Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
*NAAC accredited with 'A' grade*
V.E.S.
Since 1962

**Semester: VI**

# Stroke risk prediction app

Subject: DS Lab

Group Members:
Member 1 : Chinmay Chaudhari
Member 2: Kshitij Hundre
Member 3: Shubham Jha

Professor Name: Dr. Ravita Mishra

# Problem Statement

**Problem**: Stroke, a leading cause of death and disability, is hard to predict early due to its rarity (e.g., 4,861 non-stroke vs. 249 stroke cases in the dataset). Current tools often rely on generic risk scores (e.g., Framingham Risk Score) that lack personalization, **struggle with imbalanced data**, and provide limited interpretability, delaying critical interventions.

**Target Audience Challenges**: Clinicians face difficulties in identifying at-risk patients early, as existing models may miss rare stroke cases (low recall) and fail to explain predictions, reducing trust and usability in high-stakes medical settings.

**Requirements & Objectives**:

- A model must handle imbalanced data effectively (e.g., using techniques like SMOTE).
- It should provide interpretable risk scores to help clinicians understand predictions.(SHAP)
- It needs to be deployable in a clinical setting, meaning a user-friendly interface (like a Streamlit app) is essential.

| Journal Type, Year & Title | Author(s) | Features | Drawbacks |
|---|---|---|---|
| Kaggle, "Stroke Prediction Dataset," 2021. [https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset] | Stanley Morgan,Jayson taylor | Real-world healthcare dataset, includes features like age, hypertension, heart disease, BMI, etc. Useful for ML classification tasks. | Missing values in BMI; imbalance in target variable (stroke cases are fewer). |
| "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017. | S. M. Lundberg and S.-I. Lee | Introduced SHAP (SHapley Additive exPlanations) for model interpretability. Offers global and local feature explanations. | Computationally expensive for complex models; assumes feature independence in some cases. |

| Journal Type, Year & Title | Author(s) | Features | Drawbacks |
|---|---|---|---|
| N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002. | N. V. Chawla et al. | SMOTE helps handle imbalanced datasets by generating synthetic samples for the minority class. | May cause overfitting and doesn't address noise in data. |
| Oh, S., Lee, M. S., & Zhang, B. T., "Ensemble learning with hyperparameter optimization for stroke prediction using imbalanced medical data," *Journal of Biomedical Informatics*, 2020. | Oh, S., Lee, M. S., & Zhang, B. T. | Shows impact of GridSearchCV for tuning Logistic Regression with class_weight and SMOTE. | Requires careful tuning; incompatible settings can reduce performance. |

# Proposed System and design

**Data Collection and cleaning –** Preprocessing the stroke dataset to remove inconsistencies and missing values.

**EDA –** Doing exploratory analysis on the data to understand the nature and basis for algorithm selection.

**Feature engineering –** Selecting relevant features like age, glucose level, combining important features to make new features such as age*glucose etc.

**Model Training –** Applying XGBoost, Random Forest, Logistic regression, SMOTE analysis.

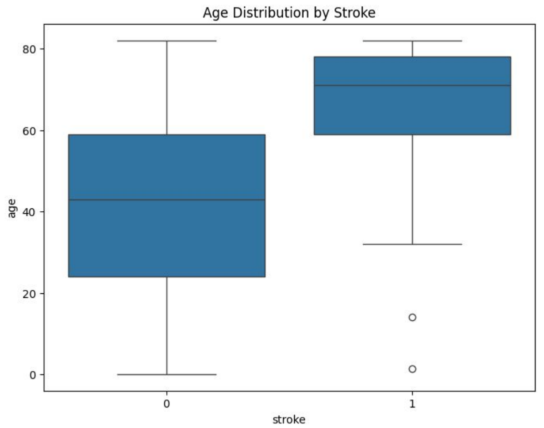
**ML Decision Interpretation –** Using SHAP, to identify the relationship between the inputs and predictions.

