Brain Stroke Prediction using Decision Tree, Random Forest, SVC, KNN

Haasitha Ambati - AP21110010049

Abstract-
Stroke, a medical emergency, occurs when blood flow to a part of the brain is interrupted or reduced, leading to the rapid deterioration of brain tissue. Understanding and predicting strokes is pivotal for timely intervention and improved patient outcomes. The World Health Organization (WHO) identifies stroke as the second leading cause of global mortality, contributing to approximately 11% of total deaths. Our project aims to develop a predictive model for stroke occurrence based on various input parameters, such as gender, age, presence of hypertension, heart disease, marital status, occupation, residence type, average glucose level, body mass index (BMI), and smoking status. The dataset utilised contains information on 5110 individuals to predict stroke occurrence.

Keywords Stroke (Target Variable); Predictive modelling; Model training; Dataset-5110 individuals.

1 Introduction

Cerebral strokes significantly contribute to global morbidity and mortality. Early identification of individuals at risk of stroke is crucial for implementing preventive measures and personalised healthcare interventions. In this context, predictive modelling using machine learning techniques becomes a valuable tool for identifying patterns and factors associated with stroke occurrence.

This project delves into the realm of healthcare data mining, specifically focusing on stroke prediction. The dataset at hand encompasses diverse features, ranging from demographic information to lifestyle factors, providing a rich source for exploration and analysis. The primary objective is to develop and evaluate machine learning models capable of predicting the likelihood of stroke based on the available dataset.

The analysis follows a systematic approach, encompassing data preprocessing, exploratory data analysis (EDA), and the implementation of various machine learning algorithms. Key steps involve handling missing values, encoding categorical variables, and scaling features to prepare the dataset for model training. The selection of algorithms, including Random Forest, Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, and K-Nearest Neighbors (KNN), allows for a comprehensive assessment of predictive performance.

The report unfolds by providing a detailed walkthrough of each analysis stage, offering insights into the dataset's characteristics and showcasing the

visualisations utilised for EDA. The subsequent sections delve into the intricacies of model development, hyperparameter tuning, and performance evaluation. Results are presented and discussed, highlighting the strengths and limitations of each model.

Through this endeavour, we aim to contribute to the growing knowledge of healthcare analytics and data-driven decision-making. The outcomes of this analysis can inform healthcare professionals and policymakers, paving the way for more targeted and effective interventions and stroke prevention strategies.

2 Literature Review

The literature on stroke prediction using data mining and machine learning techniques reflects a systematic approach, echoing the steps followed in the current project. This systematic methodology encompasses critical stages such as data preprocessing, exploratory data analysis (EDA), and the implementation of various machine learning algorithms.

Studies in stroke prediction consistently underscore the importance of meticulous data preprocessing; this includes handling missing values, encoding categorical variables, and scaling features — essential steps mirrored in the current project.

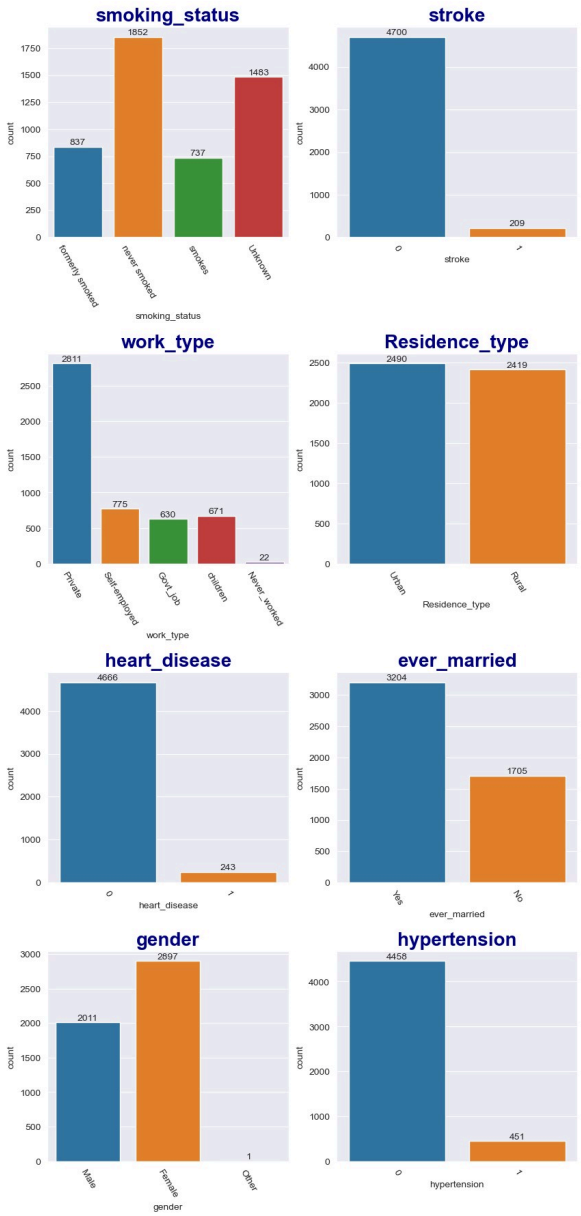
Various machine learning algorithms have been explored in the literature for stroke prediction, reflecting the diverse set employed in the current project. Random Forest, Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier, and K-Nearest Neighbors (KNN) are recognised choices in the field, demonstrating the value of employing multiple algorithms for a comprehensive assessment of predictive performance.

