Prediction of Heart Diseases Using Machine Learning Algorithms

ABSTRACT

Heart diseases have become one of the most common causes of a huge number of deaths all over the world and has emerged as the most life-threatening disease. This project describes the need of a reliable detection system that would help people to detect the disease in an early stage and start the appropriate treatment on time. Using machine learning algorithms by various researchers and applying them to large and complex medical datasets is helping the healthcare industry and the medical professionals in detecting the heart diseases. There are various models proposed in the past based on Supervised machine learning algorithms such as Logistic Regression, SVM, KNN, Naive Bayes, Decision trees and Random Forest and these models have been used and tested for their accuracy in the detection of the disease.

VISUAL ABSTRACT

Our project was focused on finding the answers to the three questions that we had in mind and the first one was identifying the risk factors, identifying the frequency of disease based on gender and identifying the survival rate based on age and gender and for that we wanted to see that how our predictor ‘Y’ gets affected by the 13 features(X) given in the dataset and it can be

seen from the correlation matrix that there was no single feature that was highly correlated with our 'target' value. The same can be seen in Figure 1.

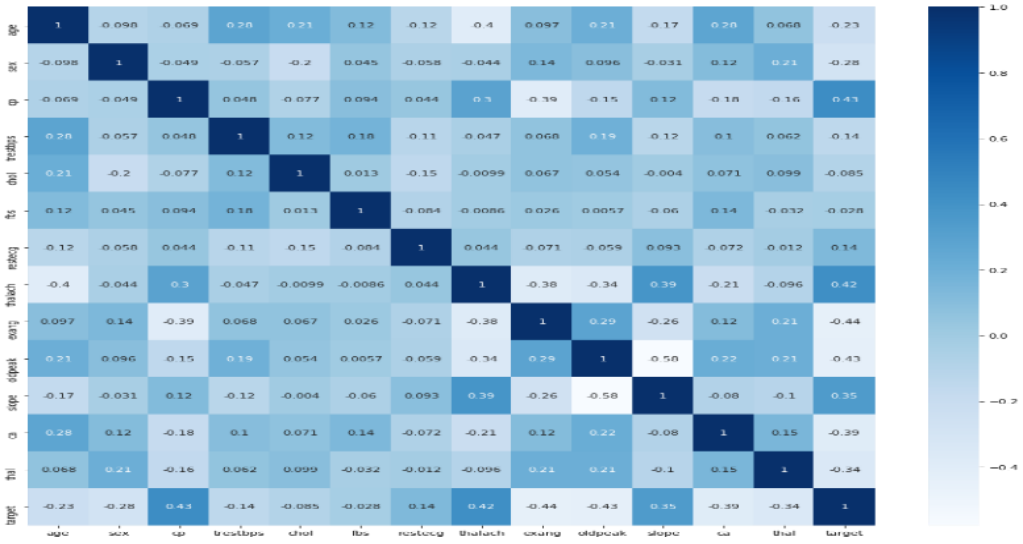

Figure 1: Correlation Matrix

We also found the total number of negative and positive cases of Heart Diseases are approximately balanced, with 45% of observations having heart disease and the remaining population not having heart disease. Age and gender do play an important role in identifying whether a person will have a heart disease or not. In our experiment the dataset consists of people with ages between 29 to 77 years and we found out that the mean age for the persons with heart disease was around 54 years and the risk of getting it increases with the increase in age and, we could identify that females are more prone to heart diseases than males and this also has a scientific reason as Women have smaller arteries than men, so heart disease develops differently, and more diffusely. Another risk factor that increases the chances of getting a heart disease are ST depression which is the slope of the peak exercise ST segment and is induced due to exercising. Figure 2 shows the total no of positive and negative cases of heart disease and figure 3 shows the impact of age on heart disease. It was also observed that Resting blood

pressure tends to increase with age regardless of heart disease and maximum heart rates tend to be lower for people without heart disease.

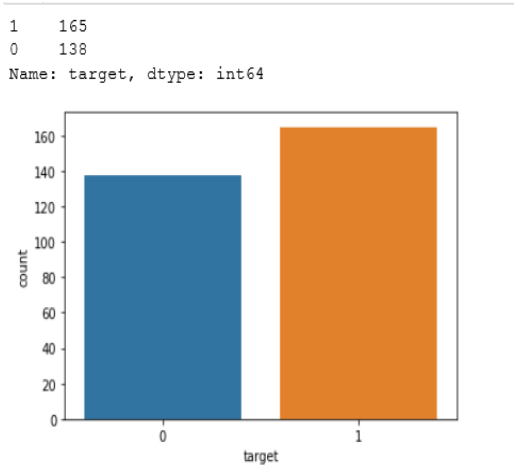

Figure 2: Total no of positive and negative cases

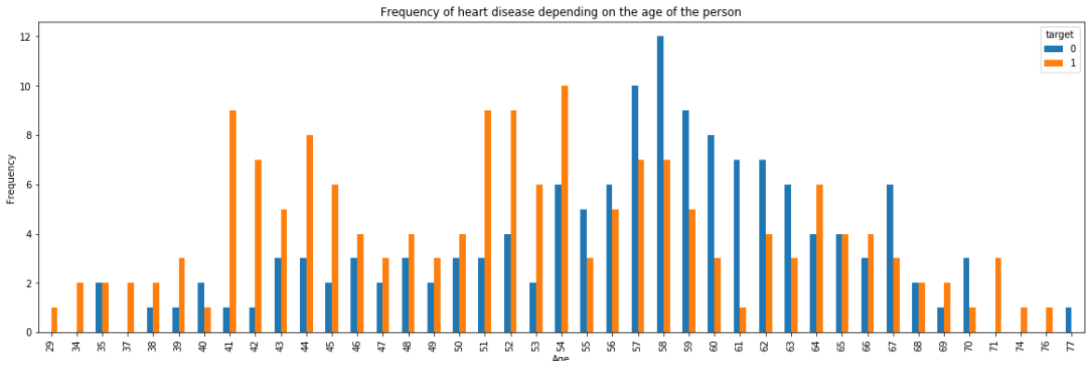

Figure 3: Frequency of heart disease depending on the age

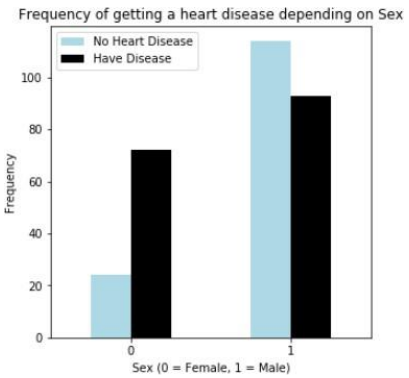

Figure 4: Frequency of getting a heart disease based on gender

INTRODUCTION

Heart diseases have become one of the most common reasons of deaths in the past few years and there are many factors that are responsible for an individual's heart health such as age, food habits, smoking, alcohol consumption, physical activity, and the type of job that the person does as stress is also a crucial factor. Machine learning is an emerging application of AI that uses different analytics and statistical techniques to improve the performance of a particular machine learning from past data. It enables a particular machine to learn from database and enhance the performance by experience. Machine learning has been successful in solving various complex problems. There are two types of ML algorithms namely "Supervised" and "Unsupervised". The Supervised ML algorithms deals with labelled datasets and the Unsupervised ML algorithm deals with the unlabelled dataset. This project is based on the supervised classification model. Classification is a supervised machine learning technique. It is a 2- step process, first step is called learning step where the model is constructed and trained by a predetermined dataset with class labels (training set) and second step is classification (testing) step where the model is used to predict class labels for given data (test data) to estimate the accuracy of classifier model. In this project we have used Logistic Regression, Naïve Bayes, KNN, Random Forest, Decision Trees and SVM algorithms. The aim of this project is to develop a machine learning model that would help in predicting the heart diseases and saving lives.

LITERATURE REVIEW

There have been a lot of experiments that have been conducted using the UCI dataset and machine learning algorithms to predict the heart diseases. Some of the related work is described in this section.

Intelligent Heart Disease Prediction System Using Data Mining Techniques

Authors: Sellappan Palaniappan, Rafiah Awang

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined"to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network.

Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design

Authors: M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur andKushal Ghosh

An Android based prototype software has been developed by integrating clinical data obtained from patients admitted with IHD (Ishchemic Heart Disease).The clinical data from 787 patients has been analyzed and correlated with the risk factors like Hypertension, Diabetes, Dyslipidemia (Abnormal cholesterol), Smoking, Family History, Obesity, Stress and existing clinical symptom which may suggest underlying non detected IHD. The data was mined with data mining technology and a score is generated. Risks are classified into low, medium, and high for IHD.

Intelligent heart disease prediction system using random forest and evolutionary approach

Authors: Akhil, Jabbar & Deekshatulu, Bulusu & Chandra, Priti.

M. Akhil jabbar, B.LDeekshatulu, and Priti Chandra propose a new algorithm that combines KNN and Genetic Algorithm for efficient classification. Optimal Solution Perform a global search for complex large and multiple modal data sets to provide genetic algorithms. It is also observed from the results that hybridization with KNN is well performed and provides great accuracy. Ankita Dewan proposed an efficient genetic algorithm mix using backpropagation for cardiac disease prediction. They concluded that neural networks are the best of all classification techniques for nonlinear data. The BP algorithm is the best classifier of the Artificial Neural Network, a common training method. Where the primary system output is compared to the desired output and the system is adjusted until the difference between the two is minimized. However, there is a drawback to being trapped in local minima.

EXECUTIVE SUMMARY

This project predicts the heart disease by exploring the algorithms namely Logistic Regression, SVM, KNN, Naive Bayes Decision trees, Random Forest. The data was divided into train and test set and We found out the accuracies of all the algorithms by applying these on our test data and recorded the accuracies. The hyperparameter tuning of KNN, SVM, Decision Trees and Random forest was also done to check and compare the different accuracies with changing parameter values. The highest accuracy of 88.59% was obtained with Logistic Regression and the total misclassified cases were less in case of LR as compared to other algorithms in the experiment. The accuracy above 70% is good and all the other models showed the accuracy more than 70% in this experiment so we can say that all the above-mentioned algorithms are appropriate for predicting the heart diseases though they can still be given priority based on the predictions of False Positives and False Negative outcomes. Table 1 shows the results of accuracies of the algorithms.

	Classifier	TP	FP	TN	FN	Precision	Recall	F1_score	Accuracy	Specificity	Sensitivity
1	Nearest Neighbors	63	10	44	9	0.86301	0.875	0.86897	0.84921	0.81482	0.875
2	Linear SVM	67	14	40	5	0.82716	0.93056	0.87582	0.84921	0.74074	0.93056
3	Decision Tree	62	11	43	10	0.84932	0.86111	0.85517	0.83333	0.7963	0.86111
4	Random Forest	54	9	45	18	0.85714	0.75	0.8	0.78571	0.83333	0.75
5	Naive Bayes	57	10	44	15	0.85075	0.79167	0.82014	0.80159	0.81482	0.79167
6	Logistic Regression	66	11	43	6	0.85714	0.91667	0.88591	0.86508	0.7963	0.91667

Table 1: Accuracy report of the six algorithms

SELCTED METHOD

Machine learning is the best approach in terms of predicting heart diseases as it allows the use of intelligent methods across different datasets to reveal useful insights. Since our experiment involves training a model based on historical dataset, ML seems to be an appropriate technology for our experiment. Our approach was to try different classifiers and try and compare which one of those works better for the given dataset of heart disease. There are several independent variables such as age, gender, resting electrocardiography were used along with a dependent variable “target” during the training phase to build the classification models. These models are then employed to forecast the dependent variable value in test dataset as accurately as possible and the accuracy was recorded for all the models.

DATA PRE-PROCESSING

The dataset has been curated from Kaggle and per the description given on Kaggle the original dataset had 76 attributes, but we have used 14 attributes that includes one “Target” variable here as the previous experiments have also been done on only 14

attributes and the desired results have been achieved. Our data is in the following formats and the features have been classified according to the data type in Table 1:

ATTRIBUTE	TYPE	DESCRIPTION
AGE	CONTINUOUS	AGE IN YEARS
SEX	BINARY	1=MALE, 0=FEMALE
CHEST PAIN(CP)	ORDINAL	VALUES:0,1,2,3,4
TRESTBPS	CONTINUOUS	RESTING BLOOD PRESSURE (IN MM HG)
CHOL	CONTINUOUS	SERUM CHOLESTROL IN MG/DL
FBS	BINARY	FASTING BLOOD SUGAR>120 MG/DL 1= TRUE; 0=FALSE
RESTECG	ORDINAL	RESTING ELECTROCARDIOGRAPHIC RESULTS (VALUES 0,1,2)
THALACH	CONTINUOUS	MAXIMUM HEART RATE ACHIEVED
EXANG	BINARY	EXERCISE INDUCED ANGINA(1=YES;0=NO)
OLDPEAK	CONTINUOUS	ST DEPRESSION INDUCED BY EXERCISE RELATIVE TO REST
SLOPE	ORDINAL	THE SLOPE OF THE PEAK EXERCISE ST SEGMENT (VALUES 0,1,2,3,4)
CA	ORDINAL	NUMBER OF MAJOR VESSELS (0-4) COLORED BY FLOUROSOPY
THAL	ORDINAL	NATURE OF DEFECT, VALUES (0-3)
TARGET	BINARY	PRESENCE OR ABSENCE OF HEART DISEASE, VALUES (1,0)

Table 2: UCI dataset description based on Data Types.

For data to be processed and cleaned it is always better to visualize and understand the data properly so that we are aware of all the attributes. The pre-processing is done to make the data noise free and to remove any null values that might affect the results. Our dataset is in different formats as discussed earlier hence to make it more understandable we applied standardization, normalization techniques on the data. The data was divided into two parts, training 60% and testing 40% and again into three parts, for core model training 60%, Hyperparameter tuning 20% and testing 20% for tuning the parameters of SVM, KNN, Random Forest and Decision Tree classifiers.

ALGORITHMS

As discussed above we have used six machine learning models for predicting the heart disease in a patient and to analyse up to optimal performance among all. The short description of each model explained in this section.

LOGISTIC REGRESSION

While generating our model the threshold was 0.5 which means that for probability values greater than 0.5, the model predicts the dependent variable to be 1 and for values less than or equal to 0.5, the model predicts the dependent variable to be 0.

DECISION TREE

Decision Tree is tree-like structure that classifies instances by sorting them based on the values of the variables. Each node in a decision tree represents a variable in an instance to be classified, and each branch represents a value that the node can assume. In the experiment the max depth of decision tress was changed 10 times for tuning and the accuracy results for all of them were obtained. We selected max depth 3 as it provided the maximum accuracy and the minimum TN and FN cases.

RANDOM FOREST

The final predictions of the random forest are made by averaging the predictions of each individual tree, which enhances the prediction accuracy for unseen data. The hyperparameter tuning was performed for the estimators of the random tree and for the final model the parameter that gave the best accuracy was chosen which was 60.

SUPPORT VECTOR MACHINES

For our model, the c value of 0.001 was selected for the SVM as it showed good accuracy and minimum FP and FN cases after performing the tuning at different c values.

K NEAREST NEIGHBOR

For our model, the hyperparameter tuning was done for the present modelling, the whole process is repeated fifteen times each with a k value starting from 1 to 15 and then the highest performing parameter was selected which in this case was k=7.

NAÏVE BAYES

A Naive Bayes classifier assumes that the presence or absence of a particular attribute of a class is independent to the presence or absence of any other attribute of that class. It is often used to compute posterior probabilities of given observations and make decisions on higher probability.

DISCUSSION OF SELECTED METHOD RESULTS

The UCI dataset had 165 positive cases and 138 negative cases of heart disease and the same can be seen below in the Figure 2. It was also noticed that none of the variable were highly correlated with each or with the 'target 'variable. Logistic regression came out as the best performing algorithm for our experiment with the accuracy of 88.59%. Logistic regression misclassified total 17 False Negatives and False Positives out of 126 total cases which is good as compared to the other algorithms that were used in this experiment. Though all the algorithms showed good accuracy which was above 70 % but we can say that based on prediction of FN and FP Logistic regression is a better model in this experiment. The following formulas were used for Recall, Precision, Accuracy and F1 score:

Recall = TP/TP+FN, Precision=TP/TP+FP, Accuracy=TP+TN/TP+TN+FP+FN

F1 Score= 2*(Precision*Recall/Precision + Recall)

(Where TP =True Negatives, TN=True Negatives, FP= False Positives, FN=False Negatives)

MODEL SELECTION CRITERIA

The model selection was done based on the comparison of accuracy results of all the six models and after analysing the FN and FP cases as in the health industry this is an important aspect when selecting a model. Since all models tend to show some error because of the noise in the

data or the incompleteness of the data hence it is important to try different models and decide based on the accuracy percentage. In our case Logistic Regression was selected as the best model for Heart disease prediction. Logistic Model is also a good choice as it is easy to understand and fast to train.

MODEL COMPARISON (TRAINING, TUNING, VALIDATION)

For the training purpose the model was split into test and training dataset in 60 and 40 ratio using train test split method. It is an important validation technique to test our model's accuracy when it is provided with an unknown data. Also, for the hyperparameter tuning the data was split int 60,20 and 20 ratios for train, test, and tuning. The Hyperparameter tuning was performed on KNN, SVM, Random Forest and Decision Trees to improve their performance with different parameter values. The KNN was tested with different neighbour values, Decision tree with different depths, Random forest with different estimator values and SVM with different optimisation(c) values and then the then the best parameter was selected based on their F1 score and then the models were tested again and finally the model that gave the best accuracy along with the lowest number of misclassifications was selected for the prediction of heart diseases. All the six models showed accuracy above 70% but the difference can be seen in the prediction FN and FP cases as that matters a lot when you are dealing with the diseases and it is always a good option to keep the counts as low as possible and hence the logistic regression was selected out of these 6 models as our best model for prediction of heart diseases.

CONCLUSIONS

Based on our experiment we can say that Machine learning is the future of healthcare industry in detecting the diseases and facilitating the task of healthcare workers in providing best care. All the above-mentioned algorithms performed well in terms of accuracy but could not perform well in terms of FN and FP cases which cannot be ignored. Logistic Regression was selected

as our final model because of the good F1 score and less misclassified cases. We found out that age is an important factor as with age the risk of getting a heart disease increases and if the trace in the ST segment is abnormally low below the baseline, this can lead to a Heart Disease. Our experiment also showed that the females are prone to heart diseases as compared to Males and hence the frequency of survival rate greatly depends on age and gender.

REFERENCES

S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.

Raihan, M. (2016). Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design. 10.1109/ICCITECHN.2016.7860213.

Akhil, Jabbar & Deekshatulu, Bulusu & Chandra, Priti. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. journal of network and innovative computing. 4. 175-184.

Sonam Nikhar, A.M. Karandikar."Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Advanced Engineering, Management and Science (ISSN: 2454-1311), vol.2, no. 6, pp.617-621,2016.

https://www.kaggle.com/ronitf/heart-disease-uci