International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT20659

Cardiovascular Disease Prediction Using Machine Learning

Digvijay Kumar
Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Article Info
Volume 6, Issue 5
Page Number : 46-54
Publication Issue :
September-October-2020

Article History
Accepted : 01 Sep 2020
Published : 12 Sep 2020

ABSTRACT

Heart-related diseases or Cardiovascular Diseases (CVDs) are the most common and main reasons for a huge number of deaths in the world, not only in India but in the whole world. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. This research paper represents the various models based on such algorithms and techniques to analyze their performance. Such as Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and ensemble models which are Supervised Learning algorithms. Using various important features that are necessary for the prediction of CVDs (like a person is having CVDs or not), which we will further discuss in this paper.

Keywords : Heart-related diseases, Cardiovascular Diseases, Support Vector Machines, K-Nearest Neighbors

I. INTRODUCTION

Cardiovascular disease (CVD) is the most common disease that involves the heart and blood vessels. CVD includes coronary artery disease (CAD) such as heart attack, hypertensive heart disease, abnormal heart rhythms, peripheral artery disease.

The factors which affect the CVDs mostly are such as high blood pressure, body mass index (BMI), physical activity, diet, gender, genetics, sleep, pollution.

Blood Pressure Stages

Blood Pressure Category	Systolic mm Hg (upper #)		Diastolic mm Hg (lower #)
Normal	less than 120	and	less than 80
Elevated	120-129	and	less than 80
High Blood Pressure (Hypertension) Stage 1	130-139	or	80-89
High Blood Pressure (Hypertension) Stage 2	140 or higher	or	90 or higher
Hypertensive Crisis (Seek Emergency Care)	higher than 180	and/or	higher than 120

The main topic is the prediction using machine learning technics. Our dataset consists of 70,000 records of patient data in 12 features, such as age, gender, systolic blood pressure, diastolic blood pressure, etc.

The target class "cardio" equals 1, when the patient has cardiovascular disease, and 0, if a patient is healthy.

Our task is to predict the presence or absence of cardiovascular disease (CVD) using the patient examination results.

II. LITERATURE REVIEW

[1]Benan AKCA, uses Random forest, KNN, Naïve Bayes, SVM, Decision Tree, and neural network algorithms and analyzed the medical dataset. There is less number of features involved. So, there is a need to extract the number of features. This can be done by feature selection and extraction. He also made use of Artificial neural networks (ANN).

[2]Svetlana Ulianova, She just only analyze the cardiovascular data, in which she did data cleaning, feature extraction, and shows the various graphs and plot to visualize the medical data.

[3]Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna uses Naïve Bayes, K-means, ID3 algorithm using a Decision tree. In which they found the best accuracy from the Naïve Bayes algorithm, they also said that in future Naïve Bayes and the k-means algorithm can also be combined to gain maximum accuracy. In the Data Preprocessing part, they have done data cleaning, transformation, integration, and feature reduction.

[4]V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja uses a technique like Naïve Bayes, K-nearest neighbors, Decision tree, SVM, Random forest and Ensemble model analyze and predict the medical data. In their research they achieve very good accuracy in all the machine learning techniques, but very poor in Decision tree classifiers may be due to overfitting of data.

[5]Mohie Eldin has done data visualization, feature engineering, data cleaning and model comparison explanation like Random Forest, Decision tree, XGB classifier and KNN algorithm to perform prediction on cardiovascular data, in which he got maximum accuracy or XGB classifier around 73.5% by hyper tuning the parameter by ensuring there is no overfitting. Also, he effectively visualizes the data very well.

III. Methodology

1 Dataset Description:

All of the dataset values were collected at the moment of medical examination. I have retrieved this dataset from Kaggle (Cardiovascular disease dataset). All the data is collected from various countries' top contributors are as - INDIA, UNITED STATES, BRAZIL, RUSSIA.

The dataset consists of 70 000 records of patients data in 12 features, such as age, gender, systolic blood pressure, diastolic blood pressure, etc. The target class "cardio" equals 1, when the patient has cardiovascular disease, and it's 0 if the patient is healthy.

