International Journal of Research in Engineering, Science and Management
Volume 7, Issue 6, June 2024
https://www.ijresm.com | ISSN (Online): 2581-5792

Heart Disease Prediction Using KNN

Tulluri Sai Sriya*
Student, Department of Electronics and Communication Engineering, AI & ML, GITAM University, Hyderabad, India

***Abstract*: The purpose of this project is to build a strong machine-learning model to predict heart disease with high accuracy by taking input parameters related to weight, height, and many other important health metrics of a patient. Heart diseases are a leading cause of death and kill almost 17.6 million people every year. Such harrowing numbers indicate the necessity for actionable early detection and prevention measures. This model leverages on K -Nearest Neighbors (KNN) machine learning algorithm which classifies patients based on their risks of suffering from a heart disease. This model developed an impressive accuracy of 91%, making it a valuable tool for doctors. This particular model can help early identification of heart disease, which can be significant in saving the lives of millions of patients.**

***Keywords*: classification, cardiovascular heart disease, heart disease, K-nodes, machine learning, prediction.**

1. Introduction

Cardio-vascular diseases (CVDs) are the major cause of death worldwide. The WHO estimates that CVDs are responsible for about 17.9 million deaths annually, representing about 32 percent of all global deaths. These diseases can cause heart attacks, sudden deaths, heart failure, stroke, etc. CVDs are major health threats resulting in serious illness and disability having a high effect on the quality of life. CVD expenditures on healthcare and loss of productivity place an enormous economic burden on patients and healthcare systems. Certain CVDs, like myocardial infarctions (heart attacks) and sudden cardiac arrest, can lead to sudden death if not promptly treated, underscoring the need for immediate medical intervention. Living with CVDs can lead to psychological stress, anxiety, and depression. The fear of recurrent cardiac events and the burden of managing a chronic illness can take a toll on mental health.

Access to quality healthcare and preventive services is essential for managing CVDs, yet disparities exist, particularly in low- and middle-income countries.

The average risk factors for CVDs are raised blood pressure, raised cholesterol levels, tobacco use, excessive alcohol intake, obesity, physical inactivity, and a diet high in salt and saturated fats. Early detection is critically important as cardiovascular diseases are associated with high mortality and morbidity globally. Using machine learning, one can predict the propensity of CVDs by analyzing patient history and risk factors, hence interventions can be reached early enough. Thus, moving to ML solutions can help millions of people and also help doctors in analyzing and predicting before- hand. Machine Learning (ML) is one of the most promising technologies that presents new prospects in the horizon of health care for diagnostics, treatment and patient care. Through the analysis of patient data, ML models can predict if a patient might develop a disease or not. The data set chosen for this project has 308854 rows and 19 columns, covering all 19 different pivotal factors of 308884 patients that are helpful in determining their risk of suffering from a heart disease. The algorithm used is K-Nearest Neighbors algorithm. The model is trained on 80 percent of the data set and the remaining is used for testing the data. All computations, preprocessing, and visualizations for this project were conducted using Google Colab and Python.

2. Methodology

This study aims to predict the probability of heart disease through ML prediction, which can be beneficial for doctors and patients. To fulfill this objective, KNN (K-Nearest Neighbors) machine learning algorithm has been employed on a dataset and the results have been presented in this paper. To improve the methodology, the following steps have been undertaken: cleaning the data, eliminating irrelevant information, converting data format, implementation and evaluation.

A. Data Source

The data set used has 19 distinctive features of 308854 patients, which are listed in the table below. The class "Heart _Disease" is the target class and indicates if the patient has a heart disease (1) or not (0).

B. Data Cleaning

I started cleaning the data by first evaluating the null values in each feature. Any feature with a null value was then deleted from the dataset. I also manually examined the 19 features one by one in order to remove data with very less entries which don't contribute significantly to the prediction. Then they were eliminated from the dataset. Also, if any irrelevant data was present, it was removed. These steps were taken to improve the performance of the model.

C. Converting Data into Binary Data for Classification

Once the data was cleaned, the next step was to convert it into numerical format from categorical format, specifically to binary format of 1s and 0s, as our problem is to classify the patients into two classes: Either 1 or 0. Here, 1 represents the presence and 0 represents the absence of heart disease. Python's label encoder method facilitated this transformation.

*Corresponding author: sriyatulluri@gmail.com

D. Implementation

The dataset was split into an 80:20 ratio, with 80% of the data allocated for training the model and the remaining 20% for testing it. The K-Nearest Neighbors (KNN) algorithm was applied to the data, and its classification effectiveness was assessed using evaluation metrics such as accuracy.

1) KNN algorithm

The K-Nearest Neighbors (KNN) algorithm is a simple yet effective supervised learning method used for classification tasks. In essence, KNN operates on the principle of similarity: it assumes that similar data points are close to each other in a multi-dimensional feature space. To classify a new point, KNN dives into its 'K' nearest neighbors in a training set, done so by evaluating the distance the data points reside in from one another based on a naive preferred metric (distance metric), usually Euclidean distance. The algorithm then chooses the most common class label among its K nearest neighbors (for classification) and this label or value is assigned to the new data point. Mathematically, the classification decision can be expressed as:

$$\hat{y}=\text{mode}(y_i)$$

where y_i is the class label of i^{th} nearest neighbor
$\hat{y}$ represents the predicted class label for the new data point.

Table 1
19 Variables (Features)

Feature	Variable in data set
General health	General_Health
Check up	Checkup
Exercise	Exercise
Heart disease(Target class)	Heart_Disease
Skin cancer	Skin_Cancer
Any other cancer	Other_Cancer
Depression	Depression
Diabetes	Diabetes
Arthritis	Arthritis
Sex	Sex
Age	Age_Category
Height in cm	Height_(cm)
Weight in kg	Weight_(kg)
BMI	BMI
Patient's smoking history	Smoking_History
Patient's alcohol consumption levels	Alcohol_Consumption
Fruits intake	Fruit_Consumption
Green vegetables intake	Green_Vegetables_Consumption
Fried potato intake	FriedPotato_Consumption

This table presents the 19 variables (Features) that are a part of the data set used.

3. Results

After implementing and evaluating the model using the accuracy metric, the outcomes were promising, achieving a 91% accuracy rate.

4. Conclusion

In conclusion, this study successfully explored the application of machine learning algorithms, particularly the K-Nearest Neighbors (KNN) method, for predicting heart disease. The achieved accuracy of 91% underscores the potential of machine learning in enhancing early detection and preventive healthcare strategies. In the future, more research can be done to optimize and update the model parameters, add more features and test the model on a larger, more diverse dataset. In the end, the deployment of predictive analytics among clinical settings is a prospect for better health outcomes and resource allocation at systems level.

References

[1] Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo. (2023, Feb). Effective Heart Disease Prediction Using Machine Learning Techniques.

[2] Umarani Nagavelli, Debabrata Samanta, and Partha Chakraborty. (2022, Feb). Machine Learning Technology-Based Heart Disease Detection Models, 2022.

[3] Tsehay Admassu. (2021, Dec). Heart disease prediction model with k-nearest neighbor algorithm.

[4] Ryan Marcus Jeremy M. Lupague, Romie C. Mabborang, Alvin G. Bansil, Melinda M. Lupague, (2023, June). Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based On Multi-Dimensional Health Factors.

[5] Data-set: From Kaggle https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset