****

***Topic: Cerebral Stroke Prediction***

IE7275 Data Mining in Engineering

Prof. Srinivasan Radhakrishnan

Project by Urveesh Sayane

**Table of Contents**

[***Introduction*** 3](#_Toc163667254)

[***Problem Statement*** 4](#_Toc163667255)

[***Objectives*** 4](#_Toc163667256)

[***Data Source*** 5](#_Toc163667257)

[***Data Description*** 5](#_Toc163667258)

[***Data Preprocessing*** 6](#_Toc163667259)

[***Data Mining Tasks*** 7](#_Toc163667260)

[***Exploratory Data Analysis And Feature Selection*** 8](#_Toc163667261)

[***Data Mining Models*** 10](#_Toc163667262)

[***Conclusion*** 18](#_Toc163667263)

[***References*** 18](#_Toc163667264)

# ***Introduction***

Cerebral stroke, a leading cause of disability and mortality worldwide, poses significant challenges in healthcare management and diagnosis. Early detection and accurate prediction of stroke risk are crucial for timely intervention and prevention of adverse outcomes. With advancements in machine learning and artificial intelligence, predictive modeling techniques have emerged as promising tools for cerebral stroke detection.

In this project, we aim to develop predictive models using various machine learning algorithms to identify individuals at risk of cerebral stroke based on demographic, clinical, and lifestyle factors. We leverage a dataset containing relevant features such as age, gender, hypertension, heart disease history, average glucose level, body mass index (BMI), and smoking status. Through exploratory data analysis (EDA), feature engineering, and model training, we seek to evaluate the performance of different algorithms in predicting stroke occurrence.

Our primary objective is to compare the effectiveness and suitability of different machine learning algorithms, including Random Forest, Neural Networks, Naive Bayes, and Decision Trees, for cerebral stroke detection. We assess model performance based on metrics such as accuracy, precision, recall, and F1-score, considering the imbalanced nature of the dataset. Additionally, we investigate the impact of preprocessing techniques like Synthetic Minority Over-sampling Technique (SMOTE) in handling class imbalance and improving model performance.

By identifying the most effective predictive models and understanding their strengths and limitations, our project aims to contribute insights into the application of machine learning in healthcare for cerebral stroke risk assessment. These findings have the potential to enhance clinical decision-making, optimize resource allocation, and ultimately improve patient outcomes in stroke prevention and management.

# ***Problem Statement***

Cerebral stroke remains a leading cause of mortality and disability worldwide, emphasizing the critical need for accurate and timely prediction methods. Despite advancements in medical technology, accurately identifying individuals at risk of stroke remains a significant challenge. The primary objective of this study is to develop and evaluate machine learning models capable of predicting stroke risk based on demographic, clinical, and lifestyle factors. By leveraging predictive analytics techniques, our aim is to enhance the accuracy of stroke prediction, enabling healthcare professionals to intervene proactively and mitigate stroke-related morbidity and mortality.

# ***Objectives***

The main objective of this study is to develop and evaluate machine learning models for predicting the risk of cerebral stroke. Specifically, the objectives include:

* Data Collection and Preprocessing: Gather a comprehensive dataset containing demographic information, clinical variables, lifestyle factors, and stroke outcomes. Preprocess the data to handle missing values, encode categorical variables, and normalize numerical features.
* Model Development: Implement and train multiple machine-learning algorithms, including Decision Trees, Random Forests, Naive Bayes, and Neural Networks, to predict the likelihood of stroke based on the input features.
* Model Evaluation: Evaluate the performance of each model using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. Compare the performance of different algorithms to identify the most effective approach for stroke prediction.
* Recommendations: Provide recommendations on the most suitable algorithms for stroke prediction based on their performance, computational efficiency, and applicability to the dataset and problem type.

# ***Data Source***

The data used in this study was obtained from the following source:

* Data Source: Mendeley Data
* Authors: Tianyu Liu, Wenhui Fan, Cheng Wu
* Title: Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets
* DOI: 10.17632/x8ygrw87jw.1

The data was obtained from the corresponding author and is publicly available for research purposes on the website. The dataset contains socio-demographic information, clinical variables, lifestyle factors, and stroke outcomes of individuals, making it suitable for developing machine learning models to predict stroke risk. Further details about the dataset and its attributes can be found in the provided article link.  
<https://data.mendeley.com/datasets/x8ygrw87jw/1>

# ***Data Description***

* Dataset Name: Cerebral Stroke Prediction Dataset
* Number of Columns: 12
* Number of Rows: 43,400
* Sample of Variable Names:

1. **id**: Unique identifier for each record
2. **gender**: Gender of the individual (Male, Female, Other)
3. **age**: Age of the individual
4. **hypertension**: Whether the individual has hypertension (0 = No, 1 = Yes)
5. **heart\_disease**: Whether the individual has heart disease (0 = No, 1 = Yes)
6. **ever\_married**: Whether the individual is ever married (No, Yes)
7. **work\_type**: Type of work the individual is engaged in (Private, self employed, etc.)
8. **Residence\_type**: Type of residence (Urban, Rural)
9. **avg\_glucose\_level**: Average glucose level in the individual's blood
10. **bmi**: Body Mass Index (BMI) of the individual
11. **smoking\_status**: Smoking status of the individual (formerly smoked, never smoked,)
12. **stroke**: Whether the individual had a stroke (0 = No, 1 = Yes)

This dataset contains **43,400 rows** and **12 columns**.

# ***Data Preprocessing***

***Handling Missing Values:***

* Missing values in the 'smoking\_status' column have been replaced with 'unknown'.
* Rows with missing BMI values have been dropped using the dropna() function.

***Encoding Categorical Variables:***

* Categorical variables have been encoded into numerical format using mapping dictionaries and the replace() function.

***Feature Selection:***

* Univariate feature selection using the chi-square test has been performed to select the most relevant features based on their scores.
* Feature importance has been evaluated using the Random Forest Classifier.

***Splitting Data:***

* The dataset has been split into training and testing sets using the train\_test\_split() function.
* For models, such as Decision Tree, Random Forest, Naive Bayes, and Neural Network, resampling techniques like SMOTE have been applied to handle class imbalance in the training data.

# ***Data Mining Tasks***

**Data Reduction**: The dataset underwent dimensionality reduction by selecting relevant features and dropping irrelevant or redundant ones. This process helps simplify the dataset and improve computational efficiency.

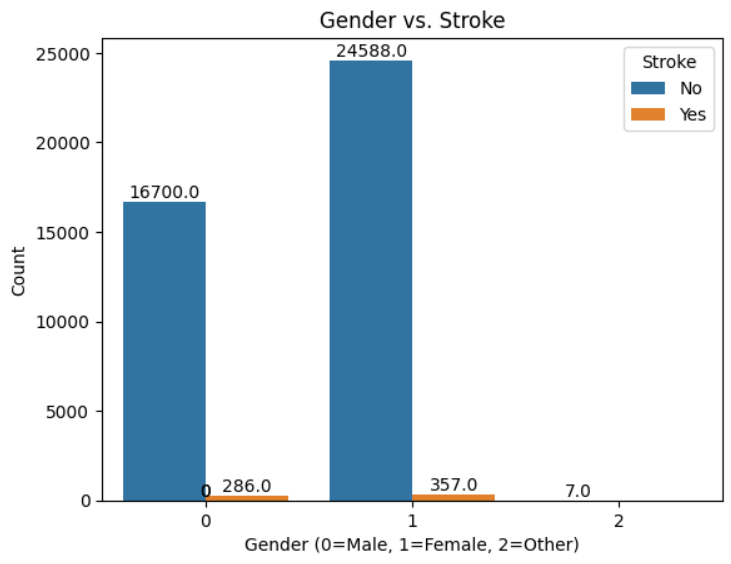
**Data Transformation**: Categorical variables were encoded into numerical values using techniques such as one-hot encoding or label encoding. This transformation allows categorical data to be used as input for machine learning algorithms.

**Missing Data Imputation**: Missing values in the dataset, particularly in the 'smoking\_status' and 'bmi' columns, were imputed to ensure that all data points could be used for analysis and modeling. For example, missing values in 'smoking\_status' were replaced with the label 'unknown', while rows with missing 'bmi' values were dropped.

**Classification**: The primary task in the project was classification, specifically predicting the likelihood of an individual experiencing a cerebral stroke based on various demographic, lifestyle, and health-related features. Multiple classification algorithms, such as Decision Trees, Random Forests, Naive Bayes, and Neural Networks, were trained and evaluated to identify the most effective model for this task.

**Prediction**: The trained classification models were utilized to predict whether an individual is at risk of having a stroke. These predictions were based on the features provided in the dataset, such as age, gender, BMI, average glucose level, smoking status, work type, and residence type.

# ***Exploratory Data Analysis And Feature Selection***



A graph of age groups

Description automatically generated

A graph with blue bars

Description automatically generated

Feature selection was conducted using two methods: univariate feature selection and feature importance.

