PACE UNIVERSITY

DATA MINING

SPRING 2024 - 20778

FINAL PROJECT

A Comparative Study of Various Classifiers on Heart Disease Dataset

Submitted by

ROHAN NIRANJAN KALPAVRUKSHA
ROSHAN NIRANJAN KALPAVRUKSHA
MEET BHANUSHALI
SRIKANTH REDDY YERUVA

A. Executive Summary

Heart disease is a significant global health concern, necessitating accurate predictive models for early detection. Our investigation encompasses dataset exploration, pre-processing steps, model training, evaluation, and comparative analysis of classifier performance. We address fundamental questions concerning classifier efficiency. Through meticulous data pre-processing, including handling missing values, encoding categorical variables, and standardizing numerical features, we ensure the dataset's suitability. Employing prominent classifiers such as KNN, Logistic Regression, Decision Tree, SVM, and AdaBoost, we conduct a rigorous comparative analysis of their performance metrics, including accuracy, precision, recall, F1-score, and Area under the ROC Curve (AUC). Our findings reveal that AdaBoost consistently outperforms other classifiers, demonstrating the highest AUC score and effectively discriminating between positive and negative heart disease cases making it the optimal choice for heart disease prediction in this study. Future research could explore ensemble methods or deep learning for improved accuracy. Expanding the feature space may also enhance performance and risk stratification.

B. Main Report

Problem Statement:

The aim is to conduct a comparative analysis of different classifiers to identify the most accurate and reliable model for predicting heart disease based on patient data. By exploring a

dataset containing information such as age, sex, chest pain type, cholesterol levels, and exercise-induced angina, among others, we seek to evaluate the performance of classifiers including KNN, Logistic Regression, Decision Tree, SVM, and AdaBoost. Through this comparative study, we aim to provide insights into the effectiveness of different classification algorithms in accurately diagnosing heart disease, ultimately contributing to improved patient care and outcomes in clinical practice.

Dataset Used:

The dataset used is:
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data
The dataset has around 15 features and around 1000 data point entries. We utilized the Heart Disease dataset, containing medical attributes such as age, sex, cholesterol levels, and electrocardiographic measurements, along with the target variable indicating the presence or absence of heart disease.

Preprocess of Data:

Handling Missing Values: Missing values were identified in multiple columns such as 'trestbps', 'chol', 'fbs', 'restecg', 'thalch', 'exang', 'oldpeak', 'slope', 'ca', and 'thal'. These missing values were replaced with appropriate strategies such as mean imputation for numerical columns and mode imputation for categorical columns. This ensured that the dataset was complete and suitable for analysis.

Encoding Categorical Variables: Categorical variables such as 'sex', 'dataset', 'cp', 'fbs', 'restecg', 'exang', 'slope', and 'thal' were encoded using LabelEncoder to convert them into numerical

format. This step was necessary for feeding the categorical data into the classification models.

Feature Scaling: The features were standardized using StandardScaler to bring them to a similar scale. Standardization helps in improving the performance of some machine learning algorithms by ensuring that all features contribute equally to the model fitting process.

Feature Selection: The top correlated features with the target variable 'num' (indicating the presence or absence of heart disease) were selected based on the correlation matrix. This helped in reducing the dimensionality of the dataset and focusing on the most relevant features for classification.

Target Variable Transformation: The target variable 'num', originally representing different levels of heart disease severity, was transformed into a binary classification problem by combining all levels of heart disease into a single category (1 for presence of heart disease and 0 for absence). This simplification facilitated the classification modelling process.

Exploratory Data Analysis (EDA): Descriptive statistics provided valuable insights into the numerical features, such as age, resting blood pressure, cholesterol levels, and maximum heart rate achieved. Furthermore, categorical variables like chest pain type, fasting blood sugar, and exercise-induced angina were analysed to understand their distributions and frequencies. Visualizations, such as count plots and scatter plots, were employed to explore relationships between variables and their potential impact on heart disease prevalence. Notably, gender-based analysis revealed differences in heart disease prevalence between males and females

Organization of Data:

The data is organized into feature vectors representing individuals' medical attributes, along with corresponding binary labels indicating the presence or absence of heart disease. The dataset is split into input features (X) and the target variable (y). The features are selected based on the previously identified relevant columns, and the target variable is defined as the presence or absence of heart disease. The dataset is then divided into training and testing sets using the 'train_test_split' function.

Results across the various different classifiers:

The results across classifiers in our study provide valuable insights into the performance of various machine learning algorithms for predicting heart disease. We evaluated five different classifiers: K-Nearest Neighbours (KNN), Decision Tree, Support Vector Machine (SVM), Logistic Regression, and AdaBoost. Each classifier was trained and tested on the pre-processed dataset, and their performance was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). AdaBoost exhibited the highest AUC-ROC score, indicating its effectiveness in discriminating between positive and negative heart disease cases. This strong performance suggests that AdaBoost effectively leverages weak learners to improve overall predictive accuracy.

While AdaBoost outperformed other classifiers, the results also provided valuable insights into the strengths and weaknesses of

each algorithm. KNN demonstrated competitive performance, particularly in terms of precision and recall, but slightly lower accuracy compared to AdaBoost. Decision Tree and Logistic Regression showed moderate performance, with Decision Tree exhibiting slightly better precision and recall, while Logistic Regression showed slightly better accuracy. SVM, despite its potential for high accuracy, demonstrated comparatively lower performance in this study.

Overall, the results underscore the importance of selecting appropriate classifiers for heart disease prediction tasks. AdaBoost emerges as the optimal choice in this study due to its consistent high performance across multiple evaluation metrics.

Metrics Used To Compare:

These metrics help evaluate the effectiveness of each classifier in correctly classifying instances into their respective categories (presence or absence of heart disease). The key metrics used for comparison include:

Accuracy: The Support Vector Machine and Logistic Regression classifiers have the highest accuracy of 0.8079710144927537

Precision: The Support Vector Machine classifier has the highest precision of 0.8146434038882551

Recall (Sensitivity): The Support Vector Machine and Logistic Regression classifiers have the highest recall of 0.8079710144927537

F1-Score: The Support Vector Machine classifier has the highest precision of 0.8089241570000033

Area under the ROC Curve (AUC): AdaBoost consistently outperformed the other algorithms in terms of Area under the ROC Curve (AUC)

Conclusion:

Our investigation into the heart disease dataset revealed several key findings and insights:

Classifier Performance: We evaluated the performance of various classifiers including KNN, Logistic Regression, Decision Tree, SVM, and AdaBoost. Among these, AdaBoost consistently outperformed others, demonstrating the highest Area under the ROC Curve (AUC) score. This indicates its superior ability to discriminate between positive and negative heart disease cases.

Data Pre-processing: Rigorous data pre-processing steps were undertaken, including handling missing values, encoding categorical variables, standardizing numerical features, and handling outliers. These steps ensured the dataset's cleanliness, proper formatting, and readiness for model training.

Feature Selection: Features with the highest correlation with the target variable 'num' were selected to train the classifiers. This helped in improving the model's predictive performance by focusing on the most relevant features.

Binary Classification: The target variable 'num', representing the presence or absence of heart disease, was transformed into binary classes (0 and 1) to facilitate binary classification. This simplified the classification task and improved model interpretability.

Future Research:

Further research in the domain of heart disease prediction can explore advanced methodologies to enhance predictive accuracy and robustness. Ensemble learning techniques such as Random Forest, Gradient Boosting, or stacking models could be investigated to leverage the collective intelligence of multiple classifiers and improve predictive performance. Additionally, deep learning models, particularly convolutional neural networks (CNNs) or recurrent neural networks (RNNs), may be explored to capture complex nonlinear relationships in the data. These models have shown promise in various healthcare applications and can potentially uncover hidden patterns and risk factors associated with heart disease. Furthermore, research efforts could focus on feature engineering and selection, including the integration of domain-specific knowledge and biomarkers, to improve model interpretability and generalizability across diverse patient populations. Collaboration with medical experts and healthcare institutions is crucial to ensure the clinical relevance and applicability of developed models. Overall, further research in this area holds the potential to advance early detection and intervention strategies for heart disease, ultimately leading to improved patient outcomes and healthcare delivery.