

**Title of the project: Brain stroke analysis and classification**

**Group: 05**

| **Name** | **ID** |
| --- | --- |
| Munib Fahmid Rafeed | 21101181 |
| Ahmed Tahlil Kadir | 21101138 |

**Table of Contents**

| **Content** | **Page No** |
| --- | --- |
| Introduction | 3-4 |
| Dataset description | 4 |

| Dataset pre-processing | 5-8 |
| --- | --- |
| Dataset splitting | 8 |
| Model training | 9-10 |
| Model testing | 10-12 |
| Conclusion | 12 |

**Abstract:**

This report suggests a thorough method for predicting brain strokes with machine learning. Due to the significant toll brain strokes take on public health worldwide, detecting and acting early are crucial for enhancing patient prognosis. The document delves into the use of machine learning for stroke diagnosis, covering aspects such as dataset, preprocessing, model training, and outcomes.

**1. Introduction**

This project focuses on leveraging machine learning to comprehend and categorize brain strokes. We collect data on various factors associated with brain strokes, such as gender, age, hypertension, heart disease, marital status, residence type, average glucose level, BMI, work type, and smoking status. Then, we employ sophisticated computer algorithms to analyze this data intelligently. We train the computer to identify patterns that suggest someone may be prone to a brain stroke. We experimented with various models to determine the most accurate one for predicting and classifying brain strokes.

**2. Motivation**

Brain strokes, often referred to as "silent killers," earn this ominous title due to their elusive nature. Their initial signs are often so subtle that even experienced medical professionals may miss them, resulting in delayed diagnoses. This delay can lead to irreversible damage to the brain, significantly impacting patients' quality of life, or, in severe cases, causing death. This bleak reality has persisted for decades, highlighting the urgent need for innovative approaches to early stroke detection.

Traditional diagnostic methods, while somewhat effective, are limited by human observation and the variability of patient symptoms. These methods rely heavily on noticeable symptoms, which may only appear in the later stages of a stroke. Additionally, with the increasing ratio of patients to doctors in many regions globally, ensuring timely and accurate diagnoses for all patients is becoming increasingly challenging.

The advent of digitization and the subsequent explosion of data present an opportunity for transformation in the medical field. The current medical landscape is abundant with health data, ranging from electronic health records to biometric data collected by wearables. However, this wealth of information remains largely untapped in traditional diagnostics.

Machine learning offers a promising solution to this challenge. By training advanced algorithms on vast datasets, it becomes possible to identify patterns and correlations that may be imperceptible to humans. This goes beyond simply complementing human observation; it enhances it to levels beyond our natural abilities. With machine learning models, it may be possible to predict the onset of a stroke based on complex patterns and numerous variables, long before visible symptoms appear. Such a shift in diagnostic methods could lead to timely interventions, significantly reducing the impact of strokes and saving lives.

Moreover, integrating machine learning into stroke diagnostics can alleviate the economic burden on healthcare systems. Early interventions could result in shorter hospital stays, fewer complications, and more effective treatments, leading to substantial cost savings in healthcare.

In essence, the drive to leverage machine learning for stroke prediction is not merely a scientific endeavor but a humanitarian one. With the potential to revolutionize stroke care, this effort aims to provide individuals with a better chance of combating one of humanity's deadliest medical adversaries.

**3. Dataset Description**

Link: Cerebral Stroke Prediction-Imbalanced Dataset | Kaggle

<https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset>

Number of Features: 12

Type of class/label: Categorical and Continuous

Number of data points: 43400

<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

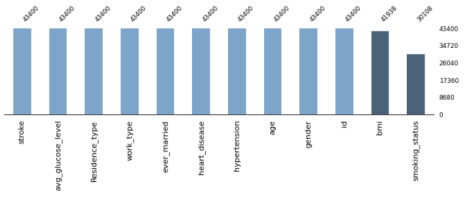
9 bmi 41938 non-null float64

10 smoking\_status 30108 non-null object

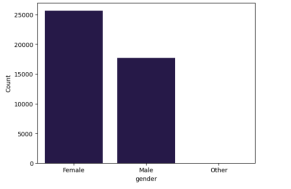
11 stroke 43400 non-null int64

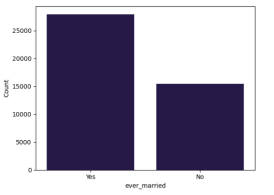
dtypes: float64(3), int64(4), object(5)

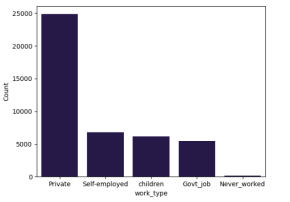
**Biasness/Balanced**

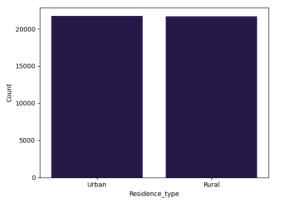
****

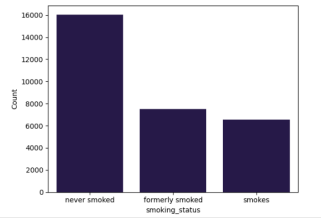
**4. Dataset Pre-processing**









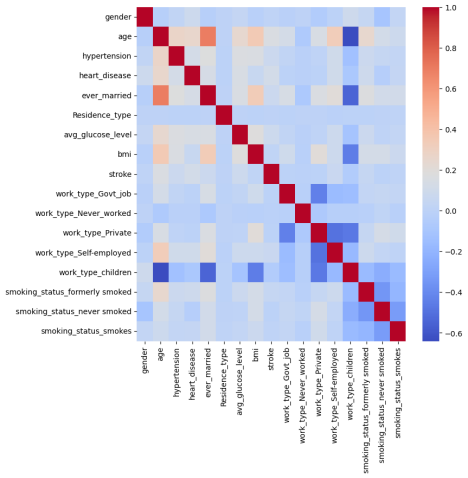


Issues such as null values and outliers were addressed:

- 5 features contained a total of 868 null values in BMI and 13292 null values in smoking status. Used mean imputation.

