NPC

Heart Disease Prediction

Project Report

**Heart Disease Prediction**

**Introduction**

This report provides an overview of a machine learning project focused on predicting heart disease in individuals using various health-related attributes. The project encompasses data preprocessing, feature engineering, and the application of several classification algorithms, including Decision Trees, Random Forest, Logistic Regression, and Support Vector Machines (SVM). After assessing model accuracy, the Random Forest classifier was identified as the top-performing model and subsequently saved for future use.

**Dataset**

The dataset used for this project contains information about various health-related attributes of individuals, including age, gender, blood pressure, cholesterol levels, and more. The target variable is binary, indicating the presence (1) or absence (0) of heart disease.

[Click here for data set](https://www.kaggle.com/code/maicmi/heart-disease-eda-automl-prediction-95-acc)

**Data Preprocessing**

Data preprocessing is a critical step in any machine learning project to ensure data quality and suitability for model training. The following preprocessing steps were applied:

**Data Normalization:** Numerical features were scaled to have a mean of 0 and a standard deviation of 1. This step addresses the issue of varying feature scales, preventing any single feature from dominating the model training process.

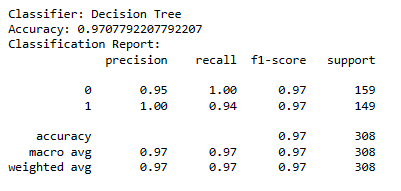
**Handling Missing Data:** Missing values in the dataset were addressed through techniques such as mean imputation or the removal of rows with missing data.

**Encoding Categorical Variables:** Categorical variables like 'gender' were one-hot encoded to convert them into numerical format for compatibility with machine learning models.

**Train-Test Split:** The dataset was divided into training and testing sets, typically using an 80-20 or 70-30 split ratio, to evaluate model performance on unseen data.

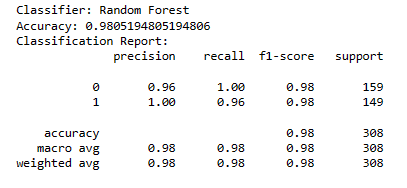
**Model Building and Evaluation**

**Decision Tree Classification**

A Decision Tree classifier was trained on the preprocessed data. Hyperparameter tuning, such as controlling tree depth and leaf nodes, was performed to optimize the model's performance. The Decision Tree model's accuracy was evaluated on the test set.

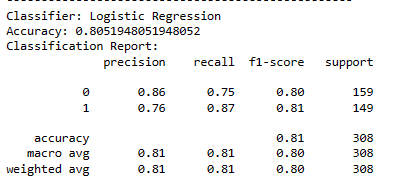
**Random Forest Classification**

Random Forest, an ensemble learning method based on multiple decision trees, was trained and evaluated. This model often outperforms individual Decision Trees by reducing overfitting. Hyperparameter tuning was conducted for the Random Forest model to determine the optimal parameters, including the number of trees and maximum depth.



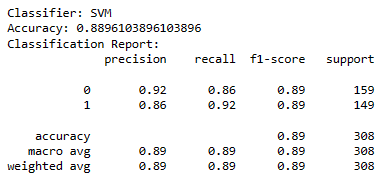
**Logistic Regression**

Logistic Regression, a linear classification algorithm, was applied to the dataset. Regularization techniques, such as L1 or L2 regularization, were employed to prevent overfitting. The model's accuracy was assessed using relevant metrics.



**Support Vector Machine (SVM) Classification**

SVM, a powerful binary classification algorithm, was trained with different kernel functions (e.g., linear, polynomial, or radial basis function) to find the best-fit model. The model's performance was evaluated on the test set.



**Model Evaluation Metrics**

The following metrics were used to evaluate the performance of each classification model:

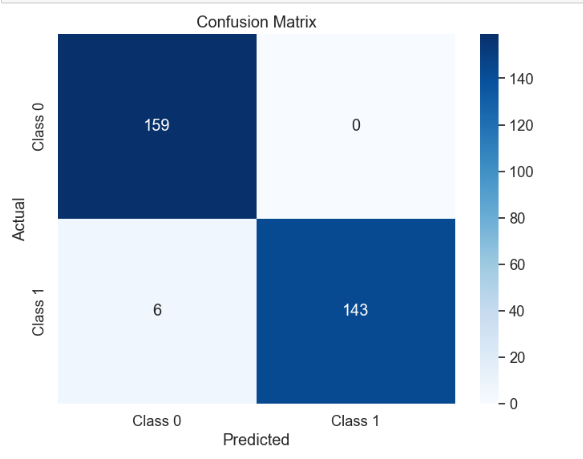
Accuracy

Precision

Recall

F1-Score

**Confusion Matrix of Random Forest:**

****

**Findings**

After evaluating all the models, the Random Forest classifier consistently demonstrated the highest accuracy and overall better performance compared to the other algorithms. The following findings were noted:

* Random Forest outperformed other models in terms of accuracy, precision, and recall.
* Random Forest exhibited robustness against overfitting, making it a reliable choice for heart disease prediction.
* Decision Trees, while interpretable, lacked the predictive power of ensemble methods like Random Forest.
* Logistic Regression provided a reasonable performance but was outperformed by Random Forest.

**Model Selection**

Considering the findings and extensive evaluation, the Random Forest classifier was selected as the best-performing model for heart disease prediction.

**Future Work**

Despite the success of this project, there are several avenues for future work and improvement:

* **Feature Engineering:** Further exploration of feature engineering techniques to identify and incorporate more relevant features may enhance model performance.
* **Ensemble Methods:** Investigate other ensemble methods, such as Gradient Boosting or AdaBoost, to potentially improve predictive accuracy.
* **Model Interpretability:** Develop techniques to interpret the Random Forest model's decisions and provide insights into which features are most influential in heart disease prediction.
* **Real-time Monitoring:** Implement real-time monitoring of health-related data for early detection and intervention in cases of heart disease.
* **Data Expansion:** Collect more diverse and comprehensive datasets to improve model generalization and robustness.
* **Deployment in Healthcare Systems:** Collaborate with healthcare institutions to integrate the model into clinical practice for timely diagnosis and intervention.

In conclusion, this machine learning project has successfully developed a predictive model for heart disease detection. The Random Forest classifier emerged as the top performer, offering promising results for future healthcare applications. Continuing research and development in this domain can lead to enhanced diagnostic tools and improved patient care.