Inspiring Excellence

CSE422 Lab Project Report

Predict the Occurrence of Stroke

Group Number: 07

Group Members:

- Md.Shamiul Islam Khan Ishrak- 20301235
- Tasnuva Hassan- 24341109

Table of Contents:

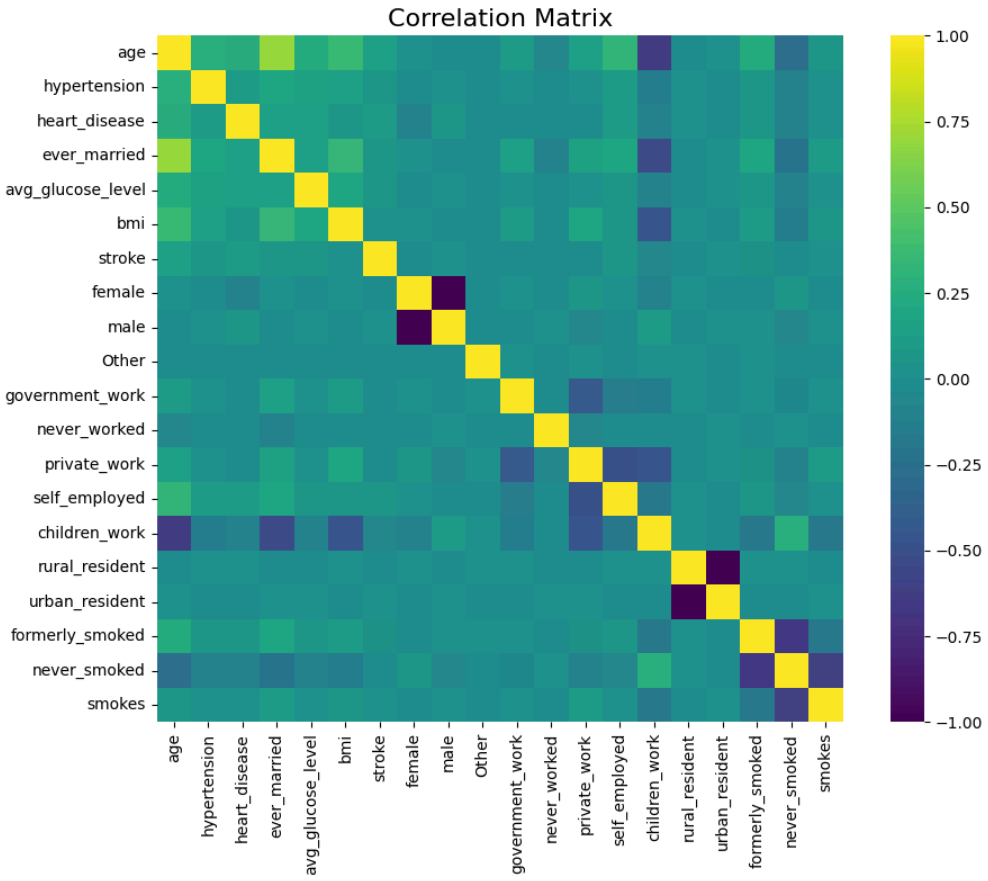
Introduction

This project aims to predict the occurrence of strokes in patients by applying machine learning techniques. The dataset includes various health and lifestyle-related features, such as age, gender, hypertension, heart disease, glucose level, BMI, and smoking status. Due to the imbalance of occurring strokes, only 783 stroke instances out of 43,400 entries; oversampling using the SMOTE technique was performed to improve model performance.

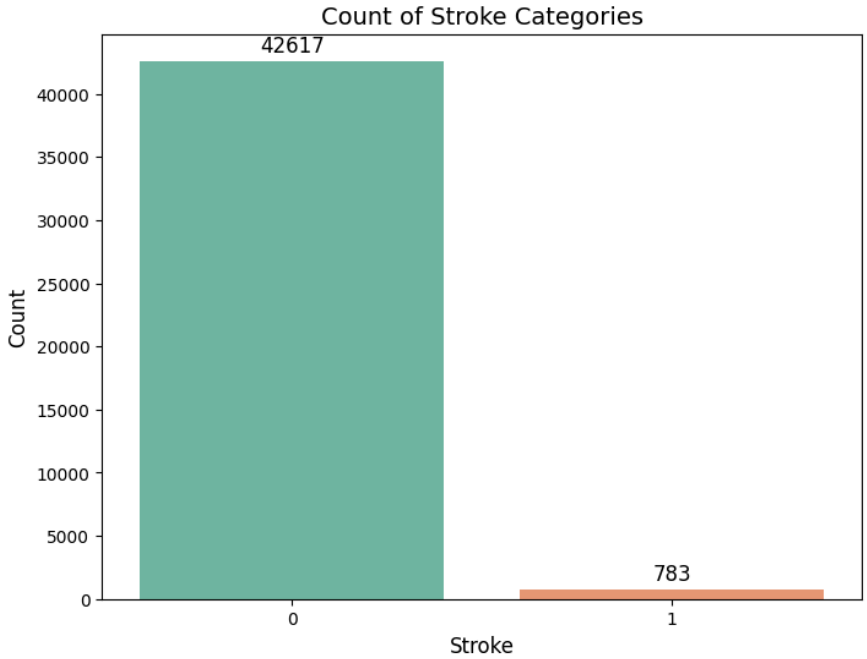
The primary focus is to develop an accurate predictive model that can assist healthcare professionals in early stroke risk assessment.

