REPORT

ON

HEART DISEASE PREDICTION PROJECT

(Predict the heart disease risk of patients)

BY

SAMUEL OSEI

AND

JOANA LAWER

1. Introduction

This report presents a predictive modeling project aimed at assessing the risk of coronary heart disease (CHD) using the Framingham dataset. The project involves data preprocessing, model building, hyperparameter tuning, and evaluation to develop a robust machine learning model for CHD prediction.

2. Dataset Description

The Framingham dataset is a well-known dataset in the field of cardiovascular research. It contains various health-related features collected from the Framingham Heart Study, a long-term, ongoing cardiovascular cohort study on residents of the town of Framingham, Massachusetts. The dataset includes demographic information, lifestyle factors, and clinical measurements for each participant.

Key features in the dataset include:

- Sex: Male or Female

- Age: Age of the participant

- Education: Level of education

- CurrentSmoker: Smoking status

- CigsPerDay: Number of cigarettes smoked per day

- BPMeds: Blood pressure medication usage

- PrevalentStroke: History of stroke

- PrevalentHyp: History of hypertension

- Diabetes: Diabetes status

- TotChol: Total cholesterol level

- SysBP: Systolic blood pressure

- DiaBP: Diastolic blood pressure

- BMI: Body mass index

- HeartRate: Heart rate

- Glucose: Glucose level

- TenYearCHD: Target variable indicating the risk of CHD within ten years

3. Machine Learning Algorithm

The machine learning algorithm used for this project is the RandomForestClassifier, an ensemble learning method based on decision trees. Random forests combine multiple decision trees to improve the model's accuracy and control overfitting. Key characteristics of RandomForestClassifier include:

- Bootstrap Sampling: Each tree is trained on a random subset of the data.

- Feature Randomness: During the construction of each tree, a random subset of features is considered for splitting nodes.

- Voting Mechanism: The final prediction is made based on the majority vote from all individual trees.

4. Performance Metrics

To evaluate the performance of the model, several metrics were used:

- Accuracy: The proportion of correct predictions out of the total number of predictions. It is a simple and intuitive measure but can be misleading for imbalanced datasets.

- Confusion Matrix: A table that outlines the performance of the model by showing the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. It provides a comprehensive view of the model's performance.

- Classification Report: This report includes precision, recall, and F1-score for each class.

- Precision: The proportion of true positive predictions among all positive predictions. It indicates the accuracy of the positive predictions.

- Recall: The proportion of true positive predictions among all actual positive instances. It indicates the ability of the model to identify positive instances.

- F1-score: The harmonic mean of precision and recall, providing a balanced measure that accounts for both metrics.

5. Results and Interpretation

The model development process included initial training, evaluation, and hyperparameter tuning using GridSearchCV. Here are the key results:

Initial Model Performance

- Accuracy: Initial accuracy was calculated, providing a baseline measure of performance.

- Confusion Matrix: Illustrated the distribution of correct and incorrect predictions.

- Classification Report: Showed precision, recall, and F1-score for both the positive and negative classes.

Hyperparameter Tuning

- GridSearchCV: A grid search over hyperparameters was conducted to find the best combination for the RandomForestClassifier. Key parameters tuned included the number of estimators, maximum depth, minimum samples split, and minimum samples leaf.

- Best Parameters: The best parameters obtained from the grid search were used to train the final model.

Final Model Performance

- Accuracy: Improved after hyperparameter tuning.

- Confusion Matrix: Showed a better distribution of predictions, indicating improved performance.

- Classification Report: Demonstrated enhanced precision, recall, and F1-score, reflecting a more balanced and accurate model.

6. Conclusion

This project successfully developed a predictive model for assessing the risk of coronary heart disease using the Framingham dataset. The RandomForestClassifier, along with appropriate preprocessing and hyperparameter tuning, provided a robust and accurate model. Key insights include:

- The importance of preprocessing steps such as imputation and scaling for numerical data and one-hot encoding for categorical data.

- The effectiveness of RandomForestClassifier in handling complex datasets with multiple features.

- The value of hyperparameter tuning in improving model performance.

Overall, this project demonstrates the feasibility and effectiveness of machine learning techniques in predicting health outcomes and can serve as a foundation for further research and development in the field of cardiovascular risk assessment.