

**CEP REPORT**

**Topic:**

Stroke Risk Predictive Analysis

**Submitted to:**

Muhammad Bux Alvi

**Submitted by:**

Muqarrish Zaib

19CS02

**3rd Year 6th Semester (2019-23)**

**B.Sc. Computer Systems Engineering**

**Faculty of Engineering**

# Abstract

I was curious to see the effect that certain features had in the possibility of a person having a stroke and how well machine learning models predicted stroke. I decided to dive deep into the background research I found, any related work I came across, the dataset I will be using, and the machine learning models that I created to predict stroke based on features. I found that for many cases, Age contributed the most to strokes, followed by BMI and Smoking Status. Conversely, factors like Ever Married, Gender, Hypertension, and Heart Disease had little to no effect on strokes.

**Contents**

[Abstract 2](#_Toc106729539)

[1 Introduction 4](#_Toc106729540)

[2 Dataset 4](#_Toc106729541)

[2.1 Columns 5](#_Toc106729542)

[3 Methodology 6](#_Toc106729543)

[3.1 Exploratory Data Analysis (EDA) 6](#_Toc106729544)

[3.2 Picking Models 10](#_Toc106729545)

[4 Methods 10](#_Toc106729546)

[4.1 Logistic Regression 10](#_Toc106729547)

[4.2 Support Vector Machines 10](#_Toc106729548)

[4.3 Random Forest 10](#_Toc106729549)

[5 Conclusion 11](#_Toc106729550)

[5.1 Overall Analysis 11](#_Toc106729551)

[5.2 Future Work 12](#_Toc106729552)

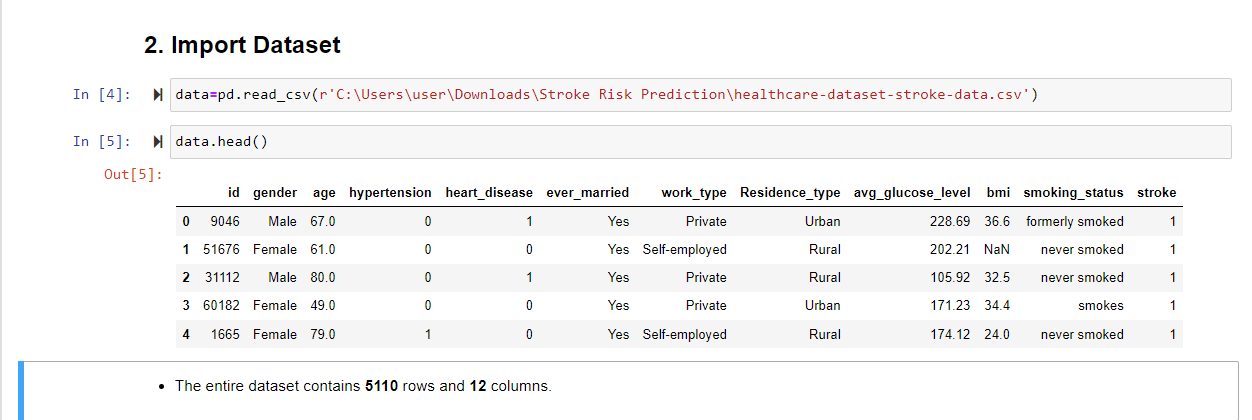
[References 13](#_Toc106729553)

# Introduction

Stroke is a disease that affects the arteries leading to and within the brain. **Stroke is the leading cause of disability worldwide**and the second leading cause of death [[1](#_bookmark55)]. A stroke happens when a person’s blood supply to their brain is interrupted or reduced, causing brain cells to die within minutes. It prevents the brain tissue from getting the oxygen and nutrients that it needs and is responsible for approximately 11% of total deaths. To see what effect certain factors had on people who suffered strokes, I decided to create Machine Learning Model: Decision Tree. I did this by not only writing our own code to create the model, but I also compared the yielded metrics to those from built-in models of Scikit-Learn, a built-in Machine Learning Python library.

