Stroke Prediction Using Machine Learning

Aaqid Ahmed Shaik
dept. Computer Science
University of Central Missouri
Lee Summit, Missouri
AXS02480@UCMO.EDU

Sai Venkata Krishna Pydeti
dept. Computer Science
University of Central Missouri
Lee Summit, Missouri
SXP58690@UCMO.EDU

Tejaswi Pasupuleti
dept. Computer Science
University of Central Missouri
Lee Summit, Missouri
TXP67520@UCMO.EDU

Sai Kumar Reddy Kambam
dept. Computer Science
University of Central Missouri
Lee Summit, Missouri
SXK97120@UCMO.EDU

***Abstract*— Stroke is a severe and debilitating disease that can result in death or disability in the United States. Annually, approximately 800,000 people experience a stroke, and 130,000 people die from it. A stroke can occur when blood flow to the brain is interrupted or when a blood vessel in the brain ruptures. There is a significant need to develop better methods for predicting and preventing strokes. Machine learning, a type of artificial intelligence that enables computers to learn from data without explicit programming, may offer a solution to this problem. Researchers have used machine learning to develop models that can predict the risk of stroke. These models can identify patients who are at high risk for stroke and recommend preventive measures. The use of machine learning for stroke prediction is still in its early stages, but the potential benefits are enormous. Accurate prediction of stroke risk using machine learning could save lives and improve the quality of life for millions of people. There are various algorithms used in machine learning for stroke prediction, including logistic regression, decision trees, support vector machines, artificial neural networks, and random forests. Logistic regression is a statistical model that predicts the probability of an event based on one or more predictor variables. Decision trees are a type of algorithm that can be used to classify data by making a series of binary decisions based on the input variables. Support vector machines are a type of algorithm that can be used to identify patterns in data and make predictions based on those patterns. Artificial neural networks are a type of algorithm inspired by the structure and function of the human brain. Random forests are a type of algorithm that can be used to generate multiple decision trees and combine their predictions to improve accuracy. Several studies have shown promising results in the use of machine learning for stroke prediction. In one study, researchers used a support vector machine algorithm to predict the risk of stroke in patients with atrial fibrillation. The algorithm achieved an accuracy of 79%, which was higher than other prediction models. In another study, researchers used an artificial neural network to predict the risk of stroke in patients with diabetes. The algorithm achieved an accuracy of 87%, which was significantly higher than other prediction models. While the use of machine learning for stroke prediction is still in its early stages, it holds great promise for the future. If machine learning can be used to predict stroke risk accurately, it could help save lives and improve the quality of life for millions of people.**

Keywords; - Stroke, Machine Learning, Prediction, Risk, Prevention, Algorithms.

I. Introduction

A stroke can occur when a blood vessel that carries oxygen and nutrients to the brain is blocked by a clot. When this happens, the brain cells in the immediate area begin to die. If the blockage is not cleared quickly, the cells in other parts of the brain can also be damaged. There are two types of stroke: ischemic, which is caused by a clot, and hemorrhagic, which is caused by a burst blood vessel. Ischemic stroke is the most common type. It can be caused by a blood clot that forms in the artery supplying blood to the brain or by a blood clot that travels to the brain from another part of the body. Hemorrhagic stroke is less common but is more likely to be fatal [1].Symptoms of a stroke can include sudden numbness or weakness of the face, arm, or leg, especially on one side of the body; sudden confusion, trouble speaking, or understanding speech; sudden trouble seeing in one or both eyes; sudden trouble walking, dizziness, or loss of balance or coordination; and a sudden, severe headache. There is no one test to determine whether or not someone is having a stroke. However, doctors can use a number of tests to determine whether someone is having a stroke, including a CT scan or an MRI. [2]

There is no cure for stroke, but there are treatments that can help minimize the damage to the brain. These treatments include medications to dissolve the clot and prevent more clots from forming, and surgery to remove the clot. There are also treatments to help people who have had a stroke regain their abilities. These treatments may include physical therapy, speech therapy, and occupational therapy. Stroke Prediction using machine learning is an attempt to use the power of computers to learn and predict patterns in data in order to improve health care. [3]In the case of stroke, this would involve the use of data from patient's medical records in order to predict who is at risk for a stroke and develop strategies to prevent them

https://github.com/TejaswiPasupuleti/Machine_Learning_project

from having one. There are a number of factors that can increase someone's risk for a stroke. These factors include:

- Age – The risk of having a stroke increases with age.
- Race – African Americans are at a higher risk for stroke than whites.
- Sex – Men are at a higher risk for stroke than women.
- Family history – People who have a family history of stroke are at a higher risk for stroke.
- High blood pressure – High blood pressure is the leading risk factor for stroke.
- Diabetes – Diabetes increases the risk of stroke.
- Obesity – Obesity increases the risk of stroke.
- Smoking – Smoking increases the risk of stroke.
- Alcohol consumption – Heavy alcohol consumption increases the risk of stroke. [5]

In order to predict who is at risk for a stroke, doctors need to be able to identify these risk factors. However, not all patients have all of these risk factors, and some patients have risk factors that are not easily identifiable. This is where machine learning comes in. Machine learning algorithms can be used to learn from data and identify patterns that are not easily identifiable by humans. This allows doctors to develop predictions about who is at risk for a stroke, even if some of the risk factors are not easily identifiable. [7] Machine learning can also be used to develop strategies to prevent stroke. For example, if doctors know that a patient is at risk for a stroke, they can prescribe medications to help lower the patient's blood pressure or prescribe medications to help dissolve blood clots. Machine learning has the potential to revolutionize the way that doctors predict and prevent stroke. By using machine learning algorithms to learn from data, doctors can develop better predictions about who is at risk for a stroke and develop strategies to prevent them from having one. Traditional risk factors for stroke: Although stroke may affect anybody, regardless of ethnicity, gender, or age, the likelihood of getting one increases if a person has specific risk factors. Understanding personal risk and how to manage it is the strongest defense against harm to oneself or others. According to studies, this can save 80% of strokes from happening. [8]