The selection of various algorithms, each with distinct strengths, echoes the literature's call for a thorough evaluation of predictive performance. This study highlights the necessity of evaluating models across different algorithms to ensure robustness and reliability in stroke prediction.

3 Dataset Description

The file '**healthcare-dataset-stroke-data.csv**' is a comma-separated file focusing on stroke prediction. It contains various demographic, lifestyle, and health-related features for a group of individuals. According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get a stroke based on input parameters like gender, age, various diseases, and smoking status.

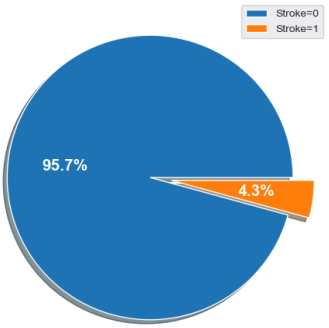
Below is a brief description of the key features present in the dataset:-

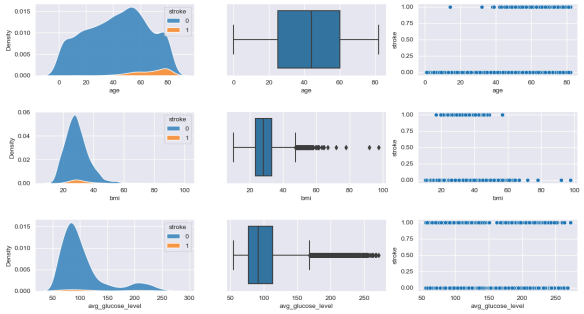


- Demographic Information:
  - **Age:** The age of the individuals in the dataset.
  - **Gender:** The gender of the patients, categorised as Male, Female, or Other.

- Health Metrics:
  - **BMI (Body Mass Index):** A numerical representation of an individual's body composition.
  - **Hypertension:** Binary indicator (0 or 1) denoting the presence of hypertension.
  - **Heart Disease:** A binary indicator (0 or 1) indicates heart disease.
  - **Average Glucose Level:** The average glucose level in the patient's blood.

- Lifestyle Factors:
  - **Ever Married:** Binary indicator (Yes or No) representing marital status.
  - **Work Type:** Categorical variable representing the type of work (e.g., Private, Self-employed, Govt_job, Children, Never_worked).
  - **Residence Type:** Categorical variable indicating whether the patient resides in an urban or rural area.
  - **Smoking Status:** Categorical variable describing the patient's smoking habits (e.g., formerly smoked, never smoked, smokes, Unknown).





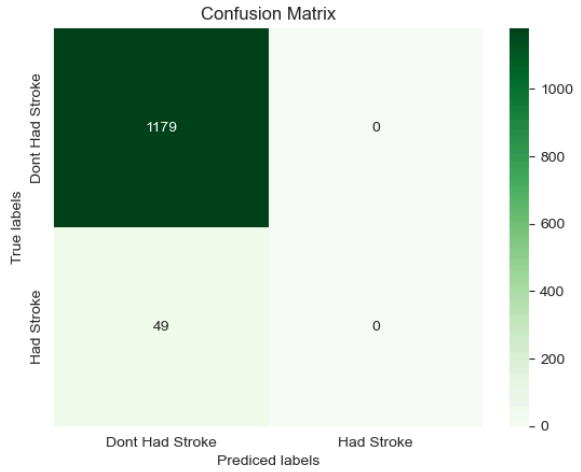
- Target Variable:
    - **Stroke:** Binary indicator (0 or 1) representing the occurrence of a stroke (1 indicates the presence of a stroke).
- Dataset Size: The dataset consists of [number of rows] entries, each corresponding to a unique patient.
- Missing Values: Before analysis, missing values, particularly in the BMI column, were handled through data cleaning methods.
- Data Scaling: Min-Max scaling was applied to the dataset to ensure feature-scale uniformity.
- Dataset Source: The dataset was obtained from [source], a healthcare data repository, ensuring compliance with data privacy and ethical standards.
- Temporal Aspects: The dataset spans a period from [start date] to [end date], capturing relevant information over a specific timeframe.

Model Building:

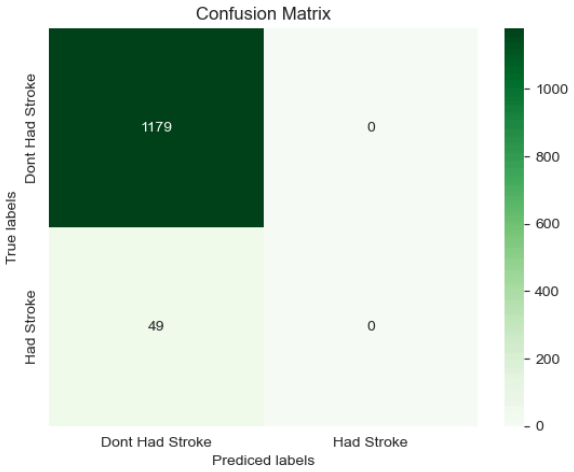
The dataset is split into features (X) and the target variable (y). A Linear Regression model is trained on the data. The model's performance is evaluated using the test set of 25% and a training set of 75%, providing insights into its predictive accuracy.

