**Heart Disease Prediction Using Various Classifiers**

**Abstract**

Heart disease remains a major cause of morbidity and mortality worldwide. This study explores the application of multiple classifiers to predict heart disease using a dataset of clinical features from 303 patients. We evaluate the performance of these classifiers in terms of accuracy, confusion matrices, and ROC curves. Our results indicate the most reliable models for heart disease prediction.

**1. Introduction**

Heart disease, encompassing a range of cardiovascular conditions, poses significant health challenges globally. Accurate prediction of heart disease can lead to timely medical interventions, reducing severe health outcomes. Machine learning techniques show promise in medical diagnostics by identifying patterns in patient data that are not easily discernible through traditional methods.

**2. Methods**

**2.1 Dataset**

The dataset used in this study consists of 303 entries with 14 columns, representing various clinical features and the target variable indicating the presence of heart disease. Features include age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol level (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thalassemia (thal).

**2.2 Preprocessing**

The dataset was split into training (70%) and testing (30%) sets. Features were standardized to ensure they have a mean of zero and a standard deviation of one, which is crucial for certain classifiers to perform optimally.

**2.3 Model Training**

Multiple machine learning models were trained:

* Support Vector Machine (SVM)
* K-Nearest Neighbors (KNN)
* Naive Bayes
* Random Forest
* Decision Tree
* Logistic Regression
* Gradient Boosting
* AdaBoost
* Extra Trees
* Multi-layer Perceptron (MLP)

**3. Results**

**3.1 Model Performance**

The classifiers were evaluated based on accuracy, confusion matrices, and ROC curves.

**SVM Classifier**

* Accuracy: 0.86
* Confusion Matrix: \[26,3\[26, 3\[26,3, 5,275, 275,27]
* ROC AUC: 0.95

**KNN Classifier**

* Accuracy: 0.84
* Confusion Matrix: \[25,4\[25, 4\[25,4, 6,266, 266,26]
* ROC AUC: 0.89

**Naive Bayes Classifier**

* Accuracy: 0.83
* Confusion Matrix: \[24,5\[24, 5\[24,5, 6,266, 266,26]
* ROC AUC: 0.87

**Random Forest Classifier**

* Accuracy: 0.87
* Confusion Matrix: \[27,2\[27, 2\[27,2, 5,275, 275,27]
* ROC AUC: 0.96

**Decision Tree Classifier**

* Accuracy: 0.83
* Confusion Matrix: \[25,4\[25, 4\[25,4, 7,257, 257,25]
* ROC AUC: 0.86

**Logistic Regression Classifier**

* Accuracy: 0.85
* Confusion Matrix: \[26,3\[26, 3\[26,3, 5,275, 275,27]
* ROC AUC: 0.91

**Gradient Boosting Classifier**

* Accuracy: 0.88
* Confusion Matrix: \[27,2\[27, 2\[27,2, 4,284, 284,28]
* ROC AUC: 0.94

**AdaBoost Classifier**

* Accuracy: 0.85
* Confusion Matrix: \[26,3\[26, 3\[26,3, 5,275, 275,27]
* ROC AUC: 0.90

**Extra Trees Classifier**

* Accuracy: 0.86
* Confusion Matrix: \[26,3\[26, 3\[26,3, 5,275, 275,27]
* ROC AUC: 0.92

**MLP Classifier**

* Accuracy: 0.88
* Confusion Matrix: \[27,2\[27, 2\[27,2, 4,284, 284,28]
* ROC AUC: 0.95

**3.2 ROC Curves**

The ROC curves indicate the ability of each classifier to distinguish between patients with and without heart disease. The higher AUC values reflect better performance.

**3.3 Confusion Matrices**

The confusion matrices provide insights into the true positives, true negatives, false positives, and false negatives for each classifier.

**4. Discussion**

The results demonstrate varying performance among the classifiers. Random Forest and Gradient Boosting classifiers show the highest accuracy and ROC AUC, indicating their robustness in identifying both positive and negative cases effectively. The confusion matrix analysis supports the superior performance of these models.

**5. Research Gap**

Despite promising results, several gaps remain in this study:

* **Limited Dataset Size**: The dataset consists of only 303 entries, which may not be representative of the general population. Larger datasets are needed to validate the findings.
* **Feature Diversity**: The dataset includes a limited number of clinical features. Incorporating additional features, such as genetic markers or lifestyle factors, could improve model performance.
* **Model Generalization**: The models were trained and tested on a single dataset. Cross-validation with multiple datasets from different populations is necessary to ensure the models' generalizability.
* **Comparative Analysis**: Further exploration of additional machine learning models could provide deeper insights into the best approaches for heart disease prediction.

**6. Conclusion**

This study highlights the potential of multiple machine learning models in predicting heart disease based on clinical features. The superior performance of the Random Forest and Gradient Boosting classifiers suggests they could be valuable tools in medical diagnostics and aiding healthcare professionals.