Features:

1. **Age | Objective Feature |**
   age in int (days)
2. **Height | Objective Feature |**
   height in int (cm)
3. **Weight | Objective Feature |**
   weight in float (kg)
4. **Gender | Objective Feature |**
   gender in categorical code | 1: female, 2: male
5. **Systolic blood pressure | Examination Feature |**
   ap_.hi in int
6. **Diastolic blood pressure | Examination Feature |**
   ap_lo in 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 | 0: non-smoker, 1: smoker
10. **Alcohol intake | Subjective Feature |**
    alco in binary | 0: non-alcoholic, 1:alcoholic
11. **Physical activity | Subjective Feature |**
    active | binary | 0: inactive, 1: active

12. **Presence or absence of cardiovascular disease | Target Variable |**

cardio | binary | 0: absent, 1: present

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	52	1	165	64.0	130	70	3	1	0	0	0	1
3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	48	1	156	56.0	100	60	1	1	0	0	0	0

2 Data preprocessing and Dimensionality Reduction

Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset. Large numbers of input features can cause poor performance for machine learning algorithms, so it becomes necessary to reduce the dimensionality of features and variables. Dimensionality reduction methods include feature selection, feature extraction, feature reduction.

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models, otherwise, it may lead to overfitting or underfitting of data, which may cause poor performance in prediction. Data preprocessing methods include various encoding methods, data normalization, scaling, etc.

Reduction: When we work on some data it may be complex or it may be difficult to understand sometime, so to make them understandable for ourselves we will reduce them to the required format so that we can visualize them properly. Here the **age** given in days, this needs to be converted into years, so it may become helpful to visualize data properly. I have done so by just dividing the ages with 365 and replaced them.

Extraction: In extraction, a new set of features is derived from the original feature set. Feature extraction involves a transformation of the features. This transformation is often not reversible as few, or maybe many, useful information is lost in the process. In the given dataset I have extracted new feature BMI (body mass index) because BMI is a very important feature for the prediction as well as for the heart patient.

Formula used

BMI = weight / ((height/100)2)**

Cleaning: Data that we want to process may be possible is not clean that it may contain noise or values missing so we can't get good results, to obtain good results and accuracy we need to eliminate all this, the process to eliminate all these are called data cleaning.

After observing the dataset I have found that weight, height, ap_hi, ap_lo are very inconsistent for the respective lowest and higher values, which may lead to overfitting and make the prediction accuracy more verse. So I overcome this situation be removing these outliers.

Transformation: This involves changing data format to one form to other that is making them most understandable by doing normalization, smoothing, and generalization, aggregation techniques on data. In this dataset, a feature like age, height, weight, ap_hi, ap_lo, BMI needs to be transformed. Which I have done so by scaling method.

3 Data Analysis and observation

1. In this graph, I have shown the distribution of ages for the target class, which shows the count for a particular age. After observing the graph we can conclude that the count of people having heart disease is greater than healthy people especially from age **56 to 64**

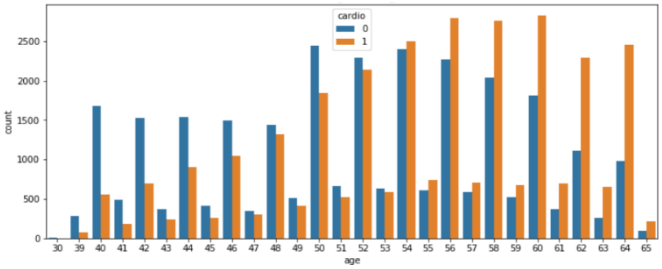


Fig 1. Age vs Count

2. In this graph, I have compared the total count of males and females which are healthy or having heart disease(CVD). We can observe that there is no major difference in people having a disease or not. But one thing is that count of the female is more than male

for both 0 and 1 values for target class (i.e. person having CVD not).

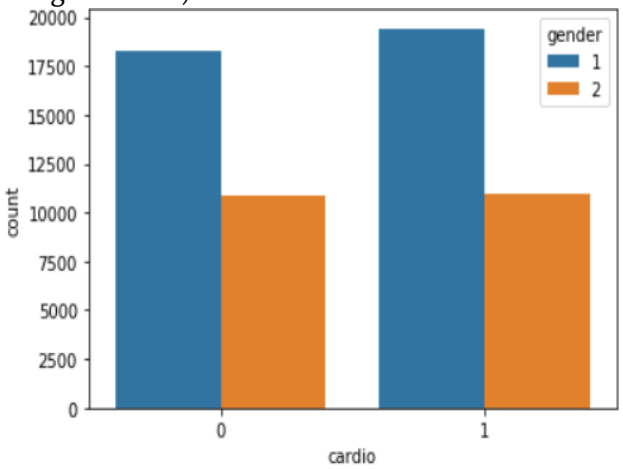


Fig 2. Cardio vs Count

3. In the below graph, it can be clearly seen that the patients with CVD have higher cholesterol and blood glucose level and, generally speaking, less active.

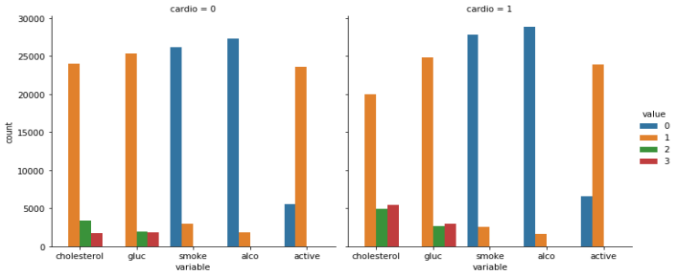


Fig 3. Variable vs Count

4. In the below graph, I have plotted this to observe the trend of the gender of a particular BMI for the target class (0 and 1). Although the trend is the same for both targets, from the graph I can clearly observe that count of people having CVD is greater after 32. Which means that large value of BMI is one of the major factor for people having heart related disease.

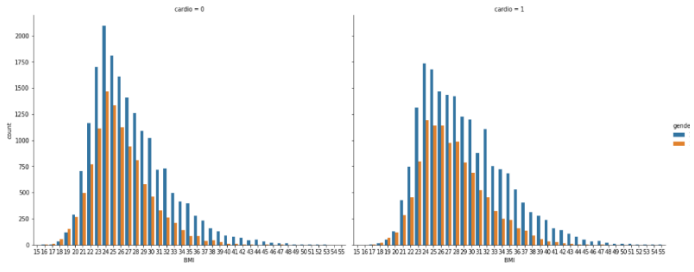


Fig 4. BMI vs gender

5. In this graph, I have compared the alcohol consumption of both males and females for the target class. But after visualizing the graph, we can say that there is no difference in alcohol consumption for both target class. But male persons drink more than females.

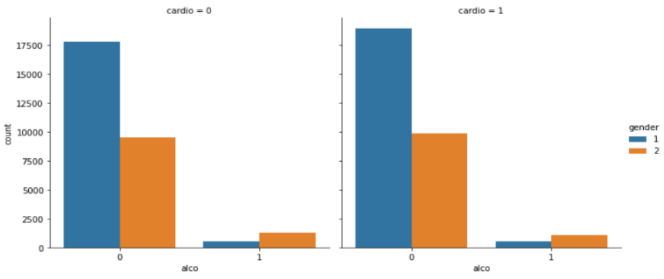


Fig 5. Alcohol vs gender

6. In the below graph, I have built the **Box plot**.

- *The bottom black* horizontal line of a blue box plot is a minimum value.
- *The first black* horizontal line of the rectangle shape of a box plot is First quartile or 25%
- *The second black* horizontal line of the rectangle shape of the box plot is the Second quartile or 50% or median.
- *The third black* horizontal line of the rectangle shape of the box plot is third quartile or 75%
- *The top black* horizontal line of the rectangle shape of the box plot is the maximum value.
- *The small diamond shape* of the box plot is outlier data or erroneous data.

In this graph, I have tried to show the comparison between Systolic pressure (ap_hi) and Diastolic pressure (ap_lo) for the target class. Which is the most important factor concerning heart disease. Here we can observe that those who are having target class equal to 1(cardio= 1) have greater value for both ap_hi and ap_lo.