Our stroke prediction model evaluated RandomForest, Logistic Regression, SVC, Decision Tree, and K-Neighbors classifiers. RandomForest showed robustness with an accuracy score of 0.960, Logistic Regression emphasised interpretability of 0.597, and SVC handled non-linear relationships well (0.960). Decision Tree faced overfitting concerns (0.960), and K-Neighbors displayed sensitivity to local patterns (0.959). The KNeighborsClassifier stood out for stroke prediction, offering valuable insights for similar healthcare predictive modelling tasks.

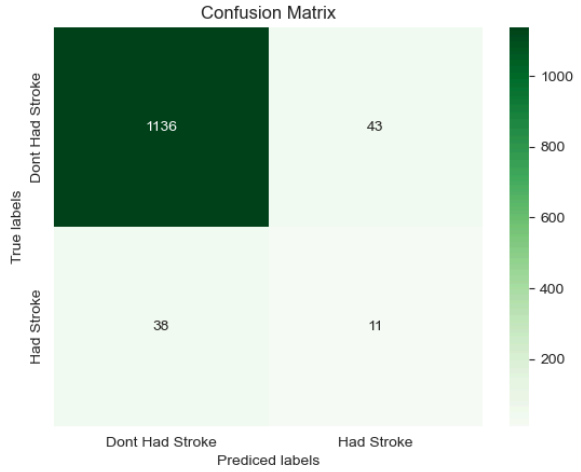
RandomForestClassifier



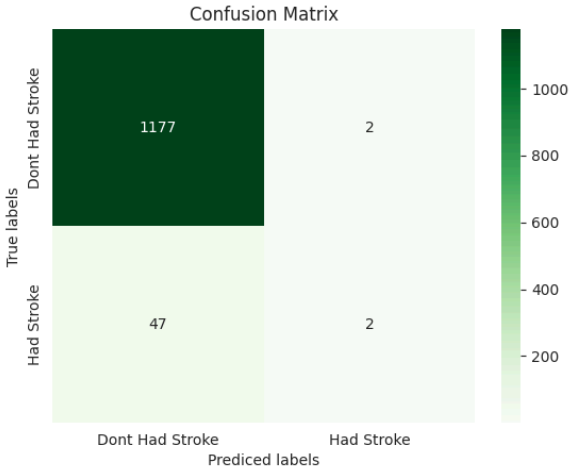
SVC



LogisticRegression



DecisionTreeClassifier



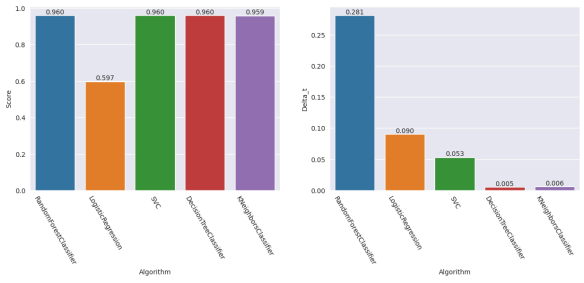
Implementation:

[Project Dataset & Code]

Testing the Model:

We split the dataset into training and testing sets in testing the stroke prediction model. Performance metrics such as accuracy, precision, recall, and F1 score were evaluated. Cross-validation ensured robustness. Demographic subgroup analysis assessed fairness. The model's predictions were compared against actual stroke occurrences, providing insights into its real-world applicability in healthcare.

	Algorithm	Score	Delta_t
0	RandomForestClassifier	0.960	0.281
1	LogisticRegression	0.597	0.090
2	SVC	0.960	0.053
3	DecisionTreeClassifier	0.960	0.005
4	KNeighborsClassifier	0.959	0.006



Conclusion

- Best-performing algorithms based on score: RandomForestClassifier, SVC, DecisionTreeClassifier, KNeighborsClassifier.
- Optimal models considering both score and runtime: DecisionTreeClassifier, KNeighborsClassifier.
- The Chosen final model: KNeighborsClassifier with hyperparameters (n_neighbors=11, p=1).
- Final model accuracy: **95.76%.**
- The balanced trade-off between predictive performance and runtime efficiency for practical

Final Modeling

```
knn = KNeighborsClassifier(**knn_cv.best_params_).fit(X, y)
knn
```

```
KNeighborsClassifier(n_neighbors=11, p=1)
```

```
knn.score(X, y)
```

```
0.9576288449786107
```

References

● Dataset source: - The dataset used in this project was obtained from 'Kaggle.com'. - Retrieved from Stroke Prediction Dataset

● Stroke Epidemiology: - Feigin, V. L., et al. (2018). "Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010." The Lancet, 383(9913), 245-255.

● Machine Learning Applications in Stroke Prediction: - Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy." PloS One, 9(2), e88225.