Paula Ndirangu

thugurip@gmail.com

Abstract

I am doing the data analysis so show the how many people with strokes depends of health and socialeconomic factors.

DATA ANALYSIS PROJECT

STROKE

ABSTRACT

Stroke occurs when the blood flow in the brain is blocked, therefore; the cells of the brain aren’t receiving the nutrients and oxygen to circulate the blood. It causes human health and financial problems for both patients and health care systems. One of the important risk factors for stroke is the health-related behaviour i.e. obesity and chronic illnesses which should be managed to prevent getting the stroke. Many machines learning has predicted the risk of stroke using data collected on radiological imaging, health factors and changing lifestyle factors. There are ways to collect and analyze data by doing Exploratory Data Analysis, Preprocessing, Statistical Analysis and Machine Learning which relates to predictive analysis. For Exploratory Data Analysis, I did crosstabulation, group by, data visualization for analyze the data. I did data preprocessing to clean the data, fill in the missing values and using capitalization. For statistical analysis, I did A/B hypothesis testing using Z and T score. Finally, I did Machine Learning which are following models: Logistic Regression, Decision Tree Regression and Random Forest Classification. The accuracy percentage of the models used in this investigation is significantly higher than the models that are used for investigation.

Chapter 1

INTRODUCTION

Stroke occurs when the blood flow in the brain is blocked, therefore; the cells of the brain aren’t receiving the nutrients and oxygen to circulate the blood. A stroke is a medical emergency that requires urgent medical attention (Tazin et al., 2021). Early detection and management are required to prevent further damage of the brain and other parts of the body. Strokes are divided in two parts: ischemic or hemorrhagic (Tazin et al., 2021).

Stroke is the a noncommunicable disease that kills approximately 11% of the population. According to World Health Organization, 15 million people suffer from strokes each year, with one person dying every 4 to 5 minutes in the affected population. (Tazin et al., 2021). According to Centers for Disease Control and Prevention (CDC), stroke is the sixth leading cause of mortality in the U.S. There are 795,000 people suffering from the effects of strokes on a regular basis. Stroke victims can experience paralysis, impaired speed or loss of vision. Stroke factors can be modified by family history of cerebrovascular disease, age, gender, race that account of 60% to 80% of people who have stroke risk (Chun et al., 2021).

Stroke can be avoided by having healthier and balanced diet, avoiding unhealthy behaviour such as binge eating and excessive alcohol drinking, maintaining normal body mass index (BMI) and average glucose level so that one maintains a healthy heart and kidney function (Tazin et al., 2021). American Heart Association recommends the assessment and tests for patients who have risk of developing blockage in their arteries, which lead into the heart attack or stroke. The growth of arterial blockages and damage to blood vessels can lead to stroke. The growth of arterial blockages and damage to blood vessels can lead to stroke which has risk factors. (Chun et al., 2021)

The purpose of this research is to create the predicting Stroke outcome using both Data Science and Machine Learning. Data Science have been studied previously as a tool for predictions about Stroke. One of them is the development of a hybrid machine learning approach for cerebral stroke prediction (Tavares 2021). Machine Learning (ML) techniques have increased in the recent years with the numerous healthcare applications. The ML models have not been adopted into the practice for stroke risk in some countries (Chun et al., 2021). The model uses the data collected during the assessment and tests for patients who have risk of developing blockage in their arteries. Patient records contain predictive factors such as patient demographic, lifestyle and existing medical condition which lead to stroke (Chun et al., 2021). Several studies have been conducted on stroke occurrences using predictive and statistical analysis. One such study had samples from 35 urine tests and 33 blood, serum and plasma tests. (Alanazi1 et al., 2021). Several studies have shown that health habits and lifestyles are important factors which help in avoiding cardiovascular diseases.

This approach can save lives and reduce the economic strain of health care services. Many machine learning models have been built to predict the risk of stroke or to automatically diagnose stroke, using predictors such as lifestyle factors or radiological imaging. However, there have been no models built using data from lab tests. (Alanazi1 et al., 2021).

Chapter 2

LITERATURE REVIEW

Femke Kremers developed machine learning prediction models for outcome of patients with acute ischemic stroke who were undergoing endovascular treatment. The purpose of the models was to improve patient management (Kremers et al., 2021). He developed models that estimate functional outcome (mRS) in patients with anterior circulation acute ischemic stroke eligible for EVT within 6.5 hours of onset. Predictive performance was evaluated. 19 models were included in this validation. Variables included in the models mainly addressed clinical and imaging characteristics at baseline. The study concluded that the model was superior in predicting functional outcome for patients with ischemic stroke after endovascular treatment within 6.5 hours.

According to Marisha Sanjay Ferme, Machine Learning (ML) delivers an accurate and quick prediction outcome and it has become a powerful tool in health settings, offering personalized clinical care for stroke patients. The aim of their work was to classify the state-of-arts on ML techniques for brain stroke into 4 categories based on their functionalities or similarity. An application of ML and Deep Learning in health care is growing. CT images are a frequently used dataset in stroke. The study showcases the contribution of various ML approaches applied to brain stroke (Sirsat, 2020).

In the study above (Sirsat, 2020, the study defined the difference between supervised learning and unsupervised learning. Supervising learning involves classification, logistics regression and naïve bayes classifier. Unsupervised learning involves groups of observations of making clusters. The process of this unsupervised learning is called clustering. Deep learning is the model with multi-processing from the raw material to processed data (Sirsat, 2020).

Tavares (2021) in his study identified stroke as the heterogenous group of disorders which is characterized as disruption of food supply (Tavares, 2021). The symptoms last for more than 24 hours. Ischemic strokes are about 85% of the cases. He also mentioned about the blooding of the weakened blood vessel. He noticed that the risk factors are related to age, gender, race and ethnicity. If someone have many cardiac diseases, obesity, poor nutrition, tobacco use and alcohol consumption, then the chances of getting of acute stroke is much higher than those living a healthier lifestyle. (Tavares, 2021)

Chapter 3

RESEARCH METHODOLOGY

Methodology

My step was to import the stroke dataset. I used seven research websites. Some of them were from National Institute of Health. The dates were mostly recently in 2020 and 2021(Kremers et al., 2021).The research was well-constructed with the references available from the authors. The research articles included: i) Stroke Magazine which showed the outcomes of patients with ischemic stroke understanding endovascular treatment, ii) “Stroke risk prediction using Machine Learning which was written by Matthew Chun and Robust Clarke. First two articles were from the PDF document and both of them were literature review. (Chun et al., 2021)

Data Collection

Data collection started when I got the data from Kaggle then I downloaded it. Then I did some changes such as minor cleaning but otherwise the dataset was very presentable. I import the dataset into DataFrame called Pandas. Then I explored the first five rows in the dataset. The dataset had about 10 metrics with the total of 43,400 patients. These metrics included the patients ’demographics (gender, age, marital status, type of work and residence type) and health records (hypertension, heart disease, average glucose level, Body Mass Index (BMI), smoking status and stroke). There were about 783 patients suffered a stroke while about 42,617 patients didn’t had stroke. (Yap, 2020)

Data Visualization

The stroke prediction dataset was used to perform the study. According to the data, there was about 5110 rows and 12 columns in the dataset. The value of the output column stroke was either 1 or 0. The probability of 0 in the output column(stroke) exceeds the possibility of 1 in the column in the dataset. (Tazin et al., 2021)There were 249 rows alone in the stroke column had value of 0. To improve accuracy, data preprocessing had to be used to show total number of people with stroke in the output column. (Tazin et al., 2021)

I started with matplotlib where you plot the scatter, histogram, line chart and bar chart. Each of the data visualization libraries showed the representation of the data of people with strokes. The data was shown was the BMI over the age of the people. BMI was very important because it showed the body mass combining the weight and the height. (Tazin et al., 2021) If someone had greater BMI then the person was overweight or obese. Seaborn was another common data visualization library because countplot and catplot represented the data using the graph of the gender over the average glucose level. Boxplot was very important because I compared variation between the data sets evaluated. It displayed the range and distribution of the data along a number line. It displayed the location of the variation and provided the indication of symmetry and skewness of the data. (Tazin et al., 2021)

