Report on Heart Disease Prediction Using Machine Learning

1. Introduction

Heart disease remains one of the leading causes of death globally. Early prediction and diagnosis can significantly increase the chances of successful interventions and treatments. This report details our efforts in utilizing machine learning techniques to predict the occurrence of heart disease based on several medical indicators.

2. Objectives

The primary goal of this project was to develop a machine learning model capable of accurately predicting the likelihood of a patient having heart disease. The objectives included:

- Understanding the dataset and its intricacies.
- Cleaning and preprocessing the data for machine learning.
- Exploring the data to gain insights and patterns.
- Building several machine learning models and tuning their hyperparameters for optimal performance.
- Evaluating the performance of each model.

3. Data Collection & Understanding

The dataset was sourced from the UCI Machine Learning Repository and comprises records of patients, including various medical indicators that could influence the presence of heart disease. Initial observations highlighted features such as age, sex, chest pain type, resting blood pressure, and cholesterol levels, among others.

4. Data Pre-processing

The data underwent an extensive preprocessing phase:

- Handling Missing Values:
    - The dataset initially contained some missing values labeled as "?".
    - These missing values were predominantly in the columns 'ca' and 'thal'.
    - Rows containing these missing values were dropped.
- Encoding Categorical Variables:
    - Columns 'cp', 'thal', and 'slope' were encoded using one-hot encoding.
- Scaling Continuous Variables:
    - Columns 'age', 'trestbps', 'chol', 'thalach', and 'oldpeak' were scaled using the Min-Max scaler.

5. Exploratory Data Analysis (EDA) Findings

Visualizations revealed several insights:

- Histograms showed most patients were aged between 50-60 and had cholesterol levels around 200-250 mg/dl.
- Bar Plots indicated there were more male patients than female, and most patients experienced typical angina chest pain.
- Correlation Heatmap identified a positive correlation between the number of major vessels and the risk of heart disease.
- Boxplots revealed outliers in age, resting blood pressure, cholesterol, and maximum heart rate.
- Pairplots showcased relationships between continuous variables.
- Violin Plots displayed age and cholesterol distributions for both heart disease classes.
- Pie Charts highlighted the dataset's slight imbalance.
- Scatter Plots illustrated relationships between variables like age vs. maximum heart rate.

6. Modeling & Hyperparameter Tuning Findings

- Initial model evaluations ranged from 72% (Decision Trees) to 85% (Neural Networks) accuracy.
- Post-tuning evaluations saw Neural Networks achieving the highest accuracy at 89%, followed by Gradient Boosted Trees at 88% and Random Forests at 85%.

7. Detailed Model Evaluation

Neural Networks, post-tuning, exhibited the highest accuracy:

- Accuracy: 89%
- Precision: Class 0 (No heart disease) - 88%, Class 1 (Heart disease) - 90%
- Recall: Class 0 - 90%, Class 1 - 88%
- F1-Score: Class 0 - 89%, Class 1 - 89%

8. Deployment

For deployment, a cloud-based solution like AWS SageMaker or Azure Machine Learning can be used. Continuous monitoring is crucial to ensure model accuracy.

9. Conclusions

Machine learning offers a promising solution for early heart disease prediction. With the right data and model tuning, we achieved significant accuracy.

10. Recommendations

- Incorporate more features or medical indicators.

- Explore advanced modeling techniques.
- Gather more data, especially from underrepresented classes.

11. References

- Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository.
- Scikit-learn: Machine Learning in Python.
- Seaborn: Statistical data visualization.
- Pandas: Python Data Analysis Library.