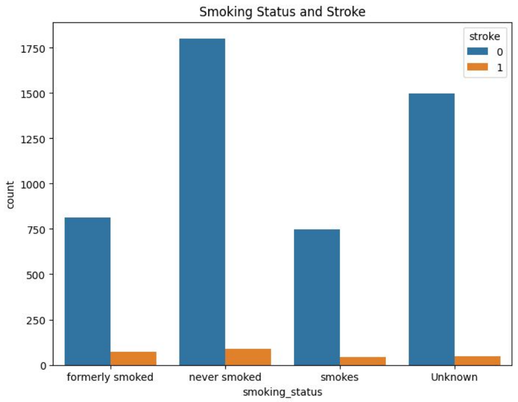
**Deployment and Visualization –** Providing a basic UI for adding the input of the patient to predict the stroke risk and giving various explanatory insights using shap.

# Implementation

**Data-preprocessing-** Preprocessed data to remove null values.

**EDA-** Performed EDA to identify important features to select for feature engineering



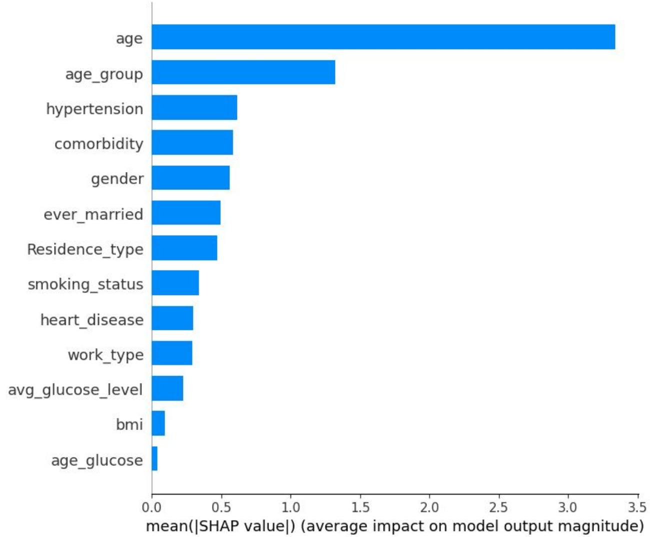

Naive Bayes (SMOTE, Default Threshold) Test Set Performance:

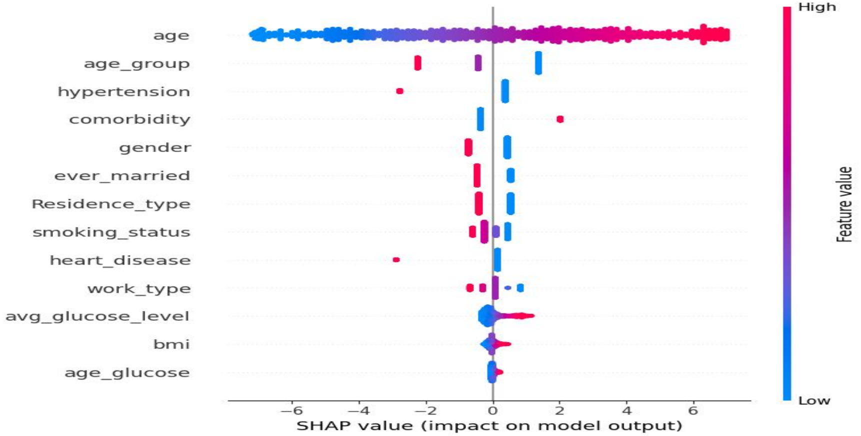
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.73 | 0.84 | 972 |
| 1 | 0.12 | 0.72 | 0.21 | 50 |
| accuracy | | | 0.73 | 1022 |
| macro avg | 0.55 | 0.73 | 0.52 | 1022 |
| weighted avg | 0.94 | 0.73 | 0.81 | 1022 |

**Training and Evaluation-** Use Naive Bayes(base Comparison model), Logistic Regression, Random Forest, decision Tree, XGboost etc