Data Preprocessing

Data preprocessing was required to remove unnecessary noise and outliers from the dataset. The stage prevents the models from functioning accordingly to that the attributes looks more uniform. The dataset had to be cleaned and prepared for model development. The id had to be removed since it was not necessary for model construction. Null values had to be filled so that the values to be complete and accurate. (Lin CH;Hsu KC;Johnson KR;Fann YC;Tsai CH;Sun Y;Lien LM;Chang WL;Chen PL;Lin CL;Hsu CY; ;, 2020)

I did label encoding because it converted from string to integer. All strings were encoded during label encoding in order to transform into numbers. The dataset used for stroke prediction was not balanced. In the data quality segment, the data had to fit for the uses in operations and decisions. The data represented the real-world settings so that the data had to be corrected in order to give the insights of the data collected. (Lin CH;Hsu KC;Johnson KR;Fann YC;Tsai CH;Sun Y;Lien LM;Chang WL;Chen PL;Lin CL;Hsu CY; ;, 2020)

I normalized 3 numerical attributes which was age, average and BMI. The classifier had ensured to have equal weight over the data. I used both SelectKBest and Chi2 in the Chi-Square in data. SelectKBest had the hyperparameter of number of features we wanted to select. (Lin CH;Hsu KC;Johnson KR;Fann YC;Tsai CH;Sun Y;Lien LM;Chang WL;Chen PL;Lin CL;Hsu CY; ;, 2020)Variance threshold was important because it included in the data as it needed to meet the threshold.

Statistical Analysis

Statistical analysis is a method of gathering data, exploring it and then representing a vast volume of data in order to examine trends and patterns in the data. I had used A/B testing which was hypothesis testing. It was an analytical method for making decisions that estimates population based on sample statistics. The population had the data collection and made generalized the population by using inferential methods. (Lin CH;Hsu KC;Johnson KR;Fann YC;Tsai CH;Sun Y;Lien LM;Chang WL;Chen PL;Lin CL;Hsu CY; ;, 2020) Inferential statistics used to describe to validate or invalidate and made conclusion. A hypothesis was the proposed explanation based on limited evidence for further investigation. There were two types of hypotheses: Null Hypothesis and Alternative Hypothesis. (Lin CH;Hsu KC;Johnson KR;Fann YC;Tsai CH;Sun Y;Lien LM;Chang WL;Chen PL;Lin CL;Hsu CY; ;, 2020)

Machine Learning

Most common diseases identified in the data was stroke. There were about three machine learning methods that predict the brain stroke occurrence. There were Random Forest, Decision Forest and Logistic Regression. (Tazin et al., 2021)

Random Forest was the classification algorithm chosen was the Random Forest classification. Random Forest were composed of numerous independent decision trees which were trained individually on a random sample of data. (Tazin et al., 2021) These trees were created during training and the decision trees outputs were collected. Each decision trees had two output classes. The final prediction was determined by the Random Forest which had most votes. (Tazin et al., 2021)

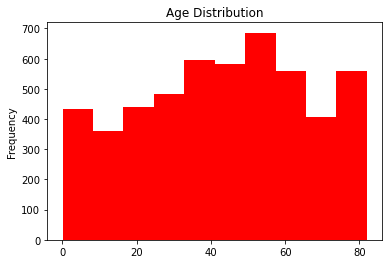
Decision Tree was when both regression and classification concerns were addressed using classification concerns were addressed using classification wit Decision Tree. Furthermore, input variables had a related output variable which was supervised learning model. The data was very segmented into the specific parameter of the method. The decision node and the leaf node were parts of the decision tree. (Tazin et al., 2021)

Logistic Regression was the flowchart for the logistic regression model. In the supervised learning approach, Logistic Regression was most commonly used Machine Learning algorithms. It was a forecasting method that used a collection of independent factors to predict a categorical dependent variable.

