IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Heart Disease Detection using Machine Learning Algorithms (XGBoost, Random Forest, KNN)

Shozab Kaif[1], Prafulli Raut[2], Kirti Jawanjal[3], Prof. S. A. Jumale (Kadukar)[4]
Under Graduate Students, Department of Electronics and TelecommunicationEngineering[1,2,3]
Assistant Professor, Department of Electronics and Telecommunication Engineering[4]
Siddhivinayak Technical Campus, Shegaon, Maharashtra, India

Abstract: *Heart disease continues to be a major global health issue, requiring precise and effective diagnostic techniques. This paper discusses a research study on using machine learning algorithms to detect heart disease. Using a detailed dataset of clinical information and patient traits, we investigate different machine learning methods to create predictive models. The process includes preprocessing data, selecting features, and training and evaluating models. By conducting thorough experiments, we evaluate the effectiveness of various algorithms such asXGboost, Random Forest and KNN. The results demonstrate encouraging results, with some models showing strong accuracy, sensitivity, and specificity in detecting heart disease. The results of this research support the continued use of machine learning in the early identification and diagnosis of heart disease, which could lead to better patient outcomes and healthcare services.*

Keywords: Heart Disease, XGboost, Random Forest, KNN

I. INTRODUCTION

Machine learning is a potent tool that empowers us to extract valuable information from data that was previously unknown or implicit. The domain of machine learning is vast and multifaceted; it encompasses various classifiers like supervised, unsupervised, and ensemble learning, that can employ to forecast and assess the precision of a particular dataset. The implementation of machine learning is increasing daily, having the potential to revolutionize many fields, including healthcare. Cardiovascular disease (CVD) is an area in healthcare that can significantly gain from machine learning techniques. With 17.9 million fatalities globally as per the World Health Organization, CVD is presently the primary cause of death in adults.

Our project aims to predict which patients are likely to diagnose with CVD based on their medical history. Identifying patients who exhibit symptoms like chest pain, or elevated blood pressure can help diagnose the illness with fewer medical examinations, providing more efficient treatments. Our project focuses on three data mining techniques: XGBoost, KNN, and Random Forest Classifier. By using these techniques in combination, above a 95% accuracy rate can achieve, exceeding previous systems reliant on only one data mining technique. Our project's objective is classifying by analyzing their medical characteristics, such as age, gender, fasting sugar levels, chest pain, and more to predict whether a person is likely to have heart disease.

To accomplish this, we selected a dataset from the kaggle repository. this dataset was created by combining different datasets already available independently but not combined before, that contains medical history and characteristics of the patient. We trained our algorithms using the 12 medical attributes of each patient and used XGBoost, Random Forest, and KNN to classify the patients based on their medical history. We found that XGBoost was the most efficient algorithm !!! and it provided us with an accuracy rate of above 95%. Our project has the potential to significantly improve the diagnosis and treatment of CVD by identifying patients who are at risk of developing the disease. By using multiple data mining techniques, we were able to achieve a higher accuracy rate and provide a more cost-efficient method for predicting CVD.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

