Georgia State University

J. Mack Robinson College of Business



Heart Stroke Prediction Analysis

Fall 2022

**By: Satish Tirumalapudi | Salma Chanbasha Nandikotkur |Sandhya Sree Agolu  
 Harini Prabha Baskar Jayanthi| Anwar Shaikh | Prathyusha Parimi**

Table of Contents

[1. Project Description 3](#_Toc84186434)

[2. Goal & Business Problem 3](#_Toc84186435)

[3. Data Preprocessing 3](#_Toc84186436)

[4. Data Cleaning](#_Toc84186437) 4

[5. Exploratory Data Analysis](#_Toc84186438) 4

[6. Data Modelling 9](#_Toc84186440)

[7. Model Comparision](#_Toc84186441) 13

[8. Conclusion](#_Toc84186441) 14

**Project Description:**

**Context**

This project is about predicting early heart strokes that helps the society to save human lives using Logistic Regression, Random Forest, KNN, Neural Networks and Ensemble Models using “**Healthcare-Dataset-Stroke-Data.csv**” Dataset from Kaggle, based on 5110 observations with 12 explanatory variables.

**Dataset**: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata>)

**Goal:**

The project is aimed at implementing a model(s) to predict early heart strokes that helps the society to save human lives.

* Clean variables, build what is needed
* Models: Logistic Regression, KNN techniques, RandomForest, Ensemble Learning & Neural Networks
* Choose the best model having best accuracy.

**Business Problem:**

Heart Stroke is one of the severe health hazards, According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Therefore, early heart stroke prediction helps the society to save human lives. Many of these strokes can be avoided by adopting healthier lifestyle, and monitoring individuals who are most at risk can significantly improve the results.  This project focuses on identifying stroke risk factors and offers suggestions for how to avoid them.

**Data Preprocessing:**

From our given dataset we know that we have been given 5110 observations, with 11 column variables, 9 of which are predictive to our outcome of **stroke** and one of which is an identification quantifier for our patients.

|  |  |
| --- | --- |
| Column name | Description |
| 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 Cleaning:**

* Replaces missing values with mean in column name BMI.
* Dropped column ID which is irrelevant.
* Checked for outliers, data entry errors.

As part of data pre-processing, we performed one hot encoding on the dataset to convert categorical variables to machine readable format (numerical values). After encoding we went ahead and checked for the imbalances in data and found asymmetry in data. We performed oversampling on the dataset using MWMOTE (Majority Weighted Minority Oversampling Technique) and balanced the data.

**Exploratory Data Analysis:**

Performed exploratory data analysis between some of the related input variables and the output variable “Stroke”.

**Stroke events vs Gender**

Male and Female show almost similar correlation with the stroke. But female gender are more likely to get a stroke when compared to male gender.

Chart, bar chart

Description automatically generated

**Stroke events vs Patient Work type**

The below bar graph shows the likeliness of stroke depending on the work type of an individual. Private job holders are most likely to get a stroke when compared to other jobs.

Chart, waterfall chart

Description automatically generated

**Stroke events vs Patient Residence type**

People living in urban area are more prone to get a stroke when compared to the people living in rural area.

Chart, bar chart

Description automatically generated

**Stroke events vs Age**

People between 65 - 75 years of age are more prone to a heart stroke

Chart, histogram

Description automatically generated

**Stroke events vs Smoking habits**

Individuals who does not smoke also shows positive correlation with the stroke event.

Chart, bar chart

Description automatically generated

**Stroke events vs hypertension**

People with hypertension are more likely to get a stroke when compared with people with no hyper tension.

Chart, bar chart

Description automatically generated

**Stroke events vs heart disease background**

Individuals without any heartdisease background are also likely to get a stroke.

Chart, waterfall chart

Description automatically generated

**Average Glucose levels**

The below graph shows the average glucose levels of an individual based on which stroke can be predicted.

Chart, histogram

Description automatically generated

**Models and their comparison:**

Being a classification problem, We have implemented the below models:

* Logistic Regression
* Classification using K-Nearest Neighbors
* RandomForest
* Neural Networks
* Ensemble method

## Logistic Regression:

The Logistic Regression model on the testing data gives an **accuracy** value of **87.07%.**

A screenshot of a computer

Description automatically generated with low confidence

**Classification using K-Nearest Neighbors:**

KNN stands for **K-Nearest Neighbors**. It is a supervised learning algorithm. It is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). We have used 11 independent features for KNN implementation. A robust implementation must consider feature engineering, data cleaning, and cross-validation.

* **K means clustering**
* K = 5
* Sampling 80% of data for training the algorithms using random sampling

We have implemented KNN with different optimal weights by changing k values and this time the **accuracy** we achieved is **87.1%.**

A screenshot of a computer

Description automatically generated with medium confidenceChart, pie chart

Description automatically generated

**Random Forest:**

Random forest is a supervised machine learning classification algorithm consisting of many decisions trees. By implementing we have got an accuracy of 94.65%

A screenshot of a computer

Description automatically generated with low confidenceChart

Description automatically generated

A picture containing chart

Description automatically generated

The varimp plot here shows the relative importance of predictors in a model.

**Neural Network:**

Neural networks are a class of machine learning algorithms used for complex patterns in datasets using multiple hidden layers and non-linear activation functions. They are also known as artificial neural networks (ANNs) or simulated neural networks (SNNs). We have implemented in our scenario and the **accuracy** we achieved for the testing set is **91.2%.**

Diagram

Description automatically generated

**Ensemble: Weighted Average**

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. We have implemented ensemble techniques with three models: Logistic Regression, Neural Network and KNN in this project. The **accuracy** we attained 94.38% for weighted average model.

A screenshot of a computer

Description automatically generated with low confidence

## Model Comparison:

## Below is the accuracy for all the five models implemented in the project:

|  |  |
| --- | --- |
| **MODEL** | **ACCURRACY** |
| Logistic Regression | 87.07 % |
| KNN | 87.1% |
| Random Forest | 94.65 % |
| Neural Networks | 91.2 % |
| Ensemble | Weighted: 94.3 % |

Random Forest performed the best with an accuracy of 94.6% followed by Ensemble and random forest with an accuracy of 94.3%.

**Conclusion:**

Age is a major risk factor for stroke. As we get older, we are more at risk to suffer a stroke. Males and females both suffer stroke at a similar rate. However, females have been shown to suffer strokes at younger ages than males. Heart problems like hypertension and heart disease greatly increase the risk of stroke. People who have been married are at a higher risk of stroke. This may be due to higher levels of stress that occur during married life. We got better accuracy while using the Random Forest Machine learning model of 94.65 To improve variance and bias we used weighted average and got an accuracy of 94.3