CMR INSTITUTE OF TECHNOLOGY
(UGC AUTONOMOUS)
(Approved by AICTE, Affiliated to JNTUH, Accredited by NBA and NAAC with A+' Grade)
Kandlakoya (V), Medchal Road, Hyderabad – 501 401

DEPARTMENT OF ARTIFICIAL INTELLEGENCE AND MACHINE LEARNING

Automated Stroke Risk Prediction Using Balanced Machine Learning Models

Under the Guidance of

DR. K. SAMPATH

Associate Professor

Department of Computer Science and Engineering (AI&ML)

Presented By

22R01A6679 – E.SRUTHI
22R01A6688 – J.SHARANYA
22R01A66C2 – T.RASHMIKA

AGENDA

- Abstract
- Literature Review
- Problem Statement
- Objective
- Existing system
- Disadvantages
- Proposed System
- Algorithms
- Requirement Specifications
- System design
- Implementation
- Results and Discussion
- Societal Needs
- References

ABSTRACT

Stroke is a serious medical condition caused by interruption of blood flow to the brain, leading to disability or death .Early prediction of stroke risk helps doctors take preventive treatment and save lives .

This work uses Machine Learning algorithms to predict whether a person may get a stroke or not . The dataset is preprocessed and balanced to improve prediction accuracy . The proposed system improves reliability and helps doctors make better medical decisions early.

LITERATURE REVIEW

- **Hager Saleh (2019):** Used multiple ML models on large data; Random Forest ≈ 90% accuracy.
- **Youngkeun Choi (2020):** Decision Tree methods achieved about 98% stroke prediction accuracy.
- **Redwanul Islam (2021):** Compared KNN, SVM, DT and XG Boost using hospital patient data.
- **Tahia Tazin (2021):** Applied SMOTE balancing; Random Forest ≈ 96% accuracy.
- **Mohammed Saidul Islam (2022):** Used EEG signals with explainable AI for stroke detection (~80% accuracy).
- **Elias Dritsas (2022):** Proposed stacking ensemble model to improve prediction performance.
- **Mridha (2023):** Developed explainable ML stroke prediction system using SHAP & LIME (~91% accuracy).

PROBLEM STATEMENT

- Stroke prediction has been widely studied using machine learning techniques, but different algorithms produce different accuracy levels on the same dataset. Many studies focus on predicting whether a patient has stroke or not, rather than analysing which algorithm performs best for the prediction task. Because of this, it becomes difficult to identify the most reliable model for stroke risk classification. There is a need to systematically evaluate multiple machine learning models using the same patient dataset.

- In addition, dataset imbalance and feature selection significantly affect prediction performance, yet not all algorithms handle them equally. Without proper comparison, selecting an appropriate model for stroke prediction becomes uncertain. Therefore, this project aims to analyse patient data and compare various machine learning algorithms to determine which model provides the highest accuracy and most reliable prediction performance.

OBJECTIVE

The goal of this study was study aimed to achieve **Three objectives**:

(i) To create a trustworthy machine learning model to predict stroke disease.

(ii) To address the severe class imbalance issue that results from the stroke patients

(iii) To interpret the model output to gain a better comprehension of the decision-making process

EXISTING SYSTEM

- In the existing system, stroke prediction relies on traditional statistical methods and conventional machine learning algorithms such as Logistic Regression, Decision Tree, and SVM. These models primarily use structured clinical data and often treat features independently, without fully capturing complex relationships among risk factors. This leads to limited interpretability, reduced generalization ability, and potential overfitting.

- Ignoring feature interactions and lacking model transparency reduces the reliability of predictions in real-world. As a result, healthcare professionals may find it difficult to understand the reasoning behind predictions, which limits trust and practical adoption.

DISADVANTAGES

•The existing system mainly **relies on traditional methods** for prediction, which are not efficient for complex medical data analysis.

•It produces **lower accuracy** in results, making the prediction less reliable.

•The process takes more time to generate outcomes, **reducing overall efficiency**.

•It does not **utilize advanced or intelligent techniques** such as machine learning for better decision making.