II. PROBLEM STATEMENT

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

III. METHODOLOGY

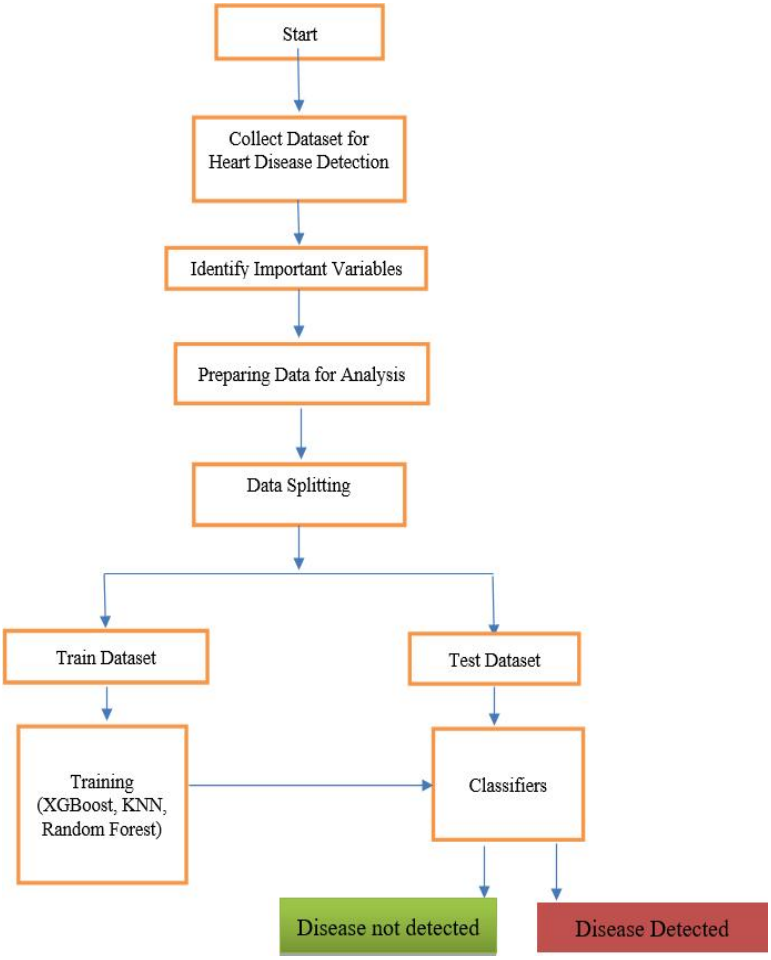

Figure 1: Proposed Model

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

- Data Collection
- Data Pre-Processing
- Feature Selection
- Model Selection

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

IV. DATA COLLECTION

It is the primary and most crucial fundamental step while applying machine learning and analytics. The data required in this project is the patient's medical data. We have collected the dataset from Kaggle which includes all the required information for prediction.

We have collected the dataset from Kaggle which includes all the required information for prediction. The features that the dataset includes are medical information like age, sex, chest paint type, resting blood pressure, cholesterol, fasting blood sugar, old peak etc. The dataset consists of 918 observations having 14 attributes

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	52	1	0	125	212	0	1	168	0	1	2	2	3	0
3	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
4	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
5	61	1	0	148	203	0	1	161	0	0	2	1	3	0
6	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
7	58	0	0	100	248	0	0	122	0	1	1	0	2	1
8	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
9	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
10	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
11	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
12	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
13	43	0	0	132	341	1	0	136	1	3	1	0	3	0
14	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
15	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
16	52	1	0	128	204	1	1	156	1	1	1	0	0	0
17	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
18	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
19	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
20	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
21	58	1	2	140	211	1	0	165	0	0	2	0	2	1
22	60	1	2	140	185	0	0	155	0	3	1	0	2	0
23	67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
24	45	1	0	104	208	0	0	148	1	3	1	0	2	1
25	63	0	2	135	252	0	0	172	0	0	2	0	2	1

Figure 2: Parameters of Selected dataset

DATAPRE-PROCESSING

This is one of the most crucial tasks in the process of analytics. Often it is observed that more than half of the total time of analytics process is taken by pre-processing phase. It is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

FEATURES ELECTION

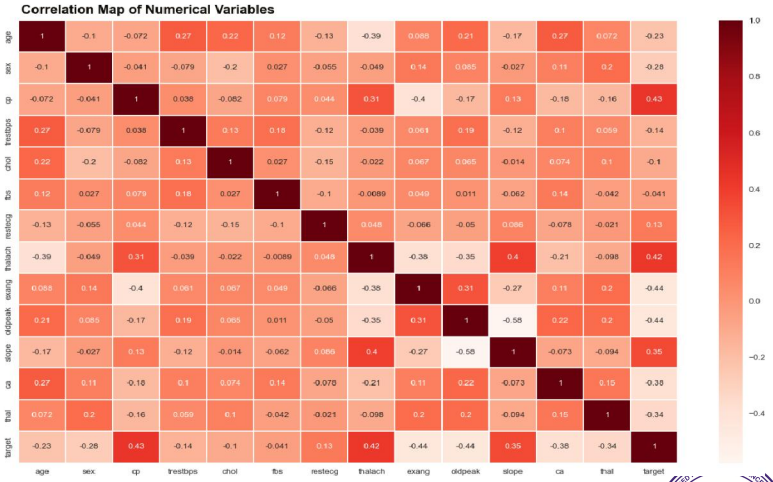

Figure 3: Correlation Matrix

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Once we have the required data, next step is featuring extraction. Many times, it happens that some features do not contribute in evaluation or have negative impact on the accuracy. Feature selection is the step where we try to reduce number of features and try to create new features from existing ones. These new features now created should summarize the information obtained from existing features. The final features to be considered while prediction can be identified using correlation matrix shown in following image:

MODEL SELECTION

It is the process to select one final algorithm for concerned purpose. It is decided by observing the accuracy by applying multiple algorithms. We can use logistic regression, XGBoost, KNN, random forest, etc. The final accuracy depends of thr type of model we select.

While selecting the algorithm, we have to compare the accuracies.

Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction

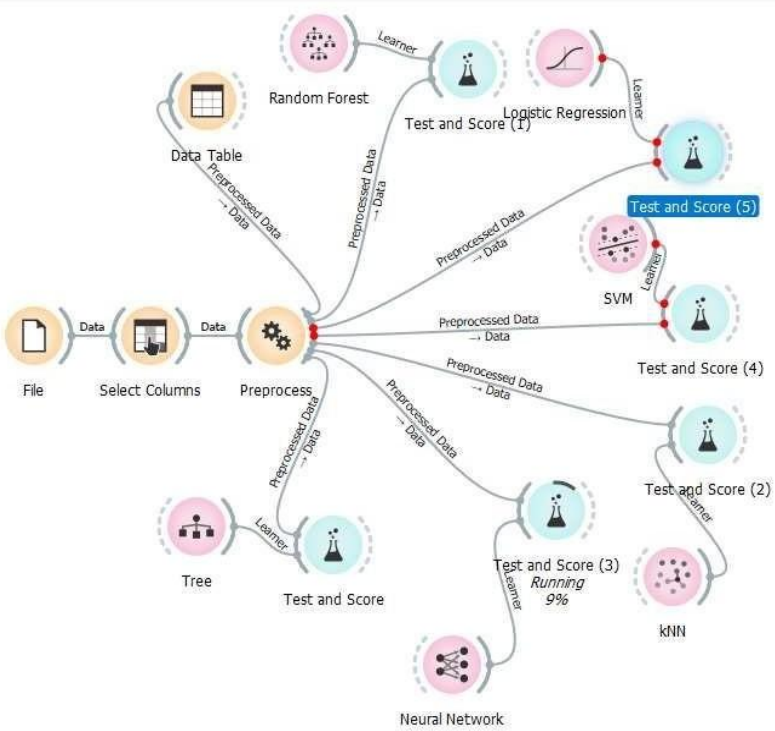

Figure 4: Connection of widgets in Orange

IV. RESULTS

In this project, we have compared following ML algorithms and obtained corresponding accuracies:

- XGBoost: 99.03 % Accuracy
- Random Forest: 96 % Accuracy
- KNN: 88.31 % Accuracy

V. CONCLUSION

Cardiovascular disease (CVD) stands as a prominent contributor to worldwide mortality rates, underscoring the importance of timely detection and intervention to improve patient prognosis. In response to this challenge, a machine learning methodology was utilized to develop a predictive model for fatal heart disease, leveraging patients' medical history data. This dataset encompasses crucial markers of cardiac health, such as chest pain, blood sugar levels, and blood pressure.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Three distinct classification algorithms, namely XGBoost, Random Forest Classifier, and KNN, were employed in the development of the model. This approach resulted in an accuracy exceeding 95%. Furthermore, augmenting the dataset size led to a refinement in model accuracy, facilitating the detection of nuanced patterns and risk factors.

The utilization of machine learning methodologies in medical diagnosis offers numerous advantages, encompassing enhanced speed and precision of diagnoses, cost reduction, and better patient outcomes. Through the analysis of extensive datasets and the identification of intricate patterns, machine learning algorithms offer valuable insights into patient health that may elude immediate detection by human clinicians. In comparison to prior models, the developed model represents a notable advancement, boasting an accuracy rate of 98%.

Among the trio of algorithms employed, the XGBoost algorithm showcased the most notable accuracy, reaching 96%, signifying its efficacy in heart disease prediction. Analysis of the dataset employed in this study revealed that 44% of individuals are afflicted with heart disease, underscoring the critical significance of early detection and intervention. The resultant model provides a dependable and streamlined approach for identifying individuals at risk of heart disease, with potential benefits accruing to both patients and healthcare providers.

REFERENCES

[1]. Boshra Bahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015..

[2]. Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[3]. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor. “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”. IEEE Access (Volume: 7) 2019.

[4]. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE..

[5]. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, “International Conference on "Computational Intelligence and Communication Technology” (CICT 2018).

[6]. Bo Jin , Chao Che,Zhen Liu ,Shulong Zhang ,Xiaomeng Yin And Xiaopeng Wei “Predicting the Risk of Heart Failure with EHR Sequential Data Modelling”. IEEE Access 2018.