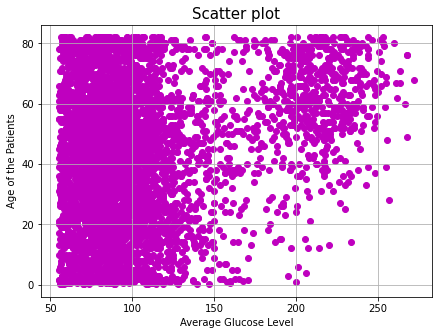
Chapter 4

DATA ANALYSIS(Results)

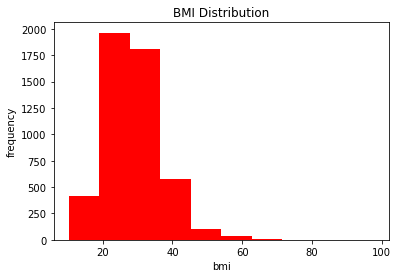
Exploratory Data Analysis



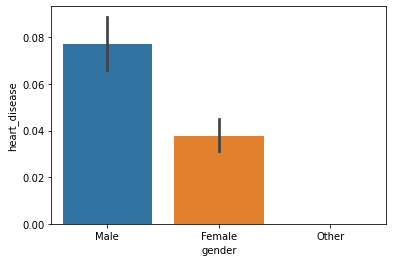
The results for importing the stroke came as a table for database. I did the information for the stroke data and it showed columns, count and data types. The description tab showed the count, mean, standard deviation, min, max and quartiles. I checked the data types of each of the columns which were integer, float and object. I started with the histogram and the results came as the age with most frequency is about 40 to 60 years. I did the scatter plot which was uniformed from 0 to 200 then the dots were scattered.



The BMI ranges which have most frequency is between 20 and 40 BMI which was normal, overweight and obese. The highest frequency is about 2000 and the lowest is about 0.5.



I did the barplot of the male and female over the rate of the heart disease. Men has more rate of heart disease than women because of the change of the lifestyle habits and social life.



I gave the group by in hypertension over age using the mean. Average age with hypertension is higher than age without hypertension. Older population has higher changes of getting high blood pressure.

hypertension

0 41.173027

1 62.244980

Name: age, dtype: float64

I did the crosstubulation on the difference between the stroke and the heart disease. More people have heart disease without the stroke than people with the stroke.

heart\_disease 0 1

stroke

0 4632 229

1 202 47

I moved on to data preprocessing. Data Preprocessing is very important to complete inconsistencies, check the outliers, check the extra spaces, fill in the missing values and drop the unnecessary columns such as “id”. I want to check if the loc is accurate according to the mean. I inserted the lstrip and it didn’t change and remained construct. I used the lambda method to put the upper string and the results came as attributes turned into the capital letters.

I tried using the bfill for the data to fill the missing values for the bmi. Unfortunately, it changed back to it was started like brought back the missing values. I used the best method by using fillna and the mode. I printed the stke data over bmi in data types, nunique and value counts. The results came when the missing value of bmi have disappointed and turned into 28.7 bmi. I also removed the ‘id’ because it doesn’t exist of showing data for people with stroke.

I tried to remove the outliers using mean and standard deviation and used filter in pos 1. Unfortunately, there was an error in the output of the data. I duplicated the data using the duplicated and sum and the result came as 0. I dropped the duplicates by cleaning as the matrix turned as: ((5110, 11), (5110, 11)).

I continued with the categorial labeling by selecting the object because categorial variables works with the encoding. I moved on with Categorial Labeling by importing sklearn.preprocessing to that the data would be active. I used the dummies to check the gender, work type and smoking status. I filled in the smoking status so they had values for people who smoking from range from 0-4.

2 1892

0 1544

1 885

3 789

Name: smoking\_status, dtype: int64

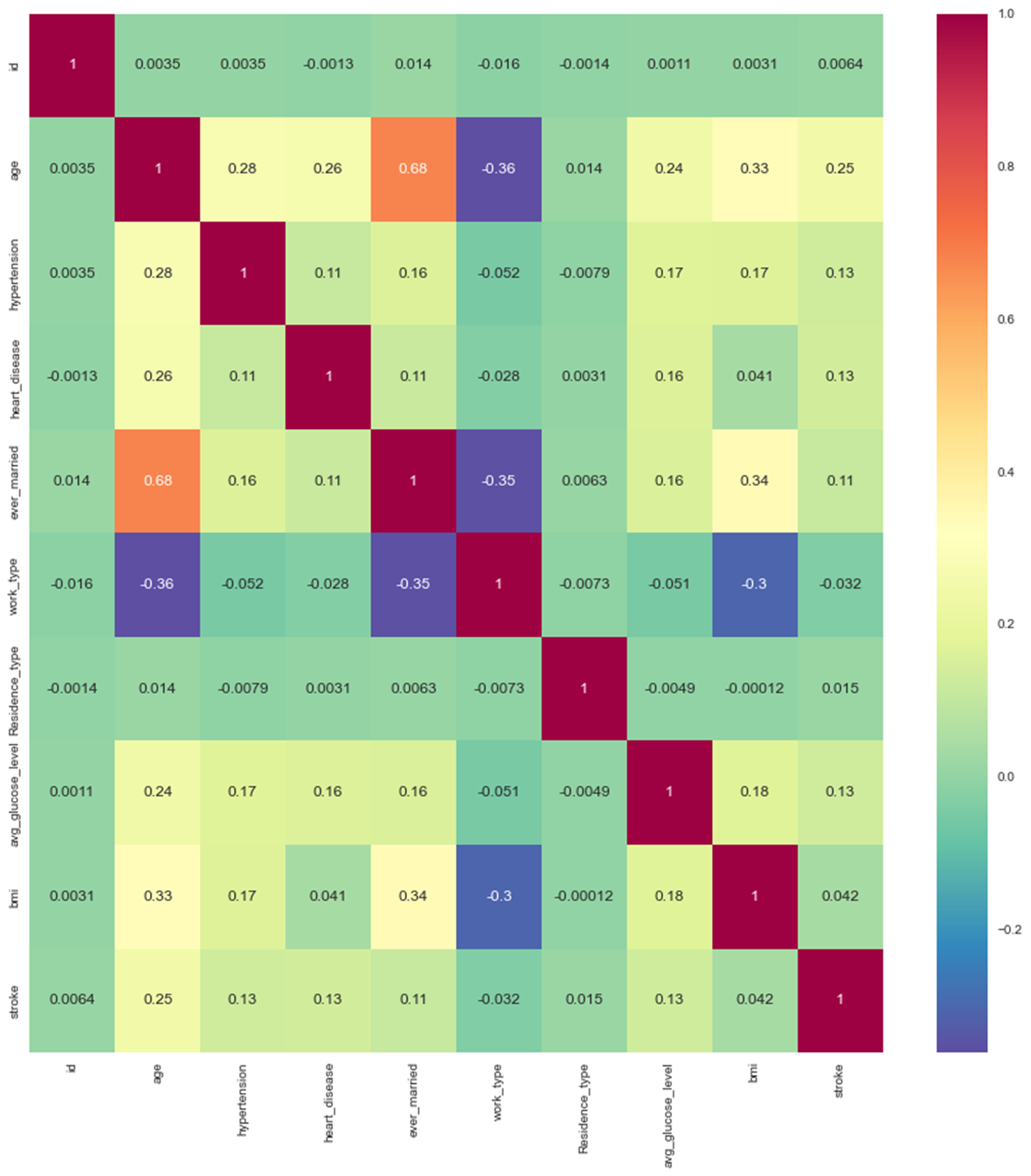
I used the feature scaling which contains the MinMaxScaler, StandardScaler and RobustScaler. I converted MinScaler into the DataFrame by transforming data and the results came to floats with decimal numbers. Same procedure as the StandardScaler and RobustScaler. I calculated the quartile by upper\_q, lower\_q and IQR.

0.0

0.0

0.0

Feature Selection is very important because it structures the data into the Machine Learning. I used from sklearn.feature\_selection import SelectKBest and from sklearn.feature\_selection import chi2 to Chi-Square and K Best. Before I started the feature selection. I did the Label Encoder and inserted the stats in three independent variables: ‘ever\_married’,’work\_type’ and ‘Residence\_type’. I did correlation which is the measurement of the strength between two variables. The result came there is the duplicates between two variables of the stroke data. One square represented the float of the variable and it can be horizontal or vertically.