Dataset Description

- Source
  - Link: https://www.kaggle.com/code/tumpanjawat/stroke-prediction-eda-resampling-xgboost/n
  - Reference: It is collected from Kaggle.
- Dataset Description
  - Features: There are a total of 12 features in the dataset. They are - id, gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and stroke history.
  - Type of Problem: Classification.
    Since the target variable is stroke, it is a classification problem. The goal is to predict one of two possible outcomes: stroke or no stroke.
  - Number of Data Points: 43,400
  - Feature Types:
    Quantitative Features: age, average glucose level, BMI.
    Categorical Features: gender, hypertension, heart disease, work type, residence type, smoking status, marital status.



- Imbalance Dataset
  - The dataset is highly imbalanced, as it contains only 783 instances of stroke out of 43,400 total instances. This means the positive class (stroke) is significantly underrepresented compared to the negative class (no stroke).



Dataset Pre-Processing

- Elimination of 'id' column.
- Null Values
    - The dataset contains missing values in the BMI column. Missing values can lead to inaccuracies during model training and affect predictive performance.
    - The solution is median imputation.In this approach, missing values in the bmi column are replaced with the median of the non-missing values.
- Categorical Values
    - The smoking_status column contains missing values. Since this is a categorical variable, it cannot be imputed using numerical measures like the mean or median.
    - The solution involves mode imputation. The mode is the most frequently occurring value in a categorical feature. By filling the missing values with the mode of the smoking_status column, you ensure that the imputed value represents the majority class, which is a reasonable guess for the missing entries.

Feature Scaling

- Importance: Standardize input characteristics using StandardScaler
- Columns Scaled: 'avg glucose level', 'bmi', and 'age'
- Visualization: Heatmap showing feature correlation