# Dataset

The dataset that I will be using for the model were taken from Kaggle [[2](#_bookmark51)]. There are 5110 entries of individuals, their relevant feature details, and whether they had a stroke or not.



## Columns

* + - 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 heart disease

*•*

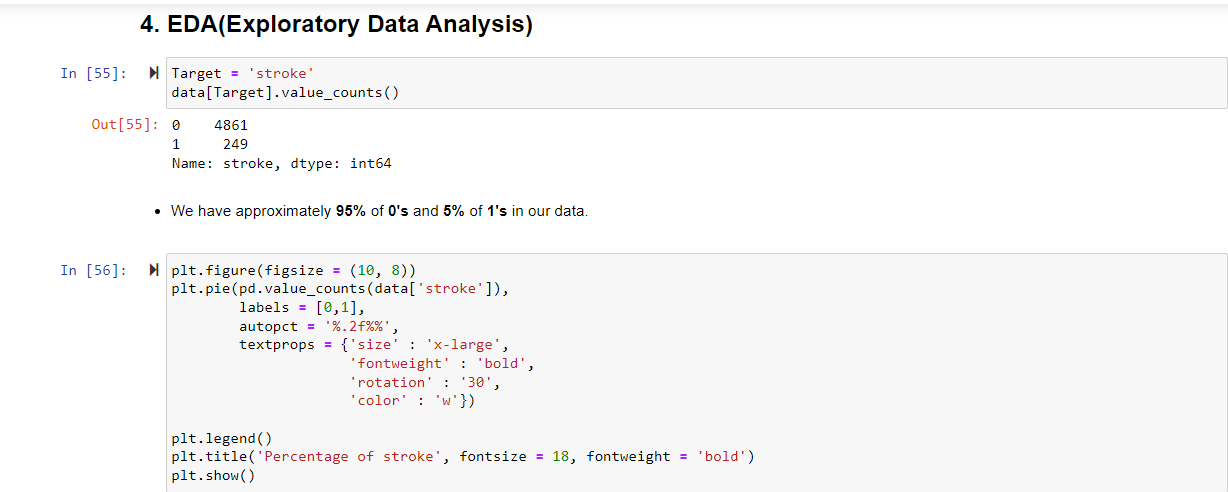
* + - ever married: “No” or “Yes”
    - work type: “children”, “Govt job”, “Never worked”, “Private” or “Self-employed”
    - Residence type: “Rural” or “Urban”
    - avg\_glucose\_level: average glucose level in blood
    - bmi: body mass index
    - smoking status: “formerly smoked”, “never smoked”, “smoked” or “Unknown”
    - stroke: 1 if the patient had a stroke or 0 if not

# Methodology

I decided to tackle the project by the help of a machine learning model. So I focused on creating the Decision Tree model, while also exploring the dataset and implement this model. Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable.

## Exploratory Data Analysis (EDA)

Before I started creating any models, I worked on the Exploratory Data Analysis (EDA) of the stroke risk dataset. For EDA, I plotted pie chart to visualize the data and the correlation between the features. Through EDA, I learned that there were some missing values for certain features, unnecessary columns, categorical variables, and most importantly, that there was an imbalance in the data in terms of the stroke feature.



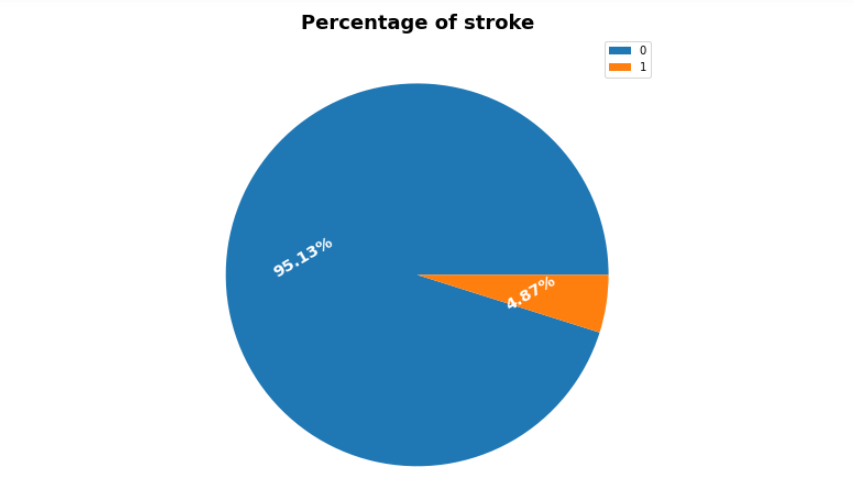
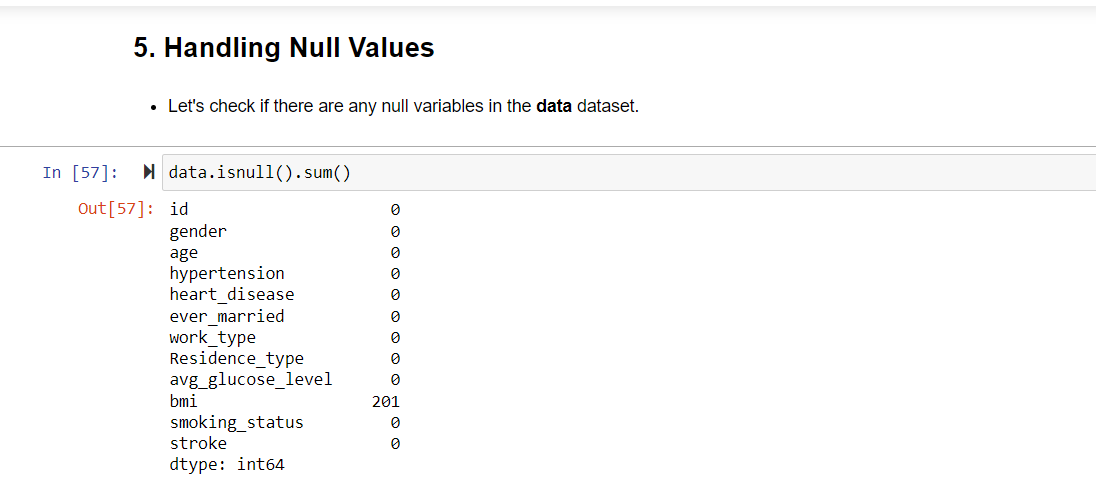
