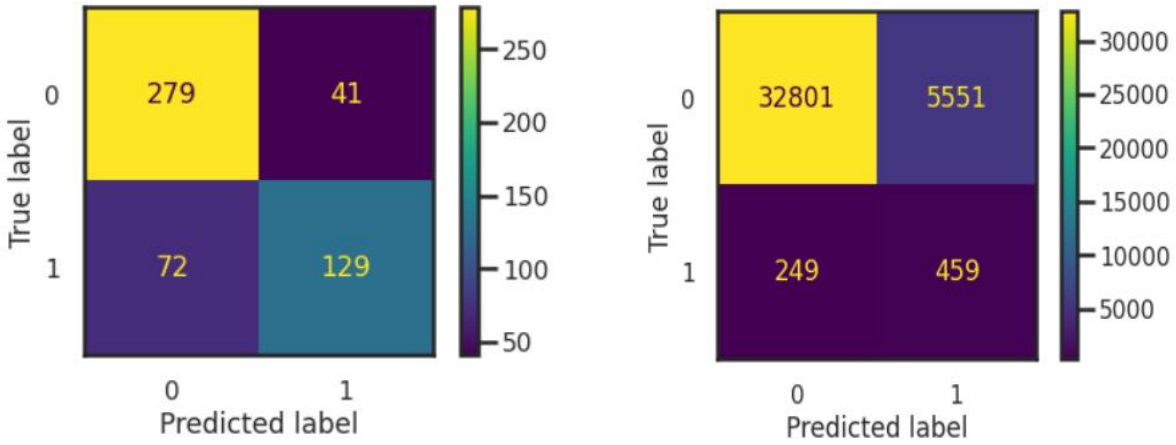
When the model was run with the entire dataset it's True Negatives were at 30788, False positives at 7564, False Negatives at 187 and True Positive at 521.

MLP classified for Prediction of Stroke:

This model optimizes the log-loss function using LBFGS or stochastic gradient descent. MLPClassifier trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters.

It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.

The confusion matrix that was created with the MLP model that had a smaller subset against the entire dataset is provided below

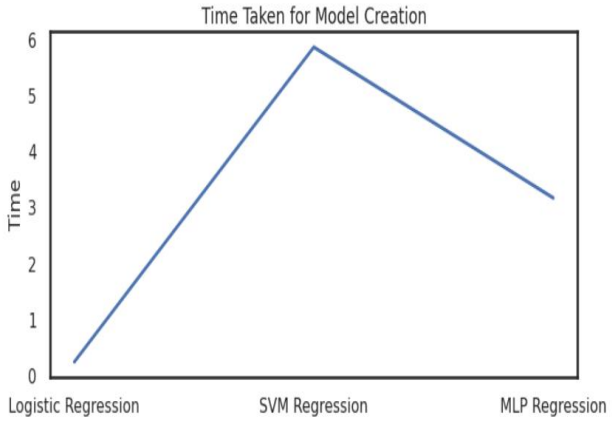


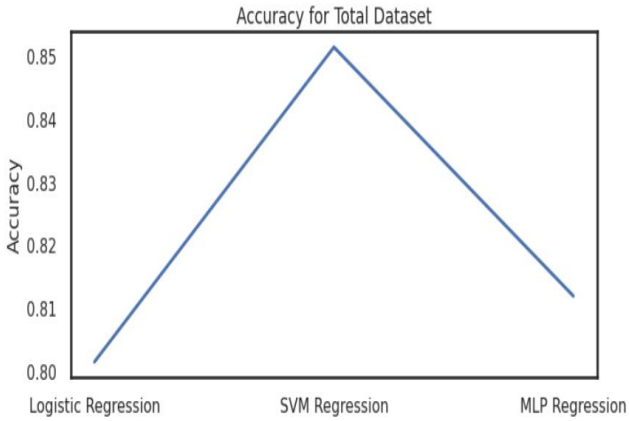
From the above we can comprehend that for the smaller subset used to train the model, it's True Negatives were at 279, False positives at 41, False Negatives at 72 and True Positive at 129.

When the model was run with the entire dataset it's True Negatives were at 32801, False positives at 5551, False Negatives at 249 and True Positive at 459.

Comparison of Time taken and Accuracy of the model:

Time taken by the Logistic Regression model was the lowest whereas the time taken by the SVM model was the highest and MLP was in between. However, SVM had a higher accuracy rate on the dataset when compared to both the MLP classifier model and Logistic Regression Model with LR being the lowest in terms of accuracy. Please find the results obtained below





Conclusion:

Stroke is one of the leading causes of deaths in America. However, the chances of survival increase drastically if early action is taken. Our goal with this project is to help identify people who are at risk of suffering from stoke based on their lifestyle and habits. Based on the results of the project, we were able to successfully implement various classification models and compare their performance (accuracy and time-taken). We found that the logistic regression model had the best all-around performance by having the lowest processing time and highest overall accuracy compared to SVM and MLP classification models.

References:

1. Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.
2. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. Circulation. 2022;145(8):e153–e639.
3. James F. Meschia, Cheryl Bushnell and Bernadette Boden-Albala, "Guidelines for the Primary Prevention of Stroke: A Statement for Healthcare Professionals from the American Heart Association/American Stroke Association", *American Heart Association*, vol. 45, pp. 00-00, 2014.
4. X. M Ding, L.H. Zhang, Y.D. Zhao, X.J. Zhang, G.Z. Bao and J.C. Ma, "Application of Logistic regression analysis in evaluation of experimental teaching effect and Its Realization on SPSS 19.0[J]", Heilongjiang Animal Husbandry and Veterinary Medicine, no. 9, pp. 261-265, 2017.