CSE422

Project Report

Title:

The analysis of a disease Cerebral Stroke and its classification

LabSection: 15

Group Number: 09

Submitted By:

MD. MOSTAFIJUR RAHMAN
ID: 22101721
MD TOUHIDUR RAHMAN
ID: 22101446

Submitted To:

Mr. Shoaib Ahmed Dipu
Lecturer
Pollock Nag
C. Lecturer

Table of Contents

Abstract:

This research study provides an effective approach for machine learning-based brain stroke prediction. Since cerebral strokes have a catastrophic effect on world health, early detection and treatment are essential to enhancing patient outcomes. This paper covers the dataset, preprocessing procedures, model training, outcomes, and the future potential of machine learning in stroke diagnosis.

1.Introduction

Brain strokes are commonly referred to as "silent killers," mostly because of their elusiveness. Even highly qualified medical professionals may fail to notice their early warning signs because they are often so subtle, which could result in delayed diagnoses. Our research is about using machine learning to understand and classify a disease Cerebral Stroke.We collect data on features related to brain strokes, including gender, age, hypertension, heart disease, marital status, housing type, average glucose level, BMI, work type, and smoking status, and apply advanced computer tools to analyse it.We train the models to identify patterns that suggest a person may be at risk for a stroke.After testing several models, we looked for the one that could predict and categorise brain strokes with the highest accuracy.

2.Dataset description

We collected this dataset from the research paper titled "A Hybrid Machine Learning Approach to Cerebral Stroke Prediction Based on Imbalanced Medical Dataset" published in the **Journal of the Operational Research Society**.
This dataset has been specially curated to facilitate cerebral stroke prediction using machine learning techniques, especially to address the challenges posed by imbalanced medical data.
Quantitative: age, avg_glucose_level, bmi, hypertension, heart_disease, stroke
Categorical: gender, ever_married, work_type, Residence_type, smoking_status
The stroke feature is highly imbalanced:
No Stroke (0): 98.20%
Stroke (1): 1.80%

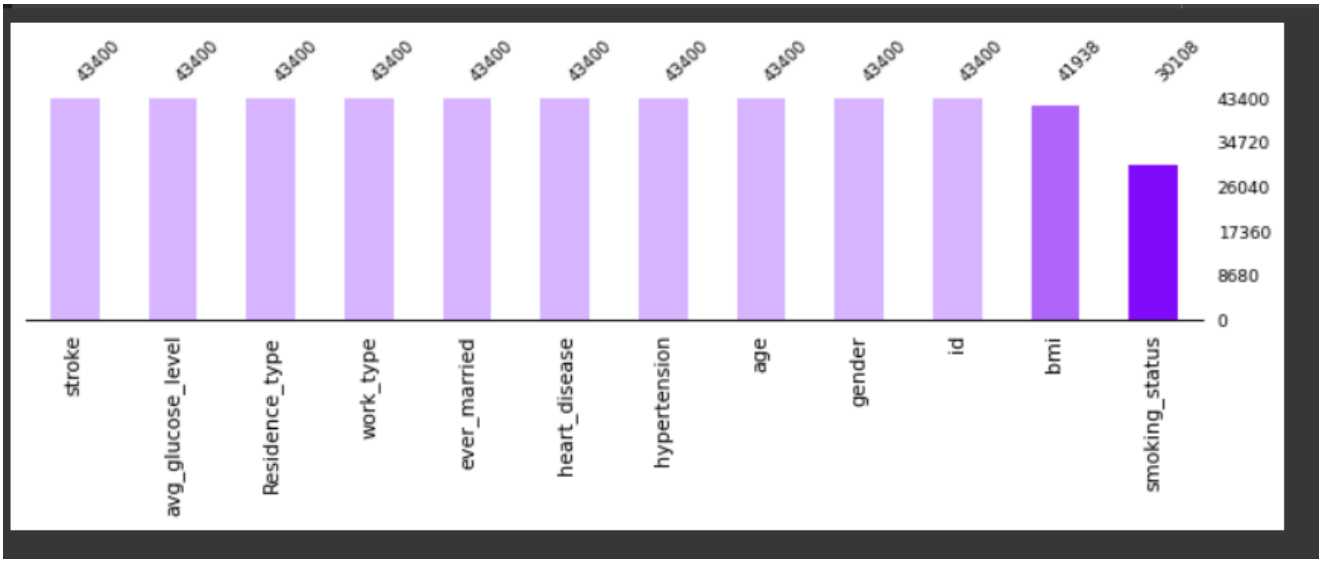
Kaggle Link of this Dataset:
https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset
Sciencedirect Link:
https://doi.org/10.1016/j.artmed.2019.101723
This dataset has medical records of **43,400 patients**, which include a variety of demographic, lifestyle, and health-related information.. It includes **12 features**, which are a mix of **categorical** and **continuous** variables.

```
print(df.shape)
```

```
(43400, 12)
```

