Stroke prediction App

ON

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

(Information Technology)

By

Chinmay Chaudhari (06)

Kshitij Hundre (18)

Shubham Jha (19)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

AIDS Lab Exp 11

Aim: Mini Project – Stroke Risk Prediction Using Machine Learning

---

Chapter 1: Introduction

1.1 Introduction

Stroke is a leading cause of death and disability worldwide. Early prediction is essential for prevention and management, but traditional medical models struggle with low recall and interpretability. This project leverages machine learning to build a predictive and interpretable system for stroke detection using clinical data.

1.2 Objectives

- To preprocess and analyze stroke-related health data
- To apply class balancing techniques like SMOTE for rare event prediction
- To train ML models including Logistic Regression, Decision Tree, and XGBoost
- To explain model predictions using SHAP (SHapley Additive Explanations)
- To deploy an interactive prediction system using Streamlit

1.3 Motivation

Medical professionals require tools that are not only accurate but also interpretable. Existing scoring systems lack transparency and struggle with imbalanced data. This project aims to provide a trustworthy, AI-based decision support system.

1.4 Scope of the Work

- Focus on supervised classification techniques
- Dataset sourced from Kaggle (Stroke Prediction Dataset)
- Visualize insights using seaborn and SHAP
- Deployable tool for real-time stroke risk evaluation

1.5 Feasibility Study

- *Technical Feasibility:* Uses Python, Pandas, Scikit-learn, SHAP, and Streamlit—all open-source and well-documented libraries.
- *Operational Feasibility:* Easily integrable into a web or mobile healthcare system.
- *Economic Feasibility:* Cost-effective due to use of free tools and datasets.

Chapter 2: Literature Survey

2.1 Introduction

Research in stroke prediction using ML is ongoing. This section compares notable models and techniques.

2.2 Problem Definition

Stroke is a rare medical condition, which creates class imbalance issues for ML models. The challenge is to build a system that can detect such rare instances while remaining interpretable and trustworthy.

2.3 Review of Literature

1. Kaggle, "Stroke Prediction Dataset," 2021

The Stroke Prediction Dataset, sourced from Kaggle, is a widely used healthcare dataset for building stroke classification models. It includes features such as age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. These attributes help in identifying key risk indicators for stroke. Despite being valuable for real-world applications, the dataset suffers from a significant class imbalance, with only 249 stroke cases out of 5,110 records. Additionally, the presence of missing values in the BMI column requires preprocessing before applying machine learning techniques.

2. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017

This foundational paper introduces SHAP (SHapley Additive exPlanations), a game theory-based approach to interpreting machine learning model predictions. SHAP values provide both local and global explanations, showing how much each feature contributed to a particular decision. This level of interpretability is especially valuable in healthcare applications, where model transparency is crucial. However, SHAP can be computationally expensive for large or complex models and may assume feature independence in certain implementations, which can limit its real-world scalability.

3. N. V. Chawla et al., “SMOTE: Synthetic Minority Over-sampling Technique,” JAIR, 2002

SMOTE is a technique developed to address the challenge of imbalanced datasets, where the number of instances in one class significantly outweighs the other. The method generates synthetic samples for the minority class by interpolating between existing examples, thereby improving the performance of classifiers that are otherwise biased toward the majority class. While SMOTE has proven to be highly effective, especially in medical datasets with rare outcomes like stroke, it may introduce noise and lead to overfitting if not combined with proper cross-validation and model tuning.

Chapter 3: Design and Implementation

3.1 Introduction

This project follows a modular data science pipeline including data loading, preprocessing, modeling, explanation, and deployment.

