EAST WEST UNIVERSITY

Project Report

Project Name: Cardiovascular disease prediction

Course Code: CSE475

Course Title: Machine Learning

Section: 02

Spring-2021

Submitted By:

Samia Sultana

2017-2-60-122

Md. Samiul Islam

2017-2-60-047

Md. Munam Kazi

2017-2-60-142

Md. Monjurul Arif

2017-3-60-007

Date of Submission: 26 May, 2021

1. **Introduction:** According to the World health organization (WHO) Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year. It is rapidly increasing day by day. So it is very important to diagnose the disease with full accuracy. In our project we want to predict whether any person has a cardiovascular disease or not by giving some medical features of the patient. If the prediction is accurate enough then it will significantly help both the patient and the hospitals. Moreover it will also help the patient to take necessary steps to cure the disease. Through our project, if we implement machine learning to predict cardiovascular disease it will also save the human resources and reduce the complicated diagnosis procedure. We are using 11 medical features of a patient to predict cardiovascular disease.

1.1 Objectives:

The main objectives of this project are:

- To implement machine learning algorithms like decision trees, random forest, knn (k nearest neighbors), SVM to predict cardiovascular disease and contribute to the medical side.
- To analyse the risk factors which have significance to the medical dataset values which may cause heart disease.
- We will analyse the medical feature selection methods.

1.2 Motivation: Nowadays machine learning has a vast effect on the technology which is helping in many spheres of our lives. It is being used in many data science applications. The prime motivation for this project is to explore the medical features of the disease and to help to create a system that can easily give results related to the disease. We will also train the dataset and process the dataset for the best results. We want to develop a computer based system by implementing machine learning algorithms which can aid to the field of medicine.

1.3 Existing works: We did a research on this topic before we tried our experiments and found a number of related works on this topic. We learned lot from those and had a clear vision to continue our work in this topic.

In 2011, Ujma Ansari [1] made use of Decision Tree model to predict heart disease and get a high accuracy of 99%, which inspires us to use a better version of Decision Tree and it is Random Forest. Unfortunately, the paper uses a dataset with 3000 instances but does not provide a reference of how they get the data. The UCI website only provides 303 instances of dataset so we doubt where the author gets 3000 instances of dataset.

[2]In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent

1.4 Necessity: Heart disease is a collective term for conditions that affect the heart. Heart disease often leads to serious cardiovascular events such as heart attacks and stroke. It has been observed to be the leading cause of death worldwide in both men and women. If we don't identify the problem at the right time, the chances of death are much higher. Our project will play an important role in identifying the problem of hard. If you tell the machine some symptoms, the machine will tell you if you have a hard disease or not. If you know you have a heart problem then you can follow the doctor's advice and you can stay saved. If we implement this project, people will be able to work very fast to solve their hard problems. Doctors will be able to give the right treatment at the right time with the help of this project. With this project, we are trying to help to detect the problem very quickly. With the right treatment at the right time doctors can give proper treatment to the patients. As a result, the death rate of people will be decreased that's why we feel the necessity to develop this project.

2. **Methodology:** To develop the project we have approached some machine learning algorithms such as decision tree algorithm, random forest, SVM (support vector machine).

Decision Tree:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules

inferred from prior data (training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Steps:

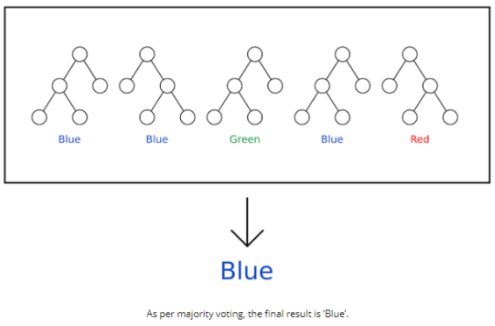
1. It begins with the original set S as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.
3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
4. The set S is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

Gain (T,X)=Entropy(T)-Entropy(T,X)

Entropy, $H(X) = -\sum_x p(x) \log p(x)$

Random Forest:

Random forest is a supervised ensemble learning algorithm that is used for both classifications as well as regression problems. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees mean more robust forest. Similarly, the random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method that is better than a single decision tree because it reduces the over-fitting by averaging the result.

