**CMP6202**

**AI and Machine Learning Project**

**2023–2024**

Individual Report

Stroke Prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Forename** | **Surname** | **Student ID** | **Instructor Name** | **ML Model(s) developed** |
| Aatif | Mahmood | 21140384 | Khalid Ismail | Naïve Bayes |
| Sohail | Rafiq | 21116869 | Khalid Ismail | Decision Tree |
| Mohammed | Alam | 21142816 | Khalid Ismail | Logistic Regression |
| Gurpreet | Singh | 21131818 | Khalid Ismail | KNN |

Table of Contents

[1 Report Introduction 3](#_Toc156174330)

[1.1 Dataset identification 3](#_Toc156174331)

[1.2 Supervised learning task identification 4](#_Toc156174332)

[2 Exploratory Data Analysis 4](#_Toc156174333)

[2.1 Question(s) identification 4](#_Toc156174334)

[2.2 Splitting the dataset 5](#_Toc156174335)

[2.3 Exploratory Data Analysis process and results 5](#_Toc156174336)

[2.4 EDA conclusions 5](#_Toc156174337)

[3 Experimental Design 5](#_Toc156174338)

[3.1 Identification of your chosen supervised learning algorithm(s) 5](#_Toc156174339)

[3.2 Identification of appropriate evaluation techniques 5](#_Toc156174340)

[3.3 Data cleaning and Pre-processing transformations 6](#_Toc156174341)

[3.4 Limitations and Options 7](#_Toc156174342)

[4 Predictive Modelling / Model Development 7](#_Toc156174343)

[4.1 The predictive modelling process 7](#_Toc156174344)

[4.2 Evaluation results on “seen” data 7](#_Toc156174345)

[5 Evaluation and further modelling improvements 7](#_Toc156174346)

[6 Conclusion 9](#_Toc156174347)

[6.1 Summary of results 9](#_Toc156174348)

[6.2 Reflection on Individual Learning 9](#_Toc156174349)

[7 References 9](#_Toc156174350)

# Report Introduction

The report aims to investigate stroke diagnosis using different machine learning models, determining stroke possibility by using a variety of different factors. “According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths” (Fedesoriano, 2021).

This report will produce clarification of implementation methods and essentially make a valid conclusion based upon the results presented by the different models which will be created in google Colab’s work environment.

## Dataset identification

Initialising the project, we started to gather information on different datasets and filter different datasets for potential suitors and to determine the area in which the project was going to focus. At a later stage confirmed as a “Stroke prediction project”. There were 3 datasets we narrowed the search to, and at the end one remained.

Our dataset is titled "Stroke Prediction Dataset" (Fedesoriano, 2021). It has many factors/attributes such as potential health risks, lifestyle choices or other personal details of patients and whether they had a stroke, these are seen in the table below. It includes 5,110 records and 12 columns (Fedesoriano, 2021).

The dataset was obtained from Kaggle, the source is confidential, the author Fedesoriano is a data scientist at Kaggle who is based in Madrid, Spain. It can be considered reliable because it has a usability score of 10, has been used over 1000 times and has over 2800 upvotes. It is 3 years old so fairly up to date.

|  |  |  |
| --- | --- | --- |
| 1 | id | A unique numerical identifier |
| 2 | gender | The gender of the user which is “Male”, “Female” and “Other” |
| 3 | age | The age of the user as an integer |
| 4 | hypertension | A binary attribute, 0 if the patient doesn't have hypertension, 1 if the patient has hypertension |
| 5 | heart\_disease | A Boolean "No" or "Yes" as to whether the patient has heart disease |
| 6 | ever\_married | A Boolean "No" or "Yes" as to whether the patient has ever been married |
| 7 | work\_type | The type of work the patient does which can be "children", "Govt\_job", "Never\_worked", "Private" or "Self-employed" |
| 8 | residence\_type | What environment they live in, either “Rural” or “Urban” |
| 9 | avg\_glucose\_level | Average glucose level in blood |
| 10 | bmi | Body mass index is a value derived from the mass and height of a person |
| 11 | smoking\_status | "formerly smoked", "never smoked", "smokes" or "Unknown” |
| 12 | stroke | A binary attribute, 1 if the patient had a stroke or 0 if not |

### Attribute information

Below are figures 1-10, which showcase the data which is in the dataset and how it is structured. These figures also aid to visualise the data contained in each column. There are assumptions that can be made looking at this data which will later be discussed.

A blue rectangular object with green text

Description automatically generated

Figure - data and visual representation for the attribute "id" (Fedesoriano,2021).

A screenshot of a graph

Description automatically generated

Figure - statistics for the attribute “gender”(Fedesoriano,2021).

A graph with blue and green bars

Description automatically generated

Figure - statistics and visual representation for the attribute "age" (Fedesoriano,2021).

A graph with a green and blue line

Description automatically generated with medium confidence

Figure - statistics and visual representation for the attribute "hypertension" (Fedesoriano,2021).

A graph with a bar and a green line

Description automatically generated with medium confidence

Figure - statistics and visual representation for the attribute "heart\_disease" (Fedesoriano,2021).

A screenshot of a survey

Description automatically generated

Figure - statistics and visual representation for the attribute "ever\_married" (Fedesoriano,2021).

A screenshot of a graph

Description automatically generated

Figure - statistics for the attribute "work\_type" (Fedesoriano,2021).

A graph with text and numbers

Description automatically generated with medium confidence

Figure - statistics for the attribute "residence\_type" (Fedesoriano,2021).

A screenshot of a graph

Description automatically generated

Figure - statistics and visual representation for the attribute "avg\_glucose\_level" (Fedesoriano,2021).

A graph with text and numbers

Description automatically generated with medium confidence

Figure - statistics for the attribute "residence\_type" (Fedesoriano,2021).

## Supervised learning task identification

The aim of this project is to explore the dataset chosen and predict how each attribute has an effect on the stroke diagnosis. Essentially, we aim to correctly predict if someone can have a stroke using the attributes as deciding factors.

### Model selection

The model chosen for the report is a classification model as *the ground truth* variable is “stroke” which is a categorical variable hence a regression model would not be practical even though with some modification to the dataset would be possible only for this type of prediction system. But in this instance as the variable would not predict a continuous numerical value it is not possible for a regression model hence classification was chosen.

A screenshot of a computer

Description automatically generated

Figure - Using "df.head()" to show the dataset and how the stroke variable is categorical with binary data in its column.

# Exploratory Data Analysis

“Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods “ (*What is exploratory data analysis?*).

## Question(s) identification

|  |  |
| --- | --- |
| Question number | Question |
| 1 | How many people have had a stroke in the dataset? |
| 2 | Do older people have a higher chance of having a stroke than younger people? |
| 3 | Do people who smoke have a higher chance of being diagnosed with a stroke than people who don’t smoke? |
| 4 | How does smoking affect the possibility of getting a stroke at different ages? |
| 5 | Does the type of work someone does affect whether they could have a stroke? |
| 6 | Does being married increase the likelihood of getting a stroke? |
| 7 | Does having a high average blood glucose level possibly lead to stroke? |
| 8 | Does having a heart disease increase the chances of having a stroke? |
| 9 | Does having a higher hypertension make it more possible to have a stroke? |

## Splitting the dataset

To split the data we used Scikit Learn’s train\_test\_split() module. This module was used to randomize the split so that no errors are made whilst training the machine learning (ml) model. The split is into two subsets called X and Y, where X is the feature and Y being the target (“stroke”). We used this method as we felt we had enough records in the dataset (approximately 5000+) which allowed us to do an 90/10 split of the dataset between the intermediary and testing dataset, which then led to the split of the intermediate dataset into the training dataset and validation dataset a 90/10 split again.

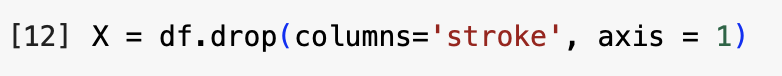


Figure - code snippet used to define X dropping the "stroke" attribute.



Figure code snippet used to define Y.

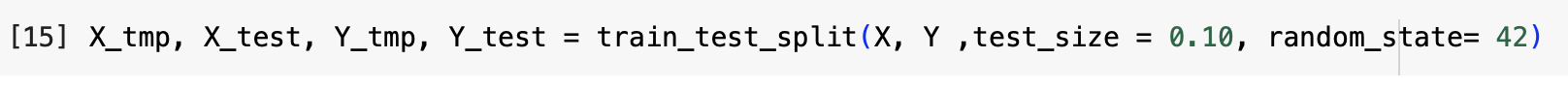


Figure - Code snippet used for the first split.



