QuillBot

Scanned on: 6:36 January 25, 2023 UTC

Overall similarity score

Results found

Total words in text

	Word count
Identical	93
Minor Changes	12
Omitted	0

Results

The results include any sources we have found in your submitted document that includes the following: identical text, minor changed text, paraphrased text.

The problem is related to healthcare applications of | Chegg... *https://www.chegg.com/homework-help/questions-and-answers/problem-...*	1%
Heart disease — heart • SLOPE *https://jolars.github.io/SLOPE/reference/heart.html* heart.html	1%
Machine Learning Technology-Based Heart Disease Detectio... *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8898839/*	1%
document *https://hal.science/hal-03628382/document*	1%

IDENTICAL

Text that is exactly the same.

MINOR CHANGES

Text that is nearly identical, yet a different form of the word is used. (i.e 'slow' becomes 'slowly')

Unsure about your report?

The results have been found after comparing your submitted text to online sources, open databases and the Copyleaks internal database. If you have any questions or concerns, please feel free to contact us atsupport@copyleaks.com

Click here to learn more about different types of plagiarism

Early Stage Cardiovascular Disease Prediction Using Machine Learning Techniques

Abstract

Machine Learning is employed across several spheres round the world. Machine Leaning plays a vital role in predicting presence/absence of movement disorders like heart diseases and a lot of. During this era, Individual area unit terribly busy and dealing difficulty so as to satisfy their materialistic wants and unable to pay time for themselves that results in physical stress and mental disturbance. Thus, Cardiovascular disease is incredibly common today. Significantly in urban areas owing to excess mental stress. As a result, Cardiovascular disease has become one among the foremost vital factors for death of men and girls. Within the medical field, predicting the heart disease has become difficult task. So, during this modern life, there is immediate need of a system which can predict accurately the chance of obtaining cardiovascular disease. Predicting a cardiovascular disease in early stage can save several people's life. The most objective of this paper is to style a robust system that works expeditiously and can ready to predict the chance of getting heart disease accurately. Machine Learning (ML) has been showing a good help in creating selections and predictions from the massive amount of knowledge created by the aid industries and hospitals. The predictions model is projected with combos of various options and a numbers of classification techniques. We'll be predicting heart diseases by using machine learning algorithms. The algorithms we'll be using like K Nearest Neighbors Classifier, Support Vector Machine, Decision Tree and Random Forest. We'll analyze prediction systems for cardiovascular disease employing a bigger variety of input attributes. The system uses medical terms like Sex, Age, Chest Pain, Cholesterol level, etc. attributes to predict the probability of patient obtaining a cardiovascular disease.

Keywords- Machine Learning, Heart Disease, Dataset, Decision Tree, Random Forest.

I. INTRODUCTION

One of the leading causes of death in globe is heart disease. According to World Health Organization research, heart disease is responsible for one out of every three fatalities worldwide [1]. The Heart is an important organ of human body It pumps blood into parts of our body. If heart doesn't work properly, the brain and various other organs can shut down and person can die within minutes. The early prediction of these kinds of disease is very important so that precaution could be taken before situation becomes more critical.

When modern technology and specialists are not available, diagnosing and treating heart disease is very difficult. Cardiovascular disease recognized by symptoms such as high blood pressure, chest ache, high cholesterol level, discomfort, difficulty in breath and so on. There are mainly two types of risk factors which are responsible for heart diseases. One category is those factors which can't be controlled such as family history, human age and gender. Another category includes those factors which are responsible for heart disease and can be controlled. Risk factors such as smoking and drinking liquor habits can be controlled [6]. Heart disease is caused by other factors also, including birth abnormalities, diabetes,

medications, and alcohol [5]. Nowadays, there are several automated methods like data processing, machine learning, deep learning, etc. for identifying diseases like cardiovascular disease. Machine Learning, a subfield of artificial intelligence, can learn from massive datasets and predict similarly previously unseen or new data based on its methods of learning or training [8]. Machine learning is like our brain, where all the learning takes place, just like we learn from their mistakes [2]. So, there is some set of training data, which we have taken from Kaggle. Machines are trained using this dataset and then creates model which takes inputs from user and make predictions.

Currently, we have a large amount of data provided by patient's electronic health records. Technology has also provided us with many methods, techniques, and models that enable data scientists and researchers to contribute to medical development. Through analytics, the data can determine the causes of the disease and the medical team's contribution by spreading awareness through prevention [13]. The heart disease can be detected by many ways, in which angiography is the most common method to detect heart diseases. However, the angiography method has some advantages. This is so expensive operation and doctors must consider many factors when diagnosing patients, which makes the doctor's job extremely difficult too [6]. These types of shortcomings encourage researchers to develop a confined method for predicting heart disease. So, there is a need to develop an automated system that can detect heart diseases on the basis of various human medical factors.

II. LITERATURE SURVEY

Xiaoming Yuan et al. [11] uses Machine Learning and Internet of Medical Things to create a model for prediction of heart disease. They first designed a Fuzzy-GBDT (gradient boosting tree) algorithm to reduce data complexity and increase the generalization of binary classification. Then, they integrated Fuzzy-GBDT with bagging to avoid overfitting. After evaluation, they got excellent accuracy and stability in both binary and multiple classification predictions. In paper [12], an integrated machine learning framework MaLCaDD (Machine Learning based Cardiovascular Disease Diagnosis) is proposed in which data balancing, feature selection and classification are targeted together for the improved and early prediction of heart disease. They achieved improved prediction accuracy through the ensemble of Logistic Regression and KNN classifiers.

Deepak Kumar Chohan and Dinesh C Dobhal [1] have used the Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naïve Bayes, Random Forest and KNN algorithms to predict Cardiovascular disease, in which Random Forest algorithm gives 98.53% accurate results which is highest among all other algorithms used. In [14] author used XGBoost algorithm, to train and evaluate models. The author presents a novel procedure to accurately detect heart diseases in real-time from the analysis of short single-lead ECGs (9-61 seconds). In [13], author created a hybrid of five models including Logistic Regression, Support Vector Machine, k-Nearest Neighbors(KNN), Decision Tree and Random Forest to classify and predict cardiovascular disease. And they got 98.18% accuracy with Random Forest by using voting ensemble technique.

Chunyan Guo et al. [15] are using the Recursion Enhanced Random Forest with an improved linear model to observe heart condition. And also planning an Artificial Neural Network with

feature choice and backpropagation learning technique for classification of disorder. In [2] Akanksha Kumari and Ashok Kumar Mehta have tried to predict cardiovascular disease using seven machine learning algorithms and tried to enhance the accuracy of weak performing algorithms using ensemble ways like AdaBoost and Voting Ensemble methodology. In [7] Mohammed Nowshad Ruhani et al. have trained their model victimization classification algorithms like Logistic Regression, Decision Tree, K-Nearest Neighbors(KNN), Naïve Bayes, Support Vector Machine, etc. although accuracy for various algorithms changes for a distinct variety of instances within the dataset, SVM shows the best performance by getting accuracy of 91%. Rather than gathering data from any online repository like Kaggle, UCI, etc. they collected dataset manually from numerous Medical Institutions. In [3] Mihir J. Gaikwad et al. developed a model to forecast cardiovascular disease using five ML algorithms area unit applied (Support Vector Machine, Random Forest, Gradient Boosting, Supply Regression and Decision Tree Classifier). The prediction of every compared to see that one is best suited to the prediction. D. P. Yadav et al. [4] have developed exploitation machine learning and have optimisation technique to help a doctor. In [5] Likitha KN et al. analyse numerous machine learning ways for predicting internal organ standing area unit gift. They applied Machine Learning algorithms and compared supported the characteristics like age, chest ache, vital sign (BP), sex, steroid alcohol and heartbeat.

Narendra Mohan et al. [6] try and do cardiovascular disease prediction, they used python and pandas activities. During this planned work, dataset is to start with divided into getting ready and testing information sets. They used four machine learning models KNN, NB, LR and RF for predict the disease in flesh on the idea of some medical parameters. In [8], Ahmed Al Ahdal et al. used six machine learning algorithms for detecting heart disease. [9] M. Snehith Raja et al. developed a reliable cardiopathy prediction system which enforced sturdy machine learning algorithmic program that is random forest algorithmic program, which gives correct result in less time. [10] Yu Lin analysed a heart disease dataset which is taken from Cleveland. Within the method of model training, six machine learning algorithms Logistic Regerssion, K-nearest Neighbors, Adaboost, CART, Random Forest, XGBoost were applied. Random Forest was the best model the surpassed the remainder of the models with outstanding score of accuracy 84.40%.

In [16], Ashir Javeed et al. highlighted the matter of overfitting within the recently planned ways for heart disease prediction and planned a unique learning system to facilitate the centre failure prediction. The created models overfit to the testing data. In order to come up with associate intelligent system that may show sensible performance on each training and testing data, author developed a unique diagnostic system. The proposed method uses random search algorithm and random forest algorithm for cardiovascular disease prediction. Senthilkumar Mohan et al. [17], proposed a method that improves the accuracy of the prediction of cardiovascular disease using machine learning techniques. They got an accuracy level of 88.7% by their prediction model. The proposed hybrid HRFLM method combines the features of Random Forest and Linear Method. In [18], Norma Latif Fitriyani et al. proposes a cardiovascular disease prediction model which consists of Density – Based Spatial Clustering of Applications with Noise (DBSCAN) to find and exclude the outliers and a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) used to balance the unbalanced training dataset. They used Extreme Gradient Boosting (XGBoost) algorithm to predict the heart disease.

III. DATASET

The dataset we'll be using for training the ML model is taken from the Kaggle, which contains 1025 records of patients along with 14 distinctive attributes, like age, sex, etc. Descriptions of the attributes are shown in Table I.

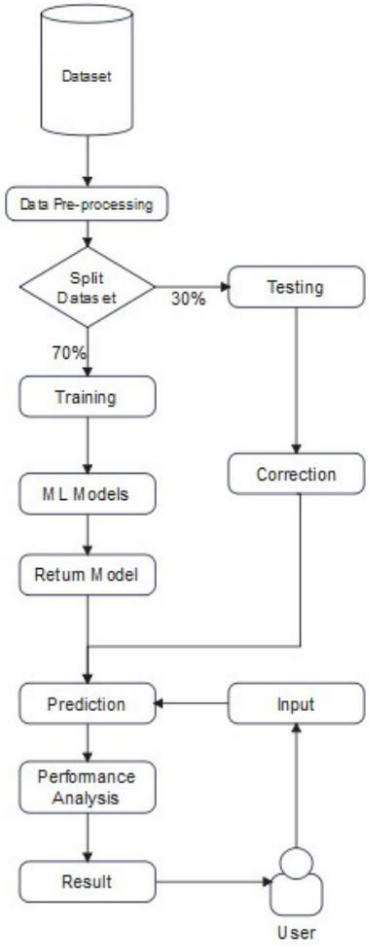
Attributes	Description	Type
age	Age of patients in years	Numeric
sex	Sex of patient: • 0 = female • 1=male	Categoric
cp	Type of chest pain: • 0 = typical angina • 1 = atypical angina • 2 = non-anginal pain • 3 = asymptomatic	Categoric
trestbps	Resting blood pressure in millimeters of mercury (mm Hg) on admission to the hospital	Numeric
chol	Serum cholesterol level in mg/dl	Numeric
fbs	Fasting blood sugar > 120 mg/dl: • 0 = no • 1 = yes	Categoric
restecg	Results of resting electrocardiogram: • 0 = normal • 1 = having ST-T wave abnormality – T wave inversions and/or ST elevation or depression of > 0.05mV • 2 = showing probable of definite left ventricular hypertrophy by Estes' criteria	Categoric
thalach	Maximum heart rate achieved	Numeric
exang	Exercise-induced angina: • 0 = no • 1 = yes	Categoric
oldpeak	ST depression induced by exercise relative to rest	Numeric
slope	The slope induced by exercise ST segment: • 0 = upsloping • 1 = flat • 2 = downsloping	Categoric
Ca	Number of major vessels (0-3) colored by fluoroscopy	Categoric
thal	The results of thallium stress test: • 1 = normal • 2 = fixed defect • 3 = reversible defect	Categoric
target	Have heart disease: • 0 = no • 1 = yes	Categoric

Table- 1:

