ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Prediction of Cardiac Disease using Supervised Machine Learning Algorithms

GAMPA DEVI PRASAD, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

LANKA JAYA SOURYA SAI, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India

MARISETTI MADHUMITHA, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India,

B. ESHWARARAO, Department of Electronics and Communication Engineering
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, India eshwar.world@gmail.com

Abstract — Millions of people throughout the world are a part of the healthcare industry, which generates a vast amount of data. The multidimensional medical datasets are being dissected by machine learning-based models, which are providing new insights. Several cutting-edge Supervised Machine Learning algorithms are employed in this study to accurately classify a cardiovascular dataset in order to provide illness predictions. According to the results, the Decision Tree classification model predicted cardiovascular illness better than other models, such as Naive Bayes and Logistic Regression. Accuracy of 73% was achieved by using the Decision Tree. Doctors may find this method useful in predicting the onset of cardiac disease and delivering timely therapy.

Keywords — Cardiovascular Disease, Naive Bayes, Decision Tree, Logistic Regression, Random Forest, SVM, KNN, Risk prediction

I. INTRODUCTION

Heart disease (CVD) is the leading cause of death in the world, according to the World Health Organization (WHO). More than 17 million people die each year from cardiovascular disease (CVD), which accounts for about a third of all fatalities worldwide. A substance called plague, which clogs up the arteries and veins that convey blood to and from the heart, is the primary cause of cardiovascular disease (CVD). Blood flow is impeded and blocked, which can lead to heart attacks and strokes. High blood pressure, poor nutrition, inactivity, high blood cholesterol levels, alcohol consumption, cigarette use, obesity, and genetic abnormalities are all risk factors for cardiovascular disease. These deaths can be prevented if early prognostication is made. The Internet of Things, on the other hand, is constantly improving the ways in which data is collected. There are gigabytes of data coming from healthcare businesses every day as a result of these improvements. Humans are unable to sift through the millions of pieces of information that are available to determine a patient's specific medical condition. As a result, Machine Learning can be used as a predictive tool to identify patterns in data.

Factors are analysed and used to determine who is most at risk of getting heart disease using the application of Machine Learning. Methods for machine learning can evaluate enormous amounts of data and detect trends that may not be visible to humans. As the volume of data processed grows, it is often more efficient and accurate. There's also no need for human assistance, which is a huge benefit. Using labelled data and output patterns, the system learns how to do a task under the guidance of supervised machine learning. Algorithms hunt for patterns in the data that correlate with desired outputs during their training. The supervised

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

learning model can predict the proper label for newly presented input data after training. The goal of this research is to identify a well-performing algorithm by comparing the classification accuracy of various supervised machine learning algorithms.

II. LITERATURE REVIEW

Machine learning techniques were used by Krumholz et al. to predict death and hospitalisation in patients with heart failure [2]. Five approaches have been applied, including LR with forwarding selection variables and LASSO regularisation variable selection, Random Forest, Gradient Descent Boosting, and SVM. Using 5-fold cross validation, a three-year follow-up was conducted. Predicting mortality had a mean C-statistic of 0.72, whereas predicting hospitalisation had a mean C-statistic of 0.76. The proposed model's results could be improved if time-to-event analysis is included.

[3] Vilasi et al. employed Machine Learning to determine the risk of cardiovascular disease (CVD) in Dialytic patients Machine Learning algorithms were tried on datasets from both Italy and the United States to see which worked best. Support Vector Machine (SVM) with RBF Kernel method produced the best results with 95.25 percent accuracy in the Italian dataset and 92.15 percent accuracy in the American dataset. In addition, the bias in the Italian dataset may affect the accuracy of forecasts.

Heart Rate Variability (HRV) can be used to predict cardiac arrest in smokers, according to Shashikant et al (HRV). Heart rate variability (HRV) [4] is a non-invasive method for assessing heartbeat regulation. To get accurate data, you need to be in the right place at the right time. We compared the results of Decision Tree, Logistic Regression, and Random Forest. Using the 10-fold validation procedure, all categorization techniques are evaluated. The accuracy of Logistic Regression was found to be 89.7%, that of Decision Tree to be 92.59%, and that of Random Forest to be 93.61%. Among these approaches, Random Forest emerged victorious.

Ahmed et al. [5] came up with the concept of 'Auto prognosis.' Models for Machine Learning are selected and tuned automatically by this system. The model was tested on data from 423,604 people. The results were compared to the 'Framingham Score,' a well-known risk prediction method. The results show that the Auto prognosis model has a 95% accuracy rate in predicting better than the Framingham Score. Other factors, such as triglycerides, inflammatory markers, and natriuretic peptides, were not taken into account when making the prognosis.

In order to overcome the missing value in the medical dataset and accurately forecast CVD, Zhou et al. proposed a learning technique [6]. Random Forest and Naive Bayes algorithms were used to predict CVD. For the most part, RF outperformed the other approaches, with 88% specificity, 87% sensitivity, and 88% accuracy.

Amin et al. [7] developed a hybrid intelligence system for predicting cardiac illness [8]. Logistic Regression, ANN, KNN, Decision Trees and Naive Bayes were employed for classification in that system. Three feature selection methods, Relief, mRMR, and LASSO, were used to identify strongly correlated characteristics that greatly influence the target variable in order to increase prediction efficiency. With the Relief feature selection algorithm, Logistic Regression with 10-fold validation achieved an accuracy of 89 percent.

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

This model's output could be improved with the use of neural network optimization techniques.

Panagiotis's et al. compared CVD established risk tool Hellenic Score with Machine Learning approaches [8]. For this investigation, researchers used a set of data called ATTICA. Hellenic Score has 85% accuracy, 20% specificity, 97.7% sensitivity, 87.7% PPV, and 58.8% NPV based on the type of classifier and training dataset. In contrast, the Machine Learning methods have 65-84 percent accuracy, 46-56 percent specificity, 67-89 percent sensitivity, 89-91 percent PPV, and 24-45 percent NPV, respectively. The best result was obtained using Random Forest. There was no examination of the link between a person's way of life and the likelihood of developing a cardiovascular disease in the research.

For the purposes of assessing three risk factors, Kang et al. used Machine Learning [9]. These are the primary causes of cardiovascular disease (CVD). HCRT-Logistic model and Logistic CART model were used to forecast the hazards. Cross-validation was used to test both models. BMI, waist-to-hip ratio, waist-to-height ratio, medical history, and other factors were taken into account. The proposed model's accuracy varies depending on the gender. Men and women alike can benefit greatly from measuring their waist circumference.

In order to predict cardiovascular disease (CVD), Stephen et al. used Machine Learning techniques. The information was gathered from 378,256 UK-based patients. Techniques including Random Forest, Logistic Regression, Neural Networks, and Gradient Boosting were used to analyse the data. The Neural Network gripped with greater precision than ever before. The intellectual nexus behind the Neural Network method, on the other hand, is difficult to understand.

According to Ashok [11], these models may predict the likelihood of heart disease by employing ANN, KNN, and SVM, logistic regression, classification trees and naive bays. Furthermore, the ROC curve was used to evaluate these approaches. Logistic Regression has the highest accuracy of 85%, with an 89.9% sensitivity and 81% specificity, making it the most accurate method. Nevertheless, the model needs to be tested on a large number of datasets before it can be considered reliable.

Aljaaf et al [12] proposed the multi-level risk assessment method. In addition to pre-existing characteristics, three new risk factors were introduced: smoking, inactivity, and obesity. Using the Decision Tree technique, we were able to accurately predict the likelihood of heart failure in 86.53% of patients. The denouement outcome of this model could be improved by implementing advanced feature selection methods.

III. MATERIALS

Description of the dataset:

This research was based on data from Kaggle [11] on cardiovascular illness. One target variable is included in a list of twelve. Table 1 shows an illustration of this. The study looked at people between the ages of 29 and 64. Their height and weight have also been recorded. 'Male and female patients were given a gender value of 1 and 0 correspondingly. A person's blood pressure is measured in two ways: systolic and diastolic. The Cholesterol

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

and Glucose levels of the patients were categorised as normal, above normal, or substantially over normal.

One's smoking and drinking habits have a significant impact on one's risk of developing heart disease. Both of these variables are denoted by a single bit of binary data. Smoker/alcohol drinkers are marked with a "1," whereas non-smokers and non-alcoholics are marked with a "0." Patients who engage in regular physical activity were given a '1' whereas those who did not were given a '0.' Whether or not a person has cardiovascular disease is what we're looking for here. It consists of a set of Boolean values. The '0' indicates a healthy heart, whereas the '1' indicates those who have been diagnosed with a cardiac condition.

T ABLE 1 FEATURE INFORMATION OF THE DATASET

S.No	Attribute Name	Description	Range of Values
1	age	Age	int (years)
2	height	Height	int (cm)
3	weight	Weight	float (kg)
4	gender	Gender	categorical code
5	ap_hi	Systolic blood pressure	int
6	ap_lo	Diastolic blood pressure	int
7	cholesterol	Cholesterol	1: normal, 2: above normal, 3: well above normal
8	gluc	Glucose	1: normal, 2: above normal, 3: well above normal
9	smoke	Smoking	binary
10	intake alco	Alcoholic	binary
11	active	Physical activity	binary
12	cardio	Presence or absence of cardiovascular disease	binary

IV. EXPERIMENTS AND RESULTS

Figure 1 depicts the functional flow of this examination. Figure 2 depicts the association between the dataset's various attributes. The dark brown colour suggests a strong positive association while the dark blue colour shows a weak negative correlation.