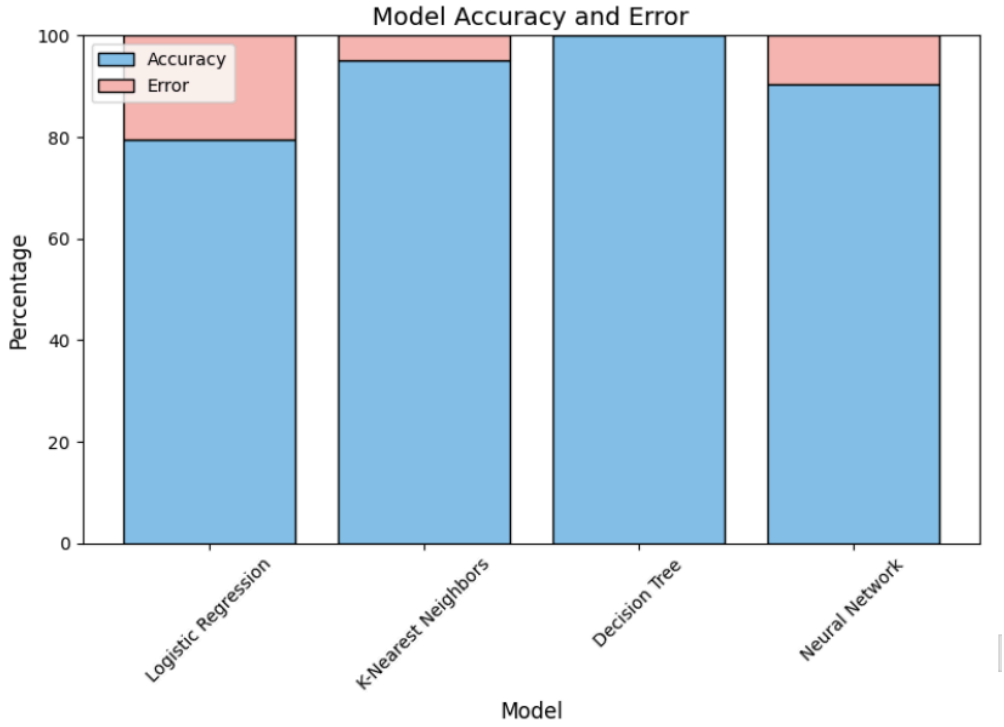
Dataset Splitting

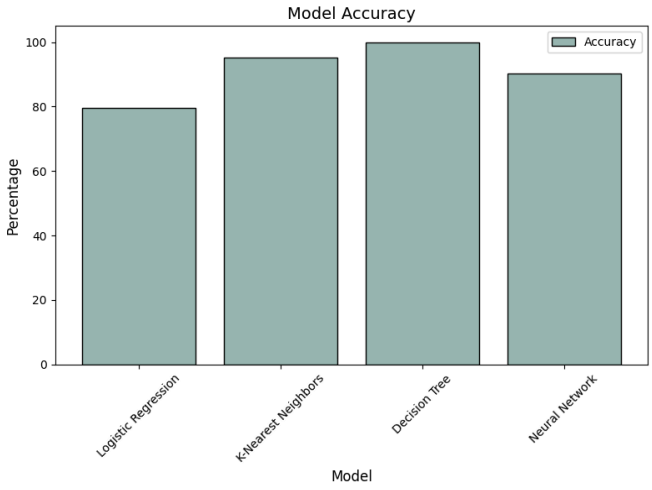
70% for training and 30% for testing (Stratified split)

