International Research Journal of Computer Science (IRJCS)
Issue 04, Volume 6 (April 2019)

ISSN: 2393-9842
www.irjcs.com

PREDICTING HEART DISEASE USING MACHINE LEARNING TECHNIQUES

D.Raghunath Kumar Babu
Computer Science & Engineering,
JNTUA College of Engineering, Pulivendula, INDIA
raghunath.d29@gmail.com

C.Usha Sree
Computer Science & Engineering,
JNTUA College of Engineering, Pulivendula, INDIA
ushasree.chowdam@gmail.com

K.Veera Vidya
Computer Science & Engineering,
JNTUA College of Engineering, Pulivendula, INDIA
reddyveeravidya@gmail.com

V.Manoj Kumar
Computer Science & Engineering,
JNTUA College of Engineering, Pulivendula, INDIA
varikutimanojkumar@gmail.com

Manuscript History
Number: **IRJCS/RS/Vol.06/Issue04/APCS10092**
Received: 04, April 2019
Final Correction: 10, April 2019
Final Accepted: 18, April 2019
Published: April 2019
Citation: D.Raghunath, C.Usha, K.Veera & V.Manoj (2019). PREDICTING HEART DISEASE USING MACHINE LEARNING TECHNIQUES. IRJCS:: International Research Journal of Computer Science, Volume VI, 149-153. **doi://10.26562/IRJCS.2019.APCS10092**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract: As heart disease is the number one killer in the world today, it is becoming one of the most difficult disease to diagnose the state of disease. If a heart disease is diagnosed early, many lives can be saved. Machine learning classification techniques can significantly benefits the medical field by providing an accurate, unambiguous and quick diagnosis of diseases. Hence, save time for both doctors and patients for prediction. We start by over viewing the machine learning & describing in brief definitions of the most commonly used classification techniques to diagnose heart disease. We used different attributes which can relate to this heart disease well to find the better method to predict and we also used classification algorithms for prediction.

Key words: Machine learning classification techniques; heart disease; Doctor; Patient; Accuracy; F-Score;

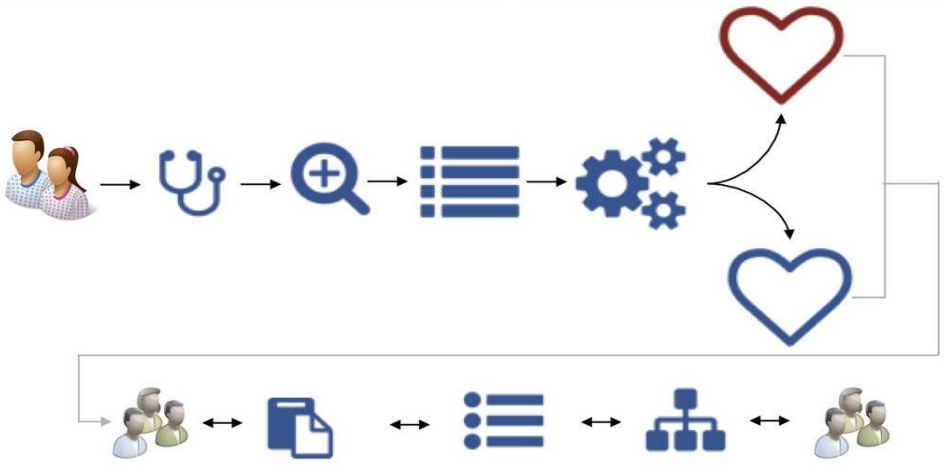
I. INTRODUCTION

Heart disease is a major cause of morbidity and mortality in the modern society. Medical diagnosis is an important but complicated task that should be performed accurately & efficiently. All doctors are unfortunately not equally skilled in every subject specialty and they are in many places a scarce resource. A system for automated medical diagnosis would enhance medical care and reduce costs and it is useful for poor people. Nowadays, diseases are increasing day by day due to life style, hereditary. Especially, heart disease has become more common now-a-days, i.e. life of people is at risk. Among these are poor diet, lack of regular exercise, tobacco smoking, alcohol or drug abuse, and high stress, Obesity, High Blood Pressure (BP), Hereditary. It is very hard for humans to derive useful information from very large amounts of data, that is why Machine Learning (ML) is widely used to analyze and process these data and diagnose problems in the healthcare by providing solutions to the problems with reducing diagnosis time and increases accuracy and efficiency. Heart diseases are the leading cause of death globally, resulted in 12.3 million (25.8%) in 1990 to an increase to 17.9 million deaths (32.1%) in 2015. It is estimated that 90% of disease is preventable using diagnosis techniques. There are many risk factors for heart diseases that we will take a closer look at. Machine Learning (ML) will help us discover different patterns and provides beneficial information from them. The main objective of this study is to find out and build the suitable machine learning (ML) technique that is computationally efficient as well as accurate for the prediction of heart disease occurrence, based on a combination of features like risk factors describing the disease. Different machine learning classification techniques will be implemented and compared upon standard performance metric such as accuracy and F-Score.

