Brain Stroke Prediction

PRML MAJOR PROJECT

Kamuju Aashish | B21AI001
Poojita Oppangi | B21CS055
Kovidh Pothireddy | B21AI026

I. INTRODUCTION

A stroke is a serious medical condition caused by the interruption of blood flow to the brain, resulting in cell death and loss of brain function. Early detection and treatment are essential for successful recovery from a stroke. In this project, we aim to predict the occurrence of stroke based on various risk factors, including age, gender, hypertension, heart disease, smoking status, and BMI. We will use machine learning algorithms to analyze the data and predict the likelihood of stroke.

II. DATA VISUALIZATION AND PREPROCESSING

Data visualization and preprocessing techniques were utilized in the Brain Stroke Prediction to prepare the data for modeling. Visualization techniques such as histograms, box plots, and scatter plots were used to understand the distribution, outliers, and relationships between different features. Correlation matrices and heatmaps were used to identify any correlations between features and the target variable. Preprocessing steps involved handling missing values, outliers, and class imbalance, along with feature engineering to extract useful information from existing features such as BMI and age group. These techniques improved the data quality, making it suitable for modeling and leading to better predictions.

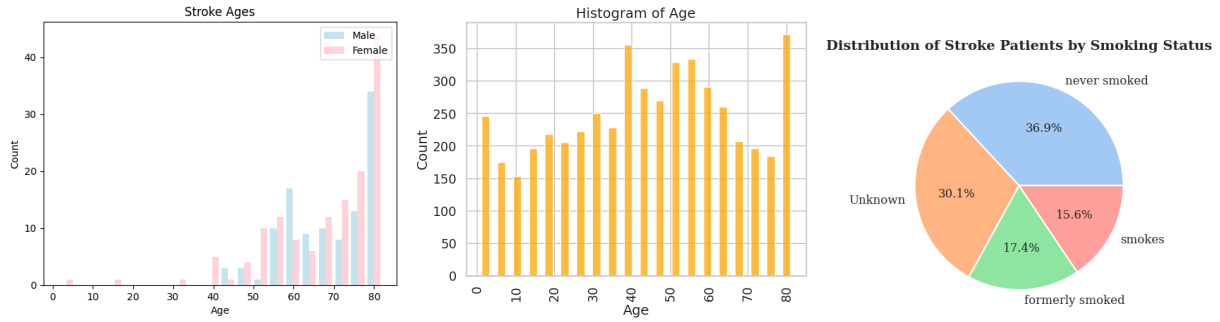

Figure 1: Data Visualisation

Preprocessing

Encoding:

The categorical variables are encoded using LabelEncoder and OneHotEncoder. The gender and ever married variables are label encoded while the work type variable is one-hot encoded to convert them into numerical values for modeling.

Replacing unknown values:

The 'smoking status' column has 'Unknown' values which are replaced with the mode value of the column.

Feature Scaling:

The numerical features are scaled using StandardScaler and MinMaxScaler to ensure that they have the same scale and to improve the performance of the model. The 'age' and 'avg glucose level' columns are standardized, while the 'bmi' column is normalized.

SMOTE:

SMOTE (Synthetic Minority Over-sampling Technique) is a powerful technique for dealing with imbalanced datasets like the brain stroke dataset, where the minority class is significantly underrepresented. It works by generating synthetic samples of the minority class, thereby balancing the dataset and improving the performance of the classifier. SMOTE helps to reduce the bias towards the majority class and increases the sensitivity of the model towards the minority class, leading to better predictions.

III. MACHINE LEARNING MODELS

Models used are:

1. Decision Tree
2. Random Forest
3. Logistic Regression
4. Support Vector Machine (SVM)
5. Multi-Layer Perceptron (MLP)
6. Gradient Boosting
7. XGBoost
8. AdaBoost
9. K-Nearest Neighbors (KNN)
10. Navie-Bayes
11. Bagging

Evaluation Metrics

Accuracy
the proportion of correct predictions among all predictions made by the model. In the case of the brain stroke dataset, accuracy represents the percentage of correctly predicted stroke and non-stroke cases.
Precision
the proportion of true positives (correctly predicted stroke cases) among all cases predicted as positive (including true positives and false positives). Precision measures how precise the model's positive predictions are in identifying stroke cases.
Recall
the proportion of true positives (correctly predicted stroke cases) among all actual positive cases (including true positives and false negatives). Recall measures how well the model can identify all actual stroke cases.
F1 score
a weighted average of precision and recall, with a range between 0 and 1, where 1 indicates the best performance. The F1 score is a balance between precision and recall and is a useful metric when there is an uneven class distribution.
ROC AUC score
the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity). The ROC AUC score represents the overall performance of the model in distinguishing between positive and negative cases, with a range between 0.5 (random guessing) and 1 (perfect performance).

Best Evaluation metrics

In the case of the brain stroke dataset, a high recall is desirable as it is important to correctly identify all actual stroke cases, while precision is also important to avoid misclassifying non-stroke cases as stroke. The ROC AUC score can be used to evaluate the overall performance of the model in predicting stroke and non-stroke cases.

IV. MODEL BUILDING

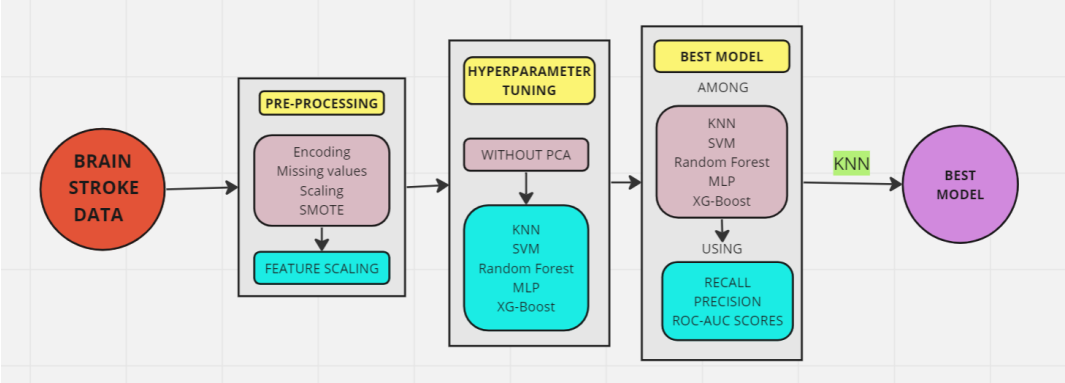

Figure 2: Pipeline

Why not PCA/LDA ?

In the case of the brain stroke dataset, the reduced number of features through PCA or LDA doesn't capture the important information needed to predict stroke accurately. This leads to a decrease in performance for models that rely on that information. Additionally, PCA and LDA assume that the data is linearly separable, which is not always the case. This also leads to a decrease in model performance when applied to non-linear datasets.
However, It's important to experiment with different techniques and evaluate their impact on model performance before making a final decision.

Figure 3: Model comparision before and after PCA

Hyperparameter tuning of models

Hyperparameter tuning is a critical step in the model evaluation process as it helps to optimize the model's performance by identifying the best combination of hyperparameters. Hyperparameters are parameters that cannot be learned from the data but can be set before the learning process. Different models have different hyperparameters, and finding the best combination of hyperparameters can significantly improve the model's accuracy and reduce overfitting.

Training and evaluating each model

By training multiple models on the same dataset and evaluating their performance on a common set of metrics, we can determine which model performs the best in terms of accuracy, precision, recall, F1 score, or other evaluation metrics.This process also allows us to identify the strengths and weaknesses of each model and select the best one for the specific problem at hand.

V. MODEL COMPARISON

The performance metrics we mentioned above are used to evaluate the performance of each model.

Model	Recall	Precision	ROC AUC Score	Accuracy	F1 Score
Decision Tree	0.919	0.795	0.899	0.840	0.851
Random Forest	0.956	0.899	0.984	0.925	0.927
Logistic Regression	0.840	0.771	0.857	0.795	0.804
Support Vector Machine	0.977	0.890	0.961	0.928	0.932
Multi-Layer Perceptron	0.955	0.894	0.966	0.921	0.924
Gradient Boosting	0.950	0.887	0.977	0.914	0.917
XGBoost	0.940	0.964	0.990	0.953	0.952
AdaBoost	0.913	0.738	0.817	0.795	0.816
K-Nearest Neighbors	0.985	0.828	0.962	0.890	0.899
Naive Bayes	0.797	0.751	0.837	0.766	0.773
Bagging(knn)	0.984	0.824	0.887	0.887	0.897

VI. BEST MODEL

As we can see , K-Nearest Neighbors has the best recall score inturn correctly identifying maximum actual stroke cases and rarely missing out on actual stroke cases. Even though XGBoost, Random Forest have optimal values of other performance metrics , we consider Recall as our prime objective for evaluation of models here.As, It is crucial to not miss any of the actual stroke cases. But, we can choose other models according to the requirement.

Contributions:

Aashish Kamuju (B21AI001): Model Training and Evaluation without hyperparameter tuning and Report
Kovidh Pothireddy (B21AI026): Model Training and Evaluation with hyperparameter tuning and without PCA, EDA and Visualizations.
Oppangi Poojita (B21CS055): Model Training and Evaluation with hyperparameter tuning and with PCA and Report

References:

"Machine Learning: A Probabilistic Perspective" by Michael Nielsen"
Scikit-learn Documentation: https://scikitlearn.org/stable/modules/classes.html
https://www.kaggle.com/code/angshumandc/brain-stroke-data-analysis
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf