Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037/HJ1: Concepts and Technologies of AI (Herald College, Kathmandu, Nepal)	⋆ Final portfolio Project	An End- to- End Machine Learning Project on Regression and Classification Task

Report of Heart Failure Prediction Dataset_Classification

Student Name: Mohammad Rashid Siddiqui

Student ID: 2413748

Module Leader: Mr. Siman Giri

Tutor: Ms. Durga Pokharel

Abstract

The project is geared toward predicting heart disease through classification techniques. It consists of a dataset admitting medical attributes such as cholesterol levels, blood pressure, age, and lifestyle factors, which are important indicators for the risk of heart disease. It involved data preprocessing, exploratory data analysis (EDA), model building using Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN), hyperparameter tuning, and feature selection. The models are evaluated based on accuracy, precision, recall, and F1 score so as to be able to show which of the models was a better performer. The results showed that KNN was the best-performing after hyperparameter tuning, making for the most reliable model for heart disease prediction. The work underscores how machine learning can augment medical decision-making in early disease detection.

1. Introduction

1.1 Problem Statement

One of the leading causes of death around the world, heart diseases are affecting millions of people every year. Early detection of heart disease allows for the prevention of severe complications and saves lives. However, traditional diagnosis methods are time-consuming and rather inaccurate. With the advent of machine learning, predictive models are on their way to assist doctors more quickly and precisely in making diagnoses. This proposal intends to classify the patient as affected or not, based on medical parameters like cholesterol, blood pressure, or age. This is expected to bring significant value for decision support in the process of developing highly accurate, dependable prediction models by employing the techniques of machine learning.

1.2 Dataset

The Heart Disease Prediction dataset was used for this study, containing the medical records of patients, along with their age, levels of cholesterol, blood pressure, and exercise-induced angina features. The dataset offers direct insights into the contributing factors to heart disease. The study operates under the UN Sustainable Development Goals (UNSDG) for Good Health and Well-Being, for the early detection and prevention of heart disease.

1.3 Objective

The primary objective of this project is to build a machine learning model that will predict whether a patient has heart disease. Through the analysis of various models, our aim is to grasp the most effective approach for classification and also gain a better understanding of how to improve prediction accuracy through hyperparameter tuning and feature selection.

2. Methodology

2.1 Data Preprocessing

In order to have the dataset clean and ready for model training, there were several preprocessing steps applied. This helps prevent unfulfilling predictions due to inaccurate treatment of missing values. Encoding techniques were used to convert categorical variables, such as sex and chest pain type, to numerical values. Also, all numerical features were standardized with StandardScaler in order to make them comparable, which generally enhances the performance of models.

2.2 Exploratory Data Analysis (EDA)

In addition to grasping the dataset and identifying major distributions, EDA was carried out. In this regard, variable distributions were visualized with histograms and pound plots, while relationships between features were revealed with a correlation heatmap. Further investigations revealed that high cholesterol and high blood pressure were strongly associated with heart disease, as such, these characteristics were good predictive factors.

2.3 Model Building

Three classification models were selected for this analysis:

• Logistic regression: A simple interpretable model that estimates the probability of heart disease.

• Decision tree: A tree-based model that splits data into different decision paths based on medical attributes.

• K-nearest neighbors: A model that classifies a patient depending on the similarity to another patient.

The dataset was split into 80% training and 20% testing in order for the model to learn on historical data before making predictions on new data.

2.4 Model Evaluation

The models were evaluated using four key metrics aimed at their performance:

· Accuracy: Measures how many predictions were correct out of the total cases.

· Precision: Determines how many predicted positive cases were really true.

· Recall: Measures how well the positive cases were identified. · F1-Score: A balance metric between precision and recall. Thus, KNN proved to be the most accurate and therefore the best model in detecting heart disease.

2.5 Hyperparameter Optimization

In order to further optimize KNN performance, hyperparameter tuning was done using GridSearchCV. The optimization process determined the best number of neighbors (k), distance metric, and weight function for classification, with results indicating superior KNN performance at k = 7 in terms of accuracy.

2.6 Feature Selection

To select a subset of the most important medical features for predicting heart disease, Recursive Feature Elimination (RFE) was performed. Cholesterol levels, resting blood pressure, and age were proven to be the three most powerful clinical features contributing to heart disease classification from the results.

3. Conclusion

3.1 Key Findings

The efficiency of hyperparameter tuning enabled KNN to achieve maximum performance with the highest accuracy and recall. The findings confirmed that high cholesterol and high blood pressure are strong indicators for assessing the risk of heart disease. The findings of this study highlight the importance of examining medical risk factors for early diagnosis and prevention of heart disease.

3.2 Final Model

The optimized KNN model was selected as the final model due to its best combination of accuracy, recall, and precision. It outperformed significantly Logistic Regression and Decision Tree models to classify the heart disease case by proper tuning.

3.3 Challenges

Some challenges were encountered throughout the course of the project. Missing values had to be handled through a cumbersome data-preprocessing process to avoid bias. Model performance was affected due to imbalanced data distribution (more characters with no disease than those with the disease). Also, feature selection and hyperparameter tuning were made for good performance.

3.4 Future Work

The work would involve training models including but not limited to Neural Networks, Random Forest classifiers, and other advanced machine learning approaches. Generalization and accuracy of the proposed methods could be advanced further in the future by increasing the dataset size through including patient data.

4. Discussion

4.1 Model Performance

Out of all models tested, KNN had performed the best, hence classifying the patients well on the bases of similarity in medical attributes. Whereas logistic regression performed poorly since, in general, it is based on a linear relationship between features, which might not fit perfectly on complex medical data.

4.2 Impact of Hyperparameter Tuning
The tuning process greatly improved the accuracy of the KNN model, indicating the importance of appropriate parameters in achieving optimal performance. Thus, KNN performed poorly without hyperparameter tuning, which reiterates the significance of choosing the right model settings.

4.3 Interpretation of Results
The findings confirmed that cholesterol levels and blood pressure are the most significant markers of heart diseases. This corroborates with the already known medical aspects. This research elucidates the effective utilization of machine learning for diagnosis of high-risk patients, thus aiding early detection of the disease by doctors.

4.4 Limitations
Although performing well, there were certain limitations. The dataset was not completely balanced and therefore could have influenced the prediction of minority classes. Missing values in patients' records may also have been responsible for variations in model accuracy. Future improvements might include gathering a complete, diverse patient database.

4.5 Future Research
In the future, further research could investigate deep learning methods to better the accuracy of classification. Besides, the model's predictability greatly solves genetic and lifestyle factors, such as family history and diet. This would thus provide a more sharpened heart disease prediction system.

Final Thoughts
The study showed that machine learning models could effectively predict heart diseases to help early diagnosis and medical decision-making. Among the various models tested, KNN with optimized hyperparameters gave the best performance and was thus selected as the most appropriate. These results underline the significant role that cholesterol and blood pressure have in heart disease risk, and they further demonstrate how machine learning can assist healthcare providers in identifying the patients at risk. Machine learning will possibly come to become the standard of preventive care in the fortune.