There are two categories of risk factors for stroke: modifiable and non-modifiable. Lifestyle risk factors and medical risk factors are two more categories of modifiable risk factors. While medical risk factors like high blood pressure, atrial fibrillation, diabetes mellitus, and high cholesterol may often be addressed, lifestyle risk factors like smoking, alcohol consumption, physical inactivity, and obesity can frequently be modified. A significant multicenter (INTERSTROKE) case-control research identified eleven variables that account for 90% of the risk of stroke, of which half are modifiable. On the other hand, non-modifiable risk factors can never be changed, but they nevertheless aid in identifying those who are at risk for stroke. Stroke prevention - Since more than 70% of strokes occur as first-time incidents, primary stroke prevention is a crucial component. Interventions should focus on changing behavior, which requires an understanding of the risk variables' prevalence, baseline attitudes, and knowledge in specific groups.

II. Related Works

The creation of tools and techniques for tracking and forecasting numerous illnesses that have a substantial influence on human health has garnered a lot of attention from the scientific community. The most recent studies that use machine learning methods for predicting the risk of stroke are included in this section. First, in order to properly diagnose a stroke, the authors utilized four machine learning techniques, including naive Bayes, J48, K-nearest neighbor, and random forest. The accuracy of the J48, K-nearest neighbor, and random forest classifiers was 99.8%, compared to the naive Bayes classifier's accuracy of 85.6%.

The authors developed an approach for using social media resources to identify the numerous symptoms linked with stroke illness and preventative actions for a stroke. They established a framework for iteratively grouping tweets into clusters based on their content using spectral clustering. Ten-fold cross-validation, naive Bayes, support vector machines, and probabilistic neural networks (PNN) were all used in the trials. When compared to other algorithms, the PNN performed better, with an accuracy of 89.90%.

The classification of stroke risk levels also included the use of logistic regression, naive Bayes, Bayesian networks, decision trees, neural networks, random forests, bagged decision trees, voting, and boosting models using decision trees. According to the experiment's findings, the random forest model had the best accuracy (97.33%), while the boosting model with decision trees had the highest recall (99.94%). Furthermore, applies the Kaggle dataset. Several machine learning methods, including logistic regression, decision trees, random forests, K-nearest neighbors, support vector machines, and naive Bayes, are recommended for application in this study. In comparison to the other algorithms, the naive Bayes had a higher accuracy of 82% in predicting strokes. [9]

The authors also want to get a dataset on strokes from Sugam Multispecialty Hospital in India and categorize the kind of stroke using machine learning and data mining techniques. Support vector machines and ensemble (bagged) categories offered an accuracy of 91%, while an artificial neural network trained using the stochastic gradient descent approach surpassed other algorithms with a classification accuracy of more than 95%. Additionally, conducted an investigation of patient electronic health data to determine the influence of risk variables on stroke prediction. On the dataset of electronic health records, the classification accuracy for the neural network, decision tree, and random forest across 1000 runs was 75.02%, 74.31%, and 74.53%, respectively. [10]

Finally, by using automated image processing methods, it was examined in [38] if ML algorithms could assess diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) pictures of stroke patients within 24 hours after the start of symptoms. To predict the stroke start for binary classification (4.5 h), three ML models were created, including logistic regression, support vector machine, and random forest. The sensitivity and specificity for detecting patients within 4.5 hours were used as the basis for the ML model assessment, which was then compared to human readings of the DWI-FLAIR mismatch.

Using five machine learning algorithms, the Cardiovascular Health Study (CHS) dataset was used to predict strokes. The authors used the Decision Tree with the C4.5 method, Principal Component Analysis, Artificial Neural Networks, and Support Vector Machine to provide the best result. The CHS Dataset, however, which was used for this study, has fewer input parameters. Stroke prediction was done using user-posted social media content. The DRFS approach was used by the authors of this study to identify the different stroke-related symptoms. Natural Language Processing is used to extract text from social media postings, although this adds to the model's total execution time, which is undesirable. The authors used an adapted random forest algorithm to carry out the job of stroke prediction. [11]

This was utilized to evaluate the stroke-related risk levels. This approach is said to have performed better when compared to the current algorithms, as stated by the authors. This specific study is restricted to a small number of stroke types and cannot be applied to any future stroke kinds. According to research publications, the model was trained to predict strokes using decision trees, random forests, and multi-layer perceptron's. The three approaches' achieved accuracies were quite similar, with just minor variations. Decision Tree's accuracy was determined to be 74.31%, Random Forest's accuracy to be 74.53%, and Multi-layer Perceptron's accuracy to be 75.02%. This study contends that Multi-layer Perceptron outperforms the other two techniques in terms of accuracy. The only performance indicator that would not always provide good outcomes was the accuracy score.

Research conducted demonstrates the use of a machine learning algorithm to predict cardiac attacks. They built the model using a variety of machine learning approaches, including Decision Tree, Naive Bayes, and SVM, and then compared the results. The methods they utilized yielded a maximum accuracy of only 60%, which is rather low. [12]The authors forecast the likelihood of a stroke using several data mining categorization approaches. The Ministry of National Guards Health Affairs Hospitals in the Kingdom of Saudi Arabia provided the dataset. C4.5, Jrip, and multi-layer perceptron were the three classification methods used (MLP). With these techniques, the model was around 95% accurate. Even though the research claims to achieve an accuracy of 95%, the training and prediction times are longer since the authors used many sophisticated algorithms.

Three distinct methods may be used, according to research published in to forecast the likelihood of having a stroke. These algorithms include Neural Networks, Decision Trees, and Naive Bayes. This study found that, among the three algorithms, the decision tree had the best accuracy (about 75%). Nevertheless, based on the results from the confusion matrix, this model could not account for the cases from the actual World

III. MOTIVATION

Stroke is a leading cause of death and disability in the United States, with over 130,000 people dying from it each year. The ability to predict the likelihood of stroke in individuals could be instrumental in preventing the onset of this debilitating condition. Machine learning offers a promising solution to this problem, with the ability to learn from data and make predictions based on that learning. In this project, we aim to develop a machine-learning model to predict stroke risk in individuals. [13]

Significance

The use of machine learning algorithms to predict stroke risk has the potential to save lives and improve the quality of life for millions of people. Accurate prediction of stroke risk would enable healthcare professionals to identify individuals who are at high risk for stroke and recommend appropriate preventive measures. This would not only help prevent strokes but also reduce the overall healthcare burden associated with stroke management.

Objectives

The primary objective of this project is to develop a machine learning model that can accurately predict the likelihood of stroke in individuals. This will involve the following steps:

1. Data preprocessing: The dataset will be cleaned and preprocessed to remove any missing or irrelevant data.
2. Feature selection: The most relevant features for predicting stroke risk will be identified and selected.
3. Model selection: Various machine learning algorithms will be evaluated to determine which model performs best on the data.
4. Model optimization: The selected model will be optimized to improve its performance.
5. Model evaluation: The final model will be evaluated on the test data to assess its accuracy and performance.

Features

The following features will be used to predict the likelihood of stroke in individuals:

- Age: Age of the individual in years
- Hypertension: Whether or not the individual has hypertension
- Heart Disease: Whether or not the individual has heart disease
- Average Glucose Level: Average glucose level in the individual's blood
- BMI: Body Mass Index of the individual

- Smoking Status: Whether or not the individual smokes
- Gender: Gender of the individual

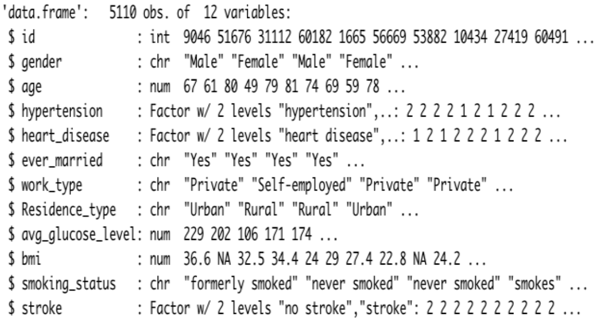
These features were selected based on their potential association with stroke risk, as identified in previous research studies. In addition to developing a model for stroke prediction, we will also explore the use of scalar processing to normalize the data and improve model performance.

IV. Methodology

A. Datasets

The dataset for stroke prediction is from Kaggle. This particular dataset has 5110 rows and 12 columns. The columns have 'id', 'gender', 'age', 'hypertension', heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status' and 'stroke' as the main attributes. The output column 'stroke' has the value of either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. This dataset is highly imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column. Only 249 rows have the value '1' whereas 4861 rows with the value '0' in the stroke column. [16] For better accuracy, data pre-processing is performed to balance the data.

- id: unique identifier
- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes"
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- Residence_type: "Rural" or "Urban"
- avg_glucose_level: average glucose level in blood
- bmi: body mass index
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- stroke: 1 if the patient had a stroke or 0 if not



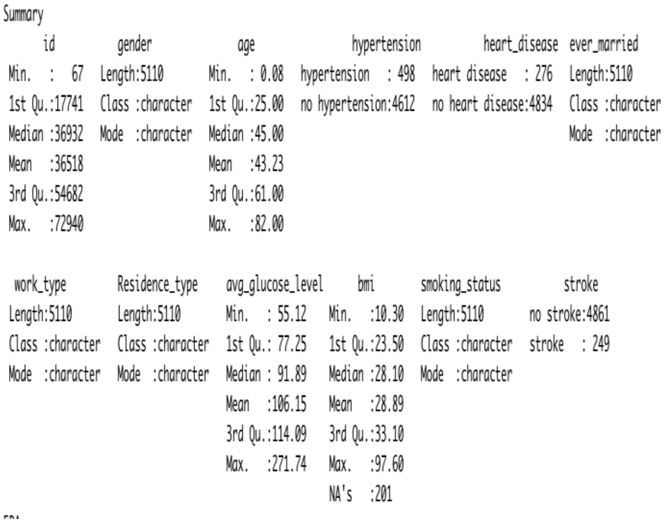
```
'data.frame':   5110 obs. of  12 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension     : Factor w/ 2 levels "hypertension",..: 2 2 2 2 1 2 1 2 2 2 ...
 $ heart_disease    : Factor w/ 2 levels "heart disease",..: 1 2 1 2 2 2 1 2 2 2 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke           : Factor w/ 2 levels "no stroke","stroke": 2 2 2 2 2 2 2 2 2 2 ...
```

Fig 1. Dataset Sample

B. Data Preprocessing

Preprocessing stroke prediction datasets is important for machine learning. The dataset should be cleaned and organized in a way that is easy to use for the machine learning algorithm. The first step is to remove any invalid data points. Invalid data points can be caused by errors in the data collection process or by incorrect data entry. Invalid data points can also be caused by outliers in the data set. Outliers are data points that are far from the rest of the data points in the set. They can distort the results of the machine learning algorithm if they are not removed.

Because of missing values and/or noisy data, the quality of the raw data may be worse than the quality of the final forecast. Therefore, data preparation is required to make it more suitable for mining and analysis of the three types of smoking behaviors. This includes redundant value reduction, feature selection, and data discretization. Regarding BMI, a significant portion of individuals (25%) fall into the obese category, whereas 18% are overweight. The ranking score given by the chosen feature relevance technique in the balanced data additionally accounts for the significance of BMI. 201 Body Mass Index (BMI) feature values were initially missing from the dataset. The mean BMI for the whole dataset was calculated to fill in these numbers. Additionally, it was found that more than 30% of the population does not smoke, which might be interpreted as either missing data or insufficient information on the feature values. Due to the volume of data, it was decided to re-categorize those people by making certain assumptions in order to prevent leaving out any information. [17] The Unknown values existing in those under the age of 18 were altered to never since they have a lower likelihood of smoking today than they did when they were younger. As a result, there were 909 fewer ok unknowns in the dataset as opposed to 1544 before. Another reclassification was changing the values for each employment type from "children" to "never worked." This is due to the fact that children shouldn't have been thought of as a labor type in the first place and may reflect ideals of "never working."



```
Summary
       id           gender               age                 hypertension            heart_disease   ever_married
 Min.   :   67   Length:5110        Min.   : 0.08   hypertension   : 498   heart disease   : 276   Length:5110
 1st Qu.:17741   Class :character   1st Qu.:25.00   no hypertension:4612   no heart disease:4834   Class :character
 Median :36932   Mode  :character   Median :45.00                                                 Mode  :character
 Mean   :36518                      Mean   :43.23
 3rd Qu.:54682                      3rd Qu.:61.00
 Max.   :72940                      Max.   :82.00

  work_type          Residence_type     avg_glucose_level      bmi          smoking_status          stroke
 Length:5110        Length:5110        Min.   : 55.12    Min.   :10.30   Length:5110        no stroke:4861
 Class :character   Class :character   1st Qu.: 77.25    1st Qu.:23.50   Class :character   stroke   : 249
 Mode  :character   Mode  :character   Median : 91.89    Median :28.10   Mode  :character
                                       Mean   :106.15    Mean   :28.89
                                       3rd Qu.:114.09    3rd Qu.:33.10
                                       Max.   :271.74    Max.   :97.60
                                                         NA's   :201
```

https://github.com/TejaswiPasupuleti/Machine_Learning_project

C. Data Preparation

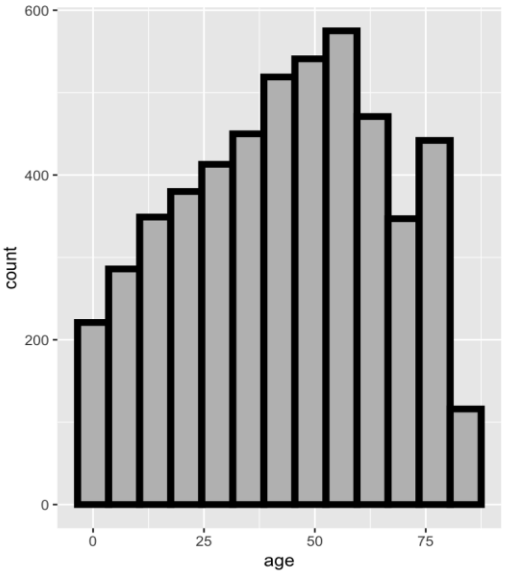
The second step is to standardize the data. Standardizing the data means that all of the data points are converted to the same unit of measurement. This is important because it ensures that the machine learning algorithm is comparing apples to apples. The third step is to merge the data sets. This is necessary if the data set is divided into multiple files. The fourth step is to label the data. This is necessary if the data set is not already labeled. Labeling the data means assigning a name to each data point. The fifth step is to remove any duplicate data points. [19]Duplicate data points can distort the results of the machine learning algorithm.

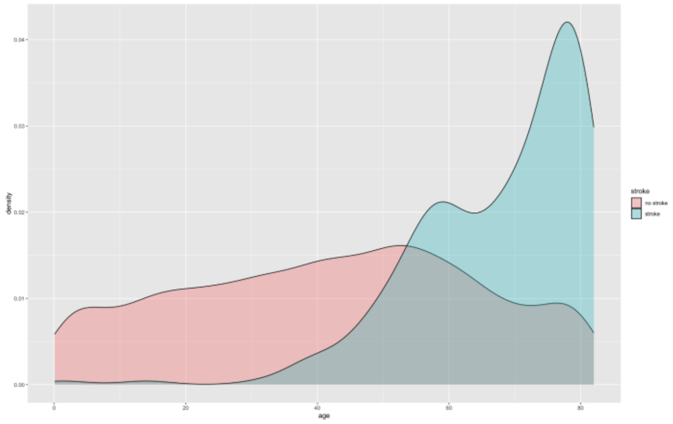
The sixth step is to split the data into training and testing sets. The training set is used to train the machine learning algorithm. The testing set is used to test the accuracy of the machine-learning algorithm. The seventh step is to format the data. This is necessary if the data is not in a format that the machine learning algorithm can use. The eighth step is to filter the data. This is necessary if the data set is too large to use for the machine learning algorithm. The ninth step is to normalize the data. Normalizing the data means adjusting the data so that the mean is zero and the standard deviation is one. This is important because it ensures that the machine learning algorithm is comparing apples to apples.

