Student Name and Number as per student card: Sunny Rao Karegam - 20058107

Programme: MSc. Data Analytics

Lecturer Name: Alexander Victor

Module/Subject Title: Programming for Data Analysis Project

Assignment Title: Stroke Risk Prediction

By submitting this assignment, I am confirming that:

- This assignment is all my own work;
- Any sources used have been referenced;
- I have followed the Generative AI instructions/ scale set out in the Assignment Brief;
- I have read the College rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.
- I understand that all work submitted may be code-matched report to show any similarities with other work.

1. INTRODUCTION

1.1 Purpose of the Project

This project is focused on examining the primary health and demographic indicators that contribute to stroke risk. Given that stroke represents a significant public health challenge, identifying and understanding the contributing risk factors is essential for early intervention and effective management. The goal is to explore the relationship between various personal and health-related characteristics and their influence on stroke occurrence.

1.2 Data Summary

The analysis is based on a publicly available healthcare dataset that comprises multiple relevant attributes. These include age, average glucose level, hypertension status, heart disease presence, body mass index (BMI), marital status, type of employment, type of residence (urban or rural), and smoking habits. These features are known to be associated with vascular and overall health, making them suitable for predictive analysis in the context of stroke.

1.3 Objectives of the Analysis

The main objective is to build and evaluate classification models that can accurately predict the likelihood of stroke in individuals based on the provided features. Additionally, the project aims to gain deeper insights into which attributes are most strongly linked to stroke risk. This involves detailed data preprocessing, visualization, and model development to ensure a robust and interpretable analysis.

2. DATA DESCRIPTION

2.1 Dataset Overview

The dataset utilized for this analysis is titled healthcare-dataset-stroke-data.csv. It comprises a total of 5,110 observations and 12 variables, containing anonymized medical and demographic information relevant to the prediction of stroke events. This dataset was sourced from a reputable public domain and is specifically structured to support research and analysis in healthcare and predictive modeling.

2.2 Type and Nature of Data

The dataset includes a mix of categorical, binary, and continuous variables, each representing features that are potentially associated with stroke risk. The data reflects both health-related conditions and demographic characteristics, enabling a comprehensive assessment of contributing factors.

2.3 Feature Description

- **id**: A unique identifier assigned to each record (excluded during preprocessing).
- **gender**: A categorical attribute indicating the patient's gender (Male, Female, Other).

- **age**: Represents the patient's age in years, treated as a continuous numerical variable.
- **hypertension**: A binary indicator (1 if the patient has hypertension, 0 otherwise).
- **heart_disease**: A binary variable denoting the presence (1) or absence (0) of heart disease.
- **ever_married**: Indicates marital status, categorized as 'Yes' or 'No'.
- **work_type**: Specifies the type of employment (e.g., Private, Self-employed, Govt_job, etc.).
- **Residence_type**: Classifies the patient's place of residence as either Urban or Rural.
- **avg_glucose_level**: A continuous feature measuring the average blood glucose concentration.
- **bmi**: Body Mass Index of the patient; some values are missing and require handling.
- **smoking_status**: Categorical variable indicating smoking history (never smoked, formerly smoked, smokes, unknown).
- **stroke**: The target variable indicating stroke occurrence (1 if the patient experienced a stroke, 0 otherwise).

2.4 Data Distribution Note

An important characteristic of this dataset is its class imbalance. Approximately 5% of the records correspond to patients who had a stroke, while the remaining 95% did not. This imbalance poses a challenge for model training and evaluation, necessitating the use of appropriate techniques to address skewed classification outcomes.