Figure - code snippet used for final split.

Test\_size was set to 0.1 signifying a 90/10 split and the random\_state parameter was left at the standard as not to make too many changes to the dataset.

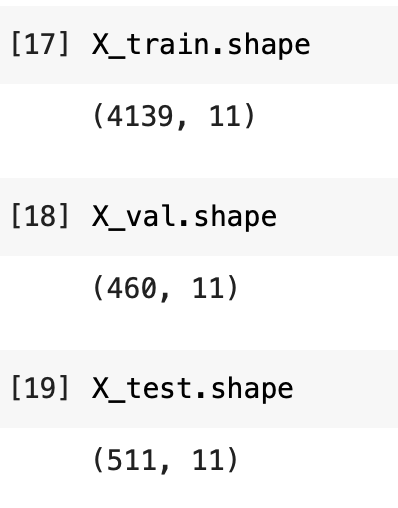


Figure - showing the values contained in each dataset after splitting.

## Exploratory Data Analysis process and results

To leave all unseen data untouched and limit other implications the training dataset alone was used for the EDA. To perform the EDA, we needed to go through some processes to better understand our knowledge on the dataset and the attributes within the dataset. External libraries were used to create visual representations and each representation was chosen in respects to the question.

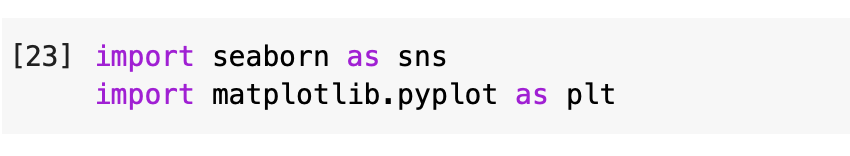


Figure - libraries imported for visual representation.

