Data Mining

Final Project Report

Heart Attack Prediction Using Machine Learning

Introduction:

Heart disease is a significant health concern globally, with heart attacks being a leading cause of mortality. Early prediction and diagnosis of heart attacks are crucial for timely intervention and improving patient outcomes. In this report, we explore the application of machine learning techniques to predict heart attacks based on clinical parameters using the Heart Attack Prediction dataset.

Dataset:

The Heart Attack Prediction dataset consists of various clinical parameters such as age, sex, cholesterol levels, blood pressure, and other medical indicators. It contains 303 samples with 13 features and a binary target variable indicating the likelihood of a heart attack.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) revealed insights into the distribution of features and their relationships with the target variable. We observed that some features, such as age and cholesterol levels, exhibited a correlation with the likelihood of a heart attack. Additionally, there were no missing values in the dataset, simplifying the preprocessing steps.

Data Preprocessing:

Before building machine learning models, we pre-processed the data by encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets. We used techniques such as StandardScaler for scaling and One-Hot Encoding for categorical variables.

Feature Extraction:

We employed various feature extraction techniques to select the most relevant features for heart attack prediction:

- **Filter-based methods:**
  - **SelectKBest:** We selected the top k features based on their correlation with the target variable.
  - **Variance Thresholding:** We removed features with low variance.
  - **Tree-based Feature Selection:** We used feature importance from tree-based models to select features.
- **Wrapper-based methods:**

- **Recursive Feature Elimination (RFE):** We recursively removed features based on their importance with respect to the model performance.
- **Lasso Regression:** We used L1 regularization to shrink coefficients, selecting important features.

- **Dimensionality Reduction:**

- **Principal Component Analysis (PCA):** We reduced the dimensionality of the data while preserving most of its variance.

Model Building:

We built classification models using eight different algorithms:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Gradient Boosting
- Neural Network

For each model, we evaluated performance using 5-fold cross-validation and reported accuracy, precision, recall, sensitivity, and specificity. We also performed grid search for hyperparameter tuning to optimize model performance.

Model Evaluation:

After evaluating the models, we found that Random Forest achieved the highest accuracy of around 85%, followed by Gradient Boosting and Neural Network, with accuracies above 80%. Other models such as Logistic Regression, SVM, and KNN also performed reasonably well, with accuracies ranging from 70% to 80%. Naive Bayes had the lowest accuracy among the models.

Explainable AI:

We employed LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to explain the predictions of our models. These techniques provided

insights into how different features influence model predictions and helped interpret the model's decision-making process.

Conclusion:

Machine learning models can effectively predict heart attacks using clinical parameters from the Heart Attack Prediction dataset. Through feature extraction techniques and model evaluation, we identified key features and algorithms that contribute to accurate predictions. The explainable AI methods further enhance our understanding of model predictions and provide insights for healthcare professionals to interpret and trust the model outcomes. **REFERENCES**

https://www.kaggle.com/datasets/m1relly/heart-attack-prediction