PROPOSED SYSTEM

- The proposed system begins with collecting and analysing the stroke dataset. Data preprocessing is carried out to handle missing values and to encode categorical attributes properly. The dataset imbalance is addressed using the SMOTE balancing technique to improve learning performance. Feature selection is then applied to remove irrelevant parameters and enhance the quality of prediction.
- After preprocessing, multiple machine learning algorithms are trained and tested using the same dataset. Performance measures such as accuracy, precision, recall and F1-score are calculated for each model. The algorithms are compared to identify the best performing model for stroke prediction. Finally, the results are presented using graphical visualization to make interpretation easier and more reliable.

ALGORITHMS

Random Forest

Builds many decision trees using the dataset and each tree gives a prediction.
Final result is decided by majority voting, giving high accuracy.

Logistic Regression

Calculates probability of stroke occurrence between 0 and 1.

Based on threshold value, it classifies stroke or non-stroke.

Support Vector Machine (SVM)

Creates an optimal boundary separating stroke and non-stroke data points.
Maximizes distance between the two classes for better classification.

K-Nearest Neighbors (KNN)

Compares a patient’s data with nearest patient records in the dataset.
Prediction is based on majority class among nearest neighbors.

ALGORITHMS

- **Naive Bayes**

Uses probability rules assuming features are independent.
Predicts stroke based on likelihood from past data patterns.

- **XGBoost**

Builds models sequentially where each new model corrects previous errors.
Provides strong prediction performance for complex datasets.

- **CatBoost**

Specially handles categorical medical features efficiently.
Improves accuracy by combining multiple weak learners into a strong model.

REQUIREMENT SPECIFICATIONS

Hardware Requirements:

A high-performance computer is required for model training.

- **Processor**: Intel i3(min)
- **RAM**: Minimum 4 GB (16 GB recommended)
- **GPU**: NVIDIA GPU (optional but recommended for Machine learning)
- **Display**: Standard monitor for visualization

Stable internet connection for dataset access

Software Requirements:

- **Operating System**: Windows 10
- **Programming Language**: Python 3.x
- **Machine Learning Libraries**: sklearn, XGBoost, tkinter, catboost
- **Data Processing**: NumPy, Pandas, Scikit-learn.
- **Backend Framework**: Flask
- **Development Tools**: VS Code