I. PROPOSED METHODOLOGY

Our proposed model for prediction of heart disease is works in very simple and easy way and it also very easy to use for users. In figure 1, we can see that first we are collecting our dataset in which some previous record of patients is available. Table 1 shows the detailed information with attribute of our dataset which is taken from Kaggle. After collecting dataset, we are processing it to see if there is no missing value present in dataset. If there is any missing value then we remove those record from our dataset and remaining records are used in pre-processing. The multi-class variable is used to see the presence or absence of cardiovascular disease. In case of patient having heart disease, the value is set to 1, otherwise value is set to 0 indicating there is no heart disease in the patient. The pre-processing of data is converting the medical records int diagnosis value. After data pre-processing we are splitting the data into two different parts: one part we use for training of our model and another part we use for testing of our model in the ratio of 70% and 30% respectively.

After training and testing we choose our model on basis of the model which gives the correct result with highest accuracy. Then after completion of model user just need to put input of each attributes, then our model will be able to give result with a best accuracy if user have heart disease or not.



II. MACHINE LEARNING ALGORITHM FOR DETECTING HEART DISEASE

The dataset taken form Kaggle, it is first pre-processed with records to check whether there are any missing or irrelevant values are present. If it's available, then these values will be deleted and replaced with right values, before pass through the classifiers to process and calculate the estimated accuracy. From among achieved results, the classifier which gives the highest accuracy will be acceptable and evaluate with the test data.

A. Heart Disease Prediction

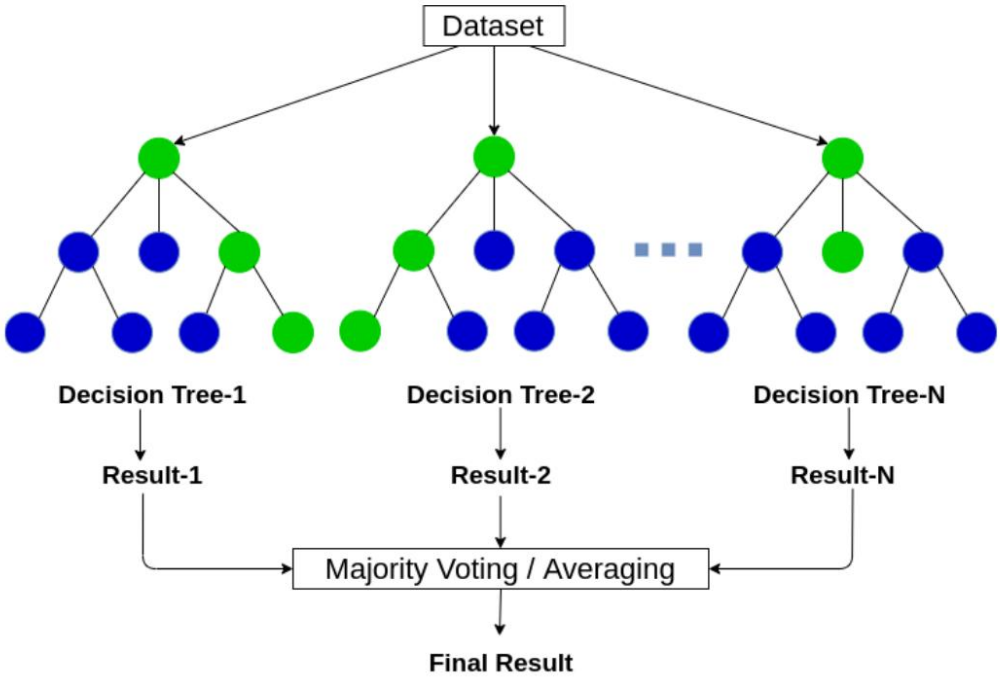
The primary goal of our model is to create a model which gives accurate result with high accuracy, which will be accomplished through use of various algorithms. A huge amount of data is generated in the medical industry, and it is very useful to use those data in early disease prediction.

B. Techniques for treating Heart Disease

The attributes of dataset such as age, sex, cholesterol level, etc. are classified using KNN, SVM, DT and RF approach. The input dataset split into two parts: training dataset and testing dataset with amount of 70% and 30% of data respectively. Then performance of trained model is evaluated using testing dataset.

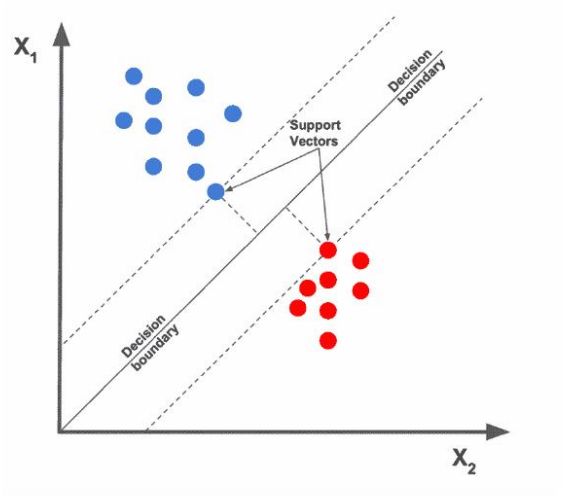
(i) Random Forest (RF):

A Random Forest contains multiple decision trees for subset of the dataset, and calculate average to improve the accuracy of the model. More number of trees gives the highest accuracy of the algorithm.



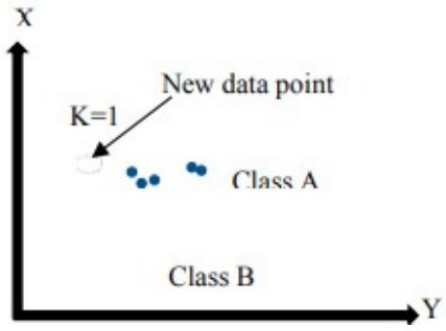
(ii) Support Vector Machine (SVM):

Like DT, SVM can also solve both classification and regression problems. But, SVM is mainly used for classification. The SVM algorithm creates a line or a hyperplane which separates the data into classes. Since we are building our model in the medical data field, the dataset can be non-linear. So, Support Vector Machine can be a good option.



(iii) K- Nearest Neighbor (KNN):

KNN can be also used for both classification and regression. KNN is very simple algorithm, its working rule is based on each data point which is labelled as neighbor. This method consists of locating the nearest k data points in the training set towards the data point in which a target value is missing and allocating the average value of the obtained data points to it.



Confusion Matrix:

It is used to determine the performance of classification models. It gives a detailed chart of the actual and anticipated outcome. The frequency of correct and incorrect predictions is represented by (n × n) matrix.

	Actual: NO	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

- True Negative (TN):MModel has predicted the disease No, and in real the person is not suffering from heart disease.
- True Positive (TP):MModel has predicted the disease Yes, and in real the person is suffering from heart disease.
- False Negative (FN):MThe model has predicted the disease No, but in real the person is suffering from heart disease.M
- False Positive (FP):MThe model has predicted the disease Yes, but in real the person is not suffering from heart disease.M

Classification Accuracy:MIt defines the how much our model predicts the result correctly. It is the ratio of total true prediction to total prediction.

$$\text{Accuracy } ¿ \frac{TP+TN}{TP+TN+FP+FN}$$

Error rate: It defines the how much our model predicts the result incorrectly. It is the ratio of total false prediction to total prediction.

$$\text{Error Rate } ¿ \frac{FP+FN}{TP+TN+FP+FN}$$

Precision: It defines the accuracy of positive prediction. It is the ratio of actual true prediction to total positive prediction.

$$\text{Precision } ¿ \frac{TP}{TP+FP}$$

Recall: The percentage of accurate results that are correctly classified is shown by recall. It is the ratio of true prediction to overall positives.

$$\text{Recall } ¿ \frac{TP}{TP+FN}$$

V. CONCLUSION

In this study, we highlighted the problem and challenges for identifying cardiovascular disease like how difficult and complicated for a doctor to identify heart disease and also as the prospective of patient, the test for identifying a heart disease is expensive. In this research, we proposed an automated system for predicting heart disease using machine learning algorithms. We collected the dataset from Kaggle which contain health record of 1025 patients with 14 attributes. These attributes have been used to train and classify using ML algorithms like K Nearest Neighbors Classifier, Support Vector Machine, Decision Tree and Random Forest. In future, we can develop a web application where people can identify if they have any heart disease or not by giving input to the model manually.

References