Thapar Institute of Engineering and Technology, Patiala

UML501 - MACHINE LEARNING
24-25 ODD SEM

Heart Stroke Prediction

Project Report

Submitted by
Devit Sah(102217044)
Anshul Mahajan(102217060)
Subgroup - 3CS2

INTRODUCTION

A heart attack is a severe medical condition that occurs when the blood supply to a part of the heart is interrupted, often due to a blood clot. This disruption leads to damage to the heart muscle, as the affected tissue dies due to inadequate blood flow. In many cases, heart attacks can be fatal, making it the leading cause of mortality worldwide, with millions of deaths reported annually.

The risk of heart attacks increases with age and is more common in males. Preventable factors, such as smoking, obesity, unhealthy diets, lack of exercise, high blood pressure, high cholesterol, diabetes, and stress, significantly contribute to heart attacks. Adopting a healthy lifestyle can help manage these risks. Key steps include quitting smoking, maintaining a healthy weight, following a balanced diet low in saturated and trans fats, engaging in regular physical activity, and managing chronic conditions. Stress and mental health also play a critical role in heart disease, making stress management equally important.

Recognizing the warning signs of a heart attack and seeking immediate medical attention can save lives. Common treatments include medications to improve blood flow, such as thrombolytic therapy, or medical procedures like angioplasty and bypass surgery. After treatment, rehabilitation programs and lifestyle modifications are often recommended to strengthen the heart and prevent future episodes.

In conclusion, a heart attack is a serious medical condition that requires awareness of its causes, symptoms, and treatments. By addressing preventable risk factors and ensuring timely medical intervention, individuals can significantly improve their heart health and reduce their risk of heart disease.

LITERATURE REVIEW

Research Paper Title: Heart Stroke Prediction using Machine Learning
Publisher: IEEE
Year: 2023
Authors: Payal Garg, Tarun Jain, Veronica Vashishtha, Virendra Tiwari, Anshul Kumar

This research paper proposes a system for monitoring and predicting the risk of heart stroke based on patients' clinical and demographic data. It explores machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree, KNN, Naive Bayes, and XGBoost, to create predictive models. The data used comprises 5,110 patient records from Kaggle, incorporating attributes such as age, gender, hypertension, blood sugar levels, BMI, smoking habits, and more.
Preprocessing techniques like normalization and feature correlation analysis were employed before training the models. The study evaluated model performance using metrics such as accuracy, precision, recall, and F1-score. Results highlighted the XGBoost classifier achieving the highest F1-score of 96%, making it the most effective for predicting heart stroke risks. This work demonstrates the potential of machine learning in early detection, aiding timely interventions and prevention strategies.

Performance Parameters	Logistic Regression	KNN	Naive Bayes	Random Forest	XGBoost
Accuracy	0.91	0.89	0.82	0.94	0.96
Precision	0.82	0.90	0.81	0.94	0.96
Recall	0.76	0.89	0.81	0.94	0.96
F1-Score	0.80	0.89	0.80	0.94	0.96

Abstract:

This project focuses on predicting heart stroke risk using machine learning techniques. Accurate prediction of stroke risk can assist healthcare providers in proactive diagnosis and personalized treatment planning. The dataset contains patient health details, including demographic, lifestyle, and medical history attributes. Various machine learning models were implemented and evaluated using metrics such as Accuracy, Precision, Recall, F1-score. KNN Classifier emerged as the best-performing model with high accuracy and robust prediction results.

Dataset Overview:

- Source: Publicly available or simulated healthcare datasets.
- Features:
  - Numerical:Age, Glucose Level, BMI, Average Blood Pressure.
  - Categorical: Gender, Smoking Status, Work Type, Residence Type.
- Target Variable:Stroke Risk (Binary: 0 = No Stroke, 1 = Stroke).
- Size:5000+ rows and 10+ columns.

Key Challenges:

1. Handling missing values in critical features like BMI and Glucose Level.
2. Addressing imbalanced classes in the dataset (low occurrence of stroke cases).
3. Ensuring model interpretability to gain actionable insights for medical use.

Methodology

Step 1: Data Loading and Initial Exploration

- The dataset was loaded using pandas from a CSV file named healthcare-dataset-stroke-data.csv
- Shape of the Dataset: The number of rows and columns were printed to understand the size of the data.

Step 2: Data Cleaning

- Handling Missing Values: Imputed missing values using median for numerical data and most frequent values for categorical data.
- Duplicate Removal: Ensured no duplicate rows were present.
- Dropped Irrelevant Features:The columns id, gender, and Residence_type were removed, as they were deemed unnecessary for analysis.

Step 2: Data Preprocessing

- Feature Categorization:
  - ➔ Numeric Features:Identified numerical columns in the dataset using data types.
  - ➔ Categorical Features: Identified columns containing object-type data (strings).
- Feature Classification:
  - ➔ Discrete Features: Features with fewer unique values (≤25) were identified for special handling, like encoding.
  - ➔ Continuous Features: Features with a high number of unique values (>25) were categorized separately for scaling and outlier detection.
- Outlier Detection and Removal:
  - ➔ Box plots were generated to visually identify outliers in continuous features.
  - ➔ A custom function (outlier_removal) was used to filter out data points beyond 3 standard deviations from the mean.

- Variance Inflation Factor (VIF):VIF was computed for continuous features to detect multicollinearity, ensuring feature selection does not introduce redundancy.

Step 4: Model Building

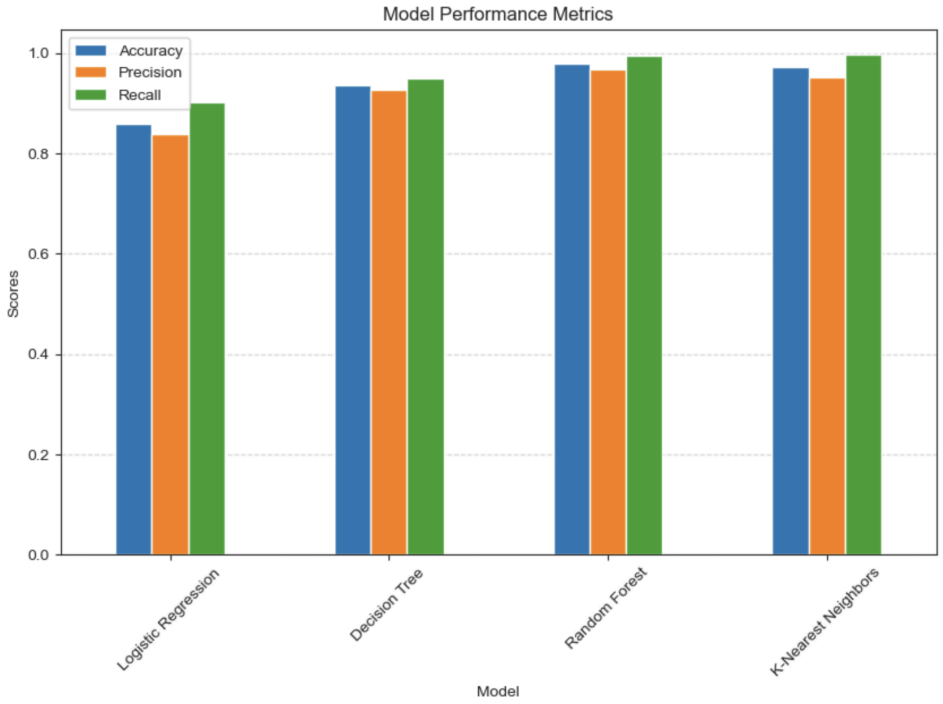
Trained and evaluated the following models:

1. Logistic Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. K-Neighbors Regressor

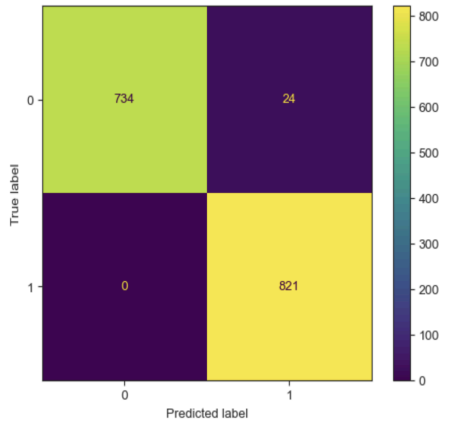
Step 5: Evaluation Metrics

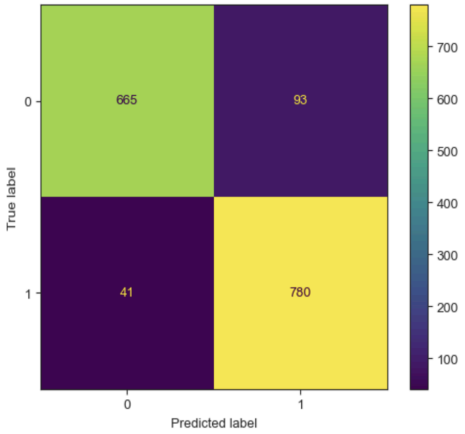
- **Accuracy**: The ratio of correctly predicted instances.
- **Precision**: Evaluates the proportion of true positives among predicted positives.
- **Recall**: Measures the ability to identify actual positive cases.
- **F1-Score**: Combines precision and recall to give a balanced metric.

6. Result



[65]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x141efa <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1





KNN

Random Forest Classifier

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.8588	0.8382	0.9026	0.8692
K-Neighbors Classifier	0.9848	0.9716	1.000	0.9856
Decision Tree Classifier	0.9386	0.9330	0.9501	0.9415
Random Forest Classifier	0.9158	0.8936	0.9513	0.9215

For the given heart stroke prediction , K-Nearest Neighbors (KNN) stands out as the best model due to its superior performance across all metrics, achieving the highest precision (97.16%) and recall (100%). This indicates that KNN is highly effective at accurately identifying both positive and negative cases, making it an excellent choice for minimizing both Type 1 (false positives) and Type 2 (false negatives) errors. Such reliability is particularly critical in a health-related context where prediction accuracy can have significant implications.