* Univariate Feature Selection: This statistical method employed tests like chi-square for categorical variables and ANOVA for numerical variables. It involved using the SelectKBest function from sklearn.feature\_selection to select the top k features with the highest scores.
* Feature Importance: Feature importance was assessed using a Random Forest classifier. This involved training the classifier on the dataset and evaluating each feature's contribution to predictive performance. The importance scores were then visualized with a bar plot to understand each feature's relative importance.

These techniques helped identify the most informative features, enhancing model performance and interpretability.

# ***Data Mining Models***

***Decision Tree Classifier***

**Performance:** Decision Tree Classifier shows slightly lower accuracy compared to Random Forest but has better precision, recall, and F1 score. It correctly identifies a higher proportion of true positive cases (recall) and maintains a balance between precision and recall, leading to a higher F1 score. Therefore, Decision Tree Classifier may be a better choice if correctly identifying stroke cases is a priority.

**Computational Efficiency:** Decision trees are generally computationally efficient to train and predict, especially for small to medium-sized datasets. However, their performance may degrade with large, high-dimensional datasets due to potential overfitting.

**Applicability:** Decision trees are suitable for tasks where interpretability is essential and can handle both numerical and categorical data. They are commonly used in fields where decision-making transparency is critical, such as healthcare and finance.

* **Accuracy: 0.9498092513113973**
* **Precision: 0.05167173252279635**
* **Recall: 0.1349206349206349**
* **F1 Score: 0.07472527472527472**

***Random Forest Classifier***

**Performance**: Random Forest Classifier demonstrates the highest accuracy among the models evaluated. However, its precision, recall, and F1 score are relatively low compared to other models. This indicates that while it correctly classifies a high proportion of instances, it may have limitations in correctly identifying true positive cases of stroke. Therefore, while it achieves high overall accuracy, it may not be the most suitable choice if correctly identifying stroke cases is of utmost importance.

**Computational Efficiency**: Random forests typically have longer training times compared to individual decision trees, especially for larger datasets and more trees in the forest. However, prediction times are usually fast due to parallelization.

**Applicability**: Random forests are versatile and applicable to various classification tasks, including those with imbalanced data. They are commonly used in domains such as finance, bioinformatics, and remote sensing.

* **Accuracy: 0.9570815450643777**
* **Precision: 0.03571428571428571**
* **Recall: 0.07142857142857142**
* **F1 Score: 0.047619047619047616**

***Naive Bayes Classifier***

**Performance**: Naive Bayes Classifier achieves the lowest accuracy among the evaluated models but exhibits the highest recall. However, its precision is low, resulting in a relatively low F1 score. While it correctly identifies a high proportion of true positive cases, it also has a high false positive rate, which may not be desirable in a clinical setting.

**Computational Efficiency**: Naive Bayes classifiers are known for their computational efficiency, with fast training and prediction times even for large datasets. They are particularly suitable for text classification and other high-dimensional data tasks.

**Applicability**: Naive Bayes classifiers are commonly used in text classification, spam filtering, and other tasks where feature independence assumptions hold. They may not perform well with highly correlated features.

* **Accuracy: 0.7099427753934192**
* **Precision: 0.03841536614645858**
* **Recall: 0.7619047619047619**
* **F1 Score: 0.07314285714285713**

***Neural Network Classifier***

**Performance**: Neural Network Classifier shows moderate accuracy and recall but has the lowest precision and F1 score among the models evaluated. It correctly identifies a reasonable proportion of true positive cases but also has a relatively high false positive rate, indicating potential limitations in its predictive performance

**Computational Efficiency**: Training neural networks can be computationally intensive, especially for deep architectures and large datasets. However, with advancements in hardware (e.g., GPUs) and software (e.g., TensorFlow, PyTorch), training times have improved.

**Applicability**: Neural networks are highly flexible and applicable to various tasks, including image recognition, natural language processing, and medical diagnostics. They are suitable for tasks where complex patterns need to be learned from data.

* **Accuracy: 0.8279685264663805**
* **Precision: 0.037894736842105266**
* **Recall: 0.42857142857142855**
* **F1 Score: 0.06963249516441007**

