STROKE PREDICTION

Group Name: Syntax

Introduction and background

From our survey, we discovered a few **major issues** that we want to address through our application as stated below:

1. 1 in 6 deaths from cardiovascular disease was due to stroke.

2. 185,000 strokes—nearly 1 in 4—are in people who have had a previous stroke.

Problem statement:

- Stroke is a major health problem that can cause death or disability. Early diagnosis and treatment can improve outcomes, but many people are not diagnosed until it is too late.

Objectives

- To identify people who are at high risk of stroke, even if they do not have any symptoms.
- To improve the diagnosis of stroke by providing additional information that could help them to make a more accurate diagnosis.
- To reduce the cost of stroke care.

Data Collection

- Source: Obtained from the Kaggle platform, specifically from the user Ruthvik PVS.
- Features: The data set contains 10 features.
- Observations: The data set contains 5110 observations with 12 attributes.

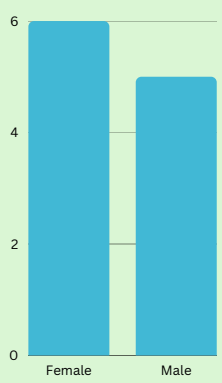
Data Preprocessing

- Replace missing values by mapping numerical mean into missing datasets
- Use mapping method to categorize categorical columns to numerical values
- Created new column called 'diabetic status' referring to the level of 'avg glucose level' column
- Oversampling method using (RandomOverSampler from imblearn) to statistically balance the data before splitting into test, train and validation data respectively.
- Standardize values to reduce deviation of data

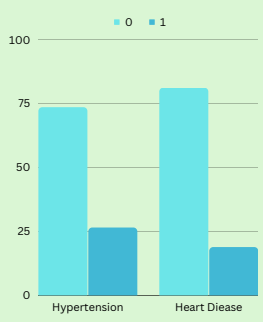
Exploratory Data Analysis

- Initial insights on the data
- To learn about the main characteristics of this dataset

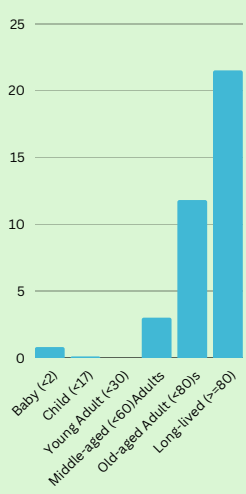
Graph of Percentage of Stroke among Gender



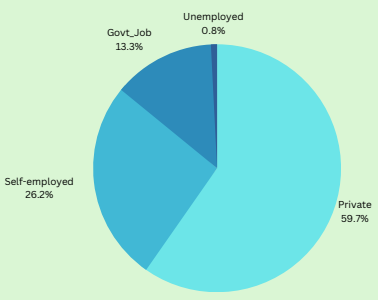
Stroke against Hypertension and Heart Disease Status



Stroke against Age Group

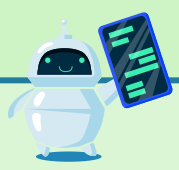


Pie chart showing Work Type of People Who Had Stroke



*When instance of stroke = 1

Machine Learning



DATA SPLITTING

- We split our data into training, validation, and testing sets
- To avoid overfitting and see whether the model can generalise well
- Partitioning of data is as follows:
  - Training data = 50%, Validation set = 16.67%, Test set = 33.33%

FEATURE SCALING

- To ensure every feature is on the same footing without any upfront importance
- Prevents features with larger magnitudes from dominating the learning process and allows for meaningful comparison between different features
- Normalization, standardization

ALGORITHM IMPLEMENTATION (ANN)

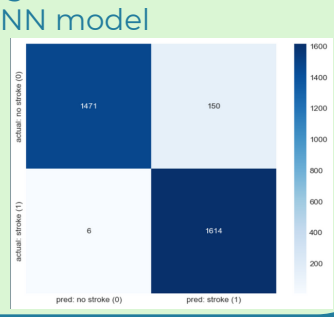
- We developed 5 different models and chose the one that performed best
- Use TensorFlow and Keras to develop neural network model
- Fitted our model to the standardized dataset with its target
- Did hyperparameter tuning to improve performance of NN model

EVALUATION METRICS

- F1 score
- Confusion matrix plot

	precision	recall	f1-score	support
0	1.00	0.91	0.95	1621
1	0.91	1.00	0.95	1620
accuracy			0.95	3241
macro avg	0.96	0.95	0.95	3241
weighted avg	0.96	0.95	0.95	3241

Train Accuracy: 0.983
Validation Accuracy: 0.954
Test Accuracy: 0.952



Demo link: