International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Prediction of Cardiovascular Disease by Applying a Combination of Principal Component Analysis with Machine Learning Techniques

Advait Shirvaikar[1], Advait Mandlik[2], Prof. Sangeeta Prasanna Ram[3]

[1,2,3]*Department of Instrumentation Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India*

Abstract - *In this era of artificial intelligence and machine learning, diagnosis of health-related disorders based on data and its analysis is gaining a lot of momentum. Nowadays, machine learning & data analysis play a pivotal role in predicting health-related disorders of various body parts of the human body. With Cardiovascular diseases being the leading cause of death at the global level for the last 20 years [9], timely prediction of proneness to cardiovascular problems becomes significant. By analyzing various parameters or features of the patient data, machine learning algorithms predict whether the patient is at a risk thereby saving time and reducing expenses. significant. Our comparative study was based on five distinct machine learning algorithms, which were Logistic Regression, Support Vector Machine, K Nearest neighbour, Random Forest & Naive Bayes in combination with Principal Component Analysis (PCA) for data selection, to predict whether a patient is prone to cardiovascular disease, based on analysis of the parameters or features of the patient data from a Kaggle dataset comprising of 70000 values and 11 features. Based on the study we did, it was found that the Random Forest algorithm is superior to the other algorithms, based on their 'sensitivity' in identifying the disease*

***Key Words*: Machine Learning algorithms, Heatmaps, Principal Component Analysis, Sensitivity, Confusion Matrix**

1. INTRODUCTION

Most diseases are related to the heart so the prediction about heart diseases is necessary and for this purpose comparative study is needed in this field. Today most patients die because their diseases are recognized at the last stage due to the lack of accuracy of the instrument, so there is a need to know about the more efficient algorithms for disease prediction. Machine Learning is one of the efficient technologies for testing, which is based on training and testing. It is the branch of Artificial Intelligence (AI) which is one of the broad areas of learning where machines emulate human abilities.

Machine Learning plays a decisive role in predicting outcomes of these diseases. The machine learning based heart disease predicting systems will be precise and will reduce the expenses. The value of machine learning technology is recognized well in the healthcare industry which has a large pool of data. It helps medical experts to predict the disease and lead to improvising the treatment.

The work done in this paper aims to recognize and predict if an individual has a heart disease or not. With the aim of a comparative study, the data was first analyzed and using feature engineering, feature selection was performed. The outcome was then predicted by 5 different machine learning algorithms. The dataset used is published by Svetlana Ulianova as in the title of Cardiovascular Disease dataset, on Kaggle [8]. This dataset. The dataset is split into 2 parts, 75% for training, whereas the remaining 25% unseen data being treated as the test set. To optimize the machine learning algorithms discussed above, the data was first cleaned, and pre-processed. This paper also explains the importance of Principal Component Analysis and its application to machine learning in improving the accuracy of the model.

2. LITERATURE REVIEW

In literature various machine learning based diagnosis techniques have been proposed by researchers to predict heart diseases. This research study presents some ongoing and existing machine learning based diagnosis techniques in order to explain the importance of the proposed work.

Avinash Golande and Pavan Kumar studied various different machine learning techniques that can be used for classification of heart disease. Research was carried out to study various such as Decision Tree, KNN and K-Means algorithms that can be used for classification and prediction and their accuracy were compared. This research came to a conclusion that the accuracy achieved by the algorithm Decision Tree was the highest, further it was deduced that it can be made efficient by combination of different methods and parameter tuning [1].

Theresa Princy. R, et al, executed a survey including different classification algorithms used for predicting heart disease. such as Naive Bayes, KNN (KNearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different numbers of attributes [2].

Nagaraj M Lutimath, Chethan C, Basavaraj S Pol performed heart disease prediction using the classification

algorithm Naive bayes and Support Vector Machine (SVM). The performance measurements used in mathematical analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM emerged as the superior algorithm over Naive Bayes when both accuracies were compared [3].

Animesh Hazara, Amit Gupta, Arkomita Mukherjee, Subrat Kumar performed "Better data mining techniques when predicting heart disease". In this paper, c4.5, k-means, decision tree, SVM, naïve bayes and all other machine learning algorithms are compared to get a better accuracy of heart disease [4].

Purusothama, 2015. In this paper the Cleveland Heart Disease Data set was used for applying different classification techniques such as decision technique, association rule, K-NN, artificial neural network, naive bayes, hybrid approach for designing a prediction model for the given data set. An accuracy of 83.66% was obtained [5].

Rizwan Khan, Apurv Garg, Bhartendu Sharma performed Heart Disease Classification of the dataset from Kaggle of Svetlana Ulianova using K- Nearest Neighbour and Random Forest. An accuracy of 86.885% was achieved [6].

Apurb Rajdhan, Milan Sai, Avi Agrawal, Dundigalla Ravi have performed Heart Disease Prediction using Machine Learning using techniques such as Decision Trees, Naive Bayes, Logistic Regression, Random Forest on the UCI Machine Learning dataset.The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease[7].

3. DATA COLLECTION AND PRE-PROCESSING

The data collection process is the first step in any machine learning based study. This process includes understanding the objective of the dataset, it's target variables and the physical interpretation of the features. The dataset in our study was taken from Kaggle's Cardiovascular Disease Dataset. After analyzing the entire dataset, our first aim was to find out the abnormalities in the dataset, and eliminate them, which included features with null values or missing values. This pre-processing part also included various steps like adding new useful features based on the existing ones, and dealing with different data types.

3.1. Dataset Collection

We have collected data from a dataset provider –Kaggle.com [8]. The dataset is published by Svetlana Ulianova as in the title of Cardiovascular Disease dataset, 2019.The dataset collected consists of 70,000 records of patient's data carries 10 features plus one target value

Table -1: The Dataset

S. N. O	Attribute Name	Description	Type
1	Age	Age of Patients in Days	Numerical
2	Height	Height of Patient in cm	Numerical
3	Weight	Weight of Patient in Kg	Numerical
4	Gender	Male OR Female	Nominal
5	Systolic BP	Blood Pressure(mmHg)	Numerical
6	Diastolic BP	Blood Pressure(mmHg)	Numerical
7	Chol	Cholesterol	Nominal
8	Glucose	Glucose	Nominal
9	Smoke	Whether a person Smokes or not	Binary
10	ID no	No of Values	Numerical
11	Target	Target Value	Binary

3.2. Data Cleaning and Manipulation

The cleaning process started by identifying the null and missing values in each feature. Since the dataset had approximately 70,000 records, we considered removing the rows with null and missing values instead of imputing them. The statistical metrics were found out, which highlighted that the standard deviation for particular features (systolic and diastolic blood pressure) was high, and hence helped us identify the outliers.

Further, we added a new feature- the Body Mass Index (BMI), using the features 'Weight' and 'Height'. Later after using the Heatmaps, we found out this BMI feature was very helpful in prediction as it was more correlated to the target variable than the 'Height' or 'Weight' features. We also filtered out BMI more than 60, as the person might be prone to a heart disease due to obesity (a feature which wasn't included in our study).

Also, since the dataset consisted of categorical features like 'Gender' which were denoted by Male - 0, Female-1, we used 'One-hot-encoding'. This technique makes sure that machine learning does not assume that higher numbers are more important. We split these categorical features into individual features, such that each feature can be represented with a binary value 0 or 1, with the help of 'pandas'- a python library for manipulating data.

4. PROPOSED MODEL

The data cleaning and preprocessing helped to reduce the inconsistent data, but to further optimize the machine learning algorithm's function, we considered reducing the number of features. This could be achieved by various statistical methods which would be further discussed below.

For a comparison-based study, we considered the 5 most versatile machine learning algorithms, namely Support Vector Machine, Random Forests, Logistic Regression, Naive Bayes and K Nearest Neighbors. Each algorithm required, or performed well on a different type of input data - SVM and KNN needed standardized data (mean =0, variance =1). Each algorithm was further optimized by GridSearch and RandomSearch, which are used to find the best hyperparameters for a particular machine learning algorithm. These algorithms were then tested using the 'Test Set' on metrics such as 'Precision', 'Accuracy', 'Recall' and 'F1-score', so that they could be easily compared.

4.1. Feature Selection

Following the data cleaning, our aim was to minimize the dimensionality of our dataset, by finding out features that did not contribute much to the target variable did not contribute much to the target variable. For this, our approach was to use Principal Component Analysis and Heatmaps, both of which dealt with preserving the variance of the dataset and finding out the features correlated to the target variable, with the aim of reducing the total number of features used.