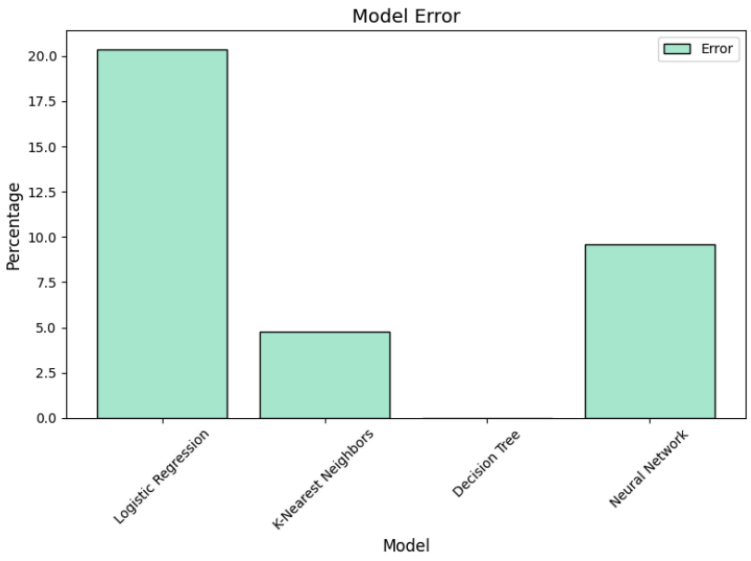
Model Training and Testing

- Logistic Regression
- KNN
- Decision Tree
- Neural Network

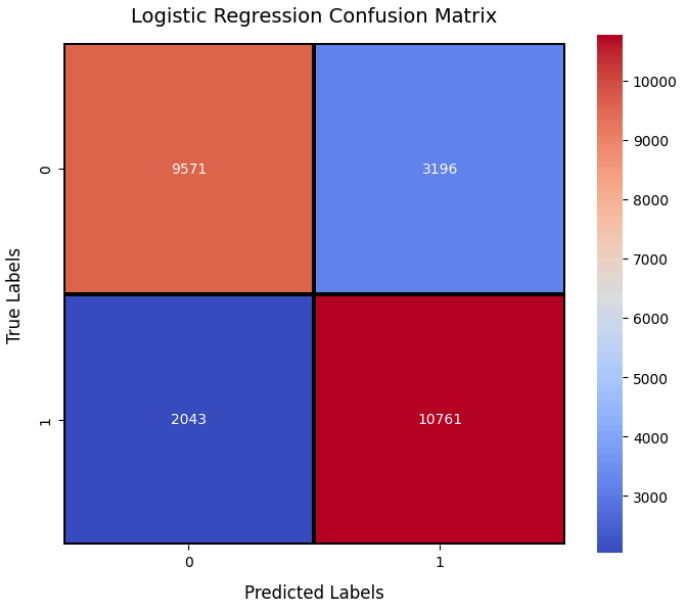
Model Selection/Comparison Analysis

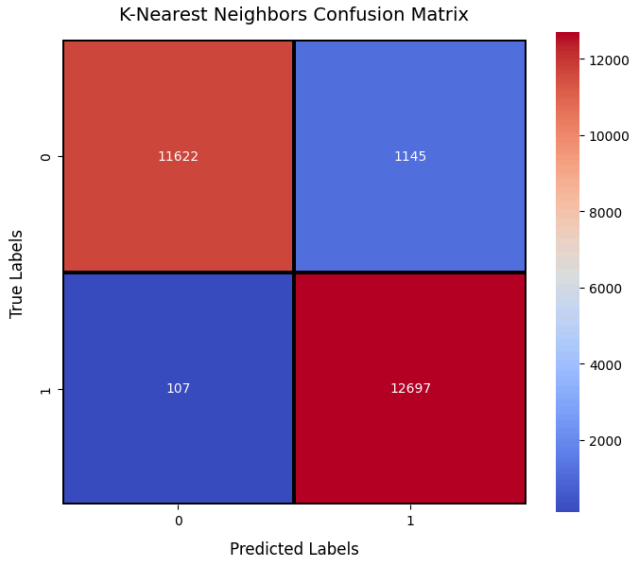


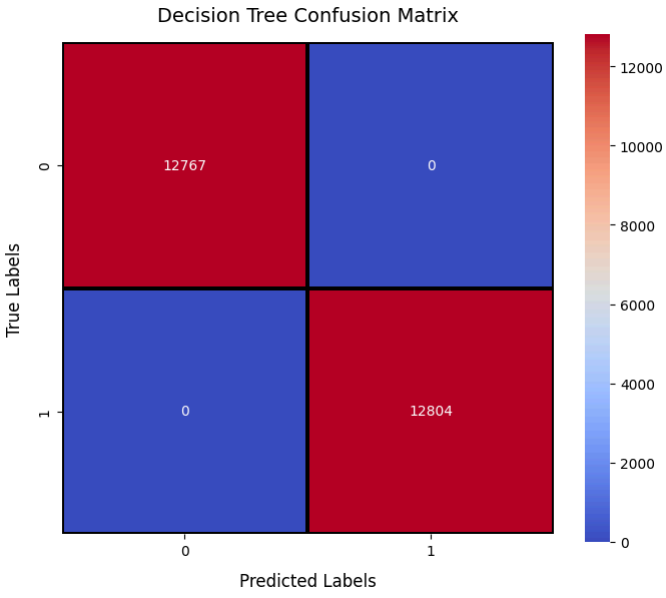


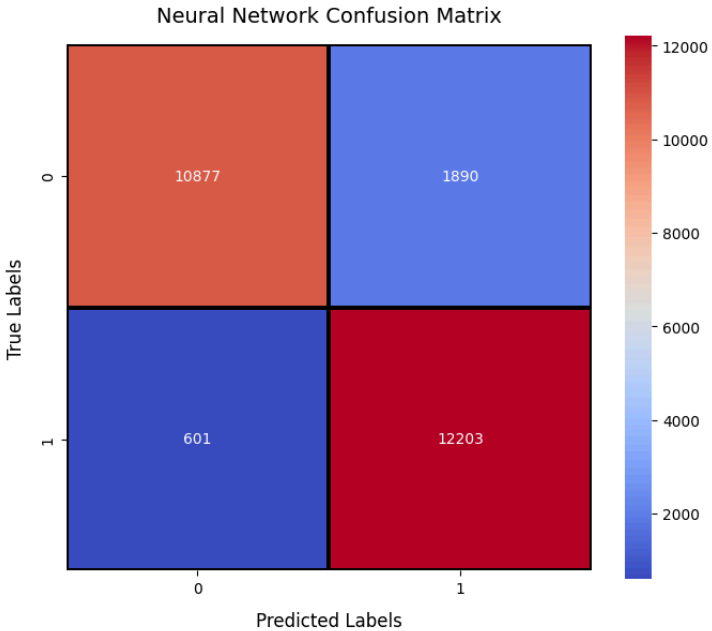


Confusion Matrix









Conclusion

- Key Steps
  The project followed a structured approach involving data preprocessing, oversampling to address class imbalances, and thorough model evaluation. These steps were pivotal in achieving high-performance metrics.
- Models
  - Logistic Regression: Achieved 79.512% accuracy, 77.101% precision, 84.044% recall, and 80.423% F1-Score.
  - KNN: Achieved 95.104% accuracy, 91.728% precision, 99.164% recall, and 95.301% F1-Score.
  - Decision Tree: Achieved 100.000% accuracy, 100.000% precision, 100.000% recall, and 100.000% F1-Score.
  - Neural Networks: Achieved 90.258% accuracy, 86.589% precision, 95.306% recall, and 90.739% F1-Score.

These results highlight the importance of model selection, feature engineering, and robust evaluation in building reliable predictive systems.