SYSTEM DESIGN

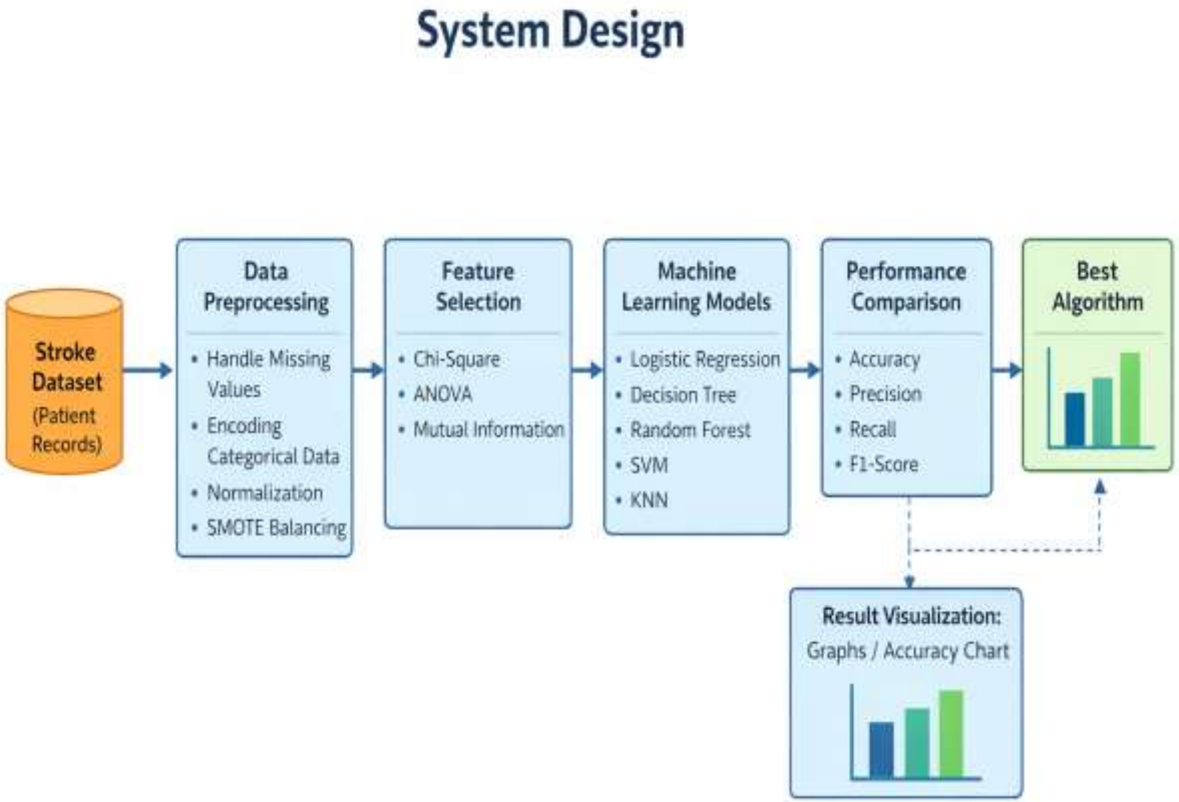


Figure: System architecture

- Dataset is loaded into the system.
- Data is preprocessed and cleaned.
- Imbalanced data is balanced using SMOTE.
- Important features are selected and cleaned.
- Multiple machine learning algorithms are trained.
- Models are tested using the same dataset Performance metrics like accuracy are calculated.
- Algorithms are compared to find the best model.
- Results are shown in graphical charts.

IMPLEMENTATION

- **Data Collection:**

  The dataset consists of **5,110 patient samples with 12 medical features**, used to train and test machine learning algorithms for stroke prediction accuracy comparison . this stroke healthcare dataset is collected from Kaggle, which contains patient medical details such as age, BMI, glucose level and other health conditions along with stroke outcome.

- **Data Pre-processing:**

  The collected data is cleaned by removing missing values, handling imbalance using SMOTE and selecting important features so the model can learn meaningful patterns.

- **Data Splitting:**

  The processed dataset is divided into training and testing sets, where 80% of data is used to train the model and 20% is used to evaluate its performance.

IMPLEMENTATION

- **Performance Evaluation:**

  Each model is evaluated using accuracy, precision, recall, F1-score and confusion matrix to measure prediction correctness.

- **Best Model Selection:**

  The algorithm with highest accuracy and least prediction error is selected as the final model for prediction.

- **Explainable AI:**

  An explainability technique is used to identify which health factors contribute most to stroke prediction.

- **Prediction:**

  Finally, new patient data is given to the trained model to predict whether the person has stroke risk or not.

- **Model Training:**

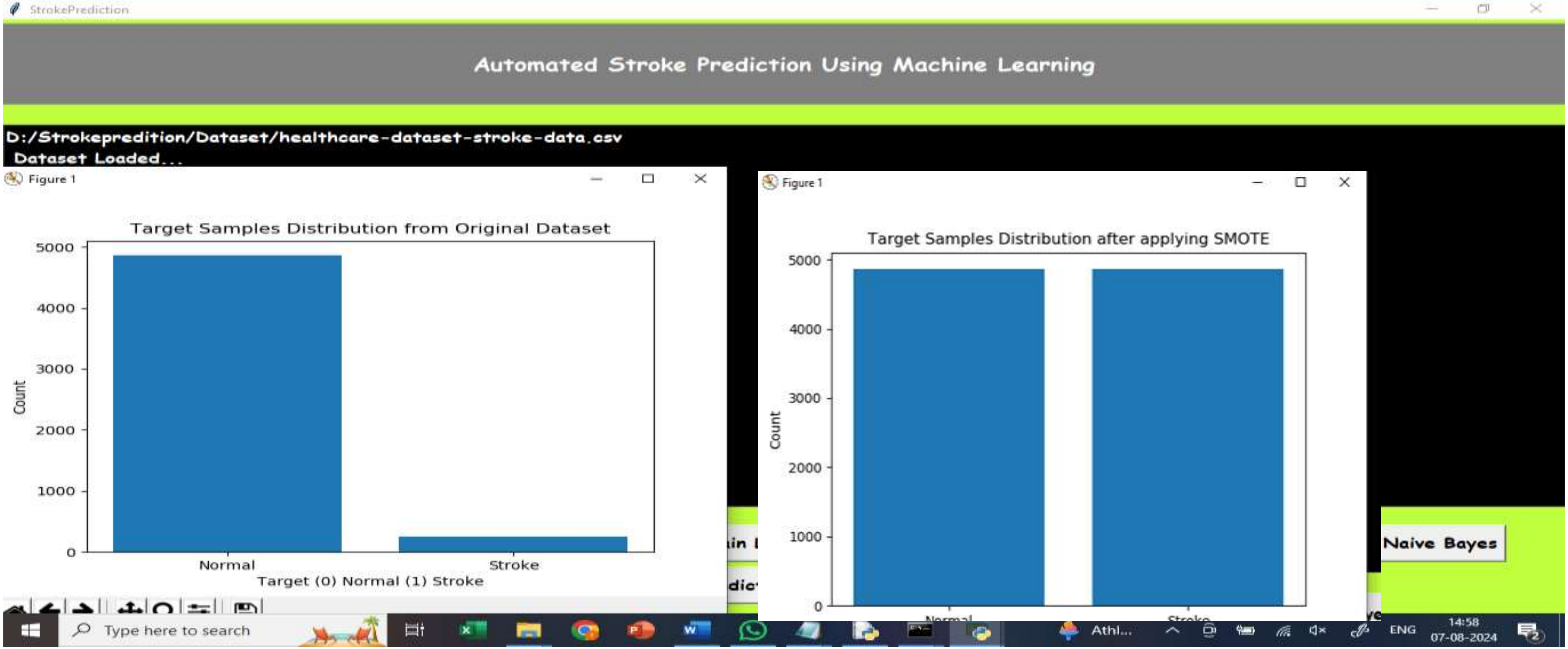
  Multiple machine learning algorithms are applied to the training data so that the system learns how to differentiate between stroke and normal patients.

SOCIETAL NEEDS

- Stroke is a major global health concern and one of the leading causes of mortality and long-term disability. Early risk identification can significantly reduce fatal outcomes and improve quality of life. However, selecting an appropriate and reliable prediction model remains a challenge in healthcare analytics.

- This project addresses that societal need by systematically comparing multiple machine learning algorithms to identify the most accurate and dependable model for stroke risk prediction. By enhancing prediction reliability and supporting data-driven healthcare decisions, the study contributes toward preventive healthcare strategies and technological advancement in medical research.

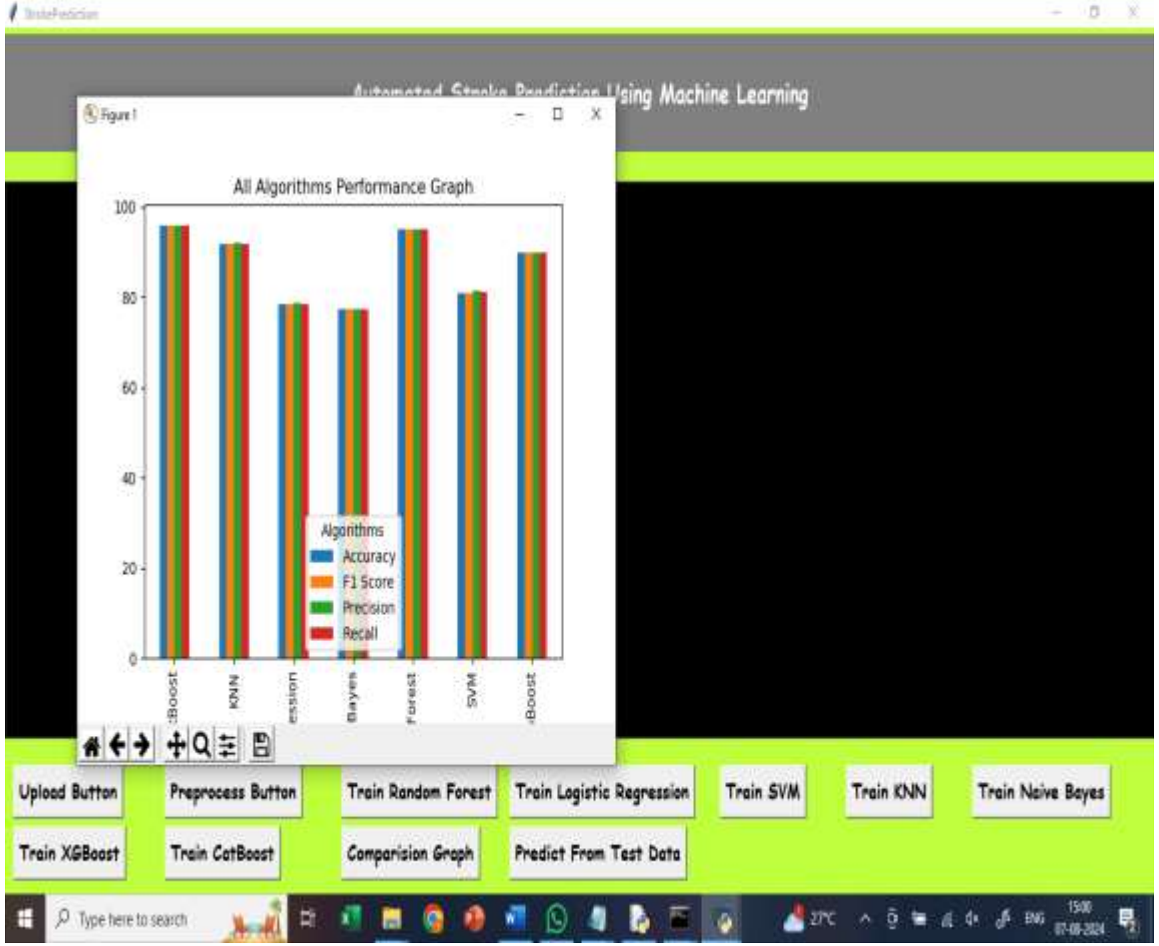
RESULTS AND DISCUSSION

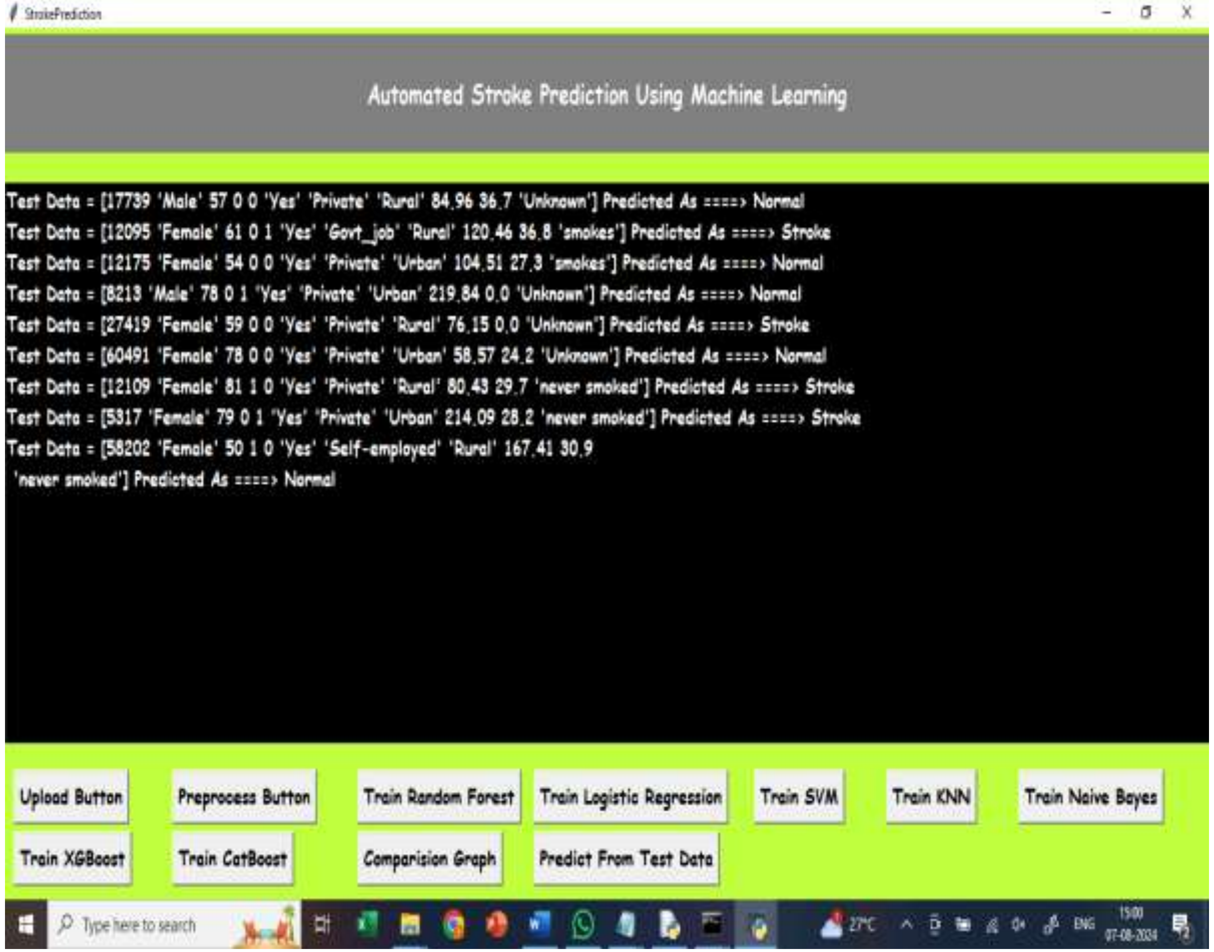


StrokePrediction
Automated Stroke Prediction Using Machine Learning
Figure 1
Target Samples Distribution from Original Dataset
5000
4000
3000
2000
1000
0
Count
Normal
Stroke
Target (0) Normal (1) Stroke
Figure 1
Target Samples Distribution after applying SMOTE
5000
4000
3000
2000
1000
0
Count
Normal
Stroke
Naive Bayes
Type here to search
Athl...
ENG
14:58
07-08-2024

```
D:/Strokepredition/Dataset/healthcare-dataset-stroke-data.csv
 Dataset Loaded...
```

RESULTS AND DISCUSSION





```
Test Data = [17739 'Male' 57 0 0 'Yes' 'Private' 'Rural' 84.96 36.7 'Unknown'] Predicted As ====> Normal
Test Data = [12095 'Female' 61 0 1 'Yes' 'Govt_job' 'Rural' 120.46 36.8 'smokes'] Predicted As ====> Stroke
Test Data = [12175 'Female' 54 0 0 'Yes' 'Private' 'Urban' 104.51 27.3 'smokes'] Predicted As ====> Normal
Test Data = [8213 'Male' 78 0 1 'Yes' 'Private' 'Urban' 219.84 0.0 'Unknown'] Predicted As ====> Normal
Test Data = [27419 'Female' 59 0 0 'Yes' 'Private' 'Rural' 76.15 0.0 'Unknown'] Predicted As ====> Stroke