Statistical Analysis

I used the method of scipy stats to test the sample from “ttest\_1samp”. I calculate the mean of the average glucose level. I did the A/B testing for the bmi over the population mean of both 32 and 39 and the results came as:

t\_score: -28.919830

pvalue: 0.000000

level of significance: 0.050000

tscore: 13.361391

pvalue: 0.000000

level of significance: 0.050000

I did the boxplot again to show the relationship of the age in range. The age range which had most range is between 43 years. I did Chi-Square by writing the chi2\_contingency. Before I proceed to the chi-square, I did the crosstubulation to know many males and females have hypertension.

gender Female Male Other

row\_0

hypertension 2994 2115 1

I did the chi-2 contingency of the p-value with the correction with was False. The alpha remained as 0.05.

level of significance-0.05,p\_value-0.00,chi\_2-1.00

I did the OLS Regression which had the results of statistics models in ols and the formula was scattered. It took long to write the formula because I need to pay much attention to details.

Dep. Variable: age R-squared: 0.692

Model: OLS Adj. R-squared: 0.663

Method: Least Squares F-statistic: 23.43

Date: Mon, 02 May 2022 Prob (F-statistic): 0.00

Time: 15:49:57 Log-Likelihood: -19368.

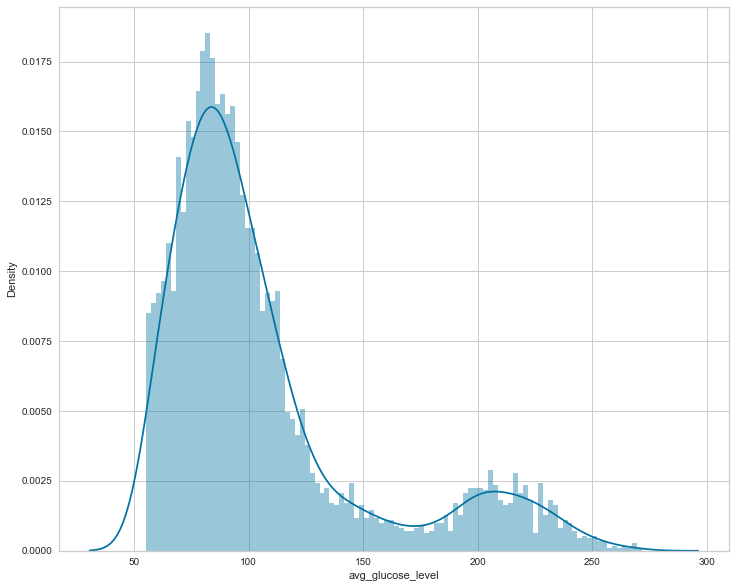
No. Observations: 4909 AIC: 3.960e+04

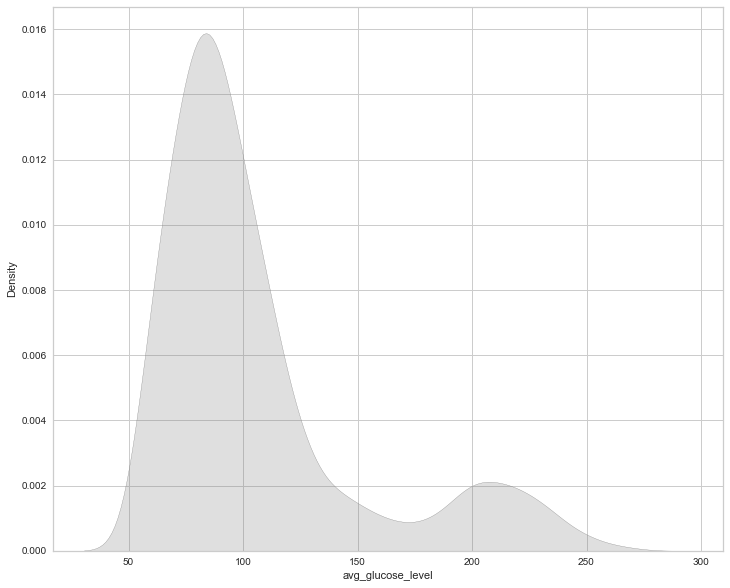
Df Residuals: 4478 BIC: 4.240e+04

Df Model: 430

Covariance Type: nonrobust

I also did the two seaborn graphs to show the distribution plot in terms of Statistical Analysis.





