HER TECH TRAIL

REPORT ON STROKE PREDICTION ANALYSIS

Using stroke indicators

A project by a group 1 HTT Cohort 4 (Data Analytics Pro track)

TABLE OF CONTENTS

1 INTRODUCTION

This report was made as a final capstone project by a group in the Data Analytics of HerTechTrail Community. The project centres on stroke prediction considering some of the key indicators.

The importance of this analysis is to help individuals who are not living with stroke identify the risk factors they already have, and how to manage them in order to prevent having it in the future.

1.1 AIMS AND OBJECTIVES

This report is provided to give actionable insights from the analysis of the data gathered concerning stroke patients.

The insights obtained will help us know the risk factors associated with stroke. Some of these factors can be modified while some cannot be modified. Risk factors for stroke that can be modified include hypertension, heart disease, glucose level, lifestyle (smoking), and body mass index. The aim of our project is to analyse an existing dataset to effectively predict stroke based on these modifiable risk factors.

This analysis will help alert individuals at risk on how to maintain a healthy lifestyle in order to avoid these risk factors.

1.2 ROADMAP OF THE REPORT

The report is in four sections representing the different phases in the analysis.

- ❖ **Background** - this section gives an insight into the project topic, it documents the nature of analysis, the target audience and how the analysis done can be useful.
- ❖ **Step specifications** - this section describes how the team approached each of the key steps of the data analysis as well as detailing the data sources.
- ❖ **Implementation and Execution** - this section focuses on the development approach, team member roles, tools and libraries, implementation process and challenges.
- ❖ **Result Reporting** - our key findings from the analysis done is documented in this section
- ❖ **Limitations** - this documents the constraints in the project
- ❖ **Conclusion -** the inferences drawn from the analysis are documented here.

2 BACKGROUND

Stroke is ranked second leading cause of death according to the World Health Organisation{WHO}. It is a major public health concern affecting millions of people worldwide and it is responsible for approximately 11% of deaths. Stroke occurs due to a decrease or blockage in the brain's blood supply when a blood vessel that carries oxygen and nutrient to the brain is either blocked by a clot or burst, when that happens part of the brain cannot get the blood and oxygen it needs which result to the death of the brain.

It is also said to be the common cause of long term disability. This remains a health burden for the individuals and the National Health care system which should not be overlooked. Stroke awareness is very important to aid its prevention.

This dataset contains records from both ischemic and hemorrhagic strokes which will assist in predicting the likelihood of a patient having a stroke based on results obtained from the analysis, timely detection and prevention has become very essential to avoid its adverse consequences.
For the purpose of this report, the target audience will be stroke enthusiasts. This is to ensure that they have the right information and resources required to educate the public.

3 STEPS SPECIFICATIONS

This section of the report documents the approach to the project. Particularly, this includes;

- **Data gathering**
  Based on a unanimous vote on the sector to carry out analysis on, we considered different secondary data sources in the health sector but settled for the stroke prediction data set gained from kaggle which is a secondary data source.
  https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

- **Data Source:**
  A data set is a structured collection of data points related to a particular subject. For this analysis, the data set was retrieved from Kaggle. Kaggle is a well known Machine Learning and Data Science community, which is made accessible for community data and code.

- **Preprocessing**
  It is a zipped file which was extracted and saved locally for easy access. We opened the file using MS Excel, because it came as a single column of text, we splitted it into its various columns.

- **Framing Questions**
  Having decided on which sector to choose the dataset from, we went through the dataset chosen individually and generated different questions for the analysis. From those questions gathered, we decided on the most relevant questions to work with.

- **In-depth analysis**

  We surveyed the data set for unusual patterns after understanding the problem it is designed to solve. Understanding the data and the information it provides was of great importance to us.

  The gathered data is a clean dataset that contains information of 5,110 patients having 12 common attributes. Out of the 12 attributes, 10 of them are input features (independent variables) including age, gender, marital status, patient identifier, work type, residence type (urban/rural), binary attribute heart disease condition indicating a patient has heart disease or not, body mass index, smoking status of patient, glucose level and binary attribute hypertension indicating a patient is suffering from hypertension or not. The 12th attribute is the binary output attribute indicating a patient suffered a stroke or not representing the dependent variable.

4 IMPLEMENTATION AND EXECUTION

- **Development approach and team member roles**

  For excellent teamwork, we divided the tasks into three main categories i.e coding and visualisation, report writing and slide decks. We employed the perks of specialisation and allowed each individual work where their strength lies. We had our deliberations prior to working on the dataset through google meet and a WhatsApp platform. This aided free flow of information between members and we were able to successfully assign roles.

- **Tools and libraries**

  For ease of collaboration, all our project files were saved on google drive.

  Our raw dataset was downloaded in a csv format and we used **Excel** to separate it into columns. After which this was imported into a **google colab notebook** for coding. **Python** was adopted as our programming language for this analysis.

  We kicked off by cleaning our data. We dropped some null values, assigned the appropriate data type where necessary and added columns to group continuous data for easy analysis.

  We import libraries such as **pandas, seaborn and matplotlib** for effective analysis and visualisation.

  For effective communication of our findings to the target audience, we came up with this report using google docs and a presentation on google slides.

- **Implementation process (achievements, challenges, decision to change something)**

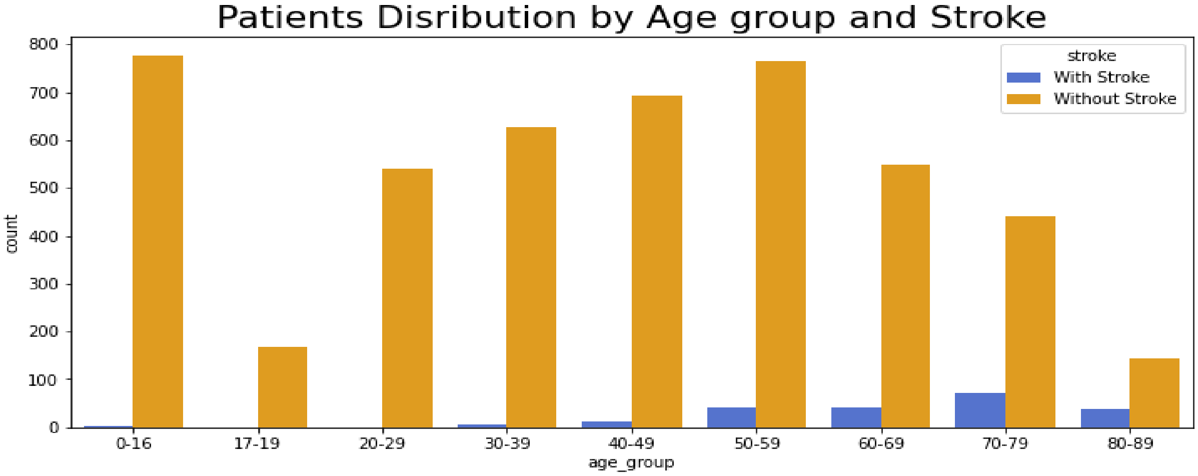
  We had a few setbacks working virtually as a result of unsaved progress made by individuals in the team due to the network. This affected the deadline set by the team but we were able to set up ad hoc meetings to still meet up with the deadline.

  The highlight of this project for us would be the use of Python for data cleaning, data wrangling, data analysis and data visualisation.

5 RESULT REPORTING

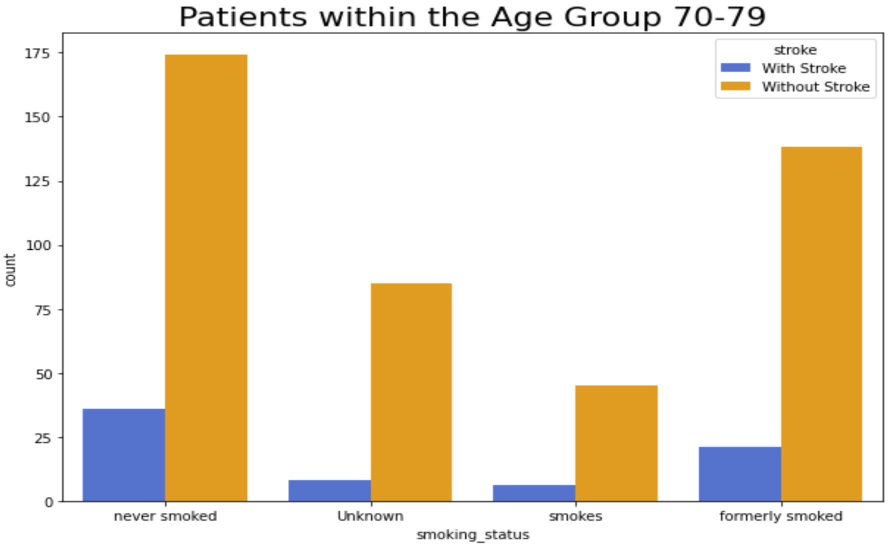
Our decision to work with the stroke dataset was borne out of our concern for the rise in death rates across the globe as a result of stroke. It is slowly becoming a worldwide phenomenon and we decided to analyse the factors that lead to this deadly ailment and subsequently come up with a detailed report on key indicators that impact on stroke .

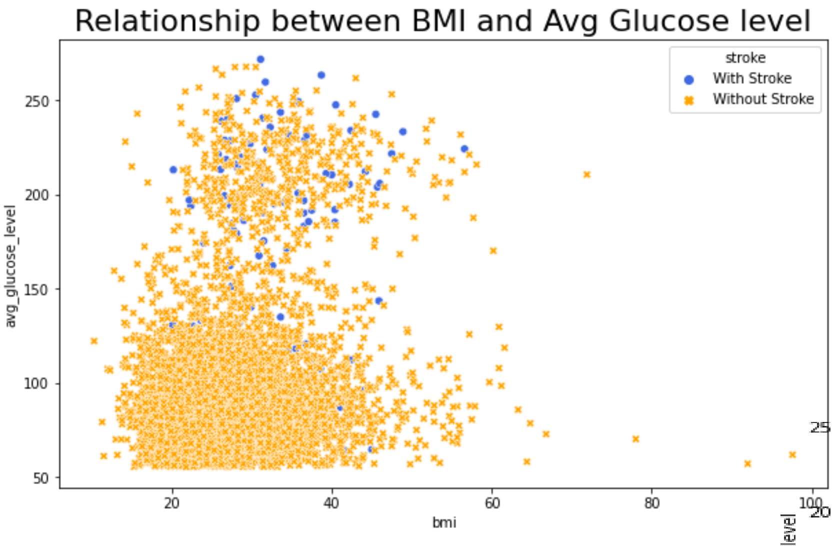
The indicators that were highlighted in the dataset employed includes but are generally not limited to gender, age, hypertension, heart disease, marital status, employment status, average glucose level and bmi.



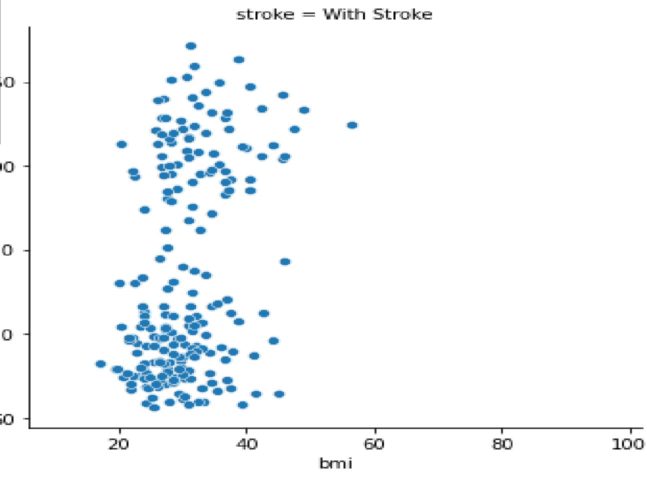
Although it’s quite clear that the age group with the highest occurrence of stroke is the 70-79 dataset, it is pertinent to pay attention to the other factors that we would be addressing once a patient is 40 years of age. Early detection has proven to be a major weapon in fighting stroke.

Sequel to our discovery on the 70 - 79 age group, we further analysed the subset to determine if the smoking history of these patients could have caused their stroke status. However, we can clearly see the category who have never smoked and those who have stopped are topping the charts. This reveals that it is not about their smoking history.

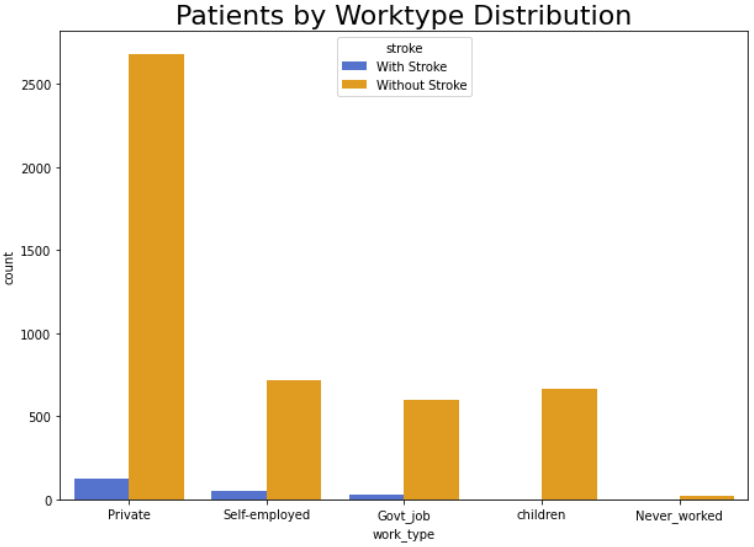




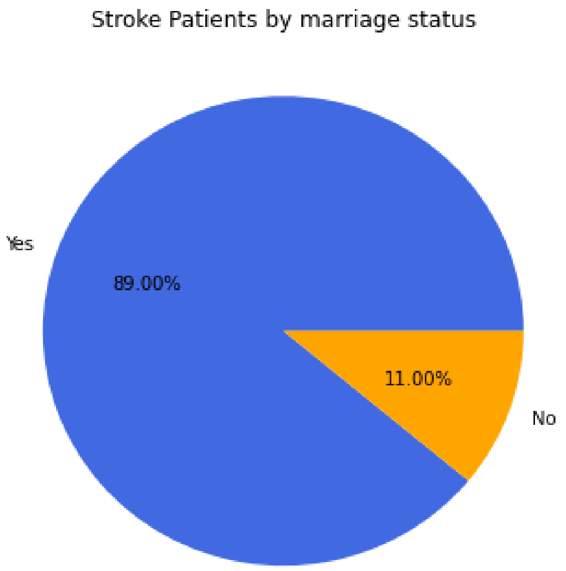
Can we find a relationship between the avg. glucose level and the BMI of a patient? Most patients with BMI between 20-40 have an average glucose level within the range of 55-100 and these categories of patients tend not to have stroke from the concentration in the figure. People with stroke tend to have a higher average glucose level.



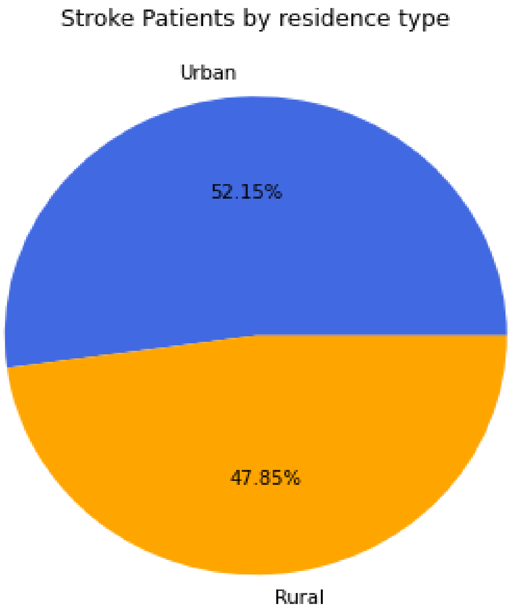
However, from further analysis, it may be difficult to conclude on the relationship. Considering the patients with stroke have similar concentration as those without stroke. It may be imperative for the patients without stroke to be cautious.



Work status: People who have never worked and children have no indication of stroke. It is fair to conclude that working partly has impacts on stroke. Although the proportion of patients with stroke in the private sector is very small compared to those without stroke in the private sector. It is important for people in the private sector to take caution and manage their working habits, prioritise their health considering that the private sector registers more stroke patients compared to the other sectors .