Test Data = [60491 'Female' 78 0 0 'Yes' 'Private' 'Urban' 58.57 24.2 'Unknown'] Predicted As ====> Normal
Test Data = [12109 'Female' 81 1 0 'Yes' 'Private' 'Rural' 80.43 29.7 'never smoked'] Predicted As ====> Stroke
Test Data = [5317 'Female' 79 0 1 'Yes' 'Private' 'Urban' 214.09 28.2 'never smoked'] Predicted As ====> Stroke
Test Data = [58202 'Female' 50 1 0 'Yes' 'Self-employed' 'Rural' 167.41 30.9
 'never smoked'] Predicted As ====> Normal
```

REFERENCES

[1] W. H. Organization, Global Health Estimates: Leading Causes of Death, World Health Organization, 2020.

[2] B. Ovbiagele and M. Nguyen-Huynh, Stroke Epidemiology: Advancing Our Understanding of Disease Mechanism and Therapy, Neurotherapeutics, 2011.

[3] S. Katan and A. Luft, Global Burden of Stroke, Seminars in Neurology, 2018. :contentReference

[4] M. A. Alkhouli et al., Machine Learning Approaches for Stroke Prediction: A Review, IEEE Access, 2020.

[5] J. Chen and X. Guestrin, XGBoost: A Scalable Tree Boosting System, KDD, 2016.

[6] S. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions (SHAP), NeurIPS, 2017.

[7]D.Molnar,Interpretable 2019.Del Pup and M. Atzori, Self-Supervised Learning in Biomedical Data, IEEE Access,2023.

THANK YOU