- ‘ID’ column was unnecessary, and was thus droped.

- Mapped ‘ever\_married’ Yes into ‘1’ and No into ‘0’. Mapped ‘Residence\_type’ ‘Urban’ into 1, ’Rural’ into 0. Using ‘one-hot encoding’ to convert categorical (non numeric) variables into a numerical format.



**5. Dataset Splitting**

A 70-30 split was maintained - 13020 for training and 30380 for testing.

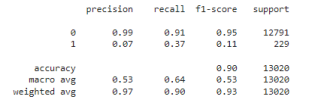
**6. Model Training**

Upon testing, the Random Forest model consistently outperformed others in predicting brain stroke from the given indicators.

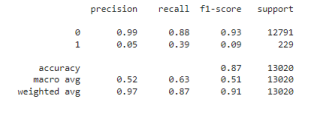
| **Model Name** | **Accuracy (%)** | **Error(%)** |
| --- | --- | --- |
| Logistic Regression | 90% | 10% |
| Decision Tree | 96% | 4% |
| KnnClassifier | 87% | 13% |
| Random forest | 98% | 2% |

From the table, we can see that the Random forest model showed the best performance with 98% accuracy and only 2% error. On the other hand, the worst performance was given by Logistic regression and Knn classifier with 90% and 87% accuracy respectively, and 10% and 13% error respectively. The Decision tree was mostly satisfying with 95% accuracy.

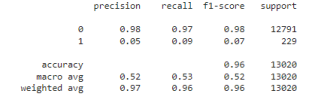
**Logistic Regression**

****

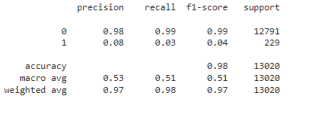
**KNN**

****

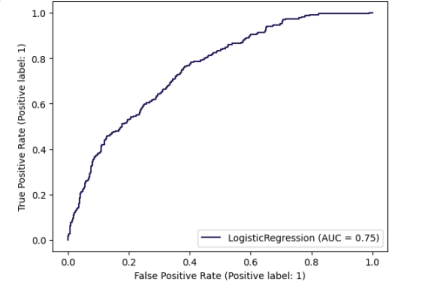
**Decision Tree**

****

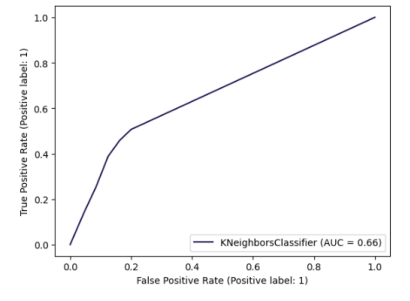
**Random Forest**

**7. Model Testing**

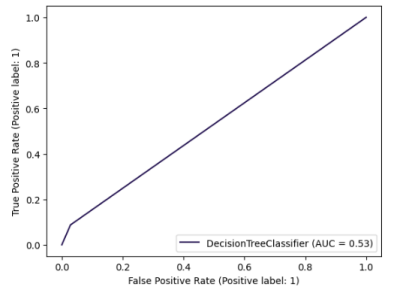
Logistic Regression:



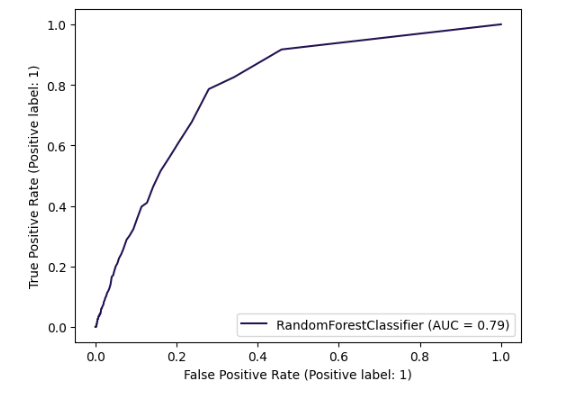
KNN:



Decision Tree:



Random Forrest:



**8. Conclusion**

In summary, the outcomes derived from employing a range of machine learning algorithms including Logistic Regression, Decision Tree, Kth Nearest Neighbor, and Random Forest Classifier to forecast brain strokes using personal key indicators are encouraging. With accuracies spanning from 87 to 98%, these models exhibit commendable efficacy in stroke prediction. Among the tested models, Random Forest yielded the highest performance, boasting an accuracy rate of 98%. Decision Tree also demonstrated promising results, achieving an accuracy of 96%. Although Logistic Regression and KNN Classifier achieved an accuracy of 90% and 87% respectively, they may benefit from additional refinement to enhance their predictive capabilities.