HEART STROKE PREDICTION

Raima Roj 222BDA10
Arya S 222BDA40
Ann Maria Joy 222BDA47

Department of Advanced Computing, St. Joseph's University
BD2P2: Advanced Statistics Lab
Dr. Jayati Kaushik

AIM

In a world where medical technology is rapidly advancing, the intersection of both these components are truly fascinating. We understood that machine learning algorithms have the potential to analyze large amounts of medical data and identify patterns that could be used to predict the risk of heart stroke. As a result, we found medical data on heart stroke prediction and eventually chose one of the datasets we believed would be ideal for the analysis and prediction.

To begin with, we performed data preparation which includes cleaning the data, handling missing values, dealing with categorical variables, and finally splitting the dataset into training and testing sets. Our dataset had a fair number of rows and columns with a few missing values. We then removed the unwanted columns and filled the remaining missing values as a part of data cleaning. In our dataset the dependent column was the 'stroke' which had categorical values '0' and '1'.
Next, we performed EDA on the columns that we identified to have a relation and found it best to visualize them in the form of plots to gain a better understanding of the relationship between each column.
Later, we performed Hypothesis Testing using T-test and One-way Anova for which we analyzed the dataset and made our assumptions of what the null and alternate hypothesis were. Finally, we split our data to train and test and used different techniques to find which model fits well for our prediction and also noticed the change in accuracy score.

DOMAIN – Health Sector

PROBLEM STATEMENT

We are predicting the chances of the occurrence of heart stroke in humans using statistics and machine-learning techniques such as Naïve Bayes and Random Forest Classifier.

INTRODUCTION

Heart is one of the major parts of the human body and any distress caused can affect all the other parts of the body. Heart disease remains a leading cause of death worldwide, and stroke is one of its major complications. According to the World Health Organization (WHO), cardiovascular diseases, including heart disease and stroke, are responsible for approximately 17.9 million deaths each year, which represents 31% of all global deaths. Heart stroke can be caused by various reasons including the unhealthy lifestyle of people. Some other reasons which cause

stroke are hypertension and smoking, while the latter mainly occurs due to the consumption of large amounts of fat. We can prevent heart issues by reducing Stress, tension and by having a healthy diet and good habits that promote heart health.
We can cure almost all diseases when they are detected at the right time, precisely diagnosis plays a very important role in the health care system. Detecting such conditions in humans is a complex task but this can be handled using datasets. Data mining techniques are used to extract meaningful insights that are hidden in the data. Machine learning is one of the popular methods for data mining. In the medical sector, ML plays a very important role in the diagnosis, detection and prediction of different diseases. The proposed work makes an attempt in making the detection of stroke at the early stages much easier. The techniques used in this work are Random Forest Classifier and Naïve Bayes, later a comparative analysis is done, in order to find out which is more accurate in accomplishing the task.

LITERATURE SURVEY

In recent years, the healthcare industry has seen a significant advancement in the field of data mining and machine learning. These techniques have been widely adopted and have demonstrated efficacy in various healthcare applications, particularly in the field of medical cardiology. The rapid accumulation of medical data has presented researchers with an unprecedented opportunity to develop and test new algorithms in this field. Heart disease remains a leading cause of mortality in developing nations, and identifying risk factors and early signs of stroke has become an important area of study. The utilization of data mining and machine learning techniques in this field can potentially aid in the early detection and prevention of heart stroke.

We searched for other projects related to the problem statement and we found some good papers regarding the same. We found that they have performed different algorithms on datasets similar to ours and obtained a comparative analysis of the different models. They used algorithms like Logistic Regression, Naive Bayes, Random Forest, Decision Tree, KNN and other models. Most of the research we read about concluded that Random Forest has the highest accuracy in prediction of the stroke. So we could not ignore that fact and included it in our prediction to know how well the Random Forest Classifier would help in giving better accuracy.

The nature of the stroke is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this, various tools & techniques are regularly being experimented to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart stroke can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart stroke or not. By obtaining the data from Kaggle, classifying them based on

certain attributes & finally analyzing them to extract the desired data we can conclude that this technique can be very well adapted to perform the prediction of heart stroke. As the well-known

quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart stroke.

The heart stroke prediction is done using classification techniques since our dataset involved a majority of categorical values, we used supervised learning in which the machine works as a supervisor that teaches to predict the output accurately, for this we focused on Naïve Bayes and Random Forest Classifier. The model has been trained using gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, and smoking_status as the independent variables and stroke as the dependent variable. We provide the visualizations to understand the relationship between each of the columns and to understand the main features that could cause heart stroke.

HYPOTHESIS / AIM OF THE WORK

The main aim of our project was to predict the occurrence of heart stroke using statistical methods and to find the main features that are important in making these predictions. As a part of Hypothesis Testing we chose numerical attributes such as age, bmi and average_glucose_level and considered:
Null hypothesis, (H0): The dependent variable 'stroke' and other independent features are related (or dependent)
Alternative hypothesis, (H1): The dependent variable 'stroke' and other independent features are not related (or not dependent).
We have performed t-test on stroke with the feature 'avg_glucose_level' and we noticed that the p-value we get is less than the alpha value, hence rejected the null hypothesis.This clearly implies that the different avg_glucose_level distribution does affect the likelihood of getting a stroke. Similarly, for the column 'bmi', from the testing, we can understand that the various various bmi values has a say in the predicition of stroke. Also on the feature 'age', since it is the driving factor in the prediction of stroke and as expected, we observe that different age distributions does affect the occurrence of stroke.
We also performed One-way Anova and observed that the p-value of all the features is less than the calculated f-statistic value, hence drawing a conclusion of rejecting the null hypothesis.

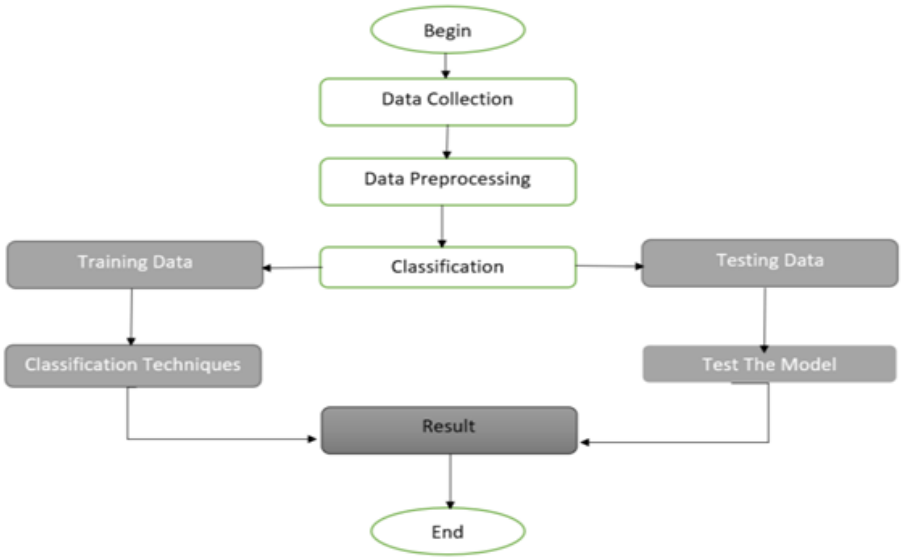
METHODS AND MATERIALS

We found the dataset that would match our requirement to predict the heart stroke from kaggle, it was the ‘train stroke.csv’ dataset for heart stroke prediction.
We used certain libraries to perform operations on the dataset, which includes:

1. Numpy and Pandas to perform the basic operations
2. Matplotlib and Seaborn for visualization(EDA)
3. Missingno for exploratory visualizations of missing data
4. Scipy for complex numerical computation
   We imported stats from scipy to perform f_oneway
5. Sklearn for selection of efficient tools for machine learning and statistical modeling.
   - From sklearn.model_selection we imported the train_test_split to split the data into training data and testing data, where training data is used to train the model and testing data is compared with the predicted output and it is checked if the model is working correctly or not.
   - From sklearn.naive_bayes we imported the GaussianNB where we compute the probability of likelihoods to check if our data is continuous.
   - From sklearn.metrics we imported accuracy_score, confusion_matrix, classification_report where the accuracy_score calculates the accuracy score for a set of predicted labels against true labels, confusion_matrix helps in understanding how many times our model has given correct or wrong output

     including the type and the classification_report is used to show the precision, recall, F1 Score, and support of the trained classification mode.
   - From sklearn.ensemble we imported the RandomForestClassifier which is created using a collection of decision trees and is used to solve classification problems

STUDY DESIGN



Data Collection- As mentioned above, the data is collected from Kaggle website. It had 43,400 rows and 12 columns. The features or attributes included: id, age, gender, hypertension, worktype, residence type, heart disease, avg level of glucose, bmi, smoking status, ever married and stroke.

Description of the dataset:
Id - Indicates a person's unique identifier.
Age- age of the person is indicated by this characteristic.
Gender- The gender of the person.
Hypertension- This feature indicates whether or not this person has high blood pressure.
Worktype- Indicates a person's job situation.
Residence type- Shows one's life situation.
Heart Disease- Depicts if a person has heart disease or not.
Avg Glucose level- This feature reflects how high a person's blood sugar is on average.
Bmi- Stands for body mass index.

Ever_married- Shows the marital status of an individual.
Smoking_status- This property indicates if a person smokes or not.
Stroke-This is our target variable(dependent), this trait reveals whether a person had a stroke or not.
Data Preprocessing- As part of this, we checked for missing values and found a significant number of missing values in bmi and smoking status,checked unique values in the dataset,we

found outliers in ‘bmi’ and filled the missing values with median. Also observed there are quite a few nan values so we grouped those values into another category called 'not known'. We noticed that there are very few values in the category 'other' so we dropped it. Since we noticed that there are differences in range of values, we normalised three columns ‘age’, ‘avg_glucose_level’,’bmi’ using min-max scaler. We converted the necessary categorical values to numerical values using replace function. Finally noticed the change in the shape of the dataset which was 43389 rows and 11 columns respectively.

EXPLORATORY DATA ANALYSIS

We performed EDA on the attributes to gain a better understanding of the relationship between these attributes and how it created a difference in predicting heart stroke. We used various types of plots using the matplotlib and seaborn to perform these visualizations.

1. We imported the missingno library to gain a quick overview of our dataset’s completeness by visualizing it through a bar chart.
2. We created a heatmap using seaborn to understand the correlation between the independent and dependent variables.
    - ❖ We noticed that the gender, residence type and the smoking status had less correlation. Also, the residence type had the least correlation therefore we can remove it since it does not make a difference in our analysis.
    - ❖ Age and heart_disease on the other had was a strong driving factor for the risk of heart stroke
    - ❖ On the contrary to our assumptions, BMI was not highly correlated to the risk of stroke. This was indeed quite surprising since we believed BMI was used to diagnose obesity and is often linked to the risk of heart stroke.
    - ❖ We also noticed that not all the attributes that show pretty fair correlation are related alone; for instance hypertension, ever_married, smoking_status are all directed to age on further understanding.
    - ❖ Another feature which is closely related to stroke is ‘. From the visualization, we can clearly understand the relation between stroke and the mentioned feature. Although there are outliers, we can conclude that, as the avg_glucose _level increases the risks of having a stroke also increases.

    - ❖ We observe that people who have been once married have a higher chance of getting a stroke compared to those who have not. Likewise, the work type of an individual also affects their probability of getting a stroke. From our data, it is clear that people who work in the private sector have a higher risk of getting strokes and people who have never worked have a very less stroke rate.