3.2 Requirements

- **Hardware:** 4GB RAM minimum
- **Software:** Python 3.x, Jupyter/Colab, Scikit-learn, Pandas, SHAP, Streamlit

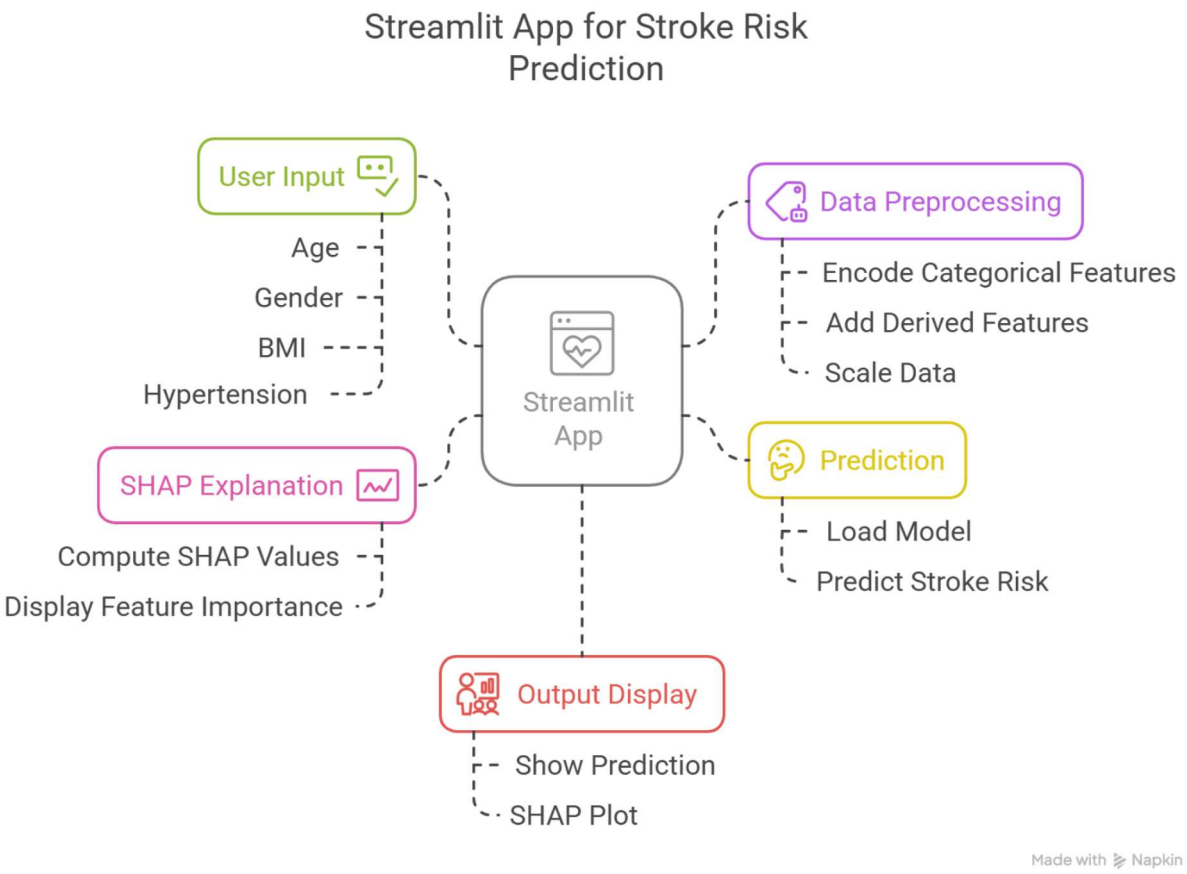
3.3 Proposed Design (CRISP-DM)

1. **Data Collection:** Kaggle's Stroke Prediction Dataset
2. **Data Cleaning:** Remove missing values, encode categorical variables
3. **EDA:** Explore relations (age, hypertension, BMI vs stroke) using bar plots and heatmaps
4. **Balancing:** Use SMOTE to synthesize samples for the minority class
5. **Modeling:** Train Logistic Regression, Decision Tree, Random Forest, XGBoost
6. **Interpretation:** Use SHAP summary plots to visualize influential features
7. **Deployment:** Build a Streamlit web app for prediction and explanation

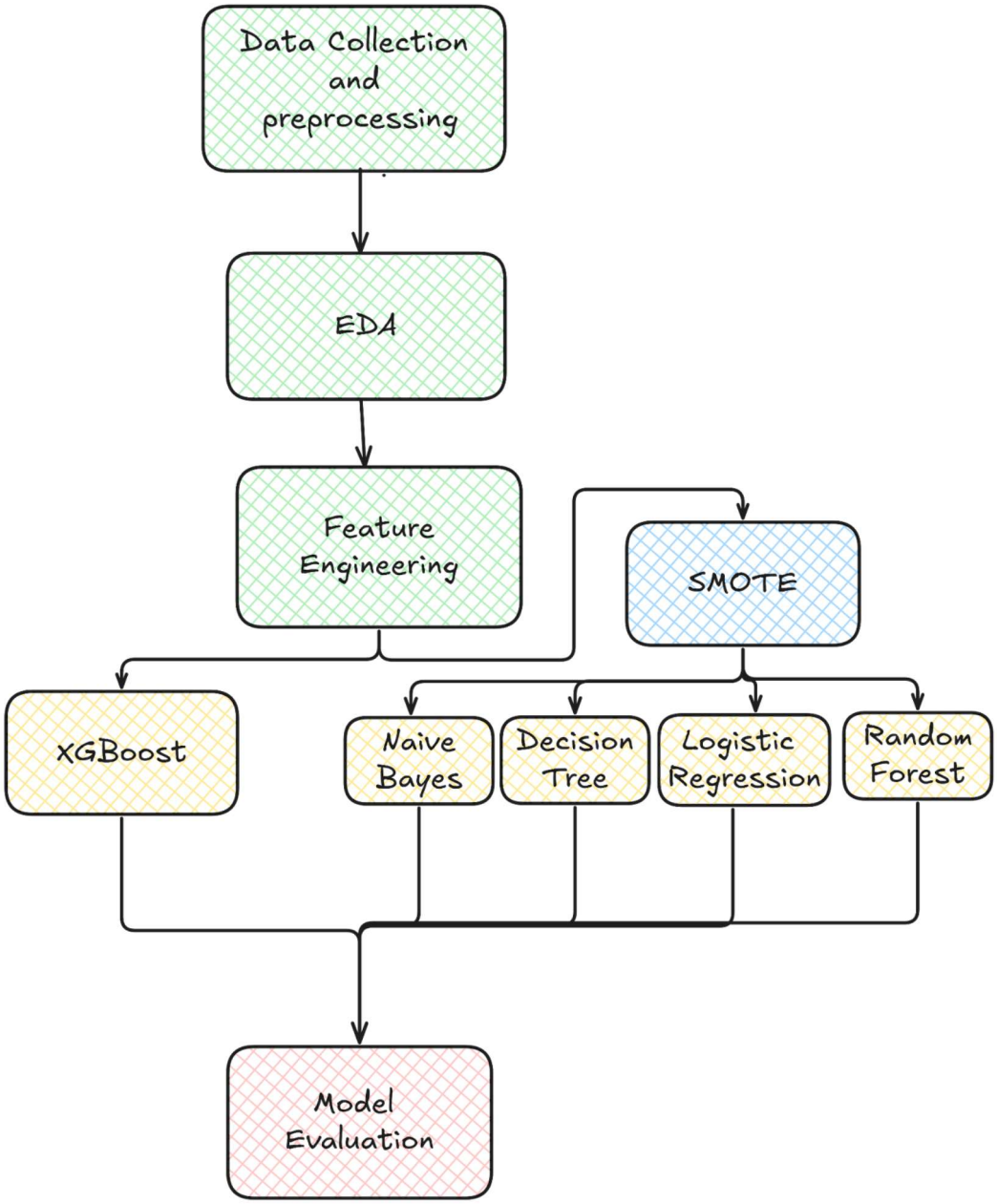
3.4 Algorithms Used

- **Logistic Regression:** High interpretability and performance post-SMOTE
- **Decision Tree:** Classification model that works on splitting on nodes that give highest information gain. moderate results on the current dataset.
- **Naive Bayes:** Mathematical model working on the principle of bayesian theorem, moderate results on the current dataset.
- **Random Forest & XGBoost:** Ensemble methods for boosting recall, moderate performance but overall very poor recall.
- **SHAP:** Used for both global and local interpretability
- **SMOTE:** Used for generating synthetic minority stroke instances

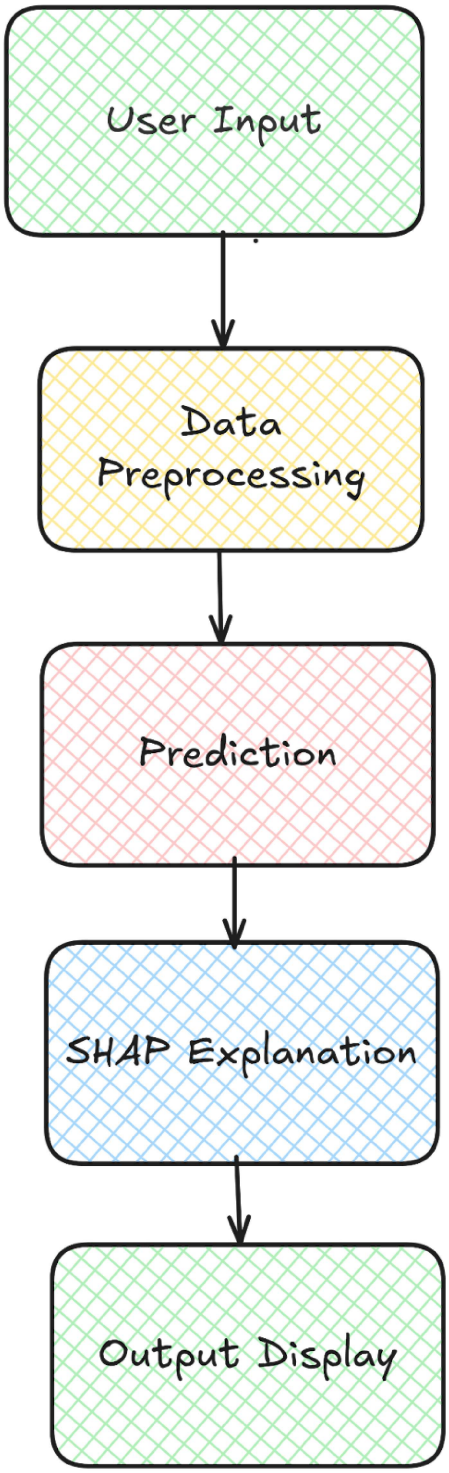
3.5 Diagrams



(Functionality of the Stroke Prediction App)



(Flow Diagram of the model training)



(Flow Diagram of the Streamlit App)

Chapter 4: Results and Discussion

4.1 Introduction

Models were evaluated using metrics such as Accuracy, Precision, Recall, and AUC. Special focus was on Recall due to class imbalance.

4.2 Evaluation Summary

To assess the effectiveness of various classification algorithms in stroke prediction, we evaluated Decision Tree, Logistic Regression (with tuned hyperparameters), Random Forest, and XGBoost models using precision, recall, f1-score, and accuracy. All models were trained and tested on a SMOTE-balanced dataset to handle the class imbalance.

Decision Tree (SMOTE, Default Threshold) Test Set Performance:

	precision	recall	f1-score	support
0	0.96	0.92	0.94	972
1	0.16	0.28	0.20	50
accuracy			0.89	1022
macro avg	0.56	0.60	0.57	1022
weighted avg	0.92	0.89	0.90	1022

- **Decision Tree (SMOTE, Default Threshold)** achieved a test set accuracy of **0.89**. It showed high precision and recall for the majority class (no stroke), with a **precision of 0.96** and **recall of 0.92.** However, for the minority class (stroke), its **f1-score was only 0.20**, reflecting poor sensitivity despite oversampling.