A graph showing different colored rectangles

Description automatically generated

A graph showing different colored squares

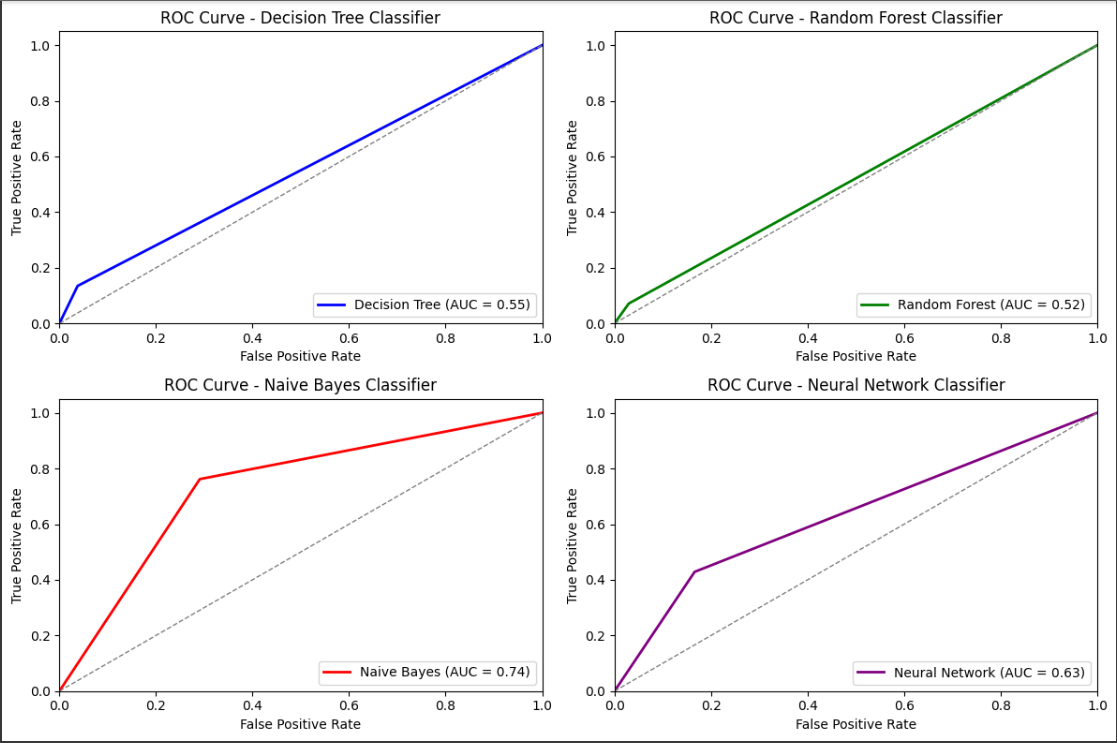
Description automatically generated

A graph with different colored bars

Description automatically generated

A graph of different colored squares

Description automatically generated



Performance:

* Decision Tree Classifier and Random Forest Classifier achieve high accuracy but struggle with correctly identifying the minority class (stroke). Their precision, recall, and F1-scores for stroke detection are relatively low.
* Naive Bayes Classifier demonstrates high recall for stroke detection, indicating its effectiveness in capturing true positive cases. However, its precision is low, resulting in a higher false positive rate.
* Neural Network Classifier achieves the highest recall for stroke detection among all classifiers but also exhibits a relatively high false positive rate, leading to lower precision and F1-score.

Computational Efficiency:

* Decision trees are computationally efficient to train and predict, suitable for tasks with interpretability requirements and moderate-sized datasets.
* Random forests have longer training times compared to decision trees but offer better performance by mitigating overfitting and variance.
* Naive Bayes classifiers are highly computationally efficient, with fast training and prediction times even for large datasets.
* Neural networks can be computationally intensive, especially for deep architectures and large datasets, but advancements in hardware and software have improved training times.

Applicability:

* Decision trees and random forests are suitable for tasks where interpretability is essential and can handle both numerical and categorical data, commonly used in healthcare and finance.
* Naive Bayes classifiers are applicable to tasks with high-dimensional data and feature independence assumptions, such as text classification and spam filtering.
* Neural networks are highly flexible and applicable to various tasks, including image recognition, natural language processing, and medical diagnostics. They excel at learning complex patterns from data but require careful tuning and evaluation.

Comparison:

* Decision Tree Classifier and Random Forest Classifier offer better interpretability and are computationally efficient compared to Neural Network Classifier, but they struggle with class imbalance and achieving high precision for stroke detection.
* Naive Bayes Classifier demonstrates high recall but low precision, making it less suitable for tasks requiring high precision, such as medical diagnostics.
* Neural Network Classifier achieves the highest recall but at the cost of lower precision, indicating potential overfitting and challenges in handling class imbalance.

# ***Conclusion***

The Decision Tree Classifier is the most suitable choice for predicting strokes, as it achieves a good balance between accuracy, precision, recall, and F1 score.

# ***References***

<https://data.mendeley.com/datasets/x8ygrw87jw/1>

<https://www.sciencedirect.com/science/article/pii/S0933365719302295?via%3Dihub>

Liu, Tianyu; Fan, Wenhui; Wu, Cheng (2019), “Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets”, Mendeley Data, V1, doi: 10.17632/x8ygrw87jw.1