International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 • Website: www.ijfmr.com • Email: editor@ijfmr.com

Predictive Modeling of Heart Disease Using Python and Machine Learning

Abhishek Upadhyay

Computer Science and Engineering , Galgotias University, Greater Noida, Uttar Pradesh, India.

Abstract

Cardiovascular disease is a sickness that can cause sudden death. It happens when the heart is not working properly due to many things. There are many factors that can affect the heart, such as obesity, high blood pressure, and cholesterol. The number of cases for death due to heart disease has been increased and there is a need for methods to help predict the disease, aid in early diagnosis, and help doc- tors treat patients medically. The current study aims to estimate the risk of heart attack based on data from patients. In practice, prediction and interpretation are the main goals of data discovery. Predictive data mining involves attributes or variables in datasets to determine unknown or future values of other factors. This definition refers to finding patterns that interpret data for human interpretation. Machine learning is now used in many fields, and healthcare is no exception. K- nearest, random forests etc. such as machine learning algorithms (classification algorithms). Medical care is about people’s lives and should be the right one. Therefore, we need to create a system that can accurately predict the disease. To give treatment for heart disease, a lot of advanced technologies are used. In medical center it is the most common problem that many of medical persons do not have equal knowledge and expertise to treat their patient so they deduce their own decision and as a result it show poor outcome and sometime leads to death. To overcome these problems predictions of heart disease using machine learning algorithms and data mining techniques, it become easy to automatic diagnosis in hospitals as they are playing vital role in this regard. Heart disease can be predicted by performing analysis on patient’s different health parameters. There are different algorithm to predict heart disease like na¨ıve Bayes, k Nearest Neighbor (KNN), Decision tree ,Artificial Neural Network(ANN).We have used different parameters to predict heart disease. Those parameters are Age, Gen- der, Cerebral palsey (CP), Gender, Cerebral palsey (CP), Blood Pressure (bp), Fasting blood sugar test (fbs) etc. In our research paper, we used built in dataset

.we have implement the five different techniques with same dataset to predict heart disease These implemented algorithm are Naive Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest .This paper investigates that which technique gives more accuracy in predicting heart disease based on health parameters. Experiment show that Na¨ıve Bayes has the highest accuracy of 88

Keywords: Naive Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest, Heart Disease

1 Introduction

Heart disease stands as a paramount global health concern, claiming more lives annu- ally than any other cause. Shockingly, an estimated 12 million individuals succumb to heart-related ailments each year. Within the United States alone, heart disease strikes with devastating frequency, snatching a life every 34 seconds. These tragic events often stem from the obstruction of blood flow to vital organs such as the heart or brain. Individuals at risk of heart disease typically exhibit elevated levels of blood pres- sure, glucose, and lipids, alongside heightened stress levels. Remarkably, these critical parameters can be readily monitored from the comfort of one's home with basic health facilities.

When delving into the realm of heart disease, one encounters a spectrum of conditions encompassing coronary heart disease, cardiomyopathy, and various cardio- vascular diseases. The term "heart disease" encompasses a multitude of afflictions affecting the heart and blood vessels, disrupting the fluid's circulation throughout the body. The pervasive impact of cardiovascular disease (CVD) reverberates through society, spawning diseases, disabilities, and untimely deaths.

The diagnosis of heart disease represents a pivotal yet intricate facet of medical practice. Medical professionals grapple with the arduous task of accurately discerning these ailments, a task further complicated by the scarcity of resources and expertise in certain regions. In such contexts, the automation of diagnostic processes emerges as a beacon of hope. Enter data mining—a powerful tool capable of unearthing hidden patterns and insights that can inform judicious decision-making in healthcare.

In the intricate tapestry of healthcare, the ability to provide accurate diagnoses and quality services hinges not only on the proficiency of individual practitioners but also on the support systems and approaches adopted by healthcare organiza- tions. Addressing the knowledge gaps among healthcare professionals who may lack specialized expertise is paramount.

Yet, one of the foremost challenges faced by existing diagnostic methods lies in their ability to draw ecise conclusions when needed most. In our pioneering approach, we harness the potential of various data mining techniques and machine learning algorithms—namely Na¨ıve Bayes, k Nearest Neighbor (KNN), Decision Tree, Artifi- cial Neural Network (ANN), and Random Forest—to predict heart disease based on key health parameters. This innovative endeavor not only promises to revolutionize diagnostic accuracy but also underscores the transformative potential of leveraging cutting-edge technologies in healthcare.

2 Methodology

The primary aim of our proposed method is to enable early detection of heart disease by accurately predicting its occurrence within a short timeframe. To achieve this objective, we employ a comprehensive array of data mining techniques and machine learning algorithms, including Na¨ıve Bayes, k Nearest Neighbor (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest. This multifaceted approach capitalizes on key health parameters to forecast the likelihood of heart disease.

Our analysis is conducted using Anaconda Navigator's Jupyter Notebook, a versatile and user-friendly platform. This open-source software empowers us to seam- lessly implement a myriad of machine learning algorithms by leveraging its extensive library of resources. Moreover, Anaconda Navigator simplifies the process of access- ing requisite libraries through its intuitive interface, thereby enhancing efficiency and productivity.

By harnessing the capabilities of Anaconda Navigator's Jupyter Notebook, we can seamlessly create live code, visualize data, process intricate datasets, and generate insightful graphs. This integrated platform serves as a dynamic tool for expeditious analysis and facilitates the exploration of diverse predictive models with unparalleled ease.

Table 1

Sl.No	Feature Name	Feature code	Description
1	Age	Age	Age of the person in years
2	Type of chest pain	chest_pain	1. Atypical angina 2. Typical angina 3. Asymptomatic 4. Non-anginal pain
3	Resting blood pressure	rest_bpress	mm Hg (92 to 200)
4	Fasting blood sugar	blood_sugar	Fasting blood sugar ¿120 mg/dl (t = true, f = false)
5	Resting electrocardiographic results	rest_electro	1. Left ventricular hypertrophy 2. Normal 3. ST-T wave abnormality
6	Maximum heart rate achieved	max_heart_rate	(82 to 188)
7	Exercise-induced angina	exercise_angina	1. Yes 2. No

2.1 Dataset for Implementation

We utilized a built-in dataset sourced from the UCI Machine Learning Repository to predict heart disease. This database comprises 14 attributes, each playing a crucial role in our predictive model. These attributes include:

Age Sex Cerebral palsy/chest pain (CP) Blood Pressure (BP) in mm Hg Choles- terol in mg Fasting blood sugar test (FBS) Resting electrocardiographic results Thalach (Maximum heart rate achieved) Exang (Exercise-induced Angina) Oldpeak (ST depression induced by exercise relative to rest) Slope (The slope of the peak exercise ST segment) CA (Number of major vessels) Thal (Reversible defect) Target (0,1) Prior to model training, we partitioned the data into training and testing sets, allocating 25 percent of the dataset for testing and 75 percent for training purposes. Additionally, we conducted data normalization to address any missing values (NaN) within the dataset, ensuring its integrity and reliability for analysis.

2.2 Algorithm's used for experiment

K Nearest Neighbor (KNN):

KNN is a machine learning algorithm commonly used for classification tasks, especially when dealing with continuous parameters. It operates by predicting the class of a data point based on the majority class among its nearest neighbors. KNN is preferred

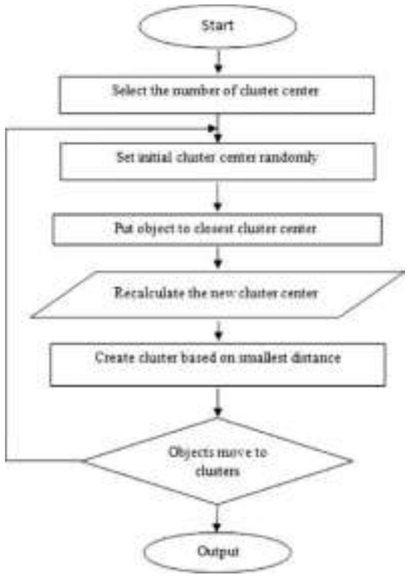

Fig. 1 Flowchart of K-means clustring -algorithm

for its simplicity and high speed, and it can handle both classification and regression problems. In our experiment, we apply KNN to classify individuals as either having heart disease or not, based on parameters such as age and sex. The algorithm does not require training data for model generation, as it utilizes all cases in the dataset and classifies new data points based on their proximity to existing data points. The KNN algorithm typically involves two stages:

- Finding the k number of instances in the dataset
- Using the k instances to determine the nearest neighbor

Artificial Neural Network (ANN):

Artificial Neural Networks mimic the structure and function of human neurons, com- prising nodes connected through directional links. Each node represents a processing unit, and the links denote causal relationships between them. ANNs find application in clinical decision-making, aiding doctors in analyzing and making decisions efficiently and accurately. A Neural Network begins with an input layer, where each node is connected to nodes in hidden layers, which may then connect to an output layer.

3 Results

In this section, we present the findings of our data analysis aimed at uncovering essential patterns for predicting heart diseases. The variables considered for heart disease prediction include age, chest pain type, blood pressure, blood glucose level, ECG during rest, heart rate, and four types of chest pain, along with the presence of exercise-induced angina.

The heart disease dataset underwent thorough pre-processing, involving the removal of unrelated records and addressing missing values. Following this pre- processing step, the heart disease dataset was subjected to analysis using the K-means algorithm.

Within our analysis, we specifically focused on four types of heart diseases: asymp- tomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain. The results were derived by considering all four types of chest pain alongside other influential variables.

Data Analysis

The histogram illustrated in Fig. 2 illustrates the distribution of ages and their cor- relation with the risk of heart disease within the targeted class. It is evident from the observation that individuals within the target class, particularly those aged between 50 to 55, exhibit a notably high risk of heart disease. This age range is particularly significant as it coincides with the onset of the development of coronary fatty streaks, a key precursor to heart disease. In Fig. 3, the visualization depicts the relationship between diabetic individuals within the target class and their maximum heart rate. The color coding highlights individuals with diabetes in blue, while those without diabetes are represented in red. Notably, among the target class with diabetes and a satisfactory heart rate, negative symptoms are observed.

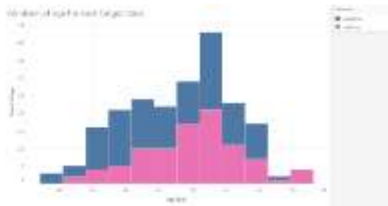

Fig. 2 Histogram of variation of age for each target class

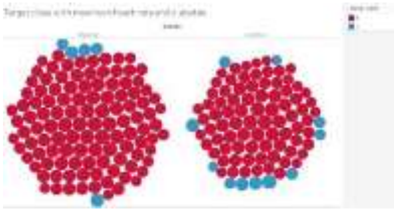

Fig. 3 Target class with maximum heart rate and diabetics

Fig. 4 Impact of blood pressure and sugar in heart disease

In Fig. 4, the illustration demonstrates the correlation between blood pressure, sugar levels, and the likelihood of heart disease. It is deduced that individuals with diabetes and elevated blood pressure are at a heightened risk of developing heart disease.

4 Conclusion

Our study primarily focused on leveraging data mining techniques in healthcare, par- ticularly in the detection of heart disease, a potentially fatal condition. By employing algorithms such as KNN, Neural Networks, Decision Tree, Naive Bayes, and Random Forest, we aimed to predict heart disease and assess their performance based on met- rics like Accuracy, True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) rates.

Through five experiments conducted on the same dataset, we compared the results of these implemented algorithms, presenting them in tabular form for easy compre- hension and comparison. Our findings revealed that Naive Bayes exhibited the highest accuracy at 88 percent, followed closely by ANN and KNN, both achieving an accu- racy of 87 percent. This underscores the potential of data mining in healthcare for early disease prediction and diagnosis.

In future endeavors, we propose further research to enhance classification accuracy by exploring advanced algorithms such as Bagging, Support Vector Machine, or deci- sion tables. It's essential to evaluate the predictive performance of each algorithm and tailor the proposed system to specific domains of interest. Additionally, incorporating additional relevant features into the algorithms could improve accuracy.
Stakeholders should consider adopting this dedicated tool to facilitate better decision-making processes. While our current implementation did not involve parame- ter adjustments, future iterations could benefit from fine-tuning parameters to enhance experiment outcomes.
Furthermore, future research could expand by incorporating more extensive datasets related to heart disease and exploring various data reduction techniques. Uti- lizing high-quality datasets devoid of inconsistencies could yield more accurate and reliable predictions of heart disease.

References

1. V. Manikantan and S. Latha, "Predicting the Analysis of Heart Disease Symp- toms Using Medicinal Data Mining Methods," *International Journal on Advanced Computer Theory and Engineering*, vol. 2, no. 2, pp. 5–10, 2013.
2. Dr. A.V. Senthil Kumar, "Heart Disease Prediction Using Data Mining prepro- cessing and Hierarchical Clustering," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 4, no. 6, pp. 07–18, 2015.
3. Uma.K, M. Hanumathappa, "Heart Disease Prediction Using Classification Techniques with Feature Selection Method," *Adarsh Journal of Information Technology*, vol. 5, no. 2, pp. 22–29, 2016.
4. Himanshu Sharma, M.A. Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 8, pp. 99–104, 2017.
5. S. Suguna, Sakthi Sakunthala.N, S. Sanjana, S.S. Sanjhana, "A Survey on Pre- diction of Heart Disease using Big data Algorithms," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 6, no. 3, pp. 371–378, 2017.
6. A.L. Bui, T.B. Horwich, G.C. Fonarow, "Epidemiology and risk profile of heart failure," *Nature Reviews Cardiology*, vol. 8, no. 1, pp. 30–41, 2011.
7. J. Mourão-Miranda, A.L.W. Bokde, C. Born, H. Hampel, M. Stetter, "Classify- ing brain states and determining the discriminating activation patterns: support vector machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005.
8. S. Ghwanmeh, A. Mohammad, A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 3, pp. 176–183, 2013.
9. Q.K. Al-Shayea, "Artificial neural networks in medical diagnosis," *International Journal of Computer Science Issues*, vol. 8, no. 2, pp. 150–154, 2011.
10. K. Vanisree, J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," *International Journal of Computer Applications*, vol. 19, no. 6, pp. 6–12, 2011.
11. I. Al Mamoon, A.S. Sani, A.M. Islam, O.C. Yee, F. Kobayashi, S. Komaki, "A proposal of body implementable early heart attack detection system," 1–4, 2013.

12. K. Patterson, “Matthias Nahrendorf,” *Circulation Research*, vol. 119, pp. 790– 793, 2016.
13. J. Soni, U. Ansari, D. Sharma, S. Soni, “Predictive data mining for medical diagnosis: An overview of heart disease prediction,” *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.
14. H.D. Masethe, M.A. Masethe, “Prediction of heart disease using classification algorithms,” in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, pp. 22–24, October 2014.
15. A. Methaila, P. Kansal, H. Arya, P. Kumar, “Early heart disease prediction using data mining techniques,” in *Proceedings of Computer Science & Information Technology (CCSIT-2014)*, vol. 24, pp. 53–59, Sydney, NSW, Australia, 2014.