4.1.1. Principal Component Analysis

PCA - PCA is an unsupervised statistical optimization technique that assesses the interrelation between a set of attributes. This technique improves the performance of the algorithm as it eliminates correlated variables that don't contribute to decision making. Using PCA, we identified the most relevant features, when it comes to a heart disease. Statistically speaking, even 95-90% of the variance contribution for a dataset highlights the most important features of that dataset. Hence, we used PCA to find out which features contributed to approximately 95% of the variance of the dataset, ensuring that the vital correlations between the features and the target variable are retained. The following shows the Scree plot of our dataset:

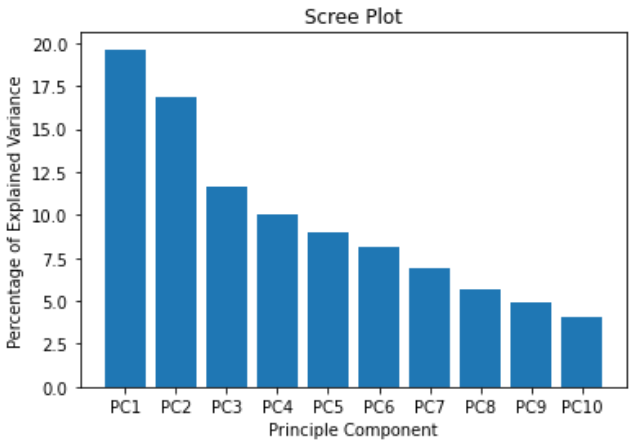

Fig -1: Principal Component Scree Plot

Here, each component (PC1, PC2, etc) displays how much a particular attribute contributes to the
total variance of the dataset (95% of the variance in our case).
The following display the features that contribute to 95% of the variance of the dataset:

	Principal Component	Feature
0	PC1	ap_hi
1	PC2	gender
2	PC3	gluc
3	PC4	alco
4	PC5	active
5	PC6	age
6	PC7	weight
7	PC8	smoke
8	PC9	cholesterol

Fig -2: PCA Feature List

4.1.2. Heatmaps

Heatmaps- Heatmaps are a visual representation of the attributes of a dataset, in the form of 2-D coloured maps. Heatmaps are used to show the relationships between the features of a dataset. The main initiative of the seaborn heatmap is to show the correlation matrix by visual information interpretation. It helps find the link between multiple features and the best features which are suited for the given model to be trained.

A correlation matrix denotes the correlation coefficients between variables at the same time. This correlation matrix is then represented with the desired colour coding, in the form of a heatmap. The correlation matrix shows how each feature is correlated on a scale of -1 to 1, 1 being positively correlated and -1 being inversely correlated.

For our model, we only considered features having at least 2% (positive or negative) correlation with the target variable, as anything lesser than that would have a negligible contribution towards the target variable.

The following shows the heatmap of our dataset:

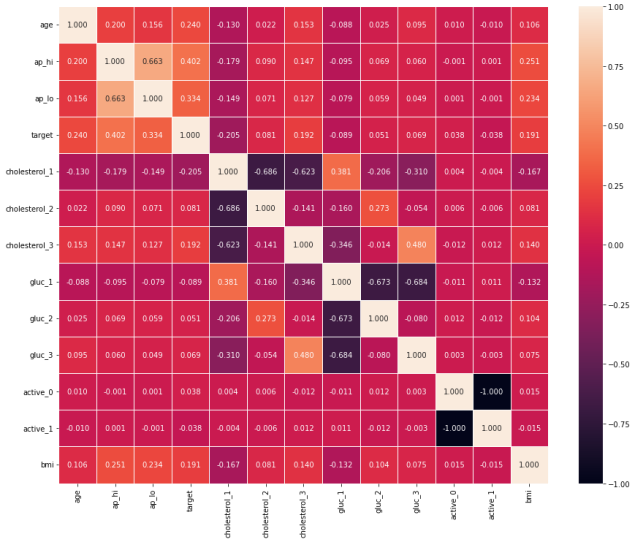

Fig -3: Heatmap

4.2. Data Prediction

Having done pre-processing of data by cleaning it and by extracting features based on feature selection, the next step was to develop a model using Machine learning algorithms such as Logistic regression, Support Vector Machine, K-Nearest neighbour, Naive Bayes and Random Forest to handle large datasets efficiently.

Logistic Regression-This technique is basically the relationship between features and the probability out of a particular event which deals with the sigmoid function which forms the basic concept of this algorithm. In logistic regression the sigmoid function fits the output of a linear equation between 0 and 1.

Support Vector Machine- SVMs revolve around the notion of a margin which divides the set into 2 parts each on either side of a hyperplane that separates two data classes. An SVM classifier creates a model that allocates various new data points to a chosen category. Using the kernel trick, SVM is used for non-linear classification. It maps inputs into high dimensional feature spaces.

K-Nearest neighbour- KNN is a non-parametric, lazy learning algorithm. Its aim is to use a database in which the given points are categorized into clusters to predict the classification of a new sample point. It is a supervised classifier that carry-out observations from within a test set to predict classification labels. KNN is one of the classification techniques used whenever there is a classification. By applying KNN in large datasets takes a long time to process.

Naive Bayes- Naïve Byes is a popular classifier which assumes no feature has any relationship or dependence to

each other. Naive Bayes classifiers are based on Bayes Theorem. The simplistic design of this model allows it to predict the data faster

Random Forests - Random forest is a very handy algorithm because the hyperparameters that it selects generally comes up with a good prediction result. It creates a tree for the data and makes predictions based on that. There are two steps in random forests, firstly design a random forest and make a prediction with the help of the classifier generated in the first stage.

5. RESULT AND ANALYSIS

The model is trained on 50986 records, having 12 features, and then tested on 16996 records. The table below shows the comparative study of different models in use, which are KNN, Random Forest, Support Vector Machine, Decision Trees, and Naive Bayes.

The dataset is first split into 2 parts, 75% is used for training, whereas the remaining 25% unseen data was treated as the test data.

Table -2: Metric Comparison

	KNN (in %)	LR (in %)	RF (in %)	Naive Bayes (in %)	SVM (in %)
Accuracy	72	70	73	66	72
Precision	72	71	73	67	73
Recall	70	70	73	66	71

KNN – K- Nearest Neighbour
LR – Logistic Regression
RF – Random Forest
SVM – Support Vector Machine

Confusion Matrix is generally plotted to get the number of records that have been classified correctly from each of the classes. It gives information about false positives and negatives. This helps in further analysis of the model. The confusion matrix for Random Forest algorithm is given to the right top hand side:

Table -3: Confusion Matrix of Random Forest

	Predicted No Heart Disease	Predicted Heart Disease
Actual No Heart Disease	6,710	1,814
Actual Heart Disease	2,721	5,691

Out of 16,936 test images, 8,524 images originally belonged to the "No Heart Disease" Class and 8,412 belonged to the "Heart Disease" Class. Out of the 8,524 "No Disease" records, 6,710 are classified correctly as "No Disease" and 1,814 are classified as "Heart Disease" records. Similarly, out of 8,412 records with Heart Disease, 5,691 are classified correctly as "Heart Disease", and 2,721 are classified as "No Heart Disease" records.

The following is the classification report of the Random Forest Classifier, on the testing set:

	precision	recall	f1-score	support
0	0.71	0.79	0.75	8524
1	0.76	0.68	0.72	8412
accuracy			0.73	16936
macro avg	0.73	0.73	0.73	16936
weighted avg	0.73	0.73	0.73	16936

Fig -4: Classification Report

0: No Heart Disease
1: Heart Disease

6. CONCLUSION

The motivation for the study was to find the most accurate Machine Learning classification algorithm for prediction of cardiovascular disease. This study compares the accuracy and recall score of 5 algorithms namely Logistic regression, K-Nearest Neighbour, Support Vector Machine, Random Forests and Naive Bayes. Data is collected from the dataset published by Svetlana Ulianova as in the title of Cardiovascular Disease dataset. After extensive pre-processing of the data, the model that provides best results for Binary classification is the Random Forest Classifier which achieves a training Accuracy of about 75% and the Testing Accuracy of about 73%. The Recall or 'Sensitivity', an important metric in Medical Analysis, of the Random Forest

Classifier model is also quite high, in comparison to the other classifier models, with a training Recall score of 74% and 73% on the Test Dataset. The conclusion can be finally drawn that machine learning is able to predict and restrict the implications done to a person's heart.

REFERENCES

[1] - Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] - Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016

[3] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019

[4] - Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee (2017) Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques. Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137- 2159.http://www.republication.

[5] - Purusothama et al, 2015. Different classification techniques to design risk prediction model for Heart Disease UCI repository

[6] - Rizwan Khan, Apurv Garg, Bhartendu Sharma (January 2021) Heart Disease Prediction using Machine Learning Techniques. IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012046 doi:10.1088/1757-899X/1022/1/012046

[7] - Apurb Rajdhan, Milan Sai, Avi Agrawal, Dundigalla Ravi (April 2020) Heart Disease Prediction using Machine Learning International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181Vol. 9 Issue 04, April-2020

[8] - Cardiovascular disease dataset https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[9]https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019#:~:text=Heart%20disease%20has%20remained%20the,nearly%209%20million%20in%202019.