Cardiovascular Disease Prediction Using Machine Learning

Sarika Simha, Sanjana Sundhar, Sonika Puvvada, Isha Atre

Professor Jorge Silva

1 Introduction

1.1 Problem/Application Motivation:

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide. Early detection and prediction of CVD risk play an important role in preventive healthcare strategies. This project will use machine learning techniques to develop a predictive model for identifying individuals at risk of cardiovascular disease based on various health-related features. We will be analyzing the Heart Disease dataset linked here to explore features that will predict individuals who are at risk for CVD.

1.2 Survey Related Work:

In the field of cardiovascular disease prediction, many studies have employed machine learning algorithms to analyze health-related data for early detection. Research indicates the importance of factors such as age, gender, blood pressure, cholesterol levels, and other parameters in predicting CVD risk.

As shown on the age-frequency graph below obtained from this dataset, individuals as young as their late 50s are at higher risk for heart disease.

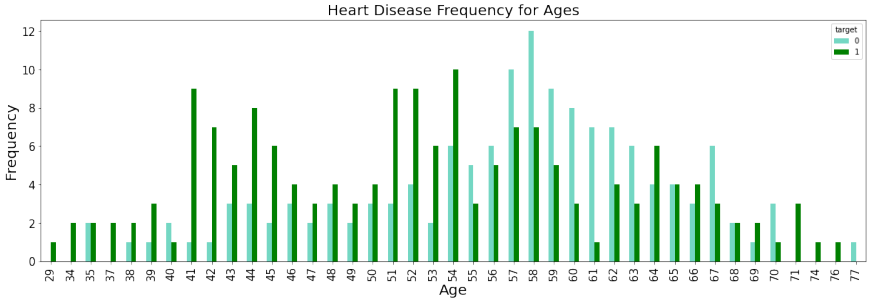

Figure 1: Age Frequency Graph

2 Methods

2.1 Approach:

Our approach involves a systematic process, starting with data exploration and preprocessing. We will convert categorical variables into numerical representations and standardize numerical features. The dataset is split into x and y training and testing sets for model training and evaluation. For modeling, we experiment on our train and test datasets with multiple machine learning algorithms: K Nearest Neighbors, Decision Trees, and Naive Bayes. The heart disease dataset contains 14 variables, ranging from age, gender, cholesterol, maximum heart rate, and target (whether the individual has the disease or not). The data has been split into x and y: x contains data excluding the target column and y contains only the target data. We will interpret the results to understand the impact of different factors leading to CVDs.

2.2 Models/Results:

2.2.1 Gaussian Naive Bayes

For our first model, we developed an instance of a Gaussian Naive Bayes classifier. We made the naive assumption that the values in our x_train dataset, the health-related factors including age, blood pressure, and cholesterol are independent from the y_train dataset- whether they have heart disease or not. The model is known to perform well which shows consistent in our accuracy value of 85.25%.

2.2.2 Decision Tree Classifier

The test accuracy of 75.41% indicates the percentage of correctly classified instances by the Decision Tree model on the provided test dataset. This value is also an indication of how well it performs on new, previously unseen instances.

2.2.3 K-Nearest Neighbors KNN

The last model explores a range of k-values and trains KNN classifiers for each k on a heart disease dataset, using input features represented by x_train and x_test and corresponding outputs in the y_train and y_test. The accuracy scores for each model on the test set are stored in the scores list. The ultimate goal is to identify the k value that maximizes the accuracy of the KNN model, providing insights into the most effective configuration for predicting heart disease. As a result, the maximum accuracy displayed was 90% with the k = 7 for predicting individuals who have heart disease based on factors, such as age and cholesterol, in the x datasets.

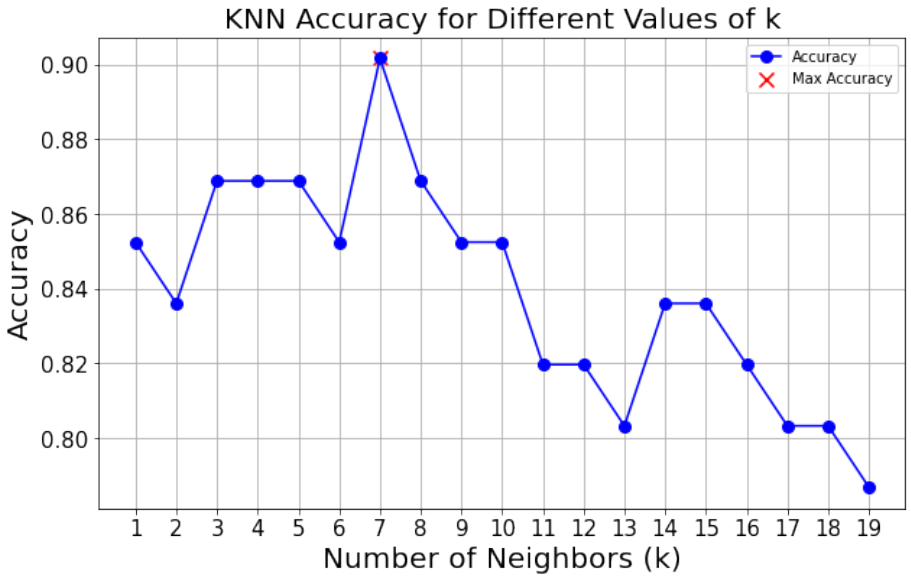

Figure 2: K-Nearest Neighbors Graph

3 Conclusion/Impacts:

Our models exhibit promising performance in predicting cardiovascular disease risk.

The graph below compares the accuracy rates for the three different methods that were used in this analysis. Looking at the graph, the K-Nearest Numbers method was the most accurate, followed by Naive Bayes, and the Decision Tree Classifier, respectively.

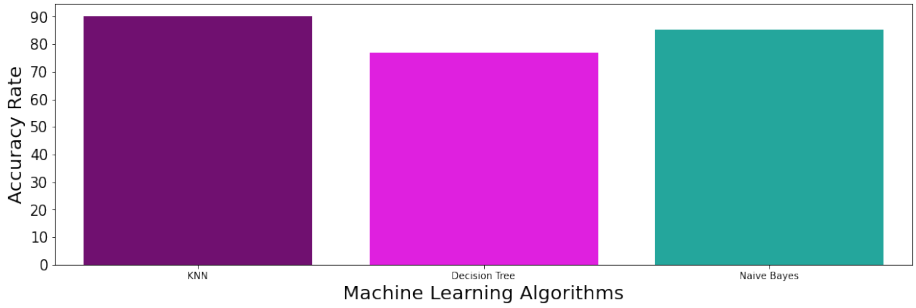

Figure 3: Model Accuracy Comparison Graph

This project contributes to the early detection of cardiovascular diseases, en-

abling timely interventions and potentially reducing the incidence of cardiovascular events. The work done in this project can be applied in real-world healthcare settings by providing a tool for clinicians to assess an individual's risk of cardiovascular disease.

Works Cited

[1] Ahmed, Rasel. "Heart Disease." Kaggle, 8 Sept. 2020, www.kaggle.com/datasets/data855/heart-disease.

[2] "Know Your Risk for Heart Disease." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 21 Mar. 2023, www.cdc.gov/heartdisease/risk_factors.htm.