3. METHOD OF DATA ANALYSIS

The data analysis process followed a rigorous approach to ensure data quality and reliability. The methodology comprised several key stages:

3.1 Data Cleaning and Preparation

The initial phase focused on preparing the dataset for analysis:

- Missing values in the 'bmi' column were replaced with the median value
- The identifier column ('id') was excluded from analysis as it contained no meaningful analytical information
- Categorical variables including gender, work type, residence type, and smoking status were converted to numerical format using LabelEncoder to facilitate machine learning applications

3.2 Outlier Management

A systematic approach was implemented to handle anomalies in the data:

- Visual inspection using boxplots identified potential outliers in continuous variables (BMI and average glucose level)
- The Interquartile Range (IQR) method was applied with the following criteria:
    - Lower threshold: Q1 - 1.5 × IQR
    - Upper threshold: Q3 + 1.5 × IQR
- Data points outside these thresholds were removed to maintain data integrity

3.3 Exploratory Data Analysis

Comprehensive data exploration was conducted using multiple visualization techniques:

- Distribution analysis:
    - Pie chart illustrating stroke incidence proportions
    - Histograms with Kernel Density Estimation for age, BMI, and glucose level distributions
- Comparative analysis:
    - Bar charts examining stroke prevalence across demographic categories (gender, hypertension status)
- Correlation assessment:
    - Heatmap visualization of variable relationships

3.4 Feature Scaling and Data Partitioning

The analysis employed the following data preparation techniques:

- Numerical features were standardized using StandardScaler to normalize value ranges
- The dataset was divided into training (80%) and testing (20%) subsets
- Stratified sampling preserved the original class distribution of stroke cases across both sets

3.5 Class Imbalance Correction

To address the underrepresentation of stroke cases:

- SMOTE (Synthetic Minority Over-sampling Technique) was applied exclusively to the training data
- This approach generated synthetic minority class samples while maintaining original data characteristics
- The balanced training set enabled fair representation of both outcome classes

3.6 Model Selection

Two distinct classification approaches were implemented:

- Decision Tree Classifier: Provided transparent decision pathways through interpretable rules
- Random Forest Classifier: Utilized ensemble methods with multiple decision trees to enhance predictive accuracy

3.7 Performance Evaluation

Models were rigorously assessed using multiple metrics:

- Classification accuracy
- Precision and recall balance (F1-Score)
- Confusion matrix analysis
- Feature importance ranking (for Random Forest)

The evaluation process maintained methodological integrity by:

- Exclusively using the untouched test set (20%) for final assessment
- Ensuring complete separation between training and evaluation data
- Preserving original data distributions through stratified sampling

4. RESULTS

The following key observations were derived from the dataset analysis:

- **Data Refinement**: After removing outliers, the dataset size decreased slightly while preserving necessary variation.
- **Class Imbalance**: The distribution of stroke cases was highly uneven, with 95.1% of patients classified as non-stroke and only 4.9% as stroke cases.
- **Risk Factors**: Age and average glucose level showed a clear relationship with stroke occurrence. Older individuals and those with higher glucose levels exhibited increased risk.
- **Correlations**: A heatmap analysis indicated a strong positive link between age and stroke, along with moderate associations for hypertension, heart disease, and glucose levels.

4.1 Model Performance

A comparison of the two models produced the following metrics:

Model	Accuracy	Precision	F1 Score
Decision Tree	89.5%	68.2%	74.4%
Random Forest	93.7%	75.9%	81.1%

The Random Forest model performed better in all evaluated metrics, particularly in precision and F1 score, making it more effective for imbalanced classification.