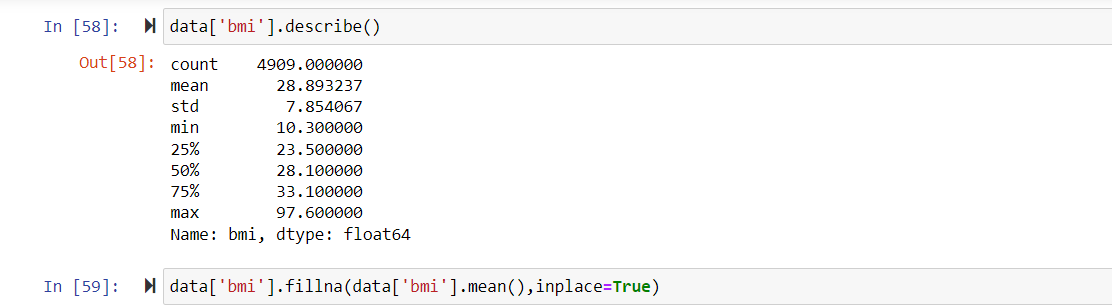
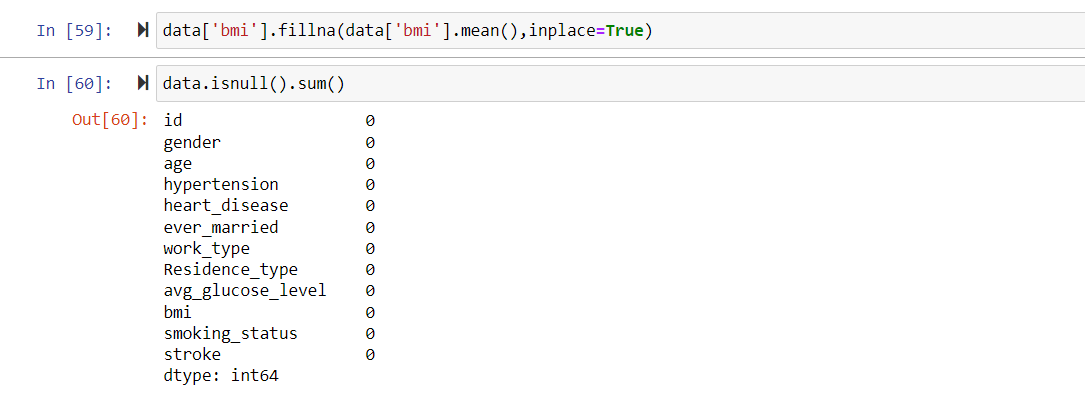