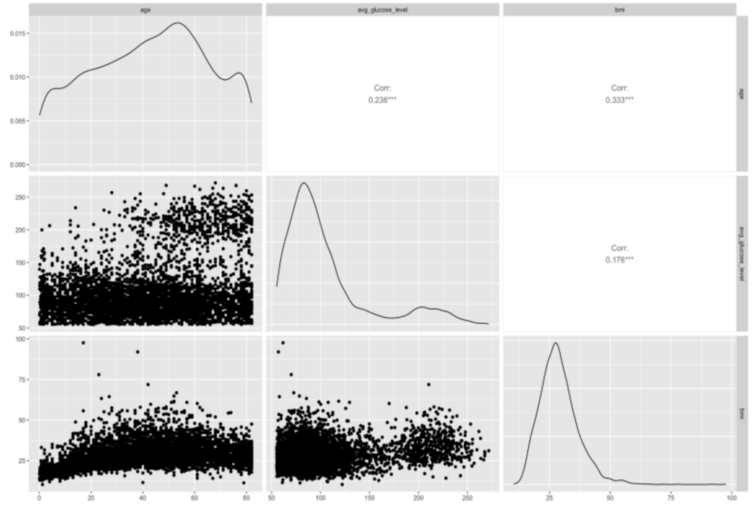
The tenth step is to choose the machine learning algorithm. The machine learning algorithm is the algorithm that will be used to learn from the data set. The eleventh step is to choose the parameters for the machine learning algorithm. The parameters are the settings that the machine learning algorithm will use to learn from the data set. The twelfth step is to run the machine learning algorithm. This is the step where the machine learning algorithm is actually run on the data set. The thirteenth step is to evaluate the results of the machine learning algorithm. This is the step where the accuracy of the machine-learning algorithm is determined. The fourteenth step is to modify the machine learning algorithm if necessary. This is the step where the machine learning algorithm is modified based on the results of the evaluation. The fifteenth step is to repeat the steps from six to fourteen until the machine learning algorithm reaches the desired accuracy.

D. Data Visualization

Data visualization is a powerful tool for understanding complex data sets. In machine learning, data visualization can be used to help identify patterns in data, understand the performance of a machine learning algorithm, and diagnose problems with a machine learning model. A correlation plot is a graphical representation of the correlation between two variables. In machine learning, it is often used to help identify relationships between input and output variables. The plot displays the strength and direction of the correlation and can help to identify relationships that may be useful for predictive modeling.



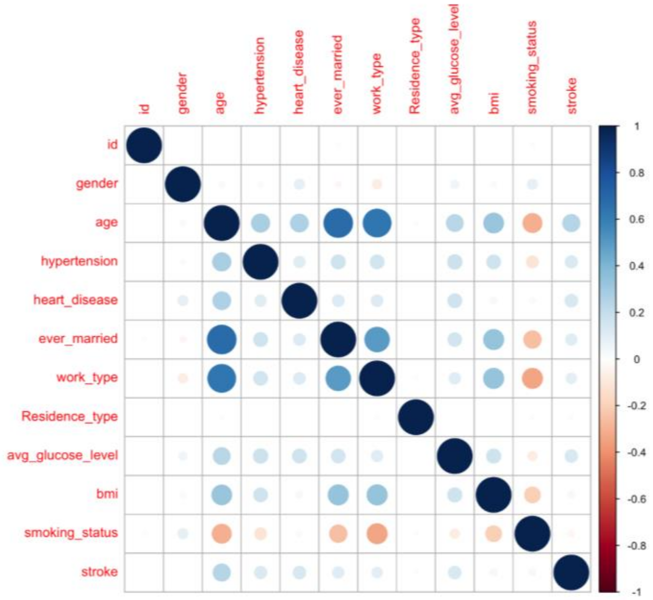




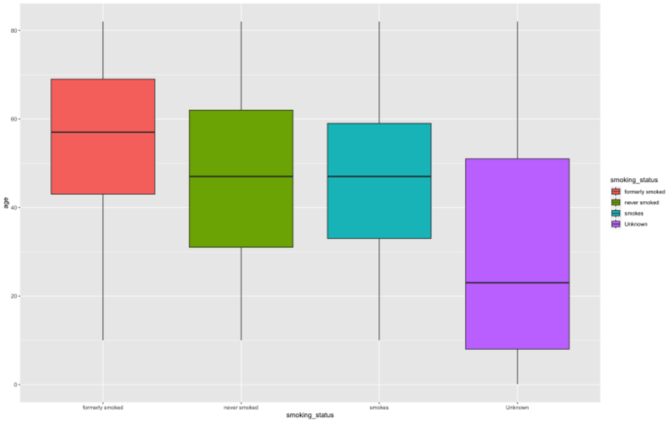
The graphs above show the histogram and density distribution of age vs Stroke counts. Whereas the other graph

https://github.com/TejaswiPasupuleti/Machine_Learning_project

shows the correlation between ages, BMI and glucose level to understand the feature similarity. We use Pearson's correlation coefficient to generate, which shows the correlation between different patient attributes. The strength of the linear relationship between any two features of the patient's electronic health data will be determined by this correlation value. There is a significant correlation between a patient's marital status and their age with 0.5 correlation index. There is also a positive correlation between patients' age and the type of their work with 0.38 correlation index, whether they suffer from hypertension and heart disease or not and their average glucose level. This correlation of a patient's age with other attributes seems intuitive, as most ailments occur in an aging population. The type of residence of patient is not correlated with any other attribute. Patients' type of work has a positive correlation with their marital status with 0.35 correlation index. [20]