### Q1

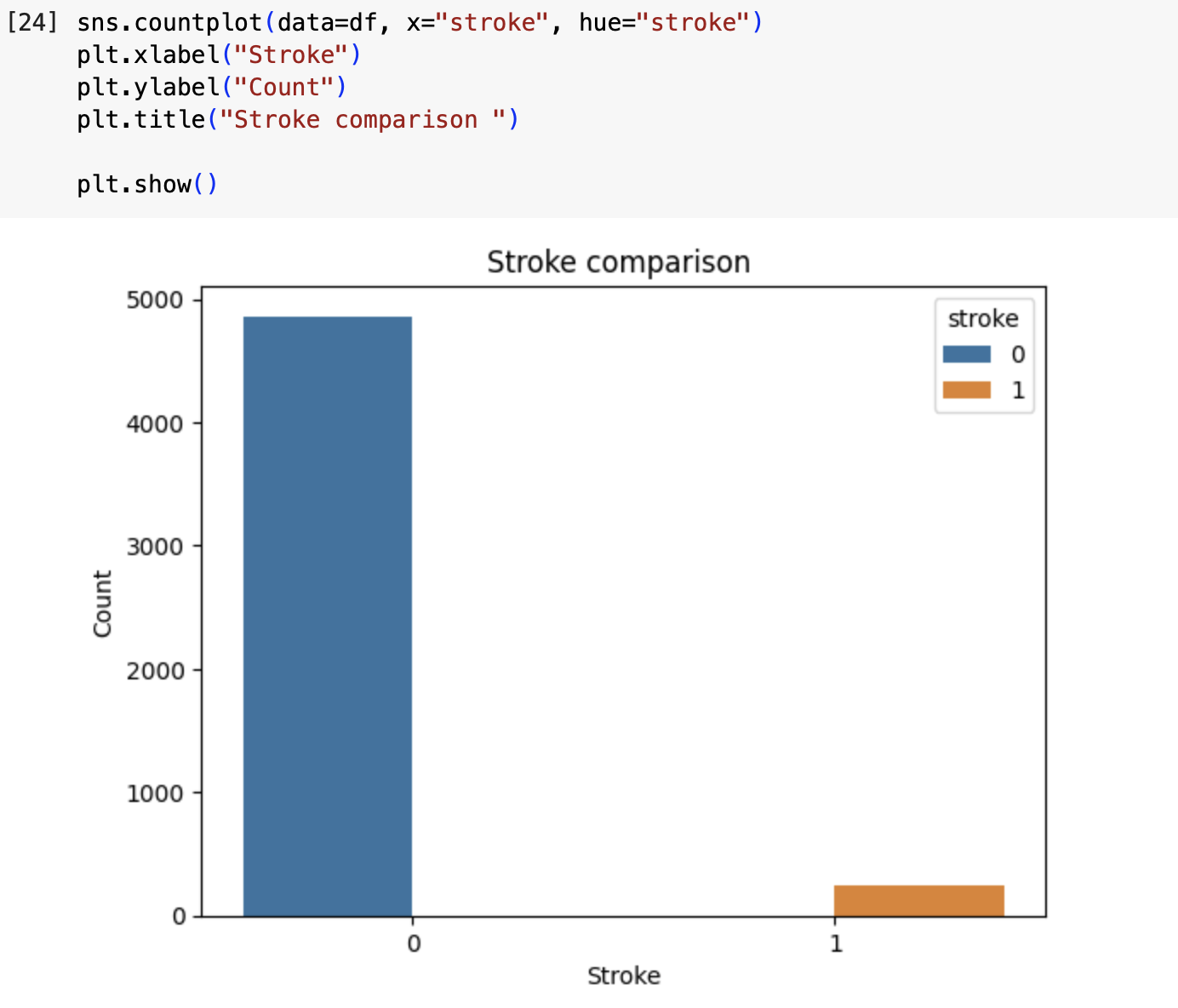


Figure - code snippet and bar chart showcasing results.

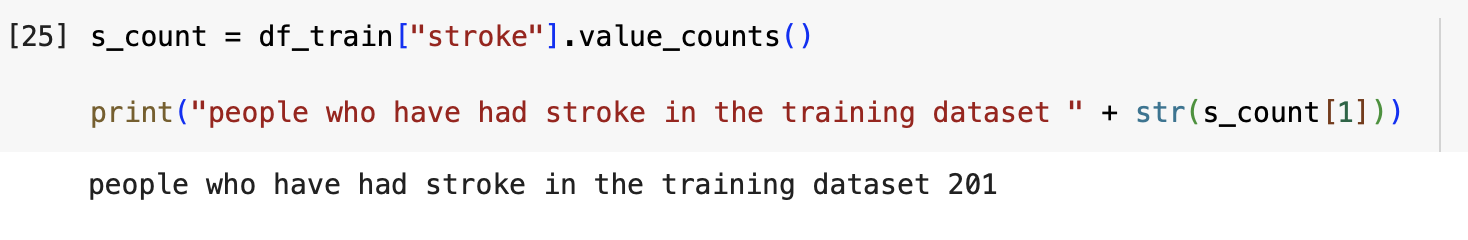


Figure - Code snippet to identify how many people in the training dataset have stroke.

### Q2

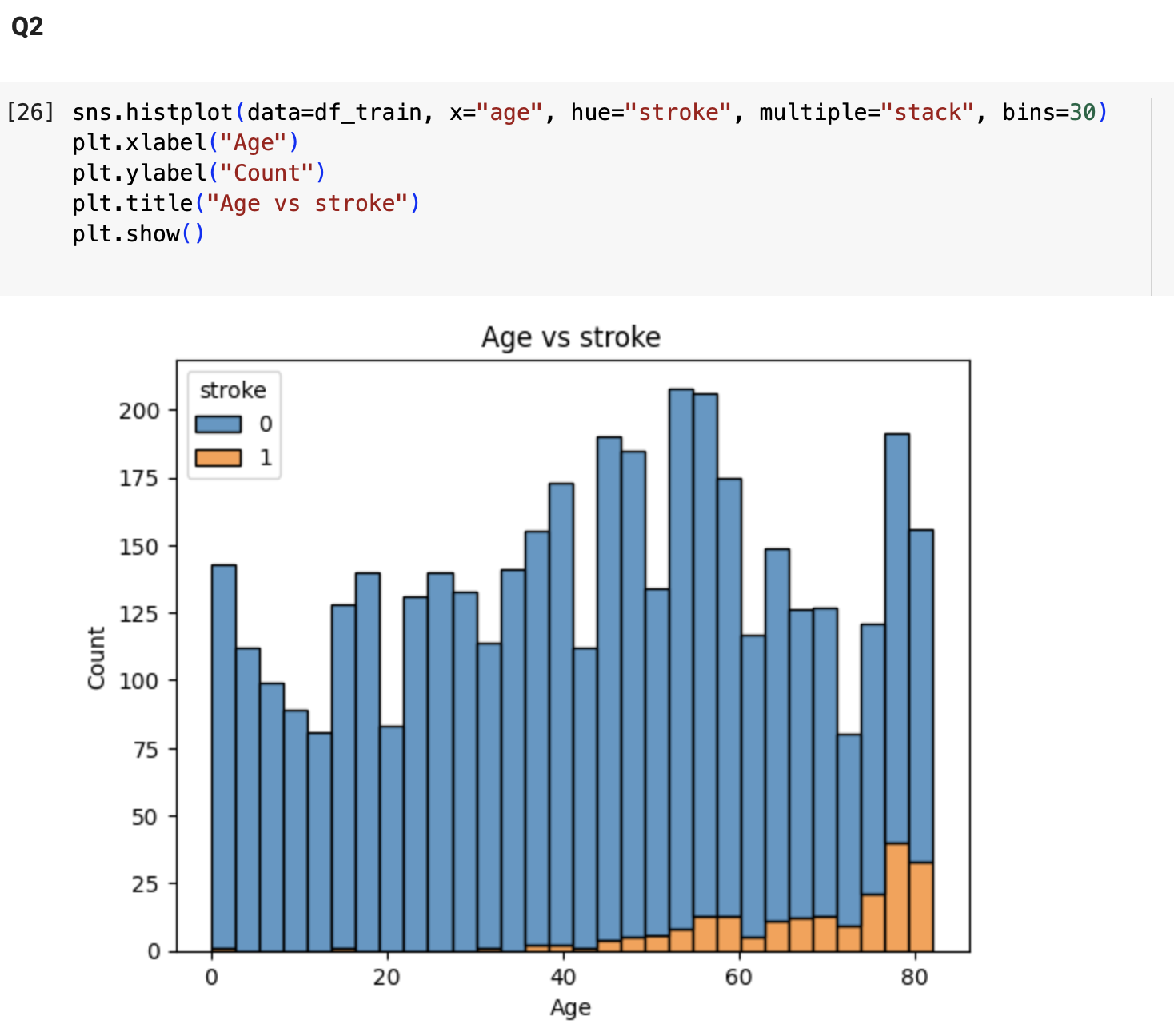


Figure - code snippet and histogram showcasing results of age vs stroke.

### Q3

A screenshot of a computer screen

Description automatically generated

Figure -- code snippet and histogram showcasing results of smoking status vs stroke.

### Q4



Figure - - code snippet and box plot showcasing results of smoking status vs age vs stroke.

### Q5

A screenshot of a computer screen

Description automatically generated

Figure - code snippet and bar chart showcasing results of different work types and how many people have had stroke all data is exclusively of people who have had a stroke..

### Q6

A screen shot of a graph

Description automatically generated

Figure - code snippet and bar chart showcasing results of marriage vs stroke.

### Q7

A graph of a bar graph

Description automatically generated with medium confidence

Figure - code snippet and histogram showcasing results for stoke diagnosed people's glucose levels.