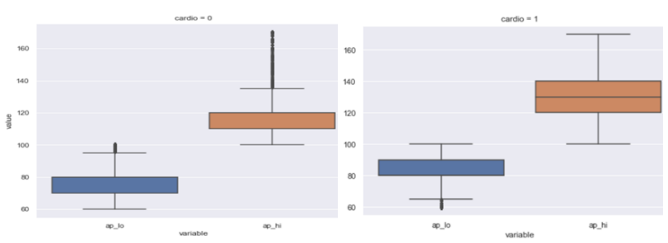


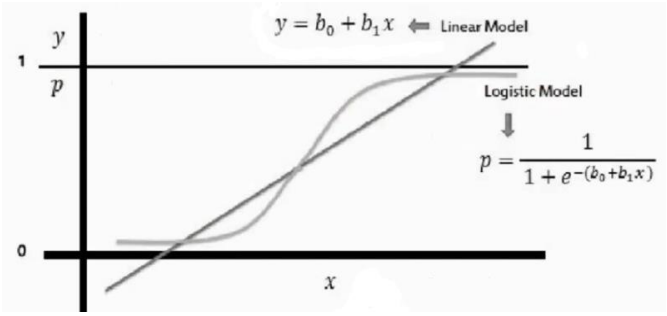
Fig 6. Blood pressure vs Cardio

Algorithms and Techniques used

1. Logistic Regression

Logistic regression is a fundamental classification technique. It belongs to the group of **linear classifier** somewhat similar to polynomial and **linear regression**. Logistic regression is fast and relatively uncomplicated, and it's convenient for us to interpret the result. Logistic regression models the probability that response falls into a specific category. A logistic regression model helps us solve, via the **sigmoid function**, for the situation where the output can take but only two values, 1 and 0. The logistic function is going to the key to using logistic regression to perform classification. We can take our linear regression solution and place it into the sigmoid function and it looks something like this

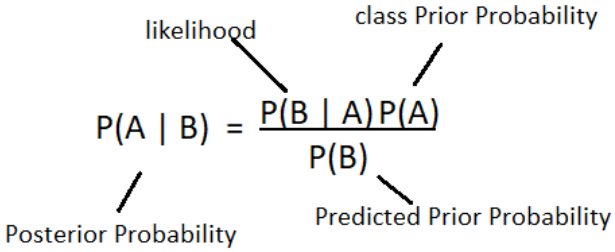
.



In Logistic regression, I got an accuracy of **71.91%** which is quite better than other algorithms. Benan AKCA got an accuracy of 72.38 which their best among all models, by implementing grid search and hyper tuning the parameter.

2. Naïve Bayes

Naive Bayes is a simple but effective classification technique which is based on the Bayes Theorem. It assumes independence among predictors, i.e., the attributes or features should be not correlated to one another or should not, in any way, be related to each other. Even if there is a dependency, still all these features or attributes independently contribute to the probability and that is why it is called Naïve.



In Naïve Bayes algorithm I got an accuracy of **70.70%** is not good concerning other algorithms. While Benan AKCA got an accuracy of 63.08%.

3. Support vector classifier

Support Vector Machine is an extremely popular supervised machine learning technique(having a pre-defined target variable) that can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they Fall.

In SVM I got an accuracy of **71.83%** for **linear** kernel, for kernel = 'rbf', gamma = .75 and c = 4 I got 70.94% and for kernel = 'poly', degree = 5, C=5 I got accuracy

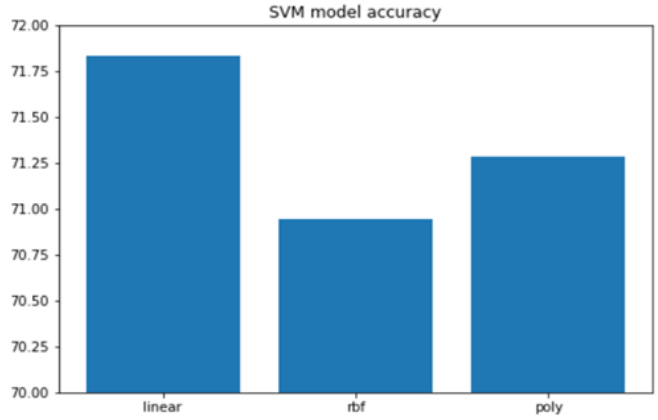


Fig 7. SVM model vs accuracy

4. RandomForestClassifier

Random Forest is also a popularly supervised machine learning algorithm. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, the Random Forest technique considers multiple decision trees before giving an output. So, it is an **ensemble** of decision trees. This technique is based on the belief that more trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality.

In this model, I got 68.46% which is worst among all models I have implemented.

5. Stacking

Stacking is also an Ensemble learning technique that uses prediction from multiple models (for example Logistic Regression, SVM, Naïve Bayes, KNN, Random forest) to build a new model. This model is used for predicting the test set.

In my model, I have implemented KNN, Random forest,naïve Bayes, and Logistic regression as a final estimator. Thus I have got 72.16% accuracy which is very good and second-highest for this dataset.

6.KNeighbors Classifier

K-Nearest Neighbour technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very little or no prior knowledge about the data distribution. To predict a new data point, the algorithm finds the closest data points in the training data set — its "nearest neighbors."

I have drowned a graph plot to calculate the best n-nearest neighbors, to make sure the algorithm perform it's best.

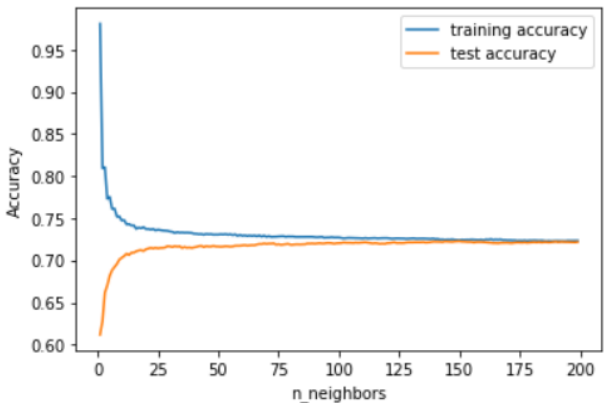


Fig 8. K-neighbors vs Accuracy

The above plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis. Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex. The best performance is somewhere around **150** neighbors.

In this model, I have got 72.28% which is **best** among all models I have implemented.

Accuracy comparison

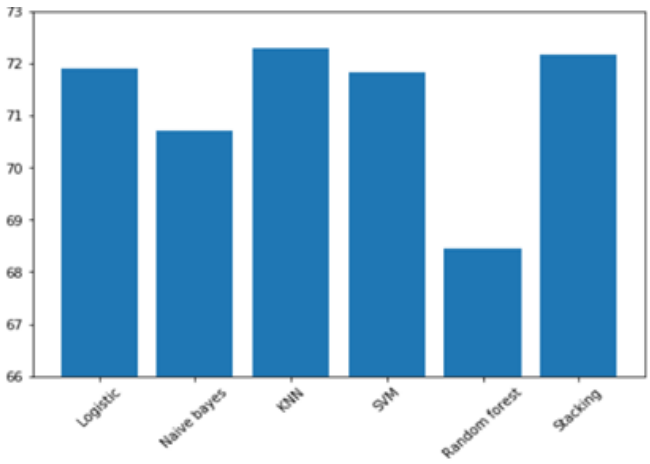


Fig 9. All model vs Accuracy

As I have already discussed that KNN has the best accuracy i.e. 72.28% then stacking algorithm i.e. 72.16%.

Model Evaluation

I have only evaluated the K-nearest neighbor(KNN) model because of its highest accuracy. It is used to

measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives, and False Negatives are used to predict the metrics of a classification model.

1. **TN / True Negative:** when a case was negative and predicted negative
2. **TP / True Positive:** when a case was positive and predicted positive
3. **FN / False Negative:** when a case was positive but predicted negative
4. **FP / False Positive:** when a case was negative but predicted positive

Confusion Matrix A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

Confusion matrix of the KNN model

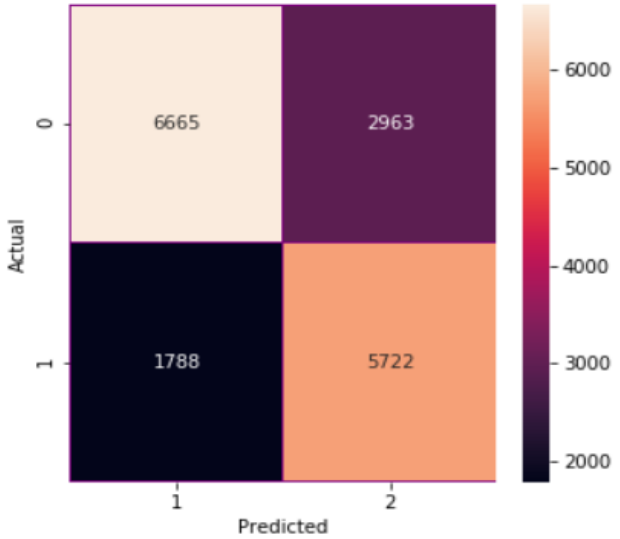


Fig 10. Predicted vs Actual value

Classification Report

Precision: Accuracy of positive predictions.
Precision = TP/(TP + FP)
Recall: Fraction of positives that were correctly identified.
Recall = TP/(TP+FN)

F1 score: What percent of positive predictions were correct F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Classification report of K-nearest neighbor

	precision	recall	f1-score
0	0.79	0.69	0.74
1	0.66	0.76	0.71

Mean Squared Error(MSE) It is an estimator measure the average of error squares i.e. the average squared difference between the estimated values and true value.

MSE for KNN model is **.277**
Mean Absolute Error(MAE) It is the difference between the measured value and "true" value.
MAE for KNN model is also **.277**

IV. CONCLUSION AND FUTURE WORK

Based on the above analysis, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart-related diseases. This paper discusses the various machine learning algorithms such as K-nearest neighbor, Support vector classifier, Random Forest, Logistic Regression, Naïve Bayes which were applied to the data set. It utilizes data such as blood pressure, cholesterol, BMI, and then tries to predict the possible coronary heart disease patient. Among all the models KNN has performed extremely well.

I also have tried to visualize the data set in every possible way to understand the fast easily.

Different Ensemble models like can also be applied XG boast, voting classifier to handle high dimensional data, and overfitting. A family history of heart disease can also be a reason for developing heart disease as mentioned earlier. So, this data can also be included for further increasing the accuracy of the model. This work will be useful in identifying possible patients who may suffer from heart disease. This may help in taking preventive measures and

hence try to avoid the possibility of heart disease for the patient.

V. ACKNOWLEDGEMENT

I have completed this work under the mentorship of Dr. Pankaj Agarwal (Professor & Head) & Ms. Sapna Yadav (Assistant Professor), Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad. I am doing an online summer internship on Machine Learning where I have learned the various Machine Learning Algorithms from both of my mentors as Course Instructors. This work is been assigned as project assignments to us.

I would like to express my special thanks to both of my mentors for inspiring us to complete the work & write this paper. Without their active guidance, help, cooperation & encouragement, I would not have my headway in writing this paper. I am extremely thankful for their valuable guidance and support on the completion of this paper.

I extend my gratitude to "IMS Engineering College" for giving me this opportunity. I also acknowledge with a deep sense of reverence, my gratitude towards my parents and member of my family, who has always supported me morally as well as economically.

VI. REFERENCES

[1]. Benan AKCA Prediction and analysis of Heart Disease using Machine Learning Algorithms (2019), Ph.D. Student at Marmara University İstanbul, Istanbul, Turkey. Link https://www.kaggle.com/benanakca/comparison-of-classification-disease-prediction

[2]. Svetlana Ulianova, Prediction and analysis of Heart Disease using Machine Learning Algorithms(2019), Data Science Student at Ryerson University Toronto, Ontario, Canada. Link https://www.kaggle.com/sulianova/eda-cardiovascular-data

[3]. Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna Prediction and analysis of Heart Disease using Machine Learning in International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366.

[4]. V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja Prediction of Heart Disease using Machine Learning Algorithms: A Survey in International Journal of Engineering & Technology, 7 (2.8) (2018) 684-687

[5]. Mohie Eldin Prediction and analysis of Heart Disease using Machine Learning Algorithms(2020), Computer Science student at Cairo University Cairo, Cairo Governorate, Egypt

[6]. Amita Malav, Kalyani Kadam, "A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K - Means", International Journal of Pure and Applied Mathematics 2018.

[7]. Amanda H. Gonsalves, Fadi Thabtah, Gurpreet Singh, Rami Mustafa A Mohammad Prediction of Heart Disease using Machine Learning Algorithms: An Experimental analysis in ResearchGate February 2020.

[8]. M.Nikhil Kumar, K.V.S Koushik, K.Deepak Department of CSE, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India Prediction Heart Diseases using Data mining and machine learning algorithms and tools. (2018)

About the Author

Digvijay Kumar is a B.Tech 3rd-year student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. He is interested in Python and Machine Learning.

Cite this article as :

Digvijay Kumar, "Cardiovascular Disease Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 46-54, September-October 2020. Available at
doi : https://doi.org/10.32628/CSEIT20659
Journal URL : http://ijsrcseit.com/CSEIT20659