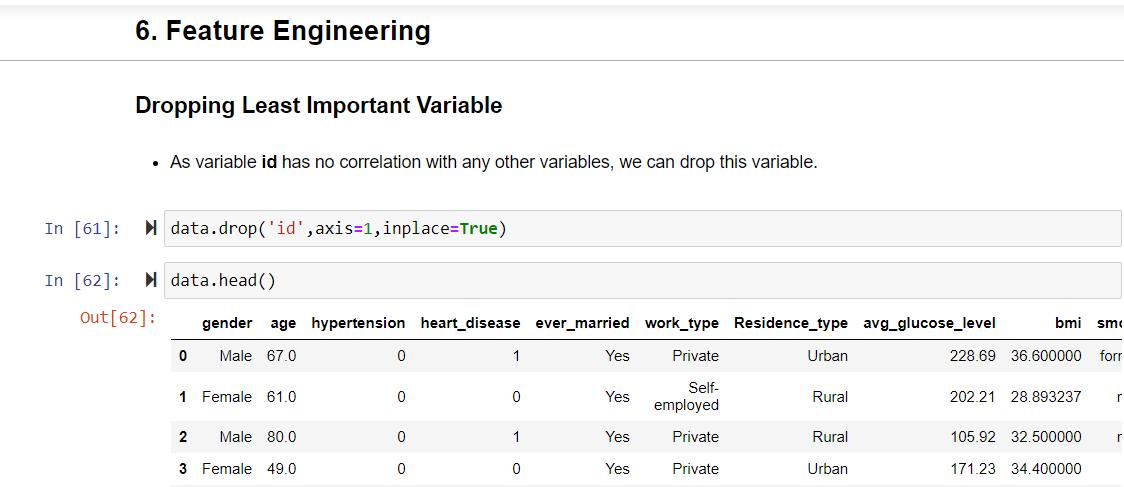
Figure 1: Stroke Percentage

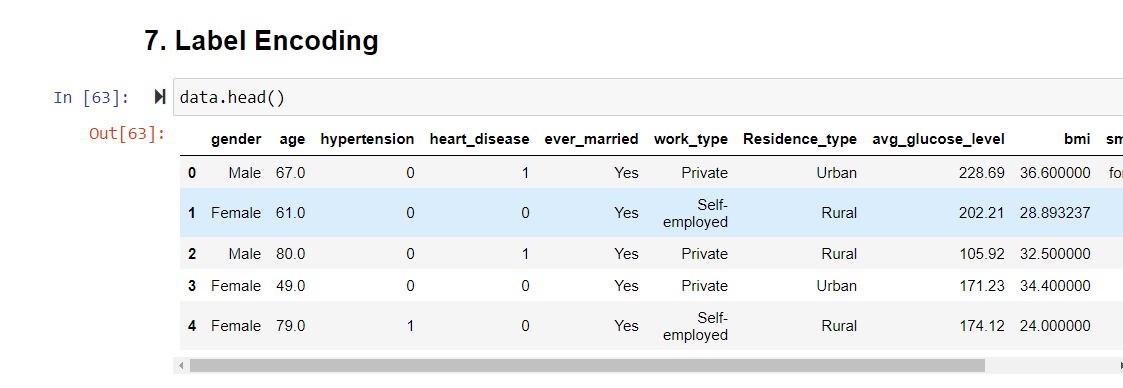
To fix these issues, I performed feature engineering [[3](#_bookmark52)]. I filled the missing values with that features average result and discarded any unnecessary columns. Then I changed categorical data to numerical data to fix the imbalance issue.

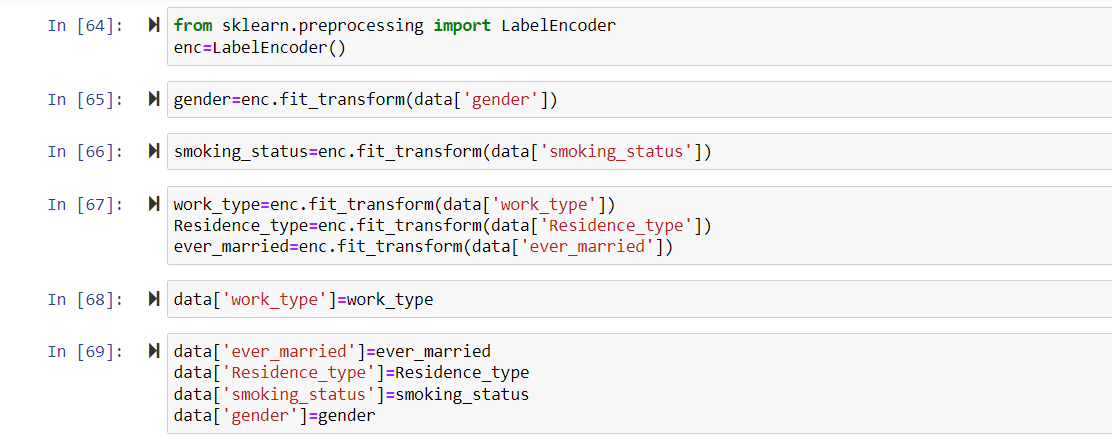


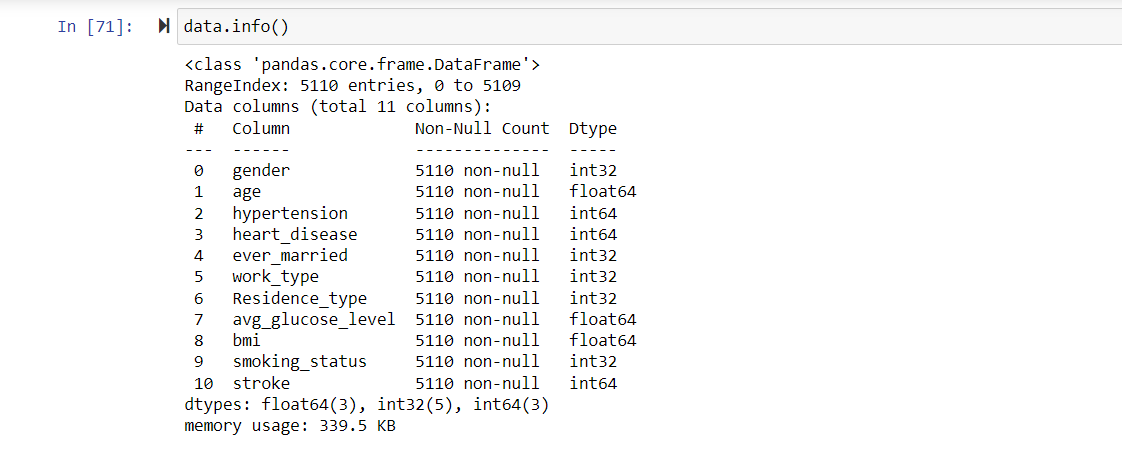












## Picking Models

With the feature engineering and (EDA) of the data, we are ready to proceed. Given the nature of the data, we are facing a supervised learning problem, more specifically a classification problem. Given the severity of the issue at hand, we wanted to compare some more basic classification algorithms, like Logistic Regression [[4](#_bookmark53)], and see how it would fare against different discriminative algorithms such as SVMs [[5](#_bookmark54)] and Decision Trees. In addition, we wanted to find out which features mattered the most in this case, which is the Random Forest algorithm was picked out of all possible Decision Tree based algorithms.

# Methods

## Logistic Regression

The Logistic Regression Model is a statistical model that is used to solve classification problems and basically models the probability of a certain class or event happening. In this case, I will be using it to predict whether a person with certain features will get a stroke or not.

## Support Vector Machines

The Support Vector Machine (SVM) is a supervised learning model that can be used to solve classification problems. The SVM model finds a hyperplane that distinctly classifies data points in an *N* -dimensional space, where *N* is the number of features. For our dataset, 10 features are being used to predict whether a person will get a stroke or not. This model chooses the best hyperplane that separates the data into positive and negative values and is the furthest away from the closest data points in order to classify the data points.

## Random Forest

Random forest, as the “forest” part of the name implies, consists of a large number of individual Decision Trees (DTs) that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. For some background, we picked this route because DTs, while still discriminative in nature, perform classification in a manner that is fundamentally different than other algorithms like Logistic Regression and SVMs.

# Conclusion

## Overall Analysis

From our data analysis people over the age of 50 are more likely to get a stroke compared to people under the age of 50. Individuals with hypertension (high blood pressure) are more likely to get a stroke. Males are more prone to a stroke compared to females. Stroke is a life-threatening medical illness that should be treated as soon as possible to avoid further complications. Development of an ML model could aid in the early detection of stroke and the subsequent mitigation of its severe consequences. Effectiveness of several ML algorithms in properly predicting stroke based on a number of physiological variables is investigated in this study.

Figure [2](#_bookmark48) shows us the importance of the different features:

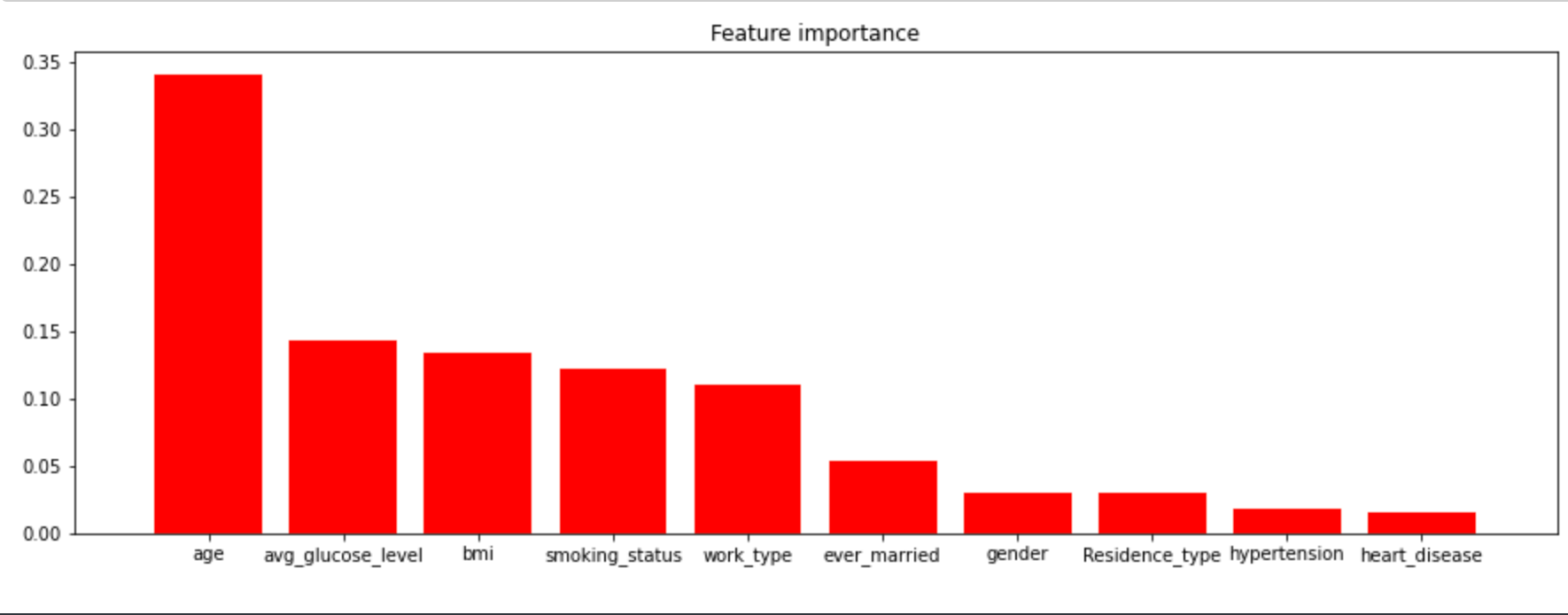


Figure 2: Feature Importance

## Future Work

For future work, I would like to explore the importance of various features to help filter features that are irrelevant or redundant. This would help improve the models which are used to predict whether a patient is likely to have a stroke. Some extra conclusions I determined were that, apart from a lot of categories that do not make sense for this context, some of the metrics are flawed. BMI in general is not a good metric to use, so it would be better to measure weight [[6](#_bookmark50)]. Also, we are unsure if smoking is related to tobacco, cannabis, crack cocaine, or any other drug that can be inflamed and inhaled. Nonetheless, we see that age is clearly the biggest contributor to one’s likelihood to suffer a stroke by far.

# References

1. World Stroke Day. October 2021. URL: <https://www.who.int/srilanka/news/detail/29-10-2021-world-stroke-day-2021-launches-the-precious-time-campaign>
2. Fedesoriano. Stroke Prediction Dataset. Technical report, Februrary 2021. URL: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
3. Aditya Khosla, Yu Cao, Cliff C. Lin, Hsu-Kuang Chiu, Jungling Hu, and Honglak Lee. An Integrated Machine Learning Approach to Stroke Prediction. Technical report, July 2010. URL: <https://people.csail.mit.edu/khosla/papers/kdd2010.pdf>
4. Andrew Ng and Tengyu Ma. CS229 Notes 1. Pages 14–17, Fall 2020. URL: <https://cs229.stanford.edu/notes2021fall/cs229-notes1.pdf>
5. Andrew Ng and Tengyu Ma. CS229 Notes 3. pages 11–17, Fall 2020. URL: <http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes3.pdf>
6. Brock Armstrong. Is Bmi an Accurate Way to Measure Body Fat? Technical report, June 2019. URL: <https://www.scientificamerican.com/article/is-bmi-an-accurate-way-to-measure-body-fat/#:~:text=Even%20though%20it%20is%20often,far%20from%20a%20perfect%20measurement.>
7. Tony Yiu. Understanding Random Forest. Technical report, June 2019. URL: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>