Predicting Heart Disease in Patients Using Machine Learning

First Large Project
Course: SAT5114
Proff: Dr. Weihua Zhou
Submitted By: Usama Ayub
Group Members: Usama Ayub & Navya Sadineni

1. Introduction:

Heart disease is a leading cause of death worldwide. According to the article Cardiovascular Diseases, in 2019, around 18 million people died from cardiovascular diseases globally. However, early detection and diagnosis can play vital role in treatment and management. Machine learning models are proved very effective in predicting heart disease based on clinical and physiological features. This project explore various statistical machine learning models to predict heart disease in patients.

2. Data Preprocessing:

The dataset heart.csv is loaded into a DataFrame. The preprocessing steps include:

- Removing duplicate rows to ensure data integrity. After removal, no duplicate rows were found.
- Discretizing the age column into categorical bins for better analysis. The age is categorized into six bins: 0-29, 30-40, 41-50, 51-60, 61-70, and 71-100.
- Dummy variables for categorical features to facilitate model training.
- Checking and handling missing values and outliers. No missing values were found, and outliers were addressed during data scaling.

3. Data Visualization:

Several visualizations are created to understand the data distribution:

- Count plots to visualize the distribution of patient genders with and without heart disease.
- Histograms to observe the distribution of numerical features like age, blood pressure, cholesterol levels, and heart rate. The distributions were approximately normal for most features.
- A heatmap to examine the correlation between features, indicating no significant multicollinearity among the features.
- Boxplots to identify outliers in numerical features. Some outliers were observed in features like 'trestbps', 'chol', and 'thalach'.

4. Feature Selection and Scaling:

Feature selection is performed with Feature Elimination (RFE) technique from wrapper method, with a Random Forest Classifier as the estimator. The selected features include 'age', 'trestbps', 'chol', 'thalach', 'oldpeak', 'cp_1', 'cp_2', 'cp_3', 'thal_2', and 'thal_3'.
The numerical features are scaled using StandardScaler to standardize the data.

5. Model Selection and Evaluation:

Three models are evaluated:

- Support Vector Machine (SVM) with a linear kernel.
- Random Forest Classifier with 10 estimators.
- AdaBoost Classifier with the SAMME algorithm.

The models are trained on the training dataset and evaluated on both the training and test datasets using accuracy and precision scores with all features and selected features respectively.

Model Performance with all Features:

Model	Performance with Train dataset		Performance with Test dataset	
	Accuracy	Precision	Accuracy	Precision
SVM	0.88796	0.86619	0.85245	0.875
RF Classifier	0.99585	0.99242	0.803278	0.8620689
AdaBoost Classifier	0.89626	0.89552	0.83606	0.89655

Model Performance with Selected Features:

Model	Performance with Train dataset		Performance with Test dataset	
	Accuracy	Precision	Accuracy	Precision
SVM	0.825726	0.8449612	0.8196721	0.84375
RF Classifier	0.99170	0.9923664	0.8196721	0.892857
AdaBoost	0.871369	0.8731343	0.7868852	0.83333

The SVM model performed well on both train and test data as we can see in tables above.

K-Fold Cross Validation:

Model	Cross Validation
	Mean Score
SVM	0.8342622950819673
RF Classifier	0.7649726775956284
AdaBoost	0.8145355191256831

SVM model has the highest mean score, indicating better performance compared to the other models in the given cross-validation setup. Therefore, SVM is the preferred choice among the three models based on the scores as shown in tables.

6. Final Model Performance:

The best hyperparameters for the SVM model are determined using RandomizedSearchCV with K-Fold CV. The tuned SVM model is evaluated using various metrics:

- Accuracy: The SVM model achieved an accuracy of approximately 85.2% on the test dataset.
- Confusion Matrix: The model correctly predicted 24 true negatives and 27 true positives, with 5 false negatives and 4 false positives.
- Classification Report: The precision, recall, and F1-score for both classes were satisfactory, with the model performing slightly better in predicting the presence of heart disease.
- Sensitivity and Specificity: The model showed a sensitivity of 85.7% and a specificity of 84.2%.
- ROC Score and ROC Curve: The area under the curve (AUC) was 85.2%, indicating good model performance.

7. Conclusion:

The project demonstrates the application of statistical machine learning models to predict heart disease in patients. The SVM model with tuned hyperparameters shows promising results in terms of accuracy and other evaluation metrics.

"Screen shots are uploaded on canvas."

GitHub Link