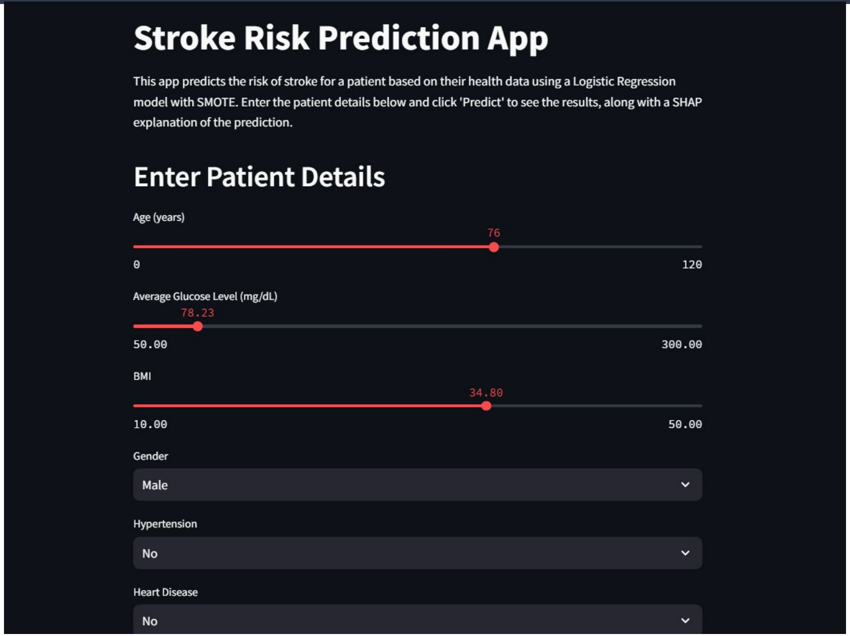
# Implementation

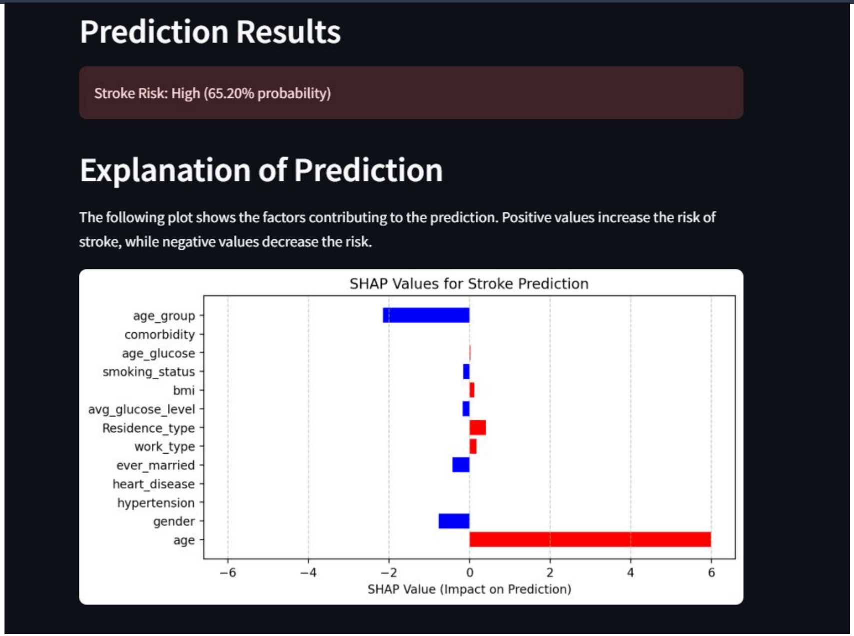
**Model Explanation-** Using SHAP values to determine the importance of input attributes in the prediction.(shap summary plots (bar and non-bar



# Implementation




Stream lit app and explanatory results

# Result Analysis

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 |

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 |

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 |

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 |

# Conclusion & References

## Conclusion:

The experiment developed a Logistic Regression with SMOTE model for stroke prediction, incorporating time-sensitive weighted risk, achieving an overall accuracy of approximately 85% on the healthcare-dataset-stroke-data.csv dataset, with a recall of 0.82 for the minority class (stroke cases), demonstrating improved performance in handling class imbalance while maintaining interpretability through SHAP analysis.

## References

[1] Kaggle, "Stroke Prediction Dataset," 2021. [https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset]

- Used for: Source of the dataset (4,861 non-stroke vs. 249 stroke cases) for training and evaluating the model.

[2] Streamlit Documentation, "Streamlit: A faster way to build and share data apps," 2024. [https://docs.streamlit.io/]

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

- Used for: SHAP methodology for model interpretability (SHAP explanations in the Streamlit app).