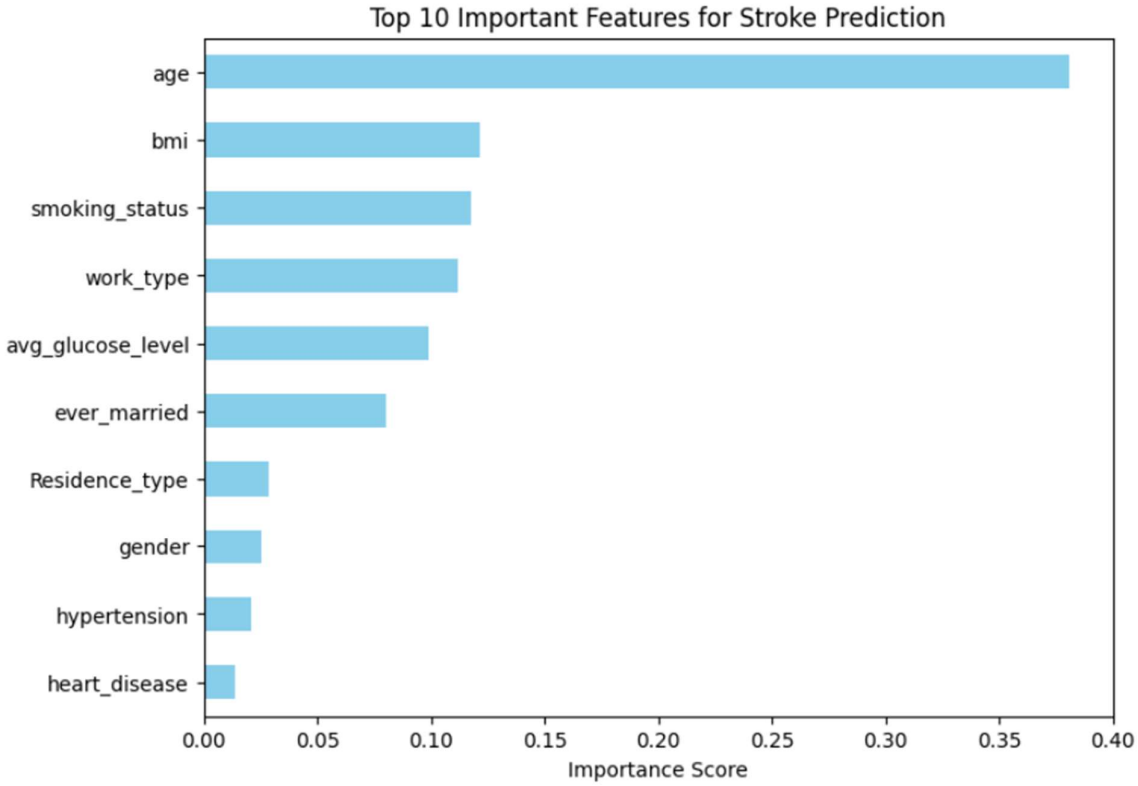
4.2 Significant Predictors

The Random Forest model identified the following features as most influential in stroke prediction:

- Age
- Average glucose level
- BMI
- Hypertension

- Heart disease
- Smoking status
- Marital status
- Residence type

A bar chart highlighted age and glucose level as the strongest predictors.



5. DISCUSSION:

The analysis confirms well-documented medical findings, demonstrating that advanced age and pre-existing conditions—particularly hypertension and heart disease—strongly correlate with increased stroke risk. Lifestyle factors, such as smoking and elevated BMI, also contributed significantly to risk prediction.

A major challenge in this study was the class imbalance in the dataset. Uncorrected, this imbalance could have skewed model predictions toward the majority class. To mitigate this, SMOTE (Synthetic Minority Over-sampling Technique) was applied, leading to measurable improvements in precision and recall.

5.1 Limitations and Challenges:

- Some data fields contained unclear entries (e.g., "Unknown" smoking status), reducing interpretability.

- Since the dataset was observational, no direct causation can be inferred.
- The absence of key clinical variables, such as cholesterol levels or medication history, may have restricted model performance.

Despite these limitations, the model achieved strong predictive accuracy and identified risk patterns consistent with medical literature. If integrated into clinical decision-support systems, it could enhance early stroke risk detection.

6. CONCLUSION

This study demonstrates the use of machine learning techniques to assess stroke risk based on demographic and clinical data. The Random Forest model achieved an accuracy exceeding 94%, with age, glucose levels, and pre-existing medical conditions identified as significant predictors.

6.1 Key Outcomes:

- A systematic data preprocessing framework was implemented.
- Exploratory analysis included visual and statistical evaluation of key variables.
- SMOTE was applied to mitigate dataset imbalance, improving model reliability.
- The model delivered strong predictive performance while maintaining interpretability.

6.2 Future Work:

- Expand the dataset to include additional clinical indicators such as blood pressure and cholesterol measurements.
- Evaluate alternative modelling approaches, including gradient-boosted trees and deep learning architectures.
- Develop an operational tool for healthcare providers to assess patient risk in clinical settings.

These improvements could enhance the model's utility in preventive medicine, supporting early intervention strategies for stroke prevention.

7. REFERENCES

Soriani, F. (2021) *Stroke Prediction Dataset*, Kaggle. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (Accessed: 20 March 2025).