II. LITERATURE SURVEY

Heart disease is a term that assigns to a very large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today's world. This paper aims at analyzing the various Machine Learning (ML) techniques introduced in recent years for heart disease prediction. Whereas previous papers has been built using Data Mining Techniques. In some papers this is given that they have been used only one technique for diagnosis of heart disease as given in Shadab et al, Carlos et al etc. but in case of other research works more than one data mining techniques are used for the diagnosis of heart disease.

III. ARCHIECTURE



3.1 DATA ANALYSIS

The objective of data analysis step is to increase the understanding of the problem from the data. There are two approaches to describe a given dataset. Summarizing and Visualizing data.

Data Exploration:

This dataset is publicly available to users for analyzing of data. In this data set, I have used 14 attributes and around 500 trained and test data to evaluate accuracy and F-Score. Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0).There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have been inserted manually.

Features Information:

- **age** - age in year
- **sex** - sex(1 = male; 0 = female)
- **chest_pain** - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- **blood_pressure** - resting blood pressure (in mm Hg on admission to the hospital)
- **serum_cholestoral** - serum cholesterol in mg/dl
- **fasting_blood_sugar** - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- **electrocardiographic** - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
- **max_heart_rate** - maximum heart rate achieved
- **induced_angina** - exercise induced angina (1 = yes; 0 = no)
- **ST_depression** - ST depression induced by exercise relative to rest
- **slope** - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 =downsloping)
- **no_of_vessels** - number of major vessels (0-3) colored by fluoroscopy
- **thal** - 3 = normal; 6 = fixed defect; 7 = reversible defect
- **diagnosis** - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

Types of features:

1. Categorical features (Has two or more categories and each value in that feature can be categorized by them): sex, chest_pain

2. Ordinal features (Variables having relative ordering or sorting between the values): fasting_blood_sugar, electrocardiographic,no_of_vessels, thal, diagnosis, , induced_angina, slope.

3. Continuous features (Variable taking values between any two points or between the minimum or maximum values in the feature column): age, max_heart_rate, ST_depression, blood_pressure, serum_cholestoral

IV. ALGORITHMS AND TECHNIQUES

1. K-Nearest Neighbors
2. Decision Trees
3. Logistic Regression
4. Gaussian Naïve Bayes
5. Support Vector Machines
6. Random Forests

1. K-Nearest Neighbors (KNN): K-Nearest Neighbors algorithm is a non-parametric method used for regression & classification. The principle behind nearest neighbor method is to find a predefined number of training samples closest in distance to the new point & predict label from known label.

Advantages:

- The K-Nearest Neighbor Classifier is a very simple classifier that works well on basic recognition problems.

Disadvantages:

- The main disadvantage occurred in this KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data & simply uses training data itself for classification.
- To predict the label of a new instance the KNN algorithm will find K closest neighbors to the new instance from the training data, the predicted class label will then be set as the common label among the K closest neighboring points.
- The algorithm must compute the distance between labels and sort all the training data at each prediction, which can be slow if there are a large number of training examples.
- This algorithm doesn't learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data.

2. Decision Trees: Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

Advantages: Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical & categorical data.

Disadvantages: Decision tree can create a complex tree that doesn't generalize well, and these trees can be unstable because small differences in the data might result in a completely different tree being generated.

3. Logistic Regression: Logistic regression is a machine learning(ML) algorithm for classification. In this algorithm, the probabilities describing the attainable outcomes of a single trial are modeled using a logistic function.

Advantages: Logistic regression is designed for classification purpose, and it is most useful for understanding the impact of several independent variables on a single outcome variable.

Disadvantages: Works only when the predicted variable is binary, assumes all predictors are independent of each other, and assumes data is free from missing values.

4. Naïve Bayes: Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as spam filtering and document classification.

Advantages: This algorithm requires a little amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely rapid compared to more sophisticated methods.

Disadvantages: Naive Bayes is known to be a bad estimator.

5. Support Vector Machine: Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into the same space & predicted to which category they belong to based on which side of the gap they fall.

Advantages: Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

Disadvantages: The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

6. Random Forest: Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Advantages: Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Disadvantages: Slow real time prediction, complex algorithm and difficult to implement.

V. METRICS

Accuracy:
It is the number of correct predictions made by the model over all kinds of predictions made

$$\text{Accuracy} = (TP+TN) / (TP+FP+FN+TN)$$

Accuracy is a good measure when the target variable classes in the data are nearly balanced.

F-score:
It is used to measure a test's accuracy and it balances the use of precision and recall to do it. It can provide a realistic measure of test's performance.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

The following posts will provide some methods to evaluate the performance of a machine learning problem:

- https://towardsdatascience.com/choosing-the-right-metric-for-evaluating-machine-learningmodels-part-2-86d5649a5428
- https://towardsdatascience.com/choosing-the-right-metric-for-machine-learning-models-part-1a99d7d7414e4

5.1 PREDICTION

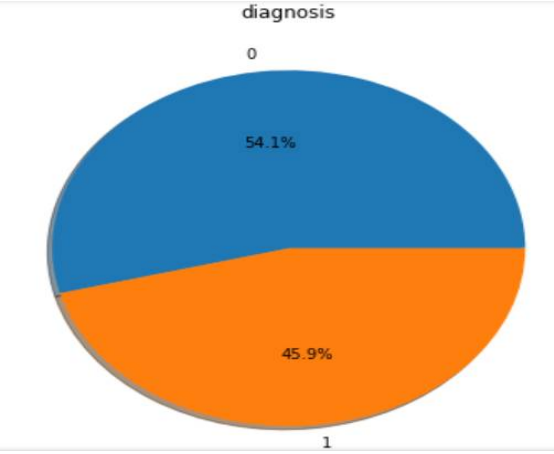
Out of the chosen algorithms we will start with KNN classification model. We will take a classifier and fit the training data. After that we will predict that by using predict (X_train). Now we will predict the accuracy of the testing data by using accuracy score (y_test, pred) and F-score by importing fbeta_score from sklearn.metrics. By doing so for, the KNN will give us the accuracy of 0.834.We will continue the same procedure on Naïve Bayes, SVM, Decision tree, Logistic Regression and Random Forest. By following the same procedure above that is fitting, predicting and finding the accuracy score and F-score we will get the accuracy score and F-score as below.

	Accuracy	F-Score
KNN	0.8344	0.8
Decision Tree	0.8344	0.9
Logistic Regression	0.827	0.79
Naïve Bayes	0.8211	0.78
SVM	0.8476	0.82
Random Forests	0.9139	0.92

From the above reports Random Forest seems to be performing well based on Accuracy and F-Score.

VI. CONCLUSION

The goal of the project was to compare different machine learning algorithms and predict if a certain person, given various personal characteristics and symptoms, will get heart disease or not. Here are the final results.



Now the distribution of target value is almost equal, so using standard metrics in further machine learning modeling like accuracy and AUC is justified. It does not come as a surprise that the more complex algorithms like SVM and Random Forests generated better results compared to the basic ones. It is worth to emphasize that in most cases hyper parameter tuning is essential to achieve robust results out of these techniques. By producing decent results, simpler methods proved to be useful as well. Machine learning (ML) has absolutely bright future in medical field. Just imagine a place where heart disease experts are not available. With just basic information about a certain patient's medical history, we may quite accurately predict whether a disease will occur or not.

	Test_Accuracy	Train_Accuracy
KNN	0.834437	0.906250
Decision Trees	0.834437	0.934659
Logistic Regression	0.827815	0.872159
Naive Bayes	0.821192	0.877841
SVM	0.847682	0.857955
Random Forests	0.913907	1.000000

REFERENCES

1. I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001. https://doi.org/10.1016/S0933-3657(01)00077-X.
2. J. Soni et al., "Intelligent and effective heart disease prediction system using weighted associative classifiers," International Journal on Computer Science and Engineering, vol. 3, no. 6, pp. 2385–2392, 2011.
3. N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26. https://doi.org/10.1145/3175684.3175703.
4. H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," International Journal of Computer Applications, vol. 168, no. 3, pp. 12–17, Jun 2017.

