Daffodil International University

Faculty of Information & Technology

Final Project Report

Title: Heart Disease Classification

Data Mining & Machine Learning Lab(CSE322)

Instructor:

Dr. Md Zahid Hasan

Associate Professor & Program Director MIS

Department of CSE

Submitted By:

Md. Shihab Mahmud Anik (202-15-3810)

Md. Raisul Islam (202-15-3813)

PC-A

Department of CSE

Date of Submission : 05-Dec, 2022

Contents

1 Abstract **2**

2 Introduction **3**

2.1 Objectives 3

2.2 Related Work 4

2.3 Problem Statement 5

3 Data-set Description **6**

3.1 Data Pre-processing & Feature Engineering 6

3.2 Data Visualization 7

3.3 Dataset Data Quality 8

4 Methodology & Investigation **9**

4.1 Logistic-regressions 9

4.2 K-nearest neighbors 9

4.3 Naive Bayes 10

4.4 Decision Tree 10

5 Result Analysis **11**

6 Conclusion **12**

1 Abstract

Machine learning utilizes artificial intelligence to handle several data science difficulties. Predicting outcomes based on known data is a popular use of machine learning. The system learns patterns from existing data sets and then applies them to a data set whose outcomes are uncertain. Classification is a potent machine learning approach that is often used for forecasting. Some categorization algorithms provide predictions with enough precision, whilst others demonstrate just limited precision. This work examines ensemble classification, a technique used to increase the accuracy of insufficient algorithms by integrating several classifications. Using a heart disease data set, experiments were conducted using this tool. Using a comparative analytical method, it was determined how ensemble approaches may be used to increase the accuracy of heart disease prediction. This research focuses not only on enhancing the accuracy of inaccurate classification systems, but also on applying the algorithm using a medical data set to show its use for early stage illness prediction. The findings of the research reveal that ensemble approaches like as bagging and boosting are efficient in enhancing the prediction accuracy of poor categorization and display good performance in identifying the risk of heart disease. The greatest accuracy for weak classifiers using ensemble classification was 7% .Adding feature selection to the procedure improves prediction accuracy.

2 Introduction

Cardiac failure is an umbrella phrase indicating abnormal heart function. Heart disease may be present at birth in certain cases. It is referred to as congenital heart disease. Heart disease that develops later in life is known as acquired heart disease. Most cardiovascular illnesses are acquired. Heart disease risk factors include age, sex, smoking, family history, cholesterol, a bad diet, high blood pressure, being overweight, not being active, and drinking too much alcohol. Cardiac disease. Some risk factors can be managed. In addition to the causes listed above, lifestyle factors such as food, inactivity, and obesity are also considered to be significant risk factors. About half of all Americans (47%) have at least one of the three major heart disease risk factors: high blood pressure, high cholesterol, and smoking. Some risk factors for heart disease, such as age and family history, cannot be managed. However, you may minimize your risk by modifying controllable variables. In general, a classification algorithm is a function that weights input information such that the output distinguishes between positive and negative values for one class. The most effective classification algorithms include Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Support Vector Machine. This study uses classification, a method of machine learning, to estimate heart disease risk based on risk variables. Using a method known as ensemble, it attempts to enhance the precision of forecasting heart disease risk.

2.1 Objectives

Supervised machine learning includes classification models. A categorization model takes in data and returns a label that indicates what category the data belongs to. Models can do binary classification, such as determining if an email is spam or not. Classification algorithms in machine learning provide predictions about the likelihood or probability that new data will fit into one of the established classes based on incoming training data. Each point's categorization is determined by how its k closest neighbors vote on its classification. It is supervised because it learns to label new points based on an existing set of labels. When deciding how to classify a new point, it considers the previously classified points that

are geographically adjacent to the target location. Algorithms designed specifically for classifying data are called classification algorithms, and their primary function is to make predictions based on the input data. The figure below illustrates how classification algorithms work.

2.2 Related Work

A heart or blood vessel condition. Tobacco use, high blood pressure, high cholesterol, a poor diet, a lack of exercise, and obesity are all risk factors for various heart diseases. Coronary artery disease (narrowed or blocked coronary arteries) is the most prevalent kind of heart disease, and it can cause chest discomfort, a heart attack, or a stroke. Congestive heart failure, irregular heartbeats, congenital heart disease (heart illness from birth), and endocarditis are some more heart disorders. Also known as cardiovascular disease.

They offer an ensemble-learning architecture comprised of two alternative neural network models, including the fundamental deep neural networks that have received the most attention for the binary classification challenge. For our heart disease binary classification challenge, they employed CNN, LSTM, GRU, BiLSTM, and BiGRU [1]. A technique called artificial neural network (ANN) can be used to predict or classify patients who will get heart disease [2]. A maximum increase of 7% accuracy for weak classifiers was achieved with the help of ensemble classification [3]. Use a hybrid categorization technique based on ReliefF and Rough Set (RFRS) to better identify cardiac conditions [4]. Additionally, employing the suggested optimal model using FCBF, PSO, and ACO, we were able to achieve a maximum classification accuracy of 99.65% [5]. Analyses of Data Mining's Potential Use in Treating Heart Disease Information will increase drastically in the near future as a result of current publishing trends and rising interest in the topic [6]. The accuracy is a measurement of the data model for finding the amount of correctly classified data using the input samples. Multiple technique used but proposed only decision tree.(fig-1)[7] .

Amongst the available classifiers, the decision tree has the highest accuracy (99.0%), followed closely by the Random forest [8]. Mortality for heart disease

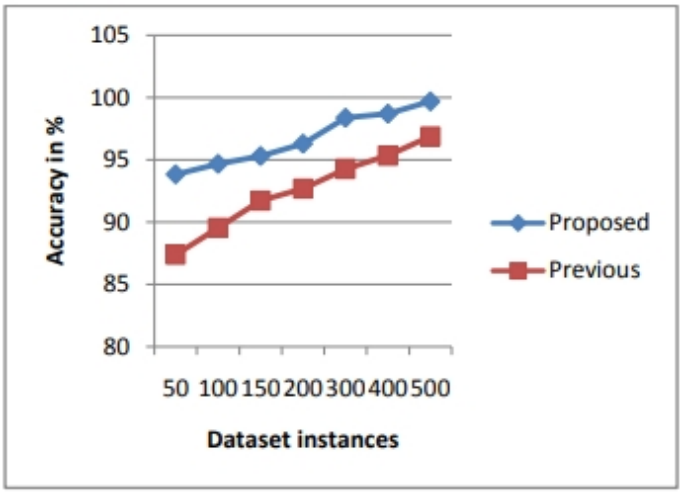

Figure 1: Accuracy [7]

is anticipated to be raised to reach 23.3 million by the year of 2030 where heart disease would stay to be the top cause of death for human [9]. On basis of best results the development of heart disease prediction system is done by using hybrid technique for classification associative rules (CARs) to achieve the prediction accuracy of 99.19% [10].

2.3 Problem Statement

Our problem basically classification. Because classification is about predicting a label and regression is about predicting a quantity. Heart disease features determine using level. If classification is about separating data into classes, prediction is about fitting a shape that gets as close to the data as possible. The object we're fitting is more of a skeleton that goes through one body of data instead of a fence that goes between separate bodies of data. The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/-category the new data will fall into.

3 Data-set Description

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge.

Data contains

1. age - age in years.
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - (fasting blood sugar $>$ 120 mg/dl) (1 = true; 0 = false)
7. restecg - resting electrocardiographic results
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest
11. slope - the slope of the peak exercise ST segment
12. ca - number of major vessels (0-3) colored by flourosopy
13. thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. target - have disease or not (1=yes, 0=no)

3.1 Data Pre-processing & Feature Engineering

Data Prepossessing first we check duplicate value then handle duplicate value by drop duplicate value. Secondly check null values, this data set hasn't any null values so there is no need to handle or drop these values.

For solving these classification problem We use fewer libraries for constructing the models. We use seaborn and matplotlib.pyplot for visualizing the dataset and constructing the algorithm to solve the classification models. We firstly check the

prediction of the targeted value and then get the accuracy of the value. After all these We consider the accuracy of the dataset confusion matrix and the test score.

3.2 Data Visualization

Firstly, We download the dataset from Kaggle. Here these dataset have 1025 rows and 14 features. Percentage of Patients Haven't Heart Disease: 48.68% .Percentage of Patients Have Heart Disease: 51.32%.The target value are has disease and not disease.Here we also see the ploting diagram of target column is;

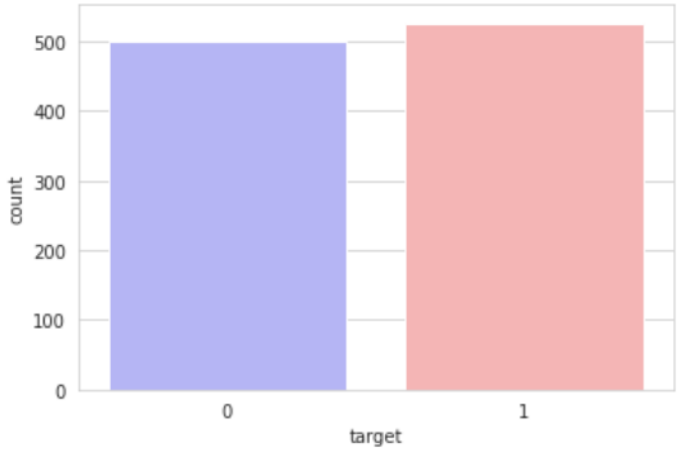

Figure 2: From these dataset we also see the Percentage of Female Patients: 30.44% and Percentage of Male Patients: 69.56%.

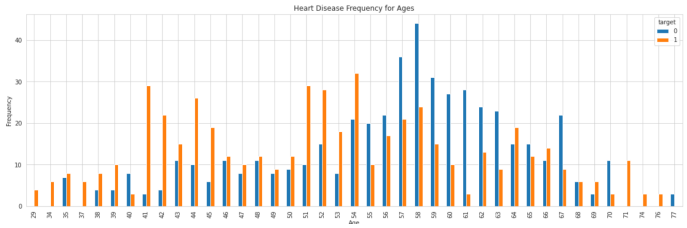

Figure 3: From these dataset We see the ploting diagram of frequency of variety age which can be disease or not disease.

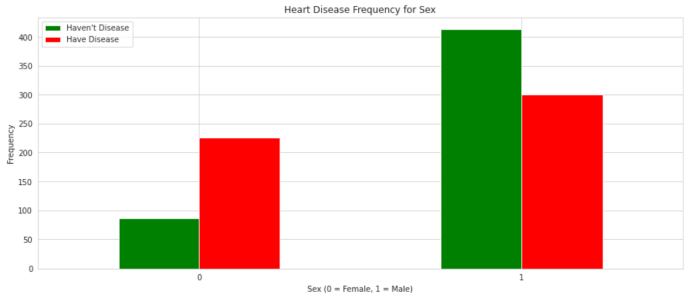

Figure 4: From these the bar chart diagram shows the frequency of disease or not disease for different sex. We say that Most of the cases the possibility of having disease would be held on the Male people.

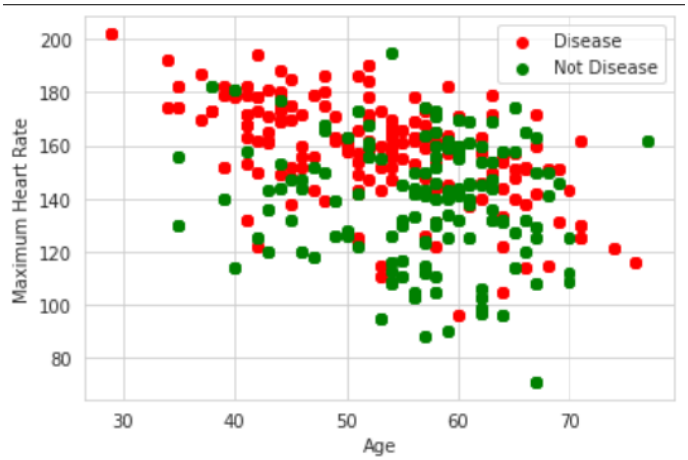

Figure 5: We see the clustering of the disease and not disease by the heart rate of any people.

3.3 Dataset Data Quality

This dataset used important data for prediction. There is no Null value, Duplicate value , Dummy value . We think data quality is good, so we expect better accuracy after using the algorithm.

4 Methodology & Investigation

Here, we use classification algorithm such as- Logistic-regressions, KNN, Naive-Bayes, Decision tree.

4.1 Logistic-regressions

Logistic regression is an example of learning with guidance. It is used to figure out or predict how likely it is that a binary (yes/no) event will happen.The categorical dependent variable is predicted using logistic regression. It is utilized, in other words, when the forecast is categorical, such as yes or no, true or false, 0 or 1. There is no middle ground between them when it comes to the projected probability or output of logistic regression. When making predictions regarding a categorical variable as opposed to a continuous one, logistic regression is used to evaluate the connection between a dependent variable and one or more independent variables.Logistic regression is more straightforward to apply and analyze, and it trains extremely quickly and efficiently. Over-fitting is possible with Logistic Regression if the number of observations is smaller than the number of features. It does not assume anything about how classes are spread out in the feature space.

4.2 K-nearest neighbors

Among the many machine learning algorithms, KNN is one of the simplest and most commonly used for categorization. The neighboring data point is used to determine the classification of this data point. New data points are sorted into categories by KNN using a similarity metric to old data. KNN uses distance measures to determine how similar each sample is to a given query, selects the K most similar instances, and then either uses a majority vote to choose a single label for a classification problem or takes an average of all the labels to get a decision. For example, if we have a data-set of tomatoes and bananas. KNN will store similar measures like shape and color. When a new object comes it will check its similarity with the color (red or yellow) and shape.

4.3 Naive Bayes

One kind of algorithm for doing classification work is called a naive Bayes classifier, and it bases its decisions on Bayes' theorem. Data points in a Naive Bayes classifier are assumed to be completely independent from one another. Spam filters, text analysis, and medical diagnosis are just a few of the many applications of naive Bayes classifiers.

Step 1: Determine the likelihood of classes given their labels.

Step 2: Determine the likelihood of each class given its attributes.

Step 3: We'll plug these numbers into the Bayes formula to get the posterior probability.

Step 4: If the input is classified as belonging to the higher probability class, then the next step is to determine which class has a greater likelihood.

4.4 Decision Tree

Data is continually separated according to a given parameter in a machine learning technique called a decision tree, which falls under the category of Supervised Machine Learning (where the input and corresponding output are explicitly defined in the training data). Decision nodes and leaves are the tree's explanatory building blocks.

Pseudocode :

i.Find the best attribute and place it on the root node of the tree.

ii.Then split the training set of the dataset into subsets. While making the subset make sure that each subset of training dataset should have the same value for an attribute.

5 Result Analysis

In this Classification problem, We have used 4 classification algorithms. These are Linear Regression, Naive Base, Decision Tree and K-nearest neighbor algorithm. For all of these algorithms we have used 30% data in the test-set and 70% data in the training set from our dataset. In the Logistic Regression Algorithm We use a logistic regression model and find the accuracy of the dataset is 80.98%. On the other hand using Naive Bayes Algorithm , we use Gaussian Naive Bayes theorem and get the accuracy of the dataset is 79.22%. After these using the Decision Tree Algorithm, We use the decision tree model library and find the score of the dataset which is 100%. From the decision tree algorithm We also see the graphical decision tree and it's very much helpful to conduct a decision from the dataset. Last of these, We use the K-Nearest Neighbor algorithm, If we chose the value of k is 1,then We find the maximum score of using KNN model is 100% and again when We chose the value of k is 23, then find the accuracy of the dataset is 71.72%. So, After above all of these, We can make the result by using the Decision Tree Algorithm and K-Nearest Neighbor Algorithm(for k=1) give the 100% accuracy.

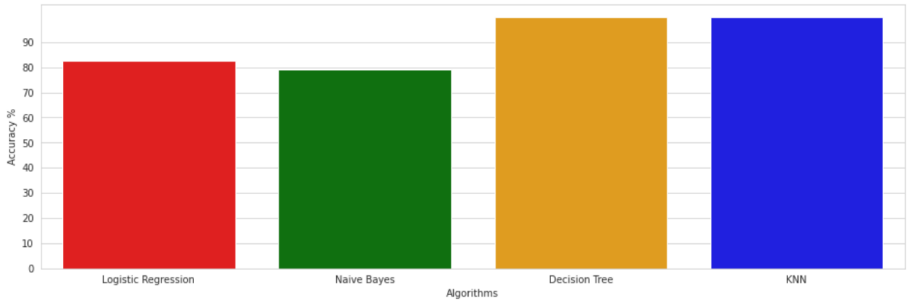

Figure 6: Accuracy of different Algorithms.

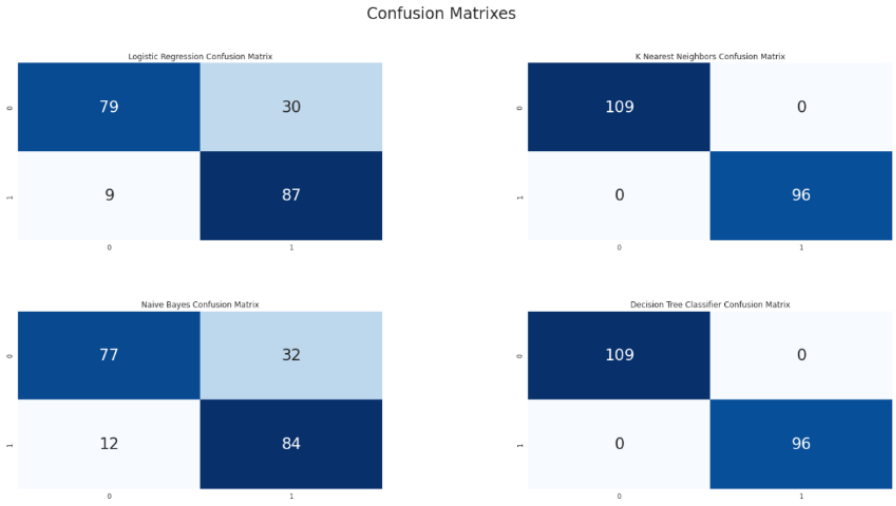

Figure 7: Here We also show the confusion matrixes for all of these using algorithms at a glance.

6 Conclusion

Heart Disease Classification mainly data mining related work. Data Mining task has two branches such as-Predictive Method another one Descriptive Method. Again predictive method has some types mainly classification, Regression, Deviation Detection. Here, we mainly use classification algorithm such as-Logistics Regression, Naive Bayes, KNN, Decision Tree classification. Here we maintain some step or rules, firstly data selection then data prepossessing again data transform to target transformed data. After transformed data we mining data and convert into some pattern. Finally, Evaluate data using algorithm and find or gain knowledge or output. We did following every step for this project. At the end of the project, We tried to compare the accuracy of all the algorithms used.

References

[1] Asma Baccouche, Begonya Garcia-Zapirain, Cristian Castillo Olea, and Adel Elmaghraby. Ensemble deep learning models for heart disease classification: A case study from mexico. *Information*, 11(4):207, 2020.

[2] M Aljanabi, Mahmoud H Qutqut, and Mohammad Hijjawi. Machine learning classification techniques for heart disease prediction: a review. *International Journal of Engineering & Technology*, 7(4):5373–5379, 2018.

[3] C Beulah Christalin Latha and S Carolin Jeeva. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16:100203, 2019.

[4] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, and Qian Wang. A hybrid classification system for heart disease diagnosis based on the rfrs method. *Computational and mathematical methods in medicine*, 2017, 2017.

[5] Youness Khourdifi and Mohamed Bahaj. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, 12(1):242–252, 2019.

[6] Gaurav Meena, Pradeep Singh Chauhan, and Ravi Raj Choudhary. Empirical study on classification of heart disease dataset-its prediction and mining. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pages 1041–1043. IEEE, 2017.

[7] Apurva Joshi, J Dangra, and M Rawat. A decision tree based classification technique for accurate heart disease classification and prediction. *Int J Technol Res Manag*, 3:1–4, 2016.

[8] Israa Ahmed Zriqat, Ahmad Mousa Altamimi, and Mohammad Azzeh. A comparative study for predicting heart diseases using data mining classification methods. *arXiv preprint arXiv:1704.02799*, 2017.

[9] M Haider Abu Yazid, Muhammad Haikal Satria, Shukor Talib, and Novi Azman. Artificial neural network parameter tuning framework for heart disease classification. In *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 674–679. IEEE, 2018.

[10] Jagdeep Singh, Amit Kamra, and Harbhag Singh. Prediction of heart diseases using associative classification. In *2016 5th International conference on wireless networks and embedded systems (WECON)*, pages 1–7. IEEE, 2016.