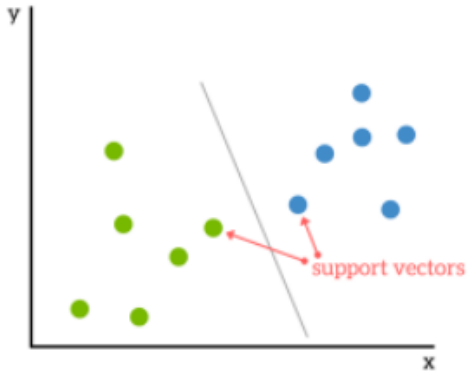


Random Forest creation pseudo code:

1. Randomly select “k” features from total “m” features where k << m
2. Among the “k” features, calculate the node “d” using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the 1 to 3 steps until “l” number of nodes has been reached
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees. Class value inherited from most trees.

Support Vector Machine:

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems. SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes. For a classification task with only two features it can be thought of a hyper plane as a line that linearly separates and classifies a set of data. So when new testing data is added, whatever side of the hyper plane it lands will decide the class that we assign to it.



3. Implementation: To implement the project we have approached with the data collection, data processing, data training and testing.

3.1 Data Collection: We have collected our dataset for the cardiovascular disease from online (kaggle.com) platform which has 700000 records of patient data with 11 features and 1 target.

Data description

There are 3 types of input features:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All the data values are in numerical form. There are no categorical values in our dataset.

age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
18393	2	168	62.0	110	80	1	1	0	0	1	0
20228	1	156	85.0	140	90	3	1	0	0	1	1
18857	1	165	64.0	130	70	3	1	0	0	0	1
17623	2	169	82.0	150	100	1	1	0	0	1	1
17474	1	156	56.0	100	60	1	1	0	0	0	0

3.2 Data Processing: We have processed the data from some of the attributes of the dataset in categories.

Data processing categories:

Age: In the dataset age is given in days, we have converted it in years and categorized in three age group. age1, age2, age3. Then we have converted these into numerical category.

BMI: we used the height and weight attribute and converted it into BMI which is measured as a medical feature of the patient.

```
df['bmi'] = (df['weight'] / (((df['height']/100)**2))).round(decimals=2)
```

BMI is categorized in 6 groups.

BMI	Weight Status
Below 18.5	underweight
18.5-24.9	Normal weight
25.0-29.9	overweight

30.0-34.9	Obesity I
35.0-39.9	Obesity II
Above 40	Obesity III

We have taken numerical values, underweight=0, Normal weight=1, overweight=2, Obesity I =3, Obesity II=4, Obesity III=5

Systolic blood pressure | Examination Feature, ap_hi :

We have categorized it in 4 groups in terms of risk factors: normal, high1, high2, emergency.

Then we have converted these into numerical category.

Diastolic blood pressure | Examination Feature, ap_lo :

We have categorized it in 4 groups in terms of risk factors: normal, high1, high2,

emergency. Then we have converted these into numerical category.

Finally to get more accuracy we have emitted unwanted data attributes which has no effect on the results.

3.3 Model Development: Our project is developed using python programming language. We have used the jupyter notebook environment to compile the project.

3.4 Results: Based on the classifiers we have used, the accuracy results are given below,

Methods	Accuracy
Decision tree	0.708641

SVM	0.721888
Random forest	0.714135

4. **Conclusions:** we used some library factions to implement the project. After all the implementations we got the accuracy from the classifiers we used which are decision tree, support vector machine, random forest. We got the best accuracy from support vector machine (SVM) which is 72.3%. there is a limited property of the dataset which is why we can get less accuracy. Moreover to improve the accuracy we can improve the dataset which can give the best results. The data attributes can also be effected in recent times for the patients. There are some limitations like we could not implement more classifiers to predict the result because of the time constraint. However, the project can be improved with large dataset.

References:

[1]. Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48

[2] . M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms"