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



Contents

[What is your dataset, problem domain? 3](#_Toc87028596)

[Is your model classification or regression? 3](#_Toc87028597)

[Did you have any missing, corrupt, or misleading data? 3](#_Toc87028598)

[If so, how did you cope it? 3](#_Toc87028599)

[Have you omitted some data? If so, why? 3](#_Toc87028600)

[Did you apply techniques to understand your dataset? 3](#_Toc87028601)

[What models did you use? 3](#_Toc87028602)

[How did you encode the input variables? 3](#_Toc87028603)

[What are the criteria for selecting model performance evaluation tools? 3](#_Toc87028604)

[What were your outputs? 3](#_Toc87028605)

[Did you have any problems or difficulties working with the dataset? 3](#_Toc87028606)

# 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.

# Is your model classification or regression?

# 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?