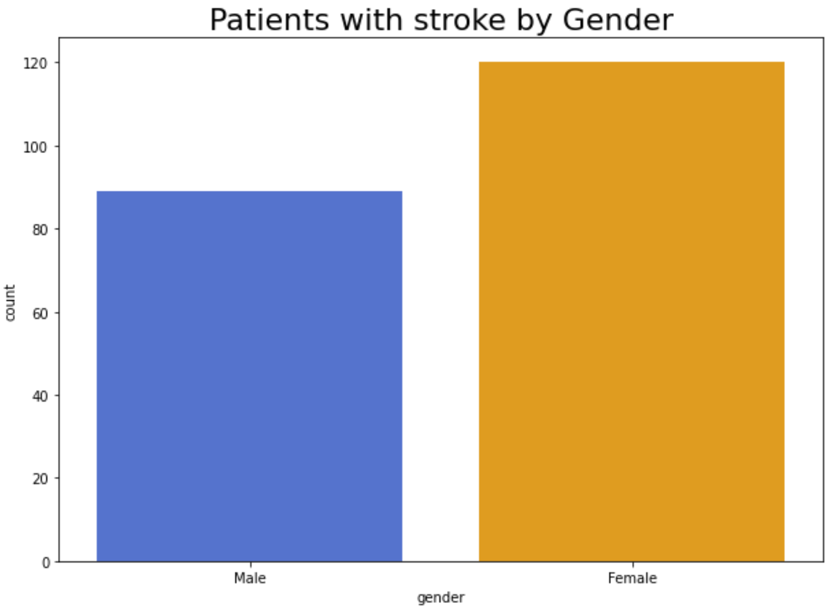


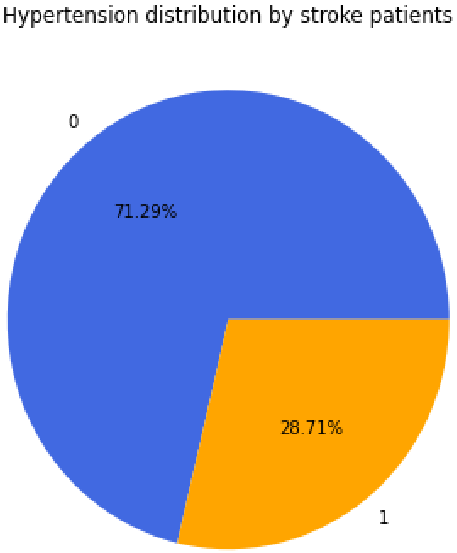
Seeking further analysis on the strictly stroke positive patients with respect to their marital status, a larger proportion (89%) of people with stroke have been married. This definitely does not mean you should shy away from getting married.



Plus, it does not seem like a patient's choice of location has an impact on their stroke status. A patient can decide on where he wants to live based on individual's preference.

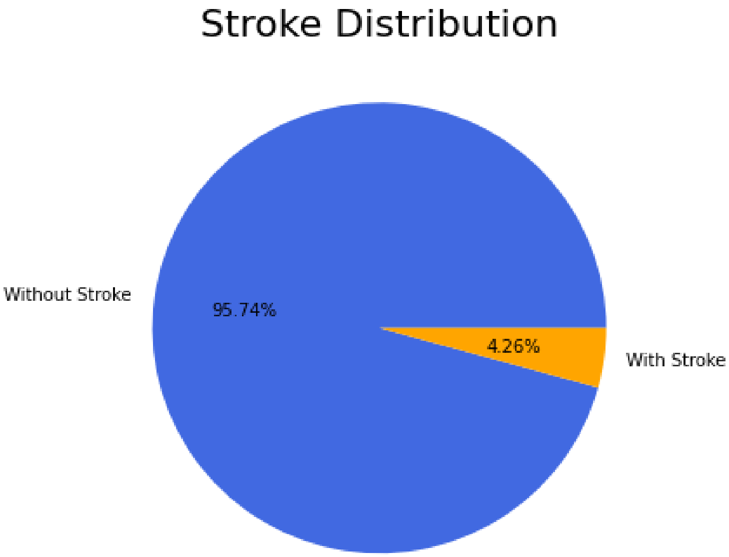
For every 3 male with stroke, there are 4 females with stroke. There's no established relationship between stroke and heart disease.





6 LIMITATIONS

- Strong imbalances exist in the dataset that is provided. Out of **4,909** individuals, only **209** had confirmed strokes, while the remaining **4,700** patient records indicated no evidence of a stroke.
- As noted in Kaggle by the author, the actual source of the data is confidential, As a result of this that, the region the data was gathered is not available. This may impact the conclusions drawn from the dataset. For instance, the climate conditions of a particular region may affect an individual's health thereby impacting on stroke. This may not be so in other regions.
- Also, the data may be specific to the respondents as such it may not be appropriate to generalise the conclusions.
- The project completion was largely constrained on time.



7 CONCLUSION

Existing research on automatic detection of stroke risk through data mining techniques faces a significant challenge in the selection of effective features as predictive cues. Similarly, efficient stroke-detection methods have been increasingly studied in recent years. The performance of the prediction model depends on the choice of key features from the high-dimensional medical dataset.

From the results, it can be deduced that the size of the database can influence the result outcome. It is difficult to conclude on the factors that impact stroke due to the limitations identified.

However, based on the analysis, the following can be concluded:

- There is a correlation between married people and the population of stroke
- Female patients with stroke are slightly more than male patients with stroke.
- Underweight patients are not likely to have stroke.
- Stroke is not common in children.