Finally, I moved to machine learning I did the Label Encoder in order to transform the independent variables. Then I drop the one variable which is “Heart\_Disease” in order test and train the Machine Learning models. I used the Supervised Learning because the models have predictive models both input and response variables. It includes the regression and classification is used most of the times.

I had to do the train/test method because to measure the accuracy of the model. It also to split the data set into two. I used 80% for training and 20% for testing so that I built and validated the model. First, I used the Logistic Regression and the results classified as: accuracy score, recall score, precision score, roc\_auc score and f1\_score. I had chosen Gradient Boosting because it was the best method of Machine Learning model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy Score | Recall Score | Precision Score | Roc\_auc score | F1\_score |
| Logistic Regression | 0.942 | 0.018 | 0.167 | 0.507 | 0.033 |
| Gradient Boosting | 0.941 | 0.0 | 0.0 | 0.497 | 0.0 |

I am also separated between the dependent and independent variable and the results came as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MSE | RMSE | R\_Squared |
| Logistic Regression | 18261 | 446392732 | 21128 | -0.00551 |
| KNeighbors | 17471 | 456416344 | 21364 | -0.028093 |

Chapter 5

CONCLUSION

The conclusion is the statistics shows more people who have stroke comes from developed countries because of the lifestyle changes such as heart disease, numerous, diabetes and high blood pressure according to UN. In the last decade, more people in the developing countries have stroke because of unhealthy eating habits and binge alcohol drinking. Stroke factors by the family health history such as cardiovascular diseases, age, weight, gender and race. I have used Exploratory Data Analysis, Data Preprocessing, Statistic Analysis and Machine Learning to get the results of the data given to draw conclusions and insights. Machine Learning is mostly used to predict if someone with the stroke capable to get other diseases or not using mostly supervised learning.

References

Alanazi, E. M., Abdou, A., & Luo, J. (2021). Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models. JMIR formative research, 5(12), e23440. https://doi.org/10.2196/23440

Lin, C. H., Hsu, K. C., Johnson, K. R., Fann, Y. C., Tsai, C. H., Sun, Y., Lien, L. M., Chang, W. L., Chen, P. L., Lin, C. L., Hsu, C. Y., & Taiwan Stroke Registry Investigators (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Computer methods and programs in biomedicine, 190, 105381. https://doi.org/10.1016/j.cmpb.2020.105381

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of healthcare engineering 7633381. https://doi.org/10.1155/2021/7633381

Matthew Chun, Robert Clarke, Benjamin J Cairns, David Clifton, Derrick Bennett, Yiping Chen, Yu Guo, Pei Pei, Jun Lv, Canqing Yu, Ling Yang, Liming Li, Zhengming Chen, Tingting Zhu, the China Kadoorie Biobank Collaborative Group, (2021) Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults, Journal of the American Medical Informatics Association, Volume 28, Issue 8, , 1719–1727, https://doi.org/10.1093/jamia/ocab068

Tavares, Jose-A. (2021). Stroke prediction through Data Science and Machine Learning Algorithms. 10.13140/RG.2.2.33027.43040.

Kremers, F., Venema, E., Duvekot, M., Yo, L., Bokkers, R., Lycklama À. Nijeholt, G., ... & MR CLEAN Registry Investigators. (2021). Outcome Prediction Models for Endovascular Treatment of Ischemic Stroke: Systematic Review and External Validation. Stroke, 29(2), 825-836.

Tavares, J. A. (2021). Stroke preiction through Data Science and Machine Learning. *Research Gate*, 1-6.

Yap, Z. C. (2020). Exploratory Data Analysis on Stroke. *Towards Data Science*, 1-5.

Sirsat, Manisha & Fermé, Eduardo & Câmara, Joana. (2020). Machine Learning for Brain Stroke: A Review. Journal of Stroke and Cerebrovascular Diseases. 29. 10.1016/j.jstrokecerebrovasdis.2020.105162.

Yeap, Z. C. (2020). Step by Step Data Analysis on Stroke. Towards Data Science, 6-7.