### Q8

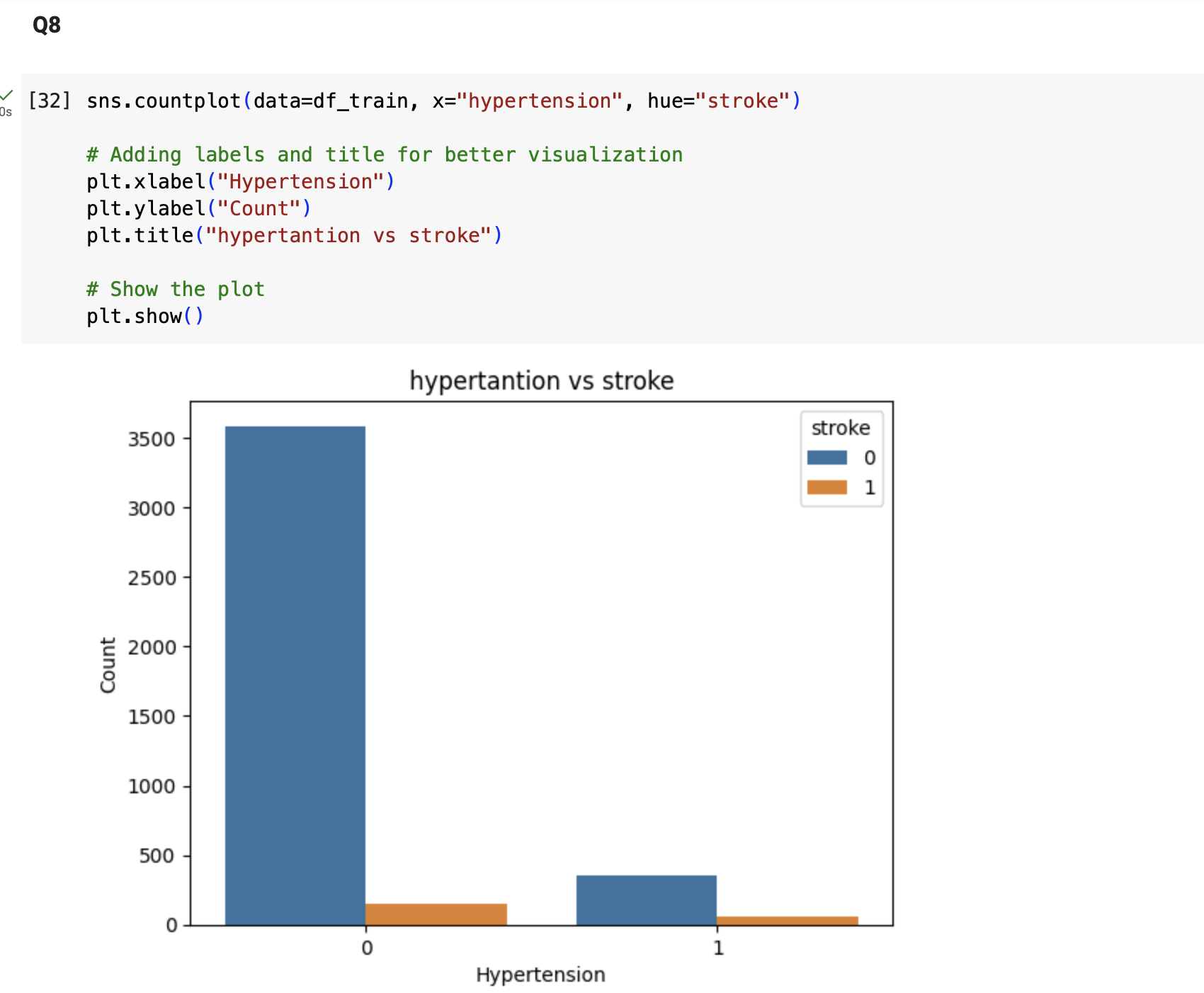


Figure - code snippet and bar chart for hypertension vs stroke.

A screenshot of a computer screen

Description automatically generated

Figure - code snippet and histogram showcasing results for stroke diagnosed participants and they're probability of also having hyper tension.

## EDA conclusions

The data is highly unbalanced as seen by the results, this quickly showed that the real-time application of this dataset would not be accurate.

|  |  |  |
| --- | --- | --- |
| Question number | Question | Answer |
| 1 | How many people have had a stroke in the dataset? | 201. |
| 2 | Do older people have a higher chance of having a stroke than younger people? | It is shown that older people are more at risk. |
| 3 | Do people who smoke have a higher chance of being diagnosed with a stroke than people who don’t smoke? | Smoking is not showing much affect as stroke participants are fairly distributed on the diagram. |
| 4 | How does smoking affect the possibility of getting a stroke at different ages? | It has a negative affect to people in higher age spectrum. |
| 5 | Does the type of work someone does affect whether they could have a stroke? | More people who are in private sector and self-employed have a risk of stroke. |
| 6 | Does being married increase the likelihood of getting a stroke? | Yes. |
| 7 | Does having a high average blood glucose level possibly lead to stroke? | No. |
| 8 | Does having a heart disease increase the chances of having a stroke? | Unexplored. |
| 9 | Does having a higher hypertension make it more possible to have a stroke? | Yes. |

# Experimental Design

## Identification of your chosen supervised learning algorithm(s)

I have chosen to experiment with the KNN model as it is the most simplistic machine learning model.

## Identification of appropriate evaluation techniques

The evaluation techniques are:

* Accuracy score
* Confusion matrix
* Classification report

## Data cleaning and Pre-processing transformations

Data cleaning and pre-processing is essential for the machine learning model as without it implementation would not be possible. Pre-processing can be seen accessing the scripts on Google Colab. It is advised to run the testing pre-processing at the end for better results.

