Stroke Prediction Using Machine Learning

Shanmukh Behara

net id: UG8788

College of Business and Economics

California State University East Bay

Hayward, United States

Sbehara2@horizon.csueastbay.edu

Vinod Kumar Reddy Parla

net id: KU9183

College of Business and Economics

California State University East Bay

Hayward, United States

vparla@horizon.csueastbay.edu

Venkat Vamsi Vemulapalli

net id: UR5173

College of Business and Economics

California State University East Bay

Hayward, United States

vvemulapalli2@horizon.csueastbay.edu

Vinesh Ramakrishnan

net id: VS8175

College of Business and Economics

California State University East Bay

Hayward, United States

vramakrishnan2@horizon.csueastbay.edu

*Abstract*— When the blood abruptly stops flowing in a part of brain stroke occurs. Brain cells die, and the person may be disabled or die. Early symptoms carry valuable information, and this information can be used to take necessary actions for a healthy life. In this project based on early symptom information, we built machine learning models to calculate long term risk of stroke occurrences. The dataset is obtained from Kaggle. Multiple algorithms are used to predict the occurrence of a stroke. We intend to find the best model that achieves with a state-of-the-art performance.

# Introduction

## Project Overview and Problem Statement

A stroke occurs when a blood vessel in the brain ruptures and bleeds, or when there's a blockage in the blood supply to the brain. The rupture or blockage prevents blood and oxygen from reaching the brain's tissues.

Every 4 minutes, someone dies of stroke. Every year, more than 795,000 people in the United States alone have a stroke. About 610,000 of these are first or new strokes. About 185,000 strokes (nearly 1 of 4) are in people who have had a previous stroke.

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally [[1](https://www.heart.org/-/media/PHD-Files-2/Science-News/2/2022-Heart-and-Stroke-Stat-Update/2022-Stat-Update-At-a-Glance.pdf)] The prevalence and mortality of stroke are still on the rise [[2](https://link.springer.com/article/10.1007/s42399-022-01156-7)]. World Stroke Day is observed every year where people are being educated about stroke and its prevention [[3](https://www.world-stroke.org/news-and-blog/news/wso-global-stroke-fact-sheet-2022)]. A timely detection and prevention of stroke has become very essential to avoid its adverse consequences.

The field of medical sciences has observed tremendous improvements due to the rise in technological advancements over time. Most importantly, the Internet of Things (IoT) has made it easier to gather the data related to healthcare because of the availability of low-cost wearable devices [[4](https://ieeexplore.ieee.org/document/7782325)] Terabytes of medical data is captured by these devices. This data can be used to obtain knowledgeable patterns using various data mining techniques. The insights gained here can be used for informed decision making in the healthcare sector thereby helping in early detection of fatal diseases and conditions and cost cutting. Machine learning (ML) acts as a complement for medical sciences thus giving rise to bioinformatics in recent years. ML models can be applied on Electronic Health Reports (EHRs) to predict the risk of having strokes on each patient efficiently.

The predictive abilities of the models depend on the features selected from the data. The EHRs of the patients collect several aspects of the Identifying Stroke Indicators. However, all the collected features in EHR may not be useful for the detection of stroke. To improve the prediction performance of the machine learning models and reduce the machine training time, only important features associated with outcome should be selected for the prediction model.

Using a publicly available dataset of patients’ records, we identify the key factors that are necessary for stroke prediction. We use correlation matrix to understand the relative importance of each input attribute. We also benchmark several popular machine-learning based classification algorithms on the dataset of patient records

In the next section, we will cover an overview of some related work and describes the dataset used in our study followed by correlation analysis and its explanation.

II. LITERATURE SURVEY

In [[5](https://www.springerprofessional.de/en/a-comparative-analysis-for-various-stroke-prediction-techniques/17841142)], stroke prediction was made on Cardiovascular Health Study (CHS) dataset using five machine learning techniques. As an optimal solution, the authors used a combination of the Decision Tree with the C4.5 algorithm, Principal Component Analysis, Artificial Neural Networks, and Support Vector Machine. But the Dataset taken for this work had a smaller number of input parameters.

In [[6](https://link.springer.com/article/10.1007/s11063-020-10279-8)], stroke prediction has been carried out from the social media posts posted by people. In this work, the authors have used the DRFS method to find the various symptoms associated with stroke disease. The usage of Natural Language Processing to extract the text from the social media posts adds up to the overall execution time of the model which is not desirable.

Research carried out in [[7](https://ieeexplore.ieee.org/document/7399556)], suggests the usage of three different algorithms to predict the possibility of stroke. These algorithms are Naïve Bayes, Decision Tree, and Neural Networks. This paper concluded that the Decision tree has the highest accuracy (about 75%) of the other two algorithms. But this model could not suit the real-world examples based on the values obtained from the confusion matrix.

In [[8](https://cs.uwaterloo.ca/~jhoey/teaching/cs793/papers/Khosla10.pdf)], the researchers have performed stroke prediction on Cardiovascular Health Study (CHS) dataset. They proposed a novel automatic feature selection algorithm that selects robust features based on their proposed conservative mean. They have combined this method with the Support Vector Machine algorithm for better efficiency. But this resulted in the generation of several vectors that tend to reduce the performance of the model.

Research in [[9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6260436/#sec2title)] proposes the prediction of thrombo-embolic stroke disease using Artificial Neural Networks. The method used for prediction was the Back-propagation algorithm. This model was able to get an accuracy of around 89%. But Neural Networks need more time to be trained and require higher processing time because of the complex structure with increasing number of neurons.

# IMPLEMENTATION

## Data Description

The data we used for this project is an example of an Electronic Medical Record, which is a collection of a patient’s information. These types of records contain information about a patient’s vitals, medical exam and diagnostic results. For our project, we used a dataset of EMR which is publicly available from Kaggle [10]. The dataset contains a total 5110 records, 11 attributes, and one output feature. The dataset discussed above is summarized in Table 1.

TABLE I. Stroke Dataset

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Type**  ***(Values)*** | **Description** |
| 1. id | Integer | A unique integer value for  patients |
| 2. gender | String literal *(Male, Female, Other)* | Tells the gender of the patient |
| 3. age | Integer | Age of the Patient |
| 4. hypertension | Integer  *(1, 0)* | Tells whether the patient has hypertension or not |
| 5. heart\_disease | Integer  *(1, 0)* | Tells whether the patient has heart disease or not |
| 6. ever\_married | String literal  *(Yes, No)* | It tells whether the patient is married or not |
| 7. work\_type | String literal *(children, Govt\_job, Never\_worked, Private, Self- employed)* | It gives different categories for work |
| 8. Residence\_type | String literal  *(Urban, Rural)* | The patient's residence type is stored |
| 9.  avg\_glucose\_level | Floating point number | Gives the value of average glucose level in blood |
| 10. bmi | Floating point number | Gives the value of the patient's Body Mass Index |
| 11. smoking\_status | String literal *(formerly smoked, never smoked, smokes, unknown)* | It gives the smoking status of the patient |
| 12. stroke | Integer  *(1, 0)* | Output column that gives the stroke status |

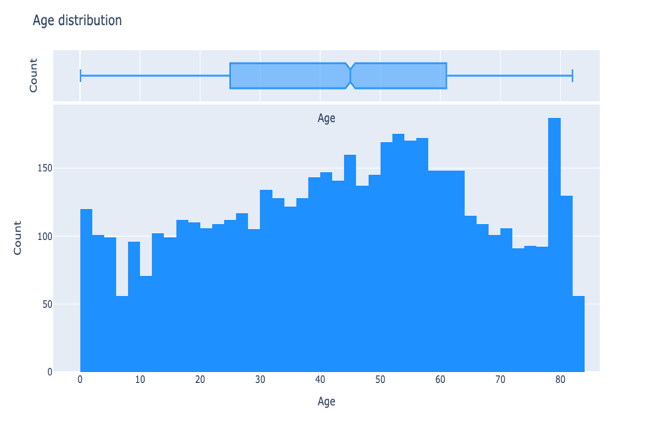
## Exploratory Data Analysis

The study of data in order to draw conclusions is known as data analysis. Data analysis may help in the creation of new research, disaster assistance, and disease outbreak management. In this part we used different data visualization techniques find the top attributes contributing to stroke, and their distribution. For example, the strongest predictor of stroke seems to be age. Looking at the overall mutual Information graph, age seems to be the only variable that is strongly correlated with stroke, BMI, Hypertension

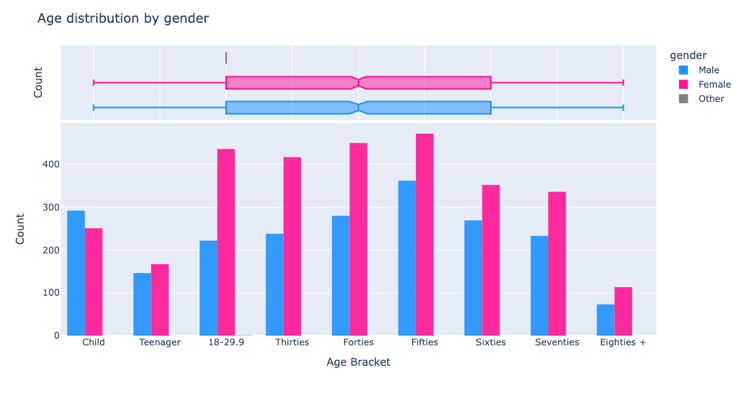
Graphical user interface, application, Teams

Description automatically generated

It’s clear that the stroke rates are significantly higher in age variable than in another variable, now when we see the distribution of data as per age, we can clearly see that the data has more people between 30-60.

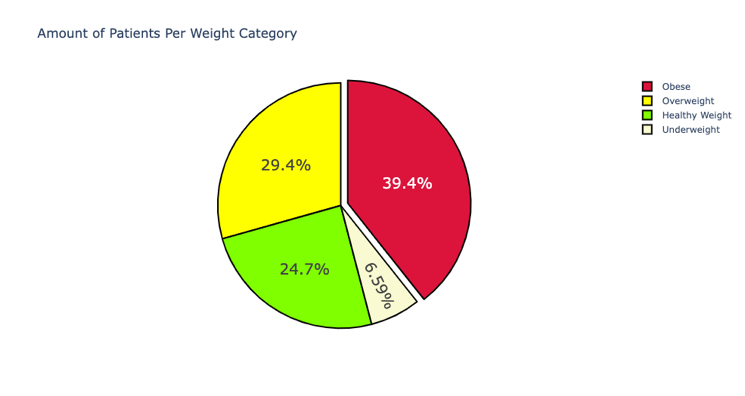


When drill down the data and see the gender distribution for the age attribute, we see that per age group, there are more females in the data than there are males with the exception being Children.



Obesity is another contributing factor for Strokes in patients. During the EDA, we could see that approximately 40% of the patients are obese and approximately 70% of the patients in our data are at least overweight.

This perhaps may lead to bias in our analysis as we don't have a fair distribution of patients per BMI range.



When we look at the BMI Distribution by gender, it makes sense that there are more women in each weight category as there are more women overall in the patient dataset. However, approximately 40% of the patients in this data are **Obese** and almost 70% of the patients in this data are *Overweight or Obese*

When we look at the distribution of Average Glucose level by gender, we know that there are more women than men in the dataset, so naturally we expect the glucose categories to be higher for the women. This is the case for Normal glucose levels, women outnumber the men significantly. It is significantly closer with respect to Prediabetes & Diabetes patients.

As part of the Exploratory Data Analysis, we have also done few multifaced comparisons to understand the correlation between BMI categories and Glucose Categories.

A key takeaway from this is that most obese patients have normal glucose levels. We can say that there are more obese patients in the data than any other group, but it's still a surprising insight to discover that a significant chunk has normal glucose levels.

Chart

Description automatically generated

It only makes sense to look at the number of strokes per BMI and age category after looking at the above correlation because we see that for every age group in the Obese bracket (except for thirties & eighties), there is a higher number of strokes. Given that the age distribution was uniform, this could *possibly* be interpreted as an attributing factor for a stroke. Indeed, we saw this upon inspection of our correlation matrix, that age is the most correlated with incidence of a stroke.

Chart, bar chart

Description automatically generated

We see that the most obese of any age group are people in their fifties, and most underweight group of patients tend to be children (12 years and below). The overweight class is roughly evenly spread out across all age brackets apart from Children and people in their eighties.

Table

Description automatically generated

When we look at the glucose spread among BMI categories, it’s interesting to see that for people in their 30's and who are Underweight, their median glucose level is roughly 168 whilst this is not the case for the rest of age brackets in the Underweight. It should be noted that the Underweight group is underrepresented in this data, so this is an insight that needs more data to make a full-blown proper conclusion. Another interesting takeaway is that if we look at the Obese category and look at people in their fifties, we see that the max glucose level is approximately 260, this would be considered an anomaly for younger people who are also Obese and younger people across other weight brackets.

Chart, bar chart

Description automatically generated

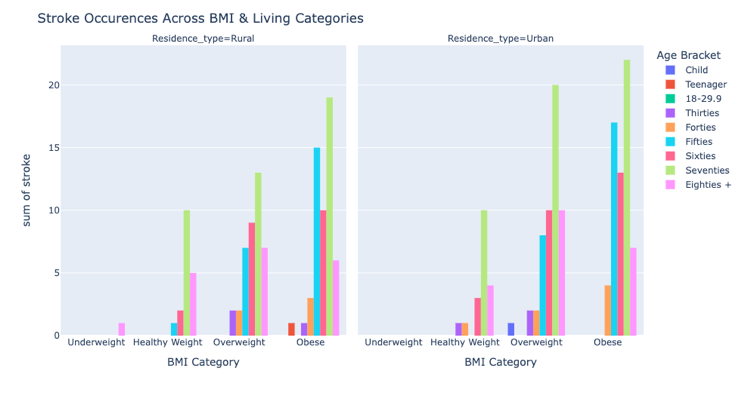
An interesting data point to look here is the number of strokes per BMI, Glucose and Age Category as it gives us a holistic view of the people who fall in the obese category, are diabetic and fall in the age group of seventies. This group leads the pack in the total number of strokes in all categories.

Chart

Description automatically generated

Comparison of stroke frequency and BMI based on the Living conditions gives us a holistic view of the impact on one’s health because of their lifestyle choices. There’s no doubt that 135 Patients who live in an urban area whilst a 114 who lived in a Rural area had a stroke. It's no surprise that the most stroke occurrences of the two areas are in the Obese and Seventies category. It’s an interesting thing to ponder on, why is it for each weight bracket the most stroke occurrences happen to a seventy-year-old, this is odd considering the uniformity of the age distribution.

Interestingly, we see that there are 5 Children who are former smokers and 1 child who currently smokes.



## Clustering and Outliers

One of the main motivations of clustering is to detect if they are any underlying patterns beyond the class labels (0 - no stroke, 1 - stroke). Some cases could have indicated maybe there are levels to how a person is diagnosed with respect to their risk of having a stroke or not. There could also be the case that patients have the symptoms and/or characteristics of someone who is going to get a stroke but has not yet had one.

We use an isolation forest to see which samples are anomalous. We compare the age and average glucose level to see what's considered normal and what's considered an anomaly.

From our visualizations we see that:

* The Outliers skew towards higher Age groups and Glucose levels. A possible explanation for this is that the number of patients who have had a stroke is approximately 5% of the entire dataset, so it makes sense as to why the model would detect these cases as anomalous. This is an example of a case where you should leave outliers in the dataset as removing them can produce inaccurate results.
* Inliers seem to be distributed more evenly across age groups with lower glucose levels. Later, we see that glucose level is one of the more important features in our dataset. This is something important to note because 95% of our data are patients who have not had a stroke and if we look at the contour plot, we see that patients who have not had a stroke tend to be spread out across age and have lower glucose levels.

Chart, scatter chart

Description automatically generated

In an informal sense, the isolation forest model has assumed that people who have not had a stroke are the inliers and who have had a stroke as outliers. This makes sense in the regard that there is a severe imbalance in the dataset and as such, no action would be taken to remove outliers. It is also a strong case to get access to more data to build better predicative models.

## Dealing with missing data

First, we must determine the type and extent of missing data before deciding the method to impute or removing them from the dataset before further processing. We see 3.9% (201 rows) of the BMI data is missing. The missing data is spread across all age categories and as such cannot be ignored due to it being a critical part in the analysis of stroke and its subsequent prediction. Deleting this data from an already small dataset may hinder accuracy metrics when trying to predict stroke or not. As mentioned earlier, approximately 3.9% (201 rows) of the BMI data is missing.

Chart

Description automatically generated

In looking for patterns with respect to the missing data, with respect to age, approximately 27% of the missing data are in their seventies, approximately 15% in their sixties, approximately 14% in their fifties, approximately 12% and approximately 11% in their forties and thirties respectively, the remaining age brackets make up the rest with less than 10% each.

In the context of stroke, approximately 80% of the missing values did not have a stroke. Similarly for heart\_disease approximately 80% did not have heart disease and it was a bit lower for hypertension with approximately 76% not having heart disease

For gender, it was a roughly equal split with approximately 52% of the data being males and the remainder being females. A similar split was seen for Residence\_type with patients living in Rural being the slight majority.

Next, we must determine the randomness of the missing data. Here, we have two cases to consider:

1. Missing at Random: The observed 𝑌 (in this case BMI) values represent a true sample in relation to the other data points but do not represent a true sample of all the possible observations of 𝑌(BMI)
2. Missing Completely Random: This means that there is no relationship between the missing data points and the other values in the dataset, the cases with the missing data are indistinguishable from cases with complete data.

Our analysis suggests that the data is MCAR, as there is a complete range of all possible BMI. There have been cases in human history [(See here)](https://en.wikipedia.org/wiki/List_of_heaviest_people) that have exceeded the BMI range in our dataset meaning that we have not had a true sample of all possible values of BMI. In this case, we would have to view our data as MAR.

## Handling Missing Data

In our model we will be using mean to ensure the missing values for the column BMI are filled in appropriately. Once we populate the missing values for the BMI columns with their means we will test to see the data set has no missing values for any columns as per the figure below

Chart, bar chart

Description automatically generated

## Data Preprocessing

#### Data Preprocessing is required before model building to remove the unwanted noise and outliers from the dataset, resulting in a deviation from proper training. Anything that interrupts the model from performing with less efficiency is taken care of in this stage. After collecting the appropriate dataset, the next step lies in cleaning the data and making sure that it is ready for model building. The dataset taken has 12 attributes, as mentioned in Table I. Firstly, the column 'id' is dropped because its existence does not make much difference in model building. Then the dataset is checked for null values and filled if any found. In this case, the column 'bmi' has null values filled with the mean of the column data. After removing the null values from the dataset, the next task is Label Encoding.

## Label Encoding

## In our dataset we see that there are five non numerical columns. The following columns are non numerical columns-

#### 'gender', 'ever\_married', 'work\_type', 'Residence\_type', 'smoking\_status'. In order to build effective predictive models it is essential that we convert the non numerical columns into integers. This ensure that the entire dataset is in integers to run our predictive analysis model.

## Handling Imbalanced Data

## Our preluminary analysis on the dataset chosen for the task of stroke prediction is highly imbalanced. Training a model using imbalanced data will lead to inaccurate prediction. To increate accuracy of the predictive models, it is critical to balance the dataset. To balance our dataset, we will use undersampling. By using the undersampling method, we can ensure that the majority (0) and minority (1) classes are balanced. Once undersampling is implemented, the resulting dataset will have 249 rows with a value of 0 (No Stroke) and 249 rows with a value of 1 (Stroke). Below graph represnts the dataset befor and after under sampling.

*Before under sampling*

Chart, bar chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

*After under sampling*

*Chart, bar chart

Description automatically generated*

## Model Imputation

#### For our model, will be Splitting the Dataset after completing data preprocessing and handling the imbalanced dataset. We will be using a 80%-20% data split for our training and validation partition and running different models to determine their accuracy. Based on our dataset we have chosen to run the below models to predict the outcome variable (Stroke) owing to its binary nature (0 or 1)

We provide a detailed analysis of various benchmarking algorithms in stroke prediction in this section. We will be benchmarking 5 popular predictive approaches – Logistics regression, Decision tree, Random Forest, KNN and SVM in this section

# Models

## Logistic regression

Since our outcome variable is binary, we will start by implementing a logistic regression model. We will use our undersampled dataset to perform the logistic regression analysis to compute the probability of a patient to have a stroke.

After performing this algorithm on the dataset, the accuracy obtained for the validation partition is 76%.

The next step is to use other models to see if we can improve the accuracy score of the validation partition by implementing other models.

Text

Description automatically generated

## Decision Tree

Decision Tree is both a classification and prediction method that performs well across a wide range of DM situations. Owing to the binary nature of the outcome variable (Stroke), we will be using regression trees to predict the probability of a patient to be diagnosed with stroke. We will build a tree like decision process with several test and applying it to the unders sampled dataset. Each node in this tree represents a test, and the branches correspond to the outcome of the test. The pruning ability of such algorithm makes it flexible and accurate, which is required in medical diagnosis.With the decision tree classification algorithm we have obtained an accuracy of 69% for the validation dataset

Classification and Support Vector Machine.

*Text

Description automatically generated*

## Random Forest

## The idea of Random Forest is to take an average of multiple estimates (models), which is more reliable than just using a single estimate. We will draw multiple random samples, with replacement, from data set. Samples with replacement means that when the sampling record is drawn from a data set and recorded in the sample, it is returned back to data before the next record is drawn.

This method is known for its flexibilty, ease of use and know for producing good results, even with minimal tuning of the hyper-parameters makes this algorithm valuable in this application. The possibility of overfitting are limited with this method.

The accuracy obtained by training the model using this particular algorithm is 76% for the validation partition

Text

Description automatically generated

## K-Nearest Neighbors Classification

One of the popular models for regression analysis is the KNN model. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.The KNN algorithm can be effective to estimate continuous variables (Stroke) based on a weighted average of the k nearest neighbors, weighted by the inverse of their distance.

Analyzing the confusion matrix for derterminig the accuracy of this model, it is foud to be 73% accurate for validation partition

*Text

Description automatically generated*

## Support vector machine:

Support Vector Machine (SVM) scales relatively well to high dimensional data. For this particular dataset, the algorithm obtained an accuracy of 78% for validation partition

Text

Description automatically generated

*Model Comparison*

|  |  |
| --- | --- |
| Model name | Accuracy |
| Logistic Regression | 76% |
| Decision Tree | 69% |
| Random Forest | 76% |
| KNN | 73% |
| SVM | 78% |

## Benchmarking

We will be comparing the accuracy of the models using original dataset vs the under sampled dataset to study our results

|  |  |  |
| --- | --- | --- |
| Model name | Accuracy (undersampled) | Accuracy (Original Dataset) |
| Logistic Regression | 76% | 94% |
| Decision Tree | 69% | 92% |
| Random Forest | 76% | 95% |
| KNN | 73% | 94% |
| SVM | 78% | 95% |

Based on the above benchmarking results, we see that the original dataset yields higher accuracy. However, owing to the highly imbalanced data (higher no of patients with stroke vs no stroke), this accuracy can be misleading.

## Conclusion

Accurate prediction of stroke can ultimately lead to saving patients life. Building a machine learning model can help in the early prediction of stroke and reduce the severe impact in the future by providing timely patient care. Our analysis suggests that SVM model has the highest accuracy of 78% and would be our recommended model to use for stroke prediction.

we can conclude that Stroke can be predicted using machine learning algorithm and effected person should be treated accordingly before it gets into final stage or worse.

We have built models for predicting stroke using machine learning algorithm. After referring various papers on internet, we decided to select five models which gives best metrics for this dataset such as Logistic Regression, Decision Trees, Random Forest, SVM and KNN. We performed Analysis on categorical features, numerical features, and multicollinearity

We performed a detailed analysis of patient’s attributes such as feature correlation choosing optimum set of features where we can find important hidden patterns and neglect unnecessary features.

After applying all the five models on the dataset we compared all models with different performance metrics and chose best model.

We can observe that most of the existing features in dataset are highly correlated to each other. There are 2 main reasons why the accuracy cannot be improved drastically are

1) Lack of discriminatory feature set.

2) Insufficient dataset.

This project can further be extended by collecting the dataset consisting of images of Brain CT scan so that the prediction of probability of stroke would be more efficient in future.

Finally, our method can be used for identifying potential risk factors for diseases without performing clinical trials. We hope that this project will motivate the application of machine learning methods in healthcare data analysis.