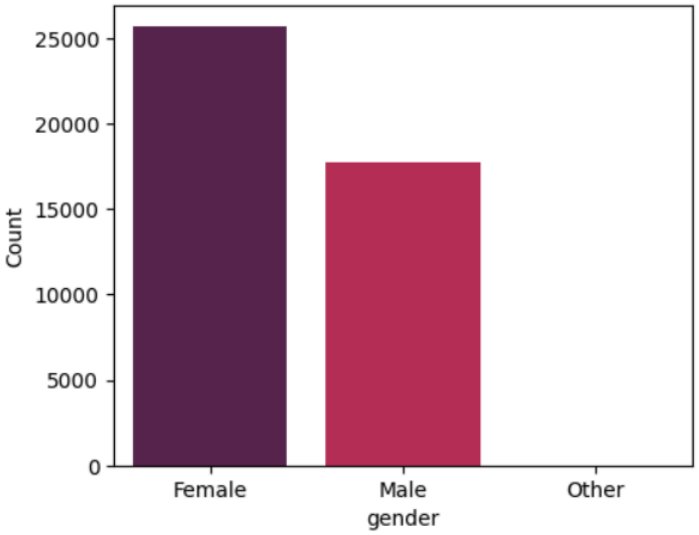
```
df.info()
```

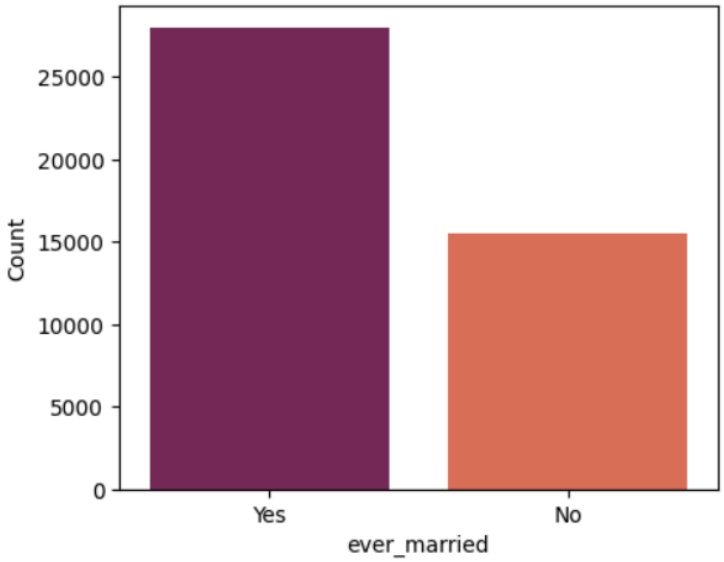
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43400 entries, 0 to 43399
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 43400 non-null  int64
 1   gender             43400 non-null  object
 2   age                43400 non-null  float64
 3   hypertension       43400 non-null  int64
 4   heart_disease      43400 non-null  int64
 5   ever_married       43400 non-null  object
 6   work_type          43400 non-null  object
 7   Residence_type     43400 non-null  object
 8   avg_glucose_level  43400 non-null  float64
 9   bmi                41938 non-null  float64
 10  smoking_status     30108 non-null  object
 11  stroke             43400 non-null  int64
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

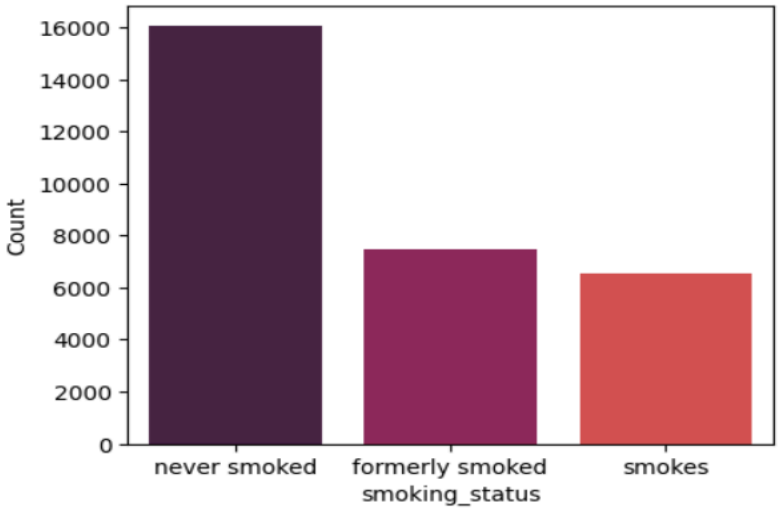


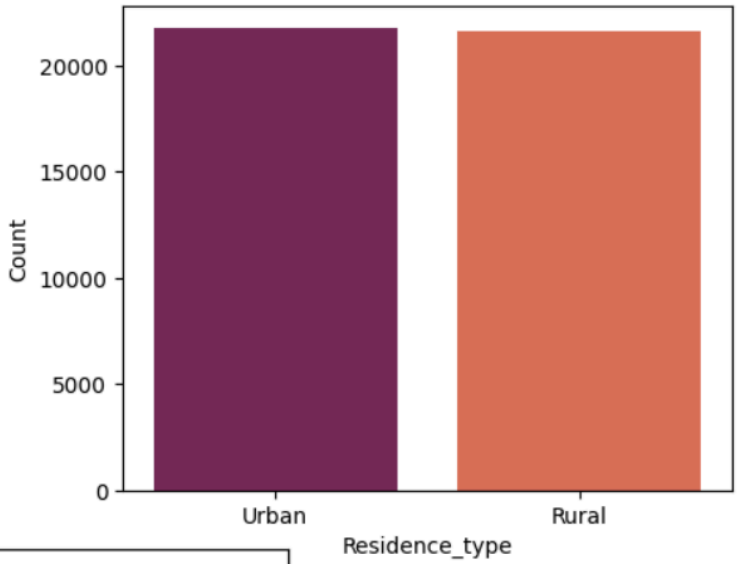
It provides a clear visual for assessing and handling missing values during data preprocessing.

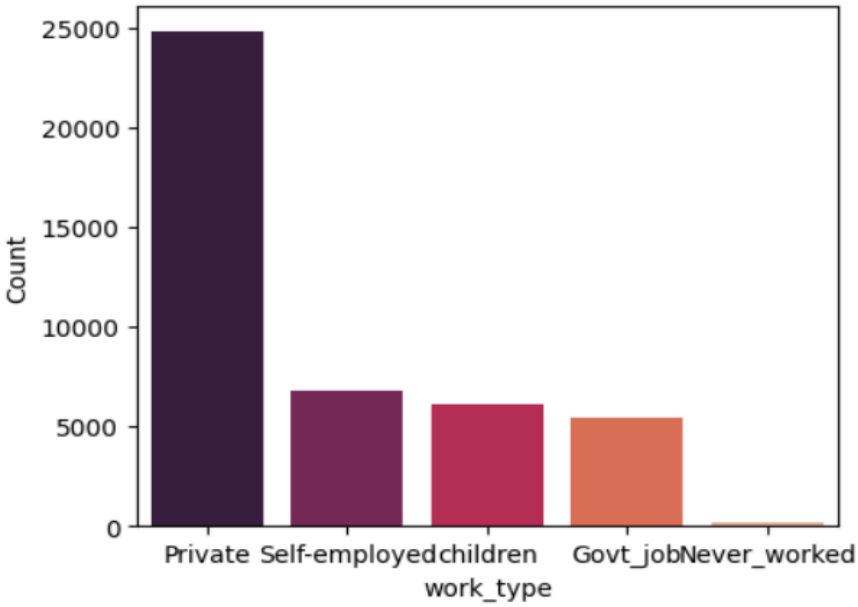
3. Dataset Pre-processing







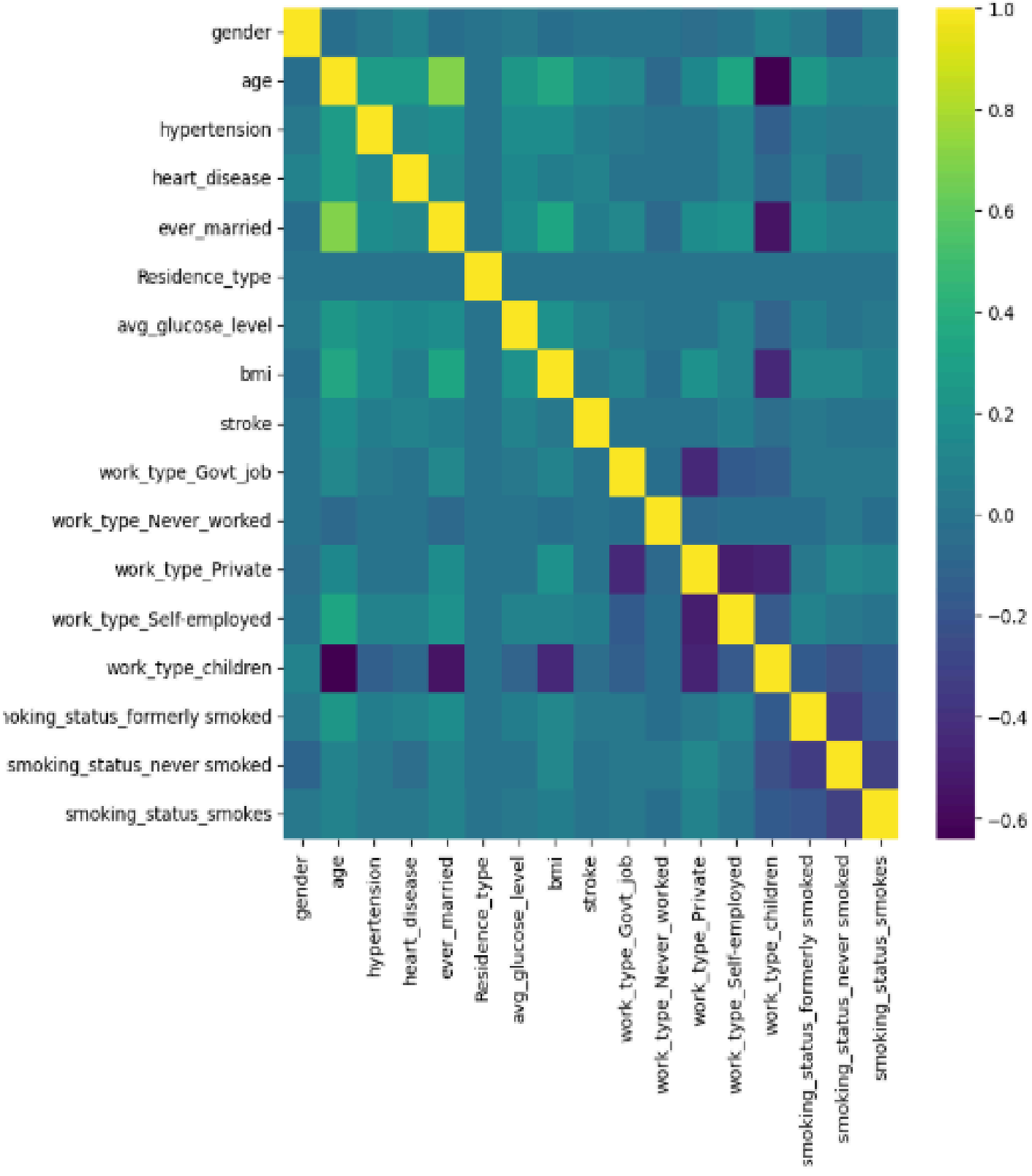




```
Null values per column:
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                 1462
smoking_status     13292
stroke                 0
dtype: int64
```