Best Hyperparameters Logistic Regression (SMOTE, Tuned) Test Set Performance:

	precision	recall	f1-score	support
0	0.98	0.83	0.90	972
1	0.17	0.70	0.27	50
accuracy			0.82	1022
macro avg	0.58	0.76	0.59	1022
weighted avg	0.94	0.82	0.87	1022

- **Logistic Regression (SMOTE, Tuned Hyperparameters)** yielded an **accuracy of 0.82**. It demonstrated relatively better balance, with a **recall of 0.70** for the stroke class and an **f1-score of 0.27**, indicating improved sensitivity compared to the Decision Tree**.** The overall weighted average f1-score was **0.87.**

Random Forest Test Set Performance:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	972
1	0.00	0.00	0.00	50
accuracy			0.94	1022
macro avg	0.48	0.49	0.48	1022
weighted avg	0.90	0.94	0.92	1022

- **Random Forest** performed well on the majority class, with a **precision of 0.95** and **recall of 0.99**, leading to a high accuracy of **0.94.** However, it failed to detect any stroke cases, with a **precision, recall, and f1-score of 0.00** for the stroke class, showing the model's bias toward the majority class**.**

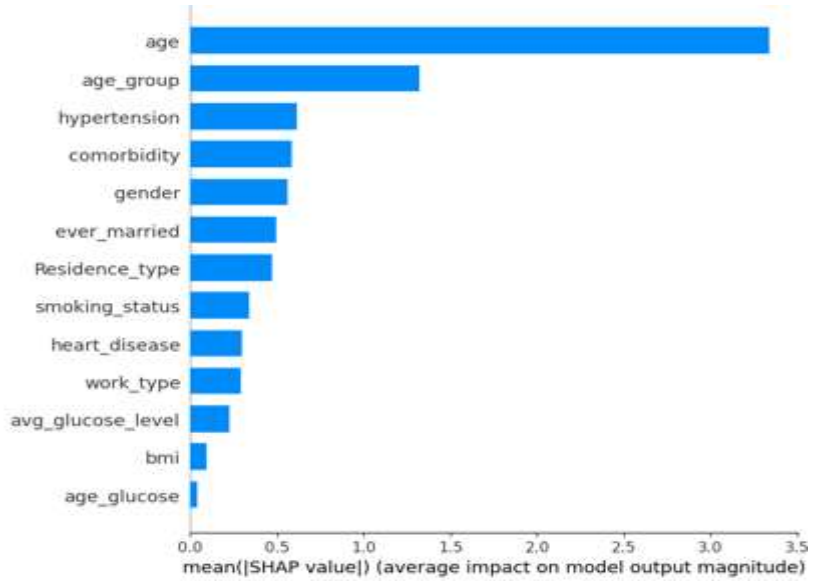
XGBoost Test Set Performance:

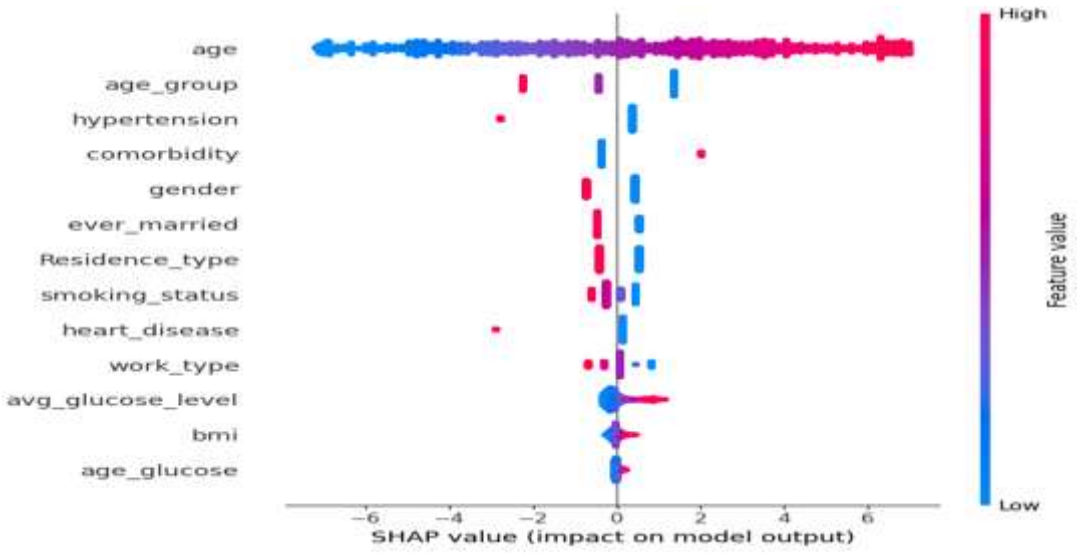
	precision	recall	f1-score	support
0	0.96	0.95	0.95	972
1	0.20	0.26	0.23	50
accuracy			0.91	1022
macro avg	0.58	0.60	0.59	1022
weighted avg	0.92	0.91	0.92	1022

- **XGBoost** achieved an accuracy of **0.91**. It maintained a strong performance on the majority class with **f1-score of 0.95**, but its ability to identify the stroke class remained limited with a **recall of 0.26** and **f1-score of 0.23**.

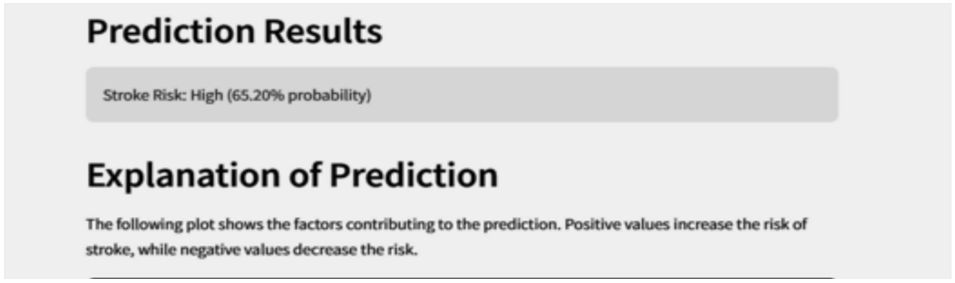
In summary, while all models performed well for the majority class, **Logistic Regression with hyperparameter tuning** provided the best balance between precision and recall for stroke prediction, making it the most promising model for early stroke detection in imbalanced medical datasets.

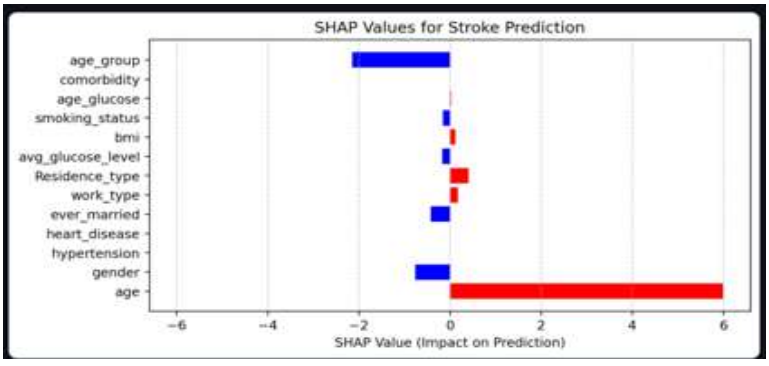
4.2.5 Visualising the key factors involved in positive stroke cases:





4.3 Screenshot/Output Section





Chapter 5: Conclusion and Future Scope

5.1 Conclusion

We successfully built a machine learning-based stroke prediction system that is both accurate and explainable. By combining SMOTE for balancing and SHAP for interpretation, our system supports clinical decision-making with confidence.

5.2 Future Scope

- Include more features like cholesterol or ECG data for better predictions
- Extend to real-time systems using health IoT or hospital dashboards
- Integrate into mobile applications for public health use

5.3 Societal Impact

- Early stroke risk prediction helps reduce mortality and improve patient care
- Builds trust among medical professionals via transparent predictions
- Scalable system that can be used in low-cost, remote diagnosis setups