Enhanced Heart Attack Prediction Using KNN, Naive Bayes, and Decision Tree Models

Shan Khanna
Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India
Corresponding author:- shank.mca@chanakyauniversity.edu.in

Karthi K.
Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India
Corresponding author:- karthik.mca24@chanakyauniversity.edu.in

Abstract

In this paper, we consider the use of three basic machine learning algorithms—K-Nearest Neighbors (KNN), Naive Bayes, and Decision Tree—to the prediction of cardiovascular disease. From freely available data, we performed data preprocessing, exploratory data analysis (EDA), model training, and model evaluation. This paper extends previous work on only the KNN algorithm by considering more than one algorithm for comparison purposes. Heatmaps and confusion matrices are an integral part of our evaluation and thus have an interpretability aspect to the results.

1. Introduction

Cardiovascular disease (CVD) is the number-one cause of death worldwide. Predictive analytics with machine learning (ML) has been successful in identifying risk candidates. This paper therefore adds to the body of current KNN research by experimenting, testing, and comparing two other alternative models, namely Naive Bayes and Decision Tree.

2. Dataset and Preprocessing

These data contain over 70,000 patient records with features such as age, gender, height, weight, blood pressure, cholesterol, glucose, and lifestyle. The most significant preprocessing tasks involved:

- Age in days to age in years conversion

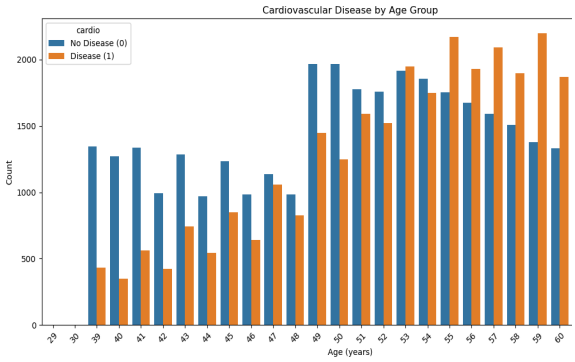
- Weight and height-based BMI calculation

- Identification and exclusion of outliers using IQR

- Normalization of continuous set of variables

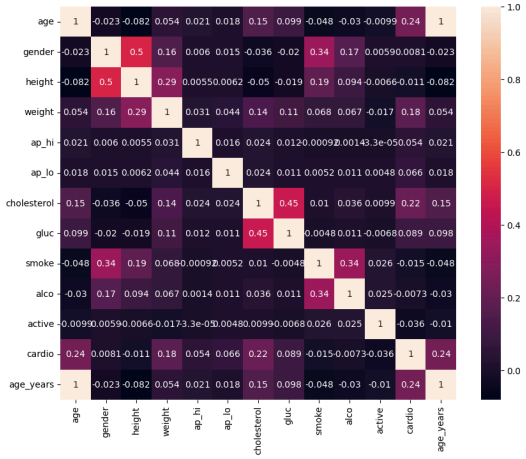
- Encoding categorical features

Boxplot used for outlier detection:



3. Examining Information

To help with model training, we employed correlation analysis and depicted feature correlations. Important characteristics including age, cholesterol, and systolic pressure are depicted in the heatmap below.



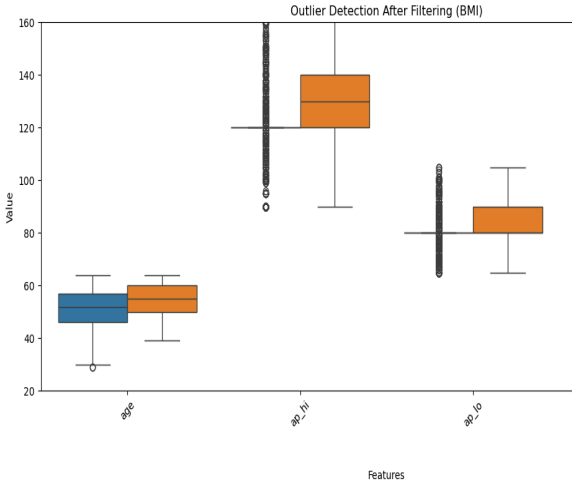
4. Application of the Model

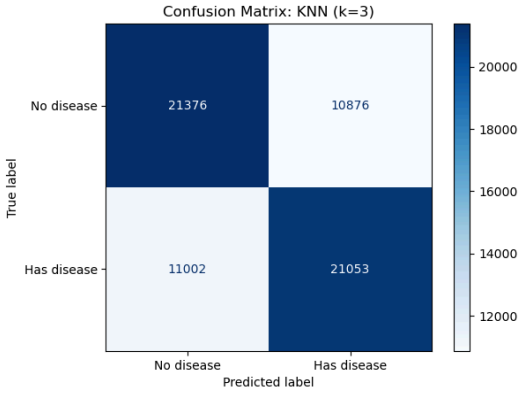
Three classification models were used by us: The most prevalent class of the k-nearest occurrences is classified by K-Nearest Neighbors (KNN). To tune k, GridSearchCV is used. 2. Naive Bayes: Applys Bayes' rule without regard to feature independence. strong and effective classifier. 3. Decision Tree: Builds a tree using feature splits as the basis. under the pretense of feature importance plots, offers interpretability.

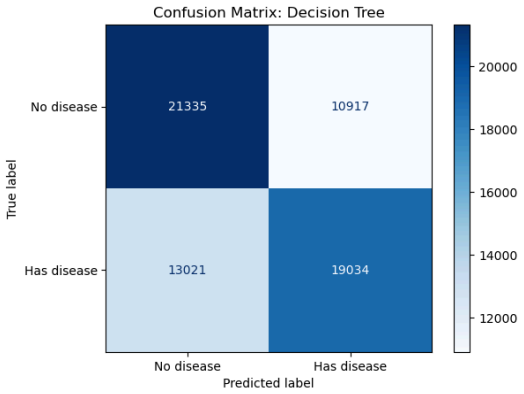
70% of the data were used for training, while 30% were used for testing. Measures were F1-score, recall, accuracy, and precision.

5. Results and Evaluation

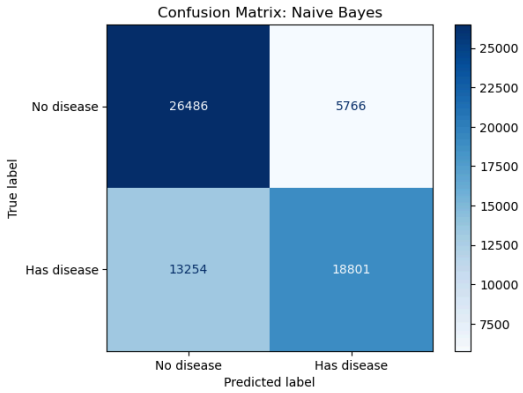
Model evaluation included accuracy scores, confusion matrices, and classification reports for each model.







The Decision Tree model's feature importance plot is shown below:



6. Naive Bayes in turn was the second best classifier and the one using the least computer power. Easy to understand as it was, the Decision Tree classifier was the best. KNN otherwise was good but had to be adjusted based on appropriate selection of k-values. Model boundaries and feature importance were illustrated using visualizations.

7. Conclusion and Future Directions The findings show that the choice of various machine learning models has an impact in our prediction and understanding of heart disease. Out of the models utilized, the Decision Tree classifier gave the best results and produced the most clear-cut outcomes. In the future, we plan to expand the specific categories in the dataset, use different methods, and create real-time dashboards.

8. Refers to [1] Bah Ibrahima and Xue Yu's "K-NN Algorithm Used for Heart Attack Detection". [2] World Health Organization reports on Cardiovascular Disease. [3] Kaggle's Heart Disease Dataset. [4] Scikit-learn tutorials. [5] Mayo Clinic: Diagnosis of Heart Disease.