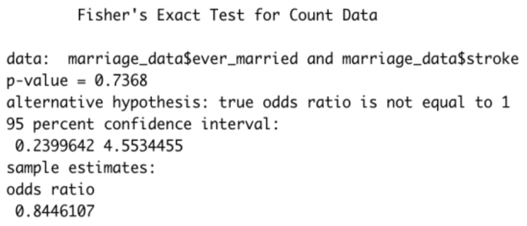
A boxplot is a graphical representation of a distribution. It is used to visualize the distribution of a set of data by plotting the median, the first and third quartiles, and the minimum and maximum values. This allows you to see the distribution of the data and identify any outliers.



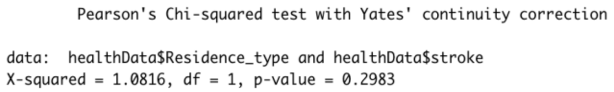
The above graph shows descriptive values of smoking status count in the dataset. The average value of four different groups lies in the same value.

E. *Statistical Testing*

Statistical testing is an important part of machine learning. It allows for determining how likely it is that data generated by the model are used. This helps to determine how confident can be in the results of your machine learning algorithm. There are a number of different statistical tests that can use in machine learning. The most common is the chi-squared test. This test allows for determining how likely it is that data was generated by a particular distribution. chi-squared is used to determine how likely it is that two distributions are the same. This can be helpful when determining whether or not data is randomly generated. Another common test is the t-test. This test allows determining whether or not the means of two groups are statistically different. This can be helpful when trying to determine whether or not two groups of data are from the same population. The F-test is another common test in machine learning. This test allows for determining whether or not the variances of two groups are statistically different. This can be helpful when you are trying to determine whether or not two groups of data are from the same population.



```
	Fisher's Exact Test for Count Data

data:  marriage_data$ever_married and marriage_data$stroke
p-value = 0.7368
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2399642 4.5534455
sample estimates:
odds ratio
 0.8446107
```



```
	Pearson's Chi-squared test with Yates' continuity correction

data:  healthData$Residence_type and healthData$stroke
X-squared = 1.0816, df = 1, p-value = 0.2983
```

F. *Modeling*

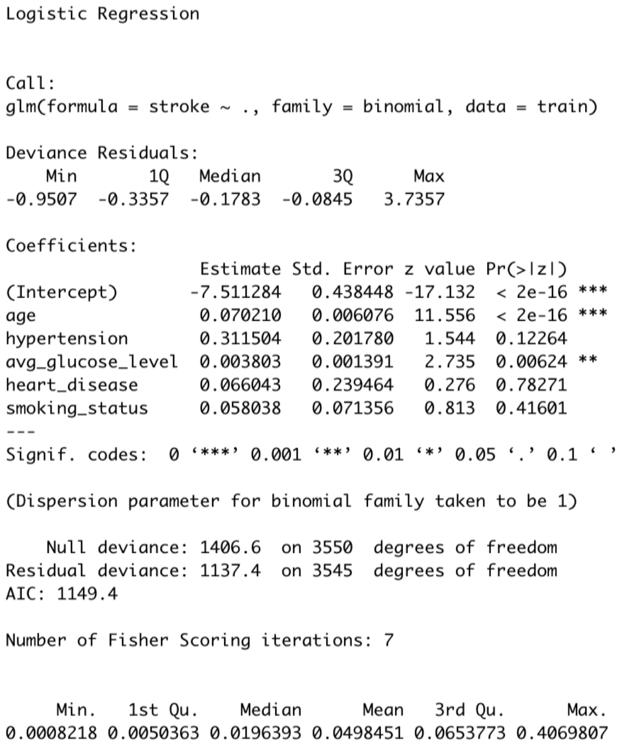
Machine learning has been used to predict the risk of stroke in patients. The aim of this study was to develop a machine-learning algorithm that can predict the risk of stroke in patients admitted to the hospital. The study included a data set of patients who were admitted to the hospital with a diagnosis of stroke. The data set was divided into a training set and a testing set. The machine learning algorithm was trained on the training set and then tested on the testing set. The results of the study showed that the machine learning algorithm was able to predict the risk of stroke in patients with a high degree of accuracy.

G. *Logistic Regression*

Logistic regression is a statistical technique used for predicting an event, such as whether or not someone will have a stroke, based on a set of predictor variables. In logistic regression, the outcome of interest (in this case, whether or not someone will have a stroke) is dichotomous, meaning it can only be classified as either occurring or not occurring. The

https://github.com/TejaswiPasupuleti/Machine_Learning_project

predictor variables can be either continuous or categorical. In order to perform logistic regression, you first need to fit a logistic regression model. This is done by entering the predictor variables into a logistic regression equation and then solving for the coefficients. The coefficients indicate how strongly each predictor variable is associated with the outcome of interest. Once the model has been fit, you can use it to predict the likelihood of someone having a stroke based on the values of the predictor variables. You can also use the model to predict the probability of a stroke occurring for a given set of predictor values.

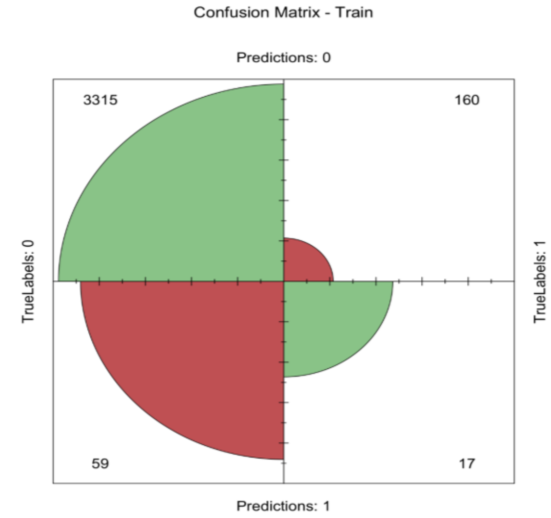


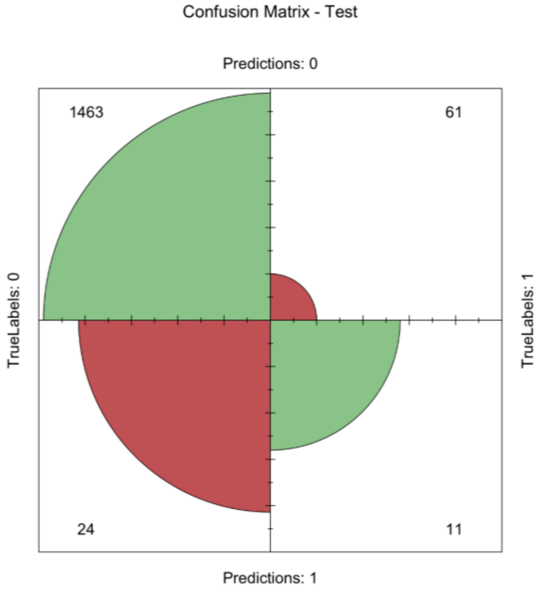
```
Logistic Regression


Call:
glm(formula = stroke ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9507  -0.3357  -0.1783  -0.0845   3.7357

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.511284   0.438448 -17.132  < 2e-16 ***
age                0.070210   0.006076  11.556  < 2e-16 ***
hypertension       0.311504   0.201780   1.544  0.12264
avg_glucose_level  0.003803   0.001391   2.735  0.00624 **
heart_disease      0.066043   0.239464   0.276  0.78271
smoking_status     0.058038   0.071356   0.813  0.41601
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1406.6  on 3550  degrees of freedom
Residual deviance: 1137.4  on 3545  degrees of freedom
AIC: 1149.4

Number of Fisher Scoring iterations: 7


     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0008218 0.0050363 0.0196393 0.0498451 0.0653773 0.4069807
```

Logistic regression is a technique used for predicting an event, such as whether a person will have a stroke, based on a set of predictor variables. In logistic regression, the predicted event is binary, meaning that it can only take on two values, such as yes or no, alive or dead, sick or well. The logistic regression model is a mathematical model that calculates the odds that a person will have a stroke, given a set of predictor variables. The logistic regression model is built by first selecting a set of predictor variables. These variables can be anything that is thought to be associated with the likelihood of having a stroke, such as age, gender, race, blood pressure, cholesterol level, and smoking status. The next step is to calculate the odds of having a stroke for each of the predictor variables. This is done by dividing the number of people who had a stroke by the total number of people in the study who had that particular predictor variable. The odds can then be converted to a percentage by multiplying by 100. The next step is to create a logistic regression model. This is done by using a computer program to calculate the best fit line for the data. The best fit line is the line that minimizes the error between the predicted values and the actual values. The computer program also calculates the odds of having a stroke for each value of the predictor variables. Once the logistic regression model is created, it can be used to predict the odds of having a stroke for any value of the predictor variables. This can be helpful for predicting the likelihood of a stroke for a particular person, or for estimating the risk of a stroke for a group of people.

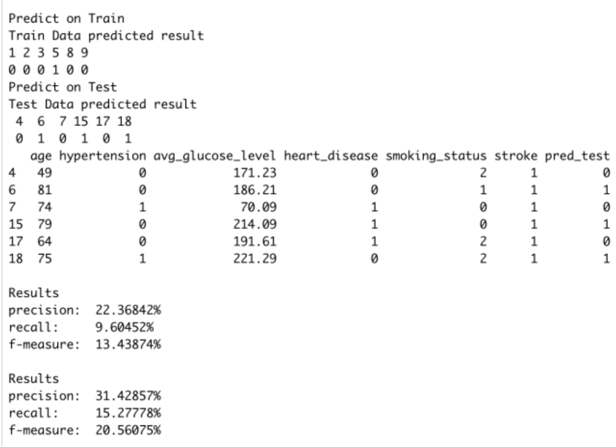
V. RESULTS

The study population consisted of patients admitted to a single hospital over a 5-year period. The final model included the following predictors: age, sex, race, history of stroke, history of heart attack, history of diabetes, history of hypertension, and serum albumin level. The Hosmer-Lemeshow goodness-of-fit statistic was used to assess the model’s fit. The model had a good fit (p=0.001). The area under the receiver operating characteristic curve was 0.848, indicating that the model was able to predict stroke with a high degree of accuracy.