There have been null value issues discovered. When we examine the datasets, we find that two features bmi and smoking status have null values. Missing bmi value can affect model training. other smoking_status missing values significantly affect the categorical analysis. Bmi missing percentage is low. so the mean imputation works well for it.
Use for work_type, Residence_type, and smoking_status. using "one-hot encoding" to convert category (non-numeric) variables into numerical values.

In this dataset some unnecessary columns like id , it does not need for data processing so we drop this column. Another ever_married column two type data yes or no .we use yes for 1 and no for 0 also Residence_type same issue as like ever married(Urban=1 and Rural=0).


gender
age
hypertension
heart_disease
ever_married
Residence_type
avg_glucose_level
bmi
stroke
work_type_Govt_job
work_type_Never_worked
work_type_Private
work_type_Self-employed
work_type_children
oking_status_formerly smoked
smoking_status_never smoked
smoking_status_smokes
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

4. Dataset splitting

There are total **43,400 patients** in the datasets.
Train set (70%) - 30380
Test set (30%) - 13020

5. Feature scaling (as required)

Feature scaling is an essential step in preparing our dataset for machine learning
The model involves normalizing a range of features to explain the Same scale by model. Feature scaling ensures quantitative features like age, avg_glucose_level, and bmi are on comparable scales, avoiding bias in gradient-based or distance-based models.

6. Model Training and Testing

We test our dataset for all models KNN, Decision Tree, Logistic Regression, Random Forest and its Random Forest gives the best accuracy for predicting stroke.

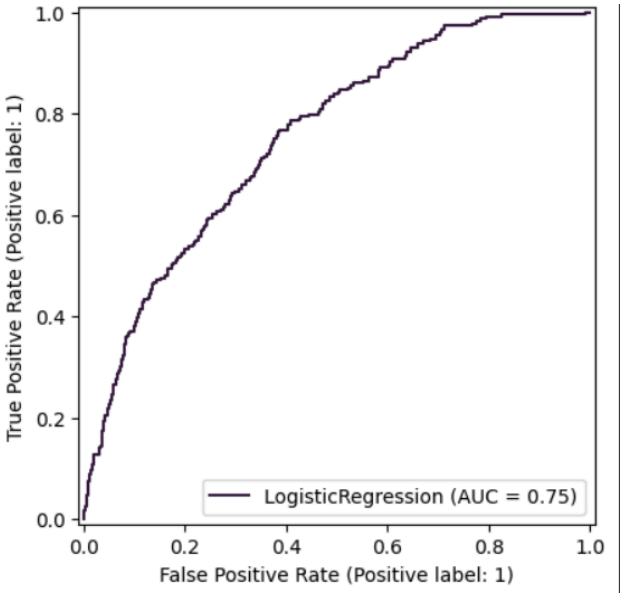
```
Model: Logistic Regression, Accuracy: 89.69%, Error: 10.31%
Model: K-Nearest Neighbors, Accuracy: 86.52%, Error: 13.48%
Model: Decision Tree, Accuracy: 95.68%, Error: 4.32%
Model: Random Forest, Accuracy: 97.74%, Error: 2.26%
```

Model	Accuracy	Error
KNN	86.52%	13.48%
Decision Tree	95.68%	4.32%
Logistic Regression	89.69%	10.31%
Random Forest	97.74%	2.26%

From this table we can see that the Random forest model gives best performance with 97.74% percentage and only 2.26% error while another KNN model gives worst performance.

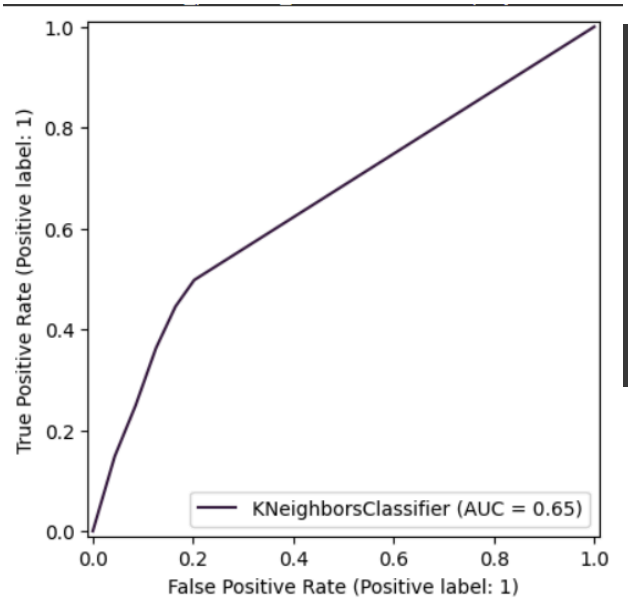
7.Model selection/Comparison analysis

Logistic Regression:



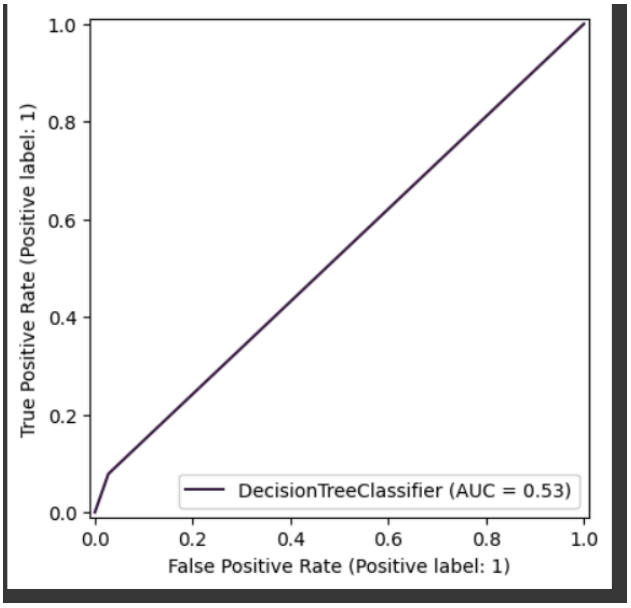
	precision	recall	f1-score	support
0	0.99	0.91	0.95	12791
1	0.07	0.37	0.11	229
accuracy			0.90	13020
macro avg	0.53	0.64	0.53	13020
weighted avg	0.97	0.90	0.93	13020

KNN:



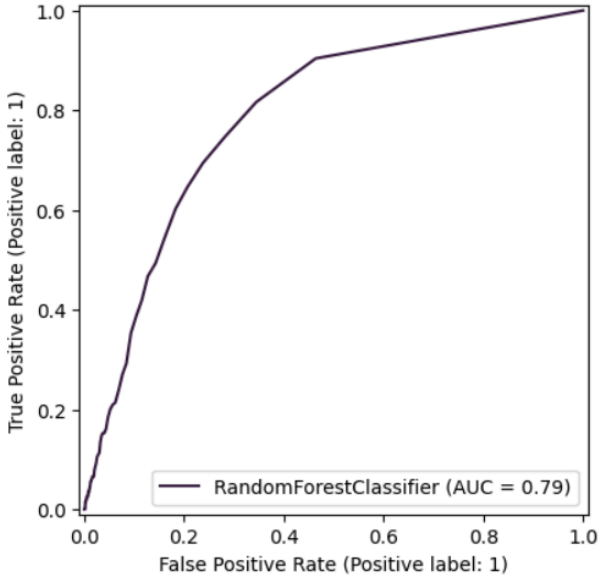
	precision	recall	f1-score	support
0	0.99	0.87	0.93	12791
1	0.05	0.36	0.09	229
accuracy			0.87	13020
macro avg	0.52	0.62	0.51	13020
weighted avg	0.97	0.87	0.91	13020

Decision Tree:

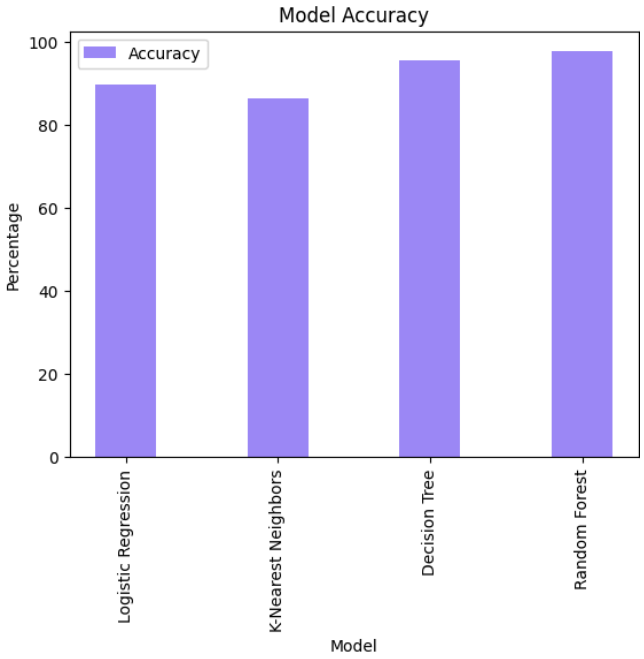


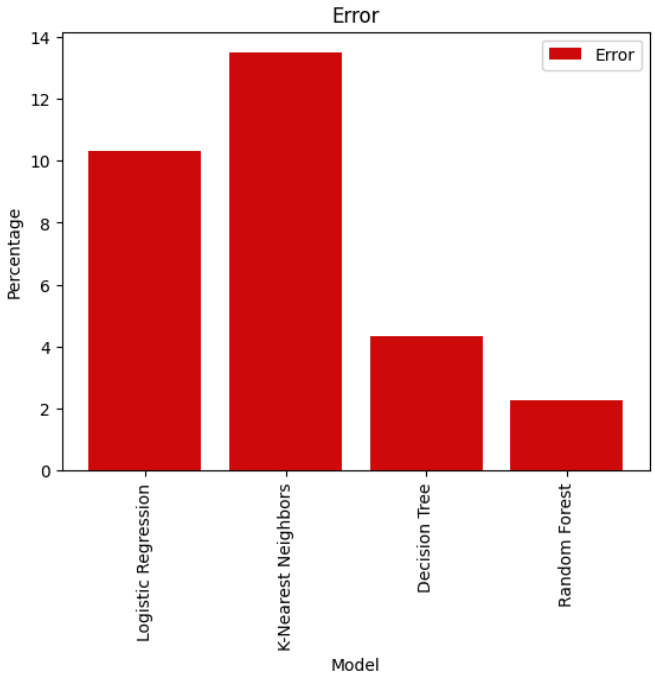
	precision	recall	f1-score	support
0	0.98	0.97	0.98	12791
1	0.05	0.08	0.06	229
accuracy			0.96	13020
macro avg	0.52	0.53	0.52	13020
weighted avg	0.97	0.96	0.96	13020

Random Forest:



	precision	recall	f1-score	support
0	0.98	0.99	0.99	12791
1	0.08	0.03	0.04	229
accuracy			0.98	13020
macro avg	0.53	0.51	0.51	13020
weighted avg	0.97	0.98	0.97	13020





8.Conclusion

In conclusion, the application of machine learning algorithms such as logistic regression, decision tree, k-nearest neighbor (KNN) and random forest to predict cerebral stroke using key individual indicators showed promising results.The models achieved accuracy rates between 90% and 97%, indicating strong predictive performance.Among these, Random Forest stood out as the most effective, achieving an impressive accuracy of 97.74%, followed by Decision Tree with 95.67%. Although Logistic Regression(89.68%) and KNN(86.52%) classifiers reached accuracies of 90%, further refinement and optimization could enhance their performance.Our accuracy has improved slightly from where we collected the datasets from the research paper. These results underscore the potential of machine learning in accurately predicting cerebral stroke outcomes.