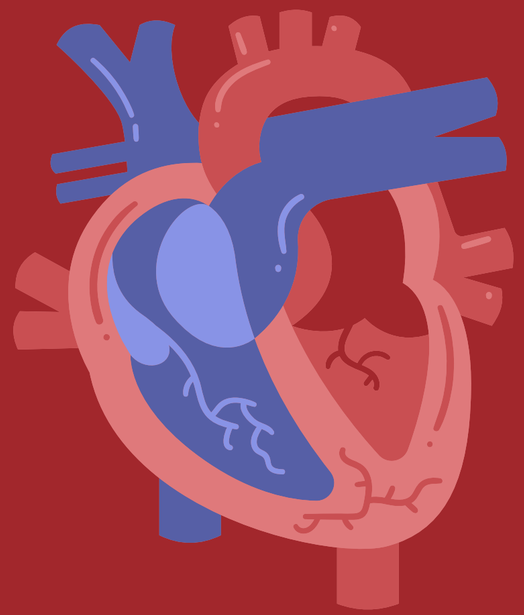
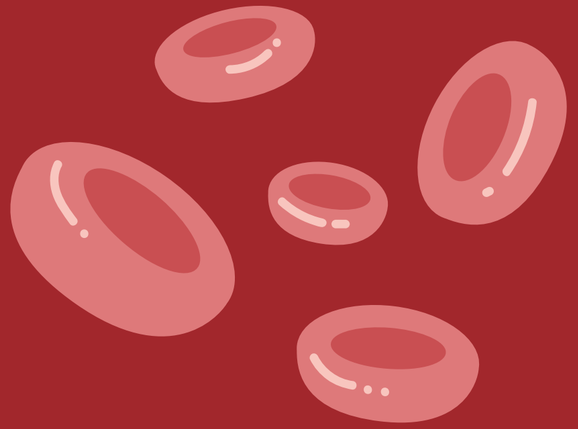
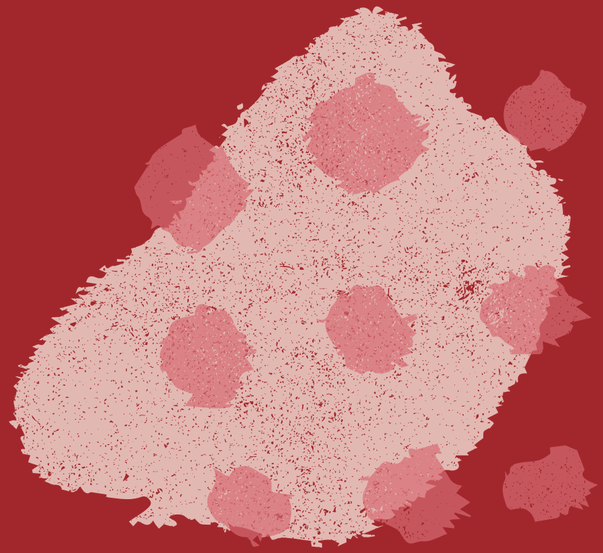
HEART STROKE PREDICTION

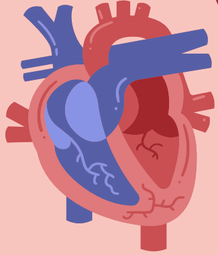






Submitted by-Devit Sah
Submitted by-Anshul Mahajan

CONTENTS



INTRODUCTION

TECHNICAL BACKGROUND OF PROJECT

The Stroke Prediction Project leverages machine learning techniques to predict the likelihood of stroke occurrences using clinical and demographic data. It integrates methods for handling imbalanced data, such as SMOTEENN, and employs hyperparameter tuning for optimized performance.
Technical concept used:
Date Pre-Processsing
Feature Engineering
Model Trainng
Imbalanced Data Handling
Evaluation Metrices

MOTIVATION

Strokes are among the leading causes of death and disability worldwide. Early detection and intervention can significantly reduce their impact. This project aims to provide a predictive solution that can assist healthcare providers in identifying at-risk individuals and implementing timely interventions.

PROBLEM STATEMENT

The model predicts the chances a person will have stroke based on symptoms like age, gender, average glucose level, smoking status, body mass index, work type and residence type.

AREA OF APPLICATION

The model can be used in:

- Healthcare Screening Programs: Prioritize high-risk individuals for preventive measures.
- Hospitals and Clinics: Support diagnostic workflows with additional insights.
- Insurance Companies: Assess risk profiles for policies and claims.

DATASET AND INPUT FORMAT

- Dataset Source: Kaggle - Stroke Prediction Dataset.
  -> Features:
  - Numerical: Age, BMI, Avg_glucose_level.
  - Categorical: Gender, Work Type, Smoking Status.
  - Target: Stroke occurrence (binary: 0 = No, 1 = Yes)
- Input Format: CSV file with preprocessed clinical and demographic features.

OBJECTIVE

Main Subjective

To develop a robust machine learning model for predicting the likelihood of stroke occurrences using clinical and demographic data. The model aims to assist healthcare providers in early identification of at-risk individuals and improve preventive care.

Sub Objective

1. Handle imbalanced datasets effectively to ensure reliable predictions for minority classes.
2. Perform feature engineering to maximize model performance.
3. Evaluate multiple classification models to identify the most accurate and efficient one.
4. Provide interpretable results for actionable insights in real-world healthcare applications.

METHODOLOGY

1

Load Dataset

heart_stroke_prediction

2

One-hot Encoding

Handling categorical variables

3

Train-Test Split

Split data into training and testing sets

4

Model Training

Logistic Regression
K-Neighbors Classifier,Random Forest

5

Model Evaluation

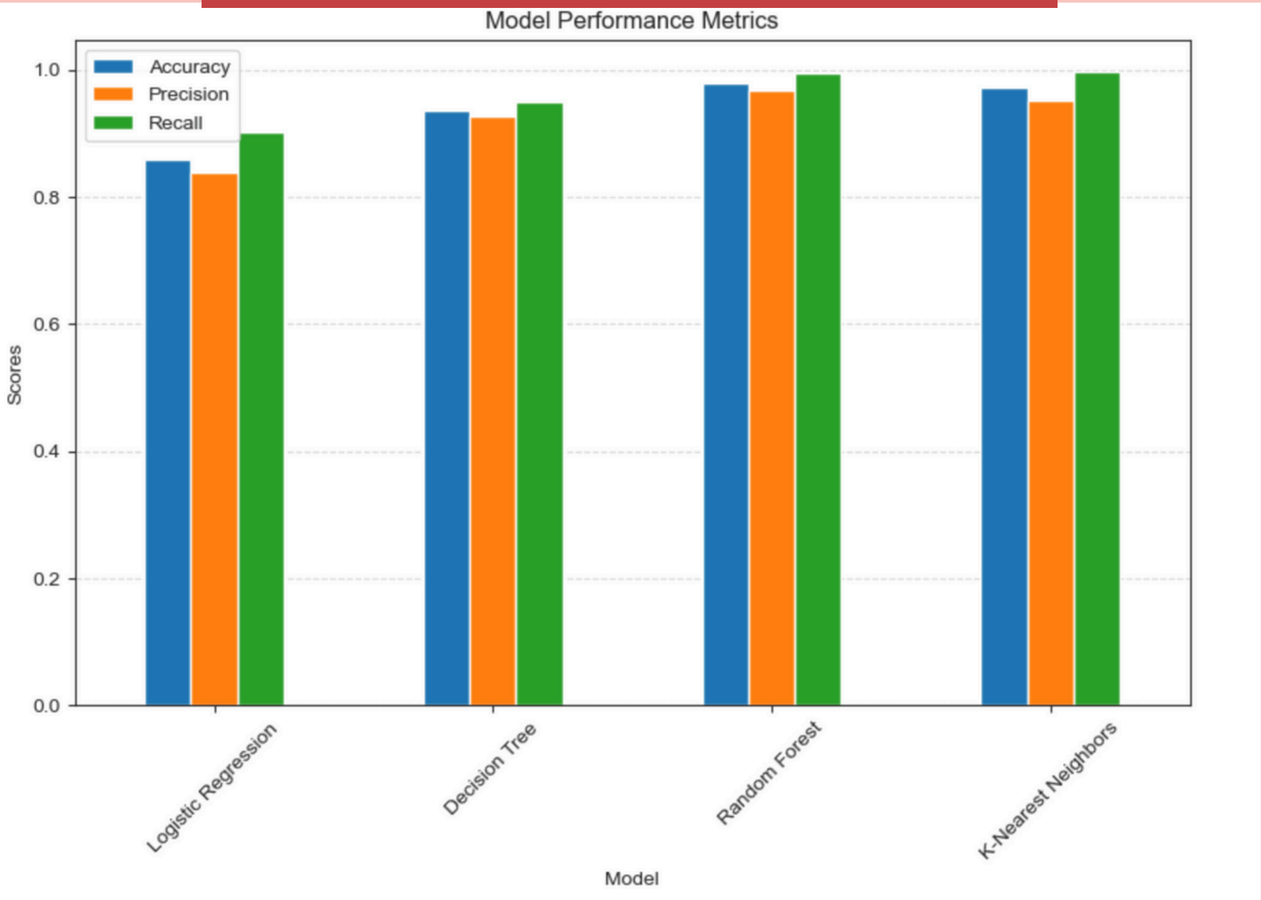
Metrics: accuracy, ConfusionMatrix, precision, recall,f1_score, roc_curve

RESULTS

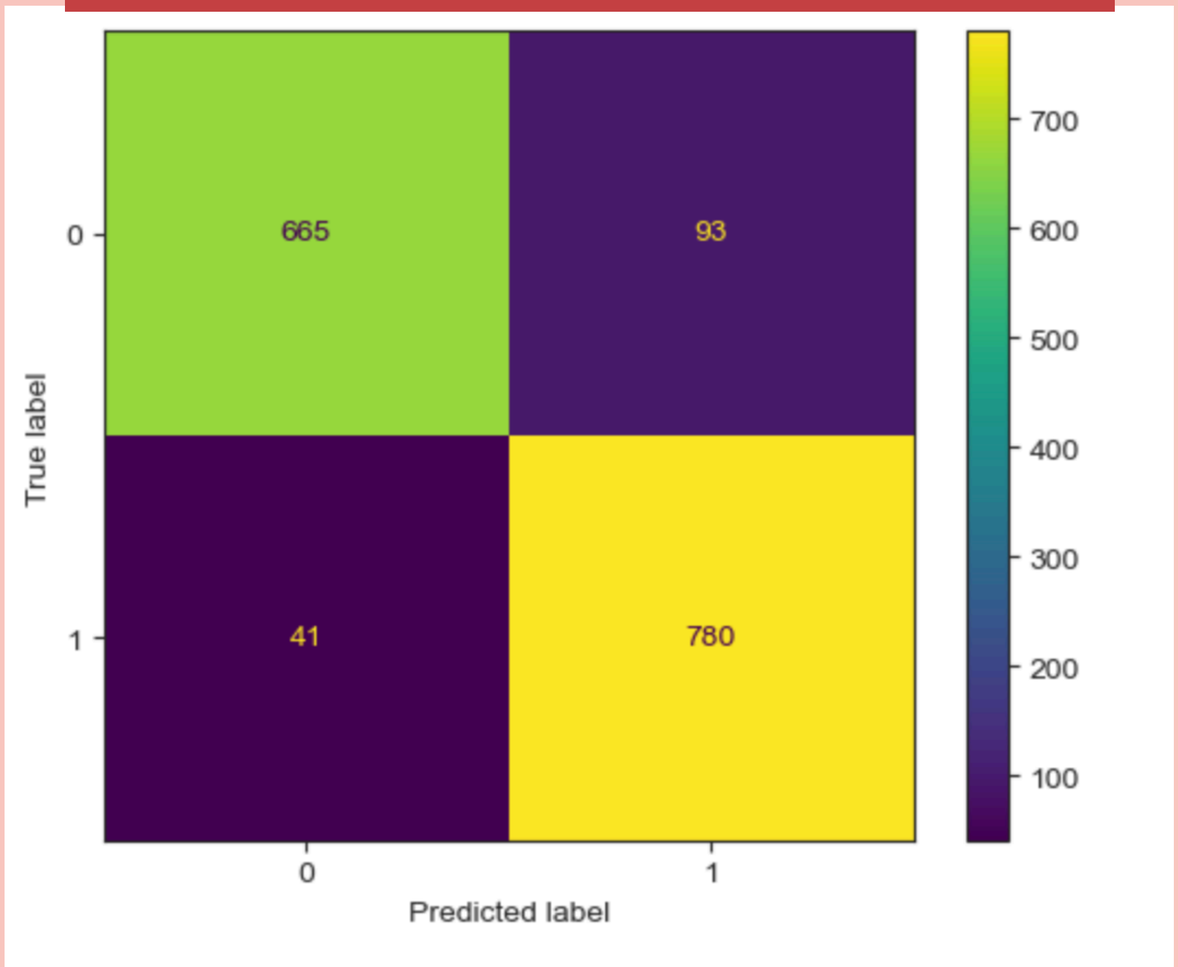
Test Cases

1. Null Data Handling:
• Verified imputation accuracy for missing BMI values.
2. Imbalanced Data:
• Tested SMOTEENN for minority oversampling and noise removal.
3. Model Accuracy:
• Tested models on unseen data to confirm reliability.

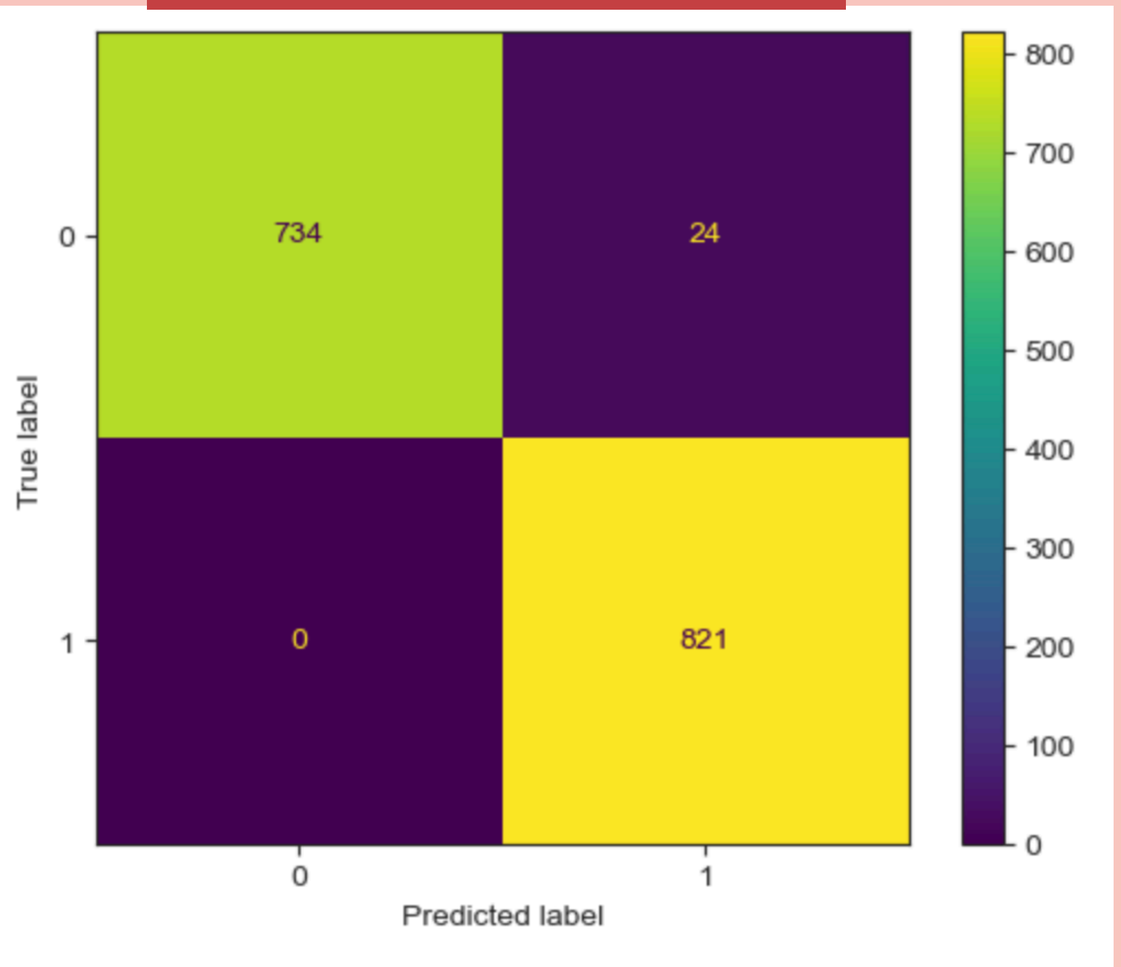
Model Performance Metrics



Confusion Matrix Random Forest



Confusion Matrix(KNN)



CONCLUSION

Justification of Objectives

The project successfully met its objectives by:

1. Identifying the best-performing model (KNN) with high accuracy(98%),precision(97%),recalll(100%)and AUC.
2. Implementing advanced preprocessing techniques to handle missing and imbalanced data.
3. Providing insights through detailed EDA and robust feature engineering.

Future Scope

1. Incorporate larger datasets with more diverse features for improved generalizability.
2. Explore advanced algorithms like deep learning for even better predictions.
3. Develop a user-friendly dashboard to visualize predictions and assist healthcare practitioners.
4. Introduce explainability techniques (e.g., SHAP values) to improve model transparency.