Handling Missing Values:

* We found Unknown values in the "Smoking\_status" column of the dataset, which we will need to address. The options discussed include imputation or removal of rows with missing values and the choice was made to remove rows with missing value as shown below due to it minimalizing the data without affecting the dataset making it more reliable.

Encoding Categorical Variables:

* We found two attributes, " ever\_married " and "work\_type", which are categorical values. The variables need to be encoded into numerical format before use in machine learning algorithms. Options include one-hot encoding and label encoding. The choice was made to use one-hot encoding.

Balancing the Target Variable:

* Our datasets target variable, “stroke” is unbalanced, with a 4861 of non-stroke patients compared to 249 stroke patients which could lead to a lack of data for the predictive model. This imbalance will need to address during the training of the model. We have chosen to oversample to balance the variable.

There are many missing values in the data which will be resolved by removing said values. We have removed 164 values for BMI and smoking\_status

## Limitations and Options

After identifying the pre-processing steps, we discussed any potential challenges that would need to be resolved. These are detailed below.

Imbalanced Dataset:

* Without addressing this issue the Model could exhibit bias towards variables with a large majority which would result in poor predictions for the outcome of a stroke.

Handling Categorical Variables:

* Encoding categorical variables may also exhibit bias, and consideration must be made towards the choice of encoding methods. If not addressed the model may not be accurate in predicting the outcome of a stroke.

Feature Selection:

* The dataset contains many attributes, of which the most relevant attributes need to be identified for our machine learning model to ensure good performance.

# Predictive Modelling / Model Development

## The predictive modelling process



Figure - importing machine learning model from sklearn libraries.

No additional parameters were added as the model was kept as simple as possible.

A screenshot of a computer program

Description automatically generated

Figure - code snippet used to fit the machine learning model.

## Evaluation results on “seen” data

A screenshot of a computer

Description automatically generated

Figure - code snippet for importing evaluation metrics and predicting the outcome.

Looking at the results the model is good as according to the accuracy score it is excellent at approx. 95%.

# Evaluation and further modelling improvements

As per the evaluation the constant k of the training model should be changed using cross validation and changed so that the best accuracy score is implemented, and the best outcome is made. Also, one hot encoding should not be used as it negatively affects the dataset and there is a huge loss of data. This is seen when pre-processing testing dataset.

Balancing the Target Variable:

Our datasets target variable, “stroke” is unbalanced, with a 4861 of non-stroke patients compared to 249 stroke patients which could lead to a lack of data for the predictive model. This imbalance could be addressed during the training of the model such as oversampling to balance the variable.

Moreover, cross validation should be used for the validation testing at the end to ensure that the model works more efficiently.

# Conclusion

## Summary of results

The results were only on seen data due to time constraints, as the data cannot be verified properly the results may be seen as non-conclusive and require more work to fully comprehend the work.

## Reflection on Individual Learning

Through this module I have learnt the fundamentals of machine learning and improved on my soft skills with the groupwork. However, I am a bit weak in team work as seen by this report and would like to further develop this. Due to time constraints and poor time management on my part this report is not complete, for my further studies I need to attend some workshops on time management and stress management as it hindered the completion of my work.

# References

Fedesoriano (2021b) *Stroke prediction dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (Accessed: 15 January 2024).

*What is exploratory data analysis?* (no date) *IBM*. Available at: https://www.ibm.com/topics/exploratory-data-analysis (Accessed: 15 January 2024).

# Bibliography

Fedesoriano (2021) *Fedesoriano: Grandmaster*, *Kaggle*. Available at: https://www.kaggle.com/fedesoriano (Accessed: 15 January 2024).

Norrving, B., Davis, S.M., Feigin, V.L., Mensah, G.A., Sacco, R.L. and Varghese, C., 2015. Stroke prevention worldwide-what could make it work. *Neuroepidemiology*, *45*(3), pp.215-220.

Thayabaranathan, T., Kim, J., Cadilhac, D.A., Thrift, A.G., Donnan, G.A., Howard, G., Howard, V.J., Rothwell, P.M., Feigin, V., Norrving, B. and Owolabi, M., 2022. Global stroke statistics 2022. *International Journal of Stroke*, *17*(9), pp.946-956.