3. We used a pie chart to visualize the frequency of our dependent variable which is stroke.
4. We used the countplot to check the frequency of occurrences of each category in a categorical variable as bars; boxplot to show the distribution of numeric data values; barplot to compare different categorical variables; displot and distplot to produce data in a histogram format; and the Facet Grid to visualize the distribution of one variable as well as the relationship between multiple variables separately within subsets of your dataset using multiple panels
    - ❖ COUNTPLOT
        - to depict how many have stroke or not, based on gender
        - to show the count of patients based on hypertension
        - to show the relationship between gender and type of smokers
        - to show the relation of residence type with respect to stroke
        - to show the relation between smoking status and stroke
    - ❖ BOXPLOT
        - to show how age and stroke are related
    - ❖ DISPLOT and DISTPLOT
        - to convey the distribution of age in our dataset
        - To display the patient's bmi
    - ❖ BARPLOT
        - to show the relation between stroke and hypertension
        - to display the relation between stroke and heart disease
    - ❖ SCATTERPLOT using FACETGRID
        - to find if combination of the variables age and avg_glucose_level has any impact on stroke.

RESULTS

Initially the dataset consisted of 43400 rows and 12 columns but after data cleaning and preprocessing it was reduced to 43389 rows and 11 columns. Since most of the attributes were categorical, outliers were removed to improve the model efficiency. The algorithms we used were Random Forest Classifier and Naive Bayes, along with which we included a few measures of performance such as accuracy, confusion matrix and classification report. The dataset was split into two parts: 75% of the data used to train the model and 25% to test the model. We also performed hypothesis testing: T-test and ANOVA are the two tests used. After performing a one-way ANOVA test we noticed that the p-value of the features bmi, avg_glucose_level and age are very less compared to the F-statistic. Hence, the null Hypothesis(H0) was rejected for stroke with respect to BMI, average glucose level and age which meant there significant relation between the dependent variable stroke and the independent variables bmi, age and avg_glucose_level. We

finally implemented the model and found the accuracy score for the models was approximately 91% for Naive Bayes and 98% for Random Forest Classifier.
The problem statement given to us was to identify the key features in making the prediction of having a stroke or not therefore on performing further analysis and prediction are age and heart_disease.

CONCLUSION

We observed that in today's world, due to our changed lifestyles, we encounter so many difficulties regarding our health. The number of heart diseases and strokes are rapidly increasing day by day. This serves as a necessity to develop a model which could predict the likelihood of an individual having a stroke. In this project, we used mainly two algorithms as mentioned, Naive bayes and Random Forest classifier and an effective algorithm was obtained out of the two. We also found that the most effective one was the Random Forest classifier with an accuracy of 98%.

Accuracies obtained using different models

MODEL	ACCURACY
RANDOM FOREST	98.08
NAÏVE BAYES	91.22

DISCUSSION / FUTURE WORK

In the proposed work, we have just trained an effective model. We can further upgrade this project by adding an interface and creating an app, accepting the user values that predict their probability of getting a stroke. We can also improve the accuracy and precision of the model by providing a much larger dataset, adding many more attributes that could possibly lead to heart stroke. This can help in producing better outcomes in the future.

KEY LEARNINGS

- One of the most important factors in predicting heart stroke using machine learning models is the selection of relevant features. Features such as age,smoking status,bmi are

some of the important factors to consider.It is important to choose the right set of features that are most relevant to the outcome to ensure accurate predictions.

- We learned how to effectively clean and preprocess the data and how important it is in improving the performance of the model. We also found that the dataset should be normalized otherwise the training model is overfitted sometimes.
- Statistical analysis is also important when a dataset is analyzed
- Another important factor is the choice of the machine learning model which is crucial for predicting heart stroke accurately. We tried with several models and found that Naive Bayes Classifier and Random Forest classifier gave the most accurate values when trained with our data. We used these two models as the predictive model to understand the chance of an individual getting a heart stroke. Both the models have been tested for accuracy. Individuals with high risk of stroke are identified and are advised to undertake necessary precautions.

REFERENCES

https://doi.org/10.1155/2021/8387680
https://ijcrt.org/papers/IJCRT2106047.pdf
https://www.ijert.org/research/comparative-analysis-and-implementation-of-heart-stroke-prediction-using-various-machine-learning-techniques-IJERTV10IS060406.pdf