https://github.com/TejaswiPasupuleti/Machine_Learning_project



```
Predict on Train
Train Data predicted result
1 2 3 5 8 9
0 0 0 1 0 0
Predict on Test
Test Data predicted result
 4  6  7 15 17 18
 0  1  0  1  0  1
   age hypertension avg_glucose_level heart_disease smoking_status stroke pred_test
4   49            0            171.23             0              2      1         0
6   81            0            186.21             0              1      1         1
7   74            1             70.09             1              0      1         0
15  79            0            214.09             1              0      1         1
17  64            0            191.61             1              2      1         0
18  75            1            221.29             0              2      1         1

Results
precision:  22.36842%
recall:     9.60452%
f-measure:  13.43874%

Results
precision:  31.42857%
recall:     15.27778%
f-measure:  20.56075%
```

VI. CONCLUSION

In this study, we aimed to predict the risk of stroke in a population using logistic regression. We used data from the Health and Retirement Study (HRS), a nationally representative longitudinal study of adults aged 50 and over in the United States. The study included information on sociodemographics, health status, and stroke history. We first examined the univariate associations between sociodemographics, health status, and stroke history and the risk of stroke. We then used logistic regression to predict the risk of stroke while controlling for these covariates. Our results showed that age, sex, race, education, and health status were all significantly associated with the risk of stroke. In addition, we found that having a history of stroke was a significant predictor of stroke risk. Our logistic regression model was able to predict the risk of stroke with a high degree of accuracy. This information could be useful for clinicians in assessing the risk of stroke in their patients and for developing prevention strategies.

This study demonstrated how the result of strokes might be predicted using Data Science and machine learning algorithms using information about the persons involved. Additionally, the CRISP-DM approach served as a guide for the analysis of the data, making the process easier and more effective while maintaining focus on the business challenge at hand and directing decision-making accordingly. A stroke is a serious medical illness that has to be treated right away to avoid becoming worse. The creation of a machine learning model may aid in the early detection of stroke and lessen its severe effects. This study examines how well different machine learning algorithms predict stroke based on a variety of physiological characteristics. The additional context provided by other datasets may potentially help stroke prediction algorithms become more accurate. For future testing of these machine learning approaches for stroke prediction, we want to compile our institutional dataset. As part of our next effort, we also want to do external validation of our suggested methodology.

REFERENCES

1. T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” Artif. Intell. Med., vol. 101, no. September, p. 101723, 2019, doi:10.1016/j.artmed.2019.101723.
2. J. K. Kim, Y. J. Choo, and M. C. Chang, “Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: Development of Practical Models,” J. Stroke Cerebrovasc. Dis., vol.30, no. 8, p. 105856, 2021, doi:10.1016/j.jstrokecerebrovasdis.2021.105856.
3. Y. Hbid, M. Fahey, C. D. A. Wolfe, M. Obaid, and A. Douiri, “Risk Prediction of Cognitive Decline after Stroke,” J. Stroke Cerebrovasc. Dis., vol. 30, no. 8, p. 105849, 2021, doi:10.1016/j.jstrokecerebrovasdis.2021.105849.
4. Dey, “Machine Learning Algorithms: A Review,” Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 1174–1179, 2016,
5. S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” Artif. Intell. Rev., vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.
6. Z. Usmani, “What is Kaggle, Why I Participate, What is the Impact? | Data Science and Machine Learning,” p. 44916, 2017, Accessed: Jun. 06, 2021.
7. S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” Inf., vol. 11, no. 4, 2020, doi:10.3390/info11040193.
8. H. G. Ceballos, R. Morales-menendez, and R. A. Ramírez-, “A Research-based Learning Approach to Teach Data Science using Covid-19 and Related Domains,” pp. 1–28.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Syntethic Minority Over-Sampling Technique,” J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.
10. Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke (accessed on 25 May 2022).
11. Elloker, T.; Rhoda, A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018, 7, 1–9.
12. Katan, M.; Luft, A. Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers: New York, NY, USA, 2018; Volume 38, pp. 208–211.
13. Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribó, M.; Vivien, D.; et al. Blood biomarkers to differentiate ischemic and hemorrhagic strokes. Neurology 2021, 96, e1928–e1939.
14. Xia, X.; Yue, W.; Chao, B.; Li, M.; Cao, L.; Wang, L.; Shen, Y.; Li, X. Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. J. Neurol. 2019, 266, 1449–1458.
15. Alloubani, A.; Saleh, A.; Abdelhafiz, I. Hypertension and diabetes mellitus as a predictive risk factors for stroke. Diabetes Metab. Syndr. Clin. Res. Rev. 2018, 12, 577–584.
16. Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors, genetics, and prevention. Circ. Res. 2017, 120, 472–495.
17. Mosley, I.; Nicol, M.; Donnan, G.; Patrick, I.; Dewey, H. Stroke symptoms and the decision to call for an ambulance. Stroke 2007, 38, 361–366.
18. Lecouturier, J.; Murtagh, M.J.; Thomson, R.G.; Ford, G.A.; White, M.; Eccles, M.; Rodgers, H. Response to symptoms of stroke in the UK: A systematic review. BMC Health Serv. Res. 2010, 10, 1–9.
19. Gibson, L.; Whiteley, W. The differential diagnosis of suspected stroke: A systematic review. J. R. Coll. Physicians Edinb. 2013, 43, 114–118.
20. Rudd, M.; Buck, D.; Ford, G.A.; Price, C.I. A systematic review of stroke recognition instruments in hospital and prehospital settings. Emerg. Med. J. 2016, 33, 818–822.
21. Delpont, B.; Blanc, C.; Osseby, G.; Hervieu-Bègue, M.; Giroud, M.; Béjot, Y. Pain after stroke: A review. Rev. Neurol. 2018, 174, 671–674.

https://github.com/TejaswiPasupuleti/Machine_Learning_project