|  |
| --- |
| **IN3062: Introduction to Artificial Intelligence Coursework**  **Stroke Prediction** |
| By  Abarna Ahilan  Rahim Ahmed  Shajeevan Chandrasekaran |



Contents

[What is your dataset, problem domain? 3](#_Toc87350793)

[Define questions and analysis tasks 3](#_Toc87350794)

[Initial investigation of the dataset and the characteristics 4](#_Toc87350795)

[of the data 4](#_Toc87350796)

[Plan as to how you might transform the data to make it useable 4](#_Toc87350797)

[Is your model classification or regression? 5](#_Toc87350798)

[Did you have any missing, corrupt, or misleading data? 5](#_Toc87350799)

[If so, how did you cope it? 5](#_Toc87350800)

[Have you omitted some data? If so, why? 5](#_Toc87350801)

[Did you apply techniques to understand your dataset? 5](#_Toc87350802)

[What models did you use? 5](#_Toc87350803)

[How did you encode the input variables? 5](#_Toc87350804)

[What are the criteria for selecting model performance evaluation tools? 5](#_Toc87350805)

[What were your outputs? 5](#_Toc87350806)

[Did you have any problems or difficulties working with the dataset? 5](#_Toc87350807)

# What is your dataset, problem domain?

The dataset used for this project will be the dataset from the website known as Kaggle (<https://www.kaggle.com/datasets>). The specific dataset used from Kaggle is the stroke dataset (<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>). The dataset provides information regarding different types of factors which can have an effect on the cause of stroke. It is a large dataset which initially contains 12 attributes and approximately 5000 observations. This means that there is a lot of data to work with.

The problem domain the project is focusing on is predicting stroke based of the different attributes that the dataset presents, and the valuable data included. There are some columns and rows that are required to be removed and edited in order to increase the accuracy of the prediction through means such as removing unnecessary data and replacing null with data acquired from the median amount. This will further increase the accuracy of the prediction regarding the problem domain.

# Define questions and analysis tasks

The domain is focused on predicting stroke based of various factors such as age, gender, diseases and if the person is a smoker or not. These factors can cause stroke. The reason for choosing stroke as the problem domain out of the similar issues is because stroke is one of the highest leading causes of death in the world. By undergoing stroke prediction, doctors can find out the chances of someone getting a stroke in the earlier stages. This will provide them with a foundation to apply more advanced checks on the patients in order to rapidly help them with recovery.

The analytical questions that are being asked in regard to the stroke prediction are:

* What factors can increase the likeliness of stroke?
* What is the cause of stroke in individuals?
* Can chances of stroke be identified at an earlier stage?
* What can people do to reduce the chances of stroke?
* How does smoking effect chances of stroke?

The objectives are as followed:

* fghf

The expected outputs are:

* fghfd
* (Expected table)

Initial investigation of the dataset and the characteristics

of the data

The dataset used includes 12 columns of data in which after discussing, we decided that most was useful data, but some were not required for the prediction as it would not affect the accuracy and precision of the prediction if it were not included. For example, the column which had data on if the person was married or not was not required as stroke is based of many individual factors and not an external factor such as being married. Similarly, the work type of the patient was not required as the data was general and not specific enough to affect the prediction. By being self-employed or working in a private field was not useful as there was no specific jobs. However, the other data were key information regarding predicting stroke such as smoking status, BMI, age, gender, and glucose levels. The characteristics of the data in the dataset is that it consists of numerical figures as well as description. For example, the BMI and if a person is a smoker or not.

# Plan as to how you might transform the data to make it useable

There were some data which were either null or missing. In order to make these useful for the prediction of stroke these data will be transformed. In some cases, there were some figures missing from the body mass index which we decided to use the median as it provides better precision and accuracy rather than using the mean. This will allow the end prediction to be more accurate. The data from the columns which are not required will be removed from the dataset as it can have a negative affluence on the prediction. The data such as the id from the dataset will be removed as it is not useful as we are only focusing on the factors which effect stroke.

**Plan as to which artificial intelligence techniques you might use and what sorts of potential observations these can lead to, and how you will evaluate these.**

**(Leave question until we have solidified what we are doing)**

**K-Fold**

**Train-test split**

**SVM (supervised learning)**

# Is your model classification or regression?

# (Leave question until we are sure of which one, we are using)

# Right now, its classification

# Did you have any missing, corrupt, or misleading data?

# If so, how did you cope it?

# Have you omitted some data? If so, why?

# Did you apply techniques to understand your dataset?

# What models did you use?

# How did you encode the input variables?

# What are the criteria for selecting model performance evaluation tools?

# What were your outputs?

# Did you have any problems or difficulties working with the dataset?

smoking, hypertension, obesity(bmi), age