This study focused on implementing a few categorization algorithms and comparing their results. The dataset was split between training and testing sections with a 70% to 30% split. CVD was predicted using Naive-Bayes, Decision Tree, Logistic Regression, Random Forest, SVM and KNN classification models.

Identification of mislabelling or prediction errors can be accomplished with the help of a confusion matrix. In this model, the anticipated and actual values are compared using four elements: the true positive, true negative, false positive and false negative, respectively. False Positive and False Negative values are the seeds of Type-I and Type-II mistakes,

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

respectively. You can quickly determine the accuracy and precision of your results by using the confusion matrix.

T ABLE 2 CONFUSION MATRIX

		Actual values	
		Positive	**Negative**
Predicted Values	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

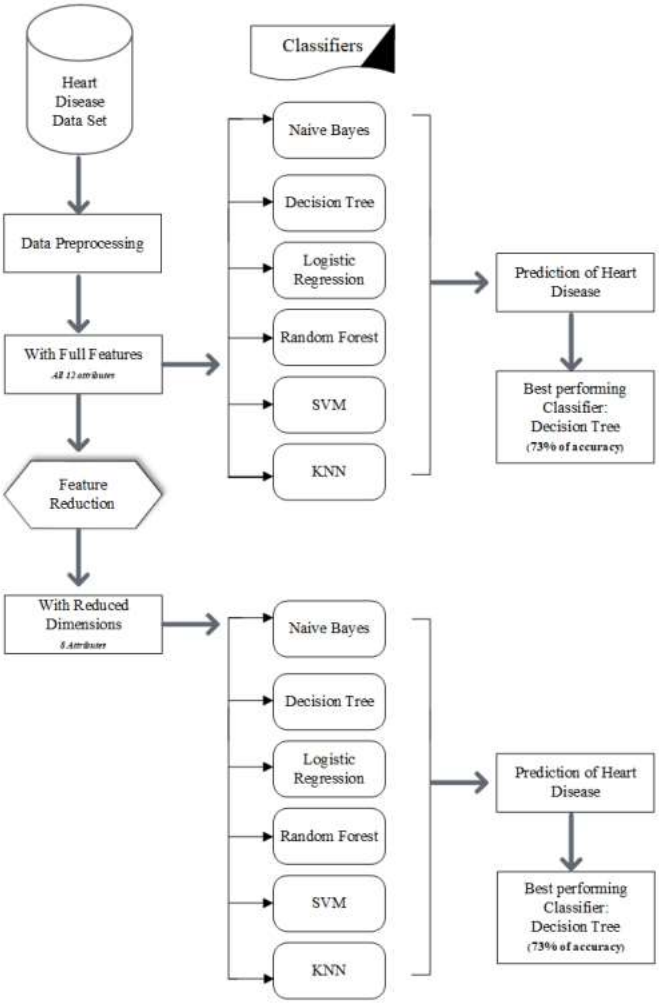


Fig. 1: Predicting CVD using supervised learning algorithms

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

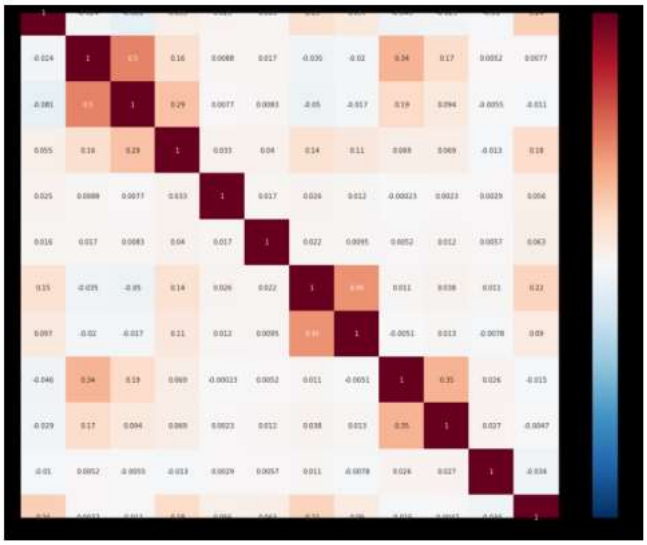


Fig. 2: Correlation between all available features

A. Prediction

Accuracy The accuracy is a measure of how well the projected values came to fruition. The accuracy of each algorithm is shown in Figure 3. (True Positive + True Negative)/Total = Accuracy In comparison to other algorithms, the decision tree method produced a 73% success rate. Random forest came in second with 71%, followed by logistic regression with 72% and SVM with 72%. The accuracy rates for the KNN and Naive-Bayes algorithms were 66% and 60%, respectively.

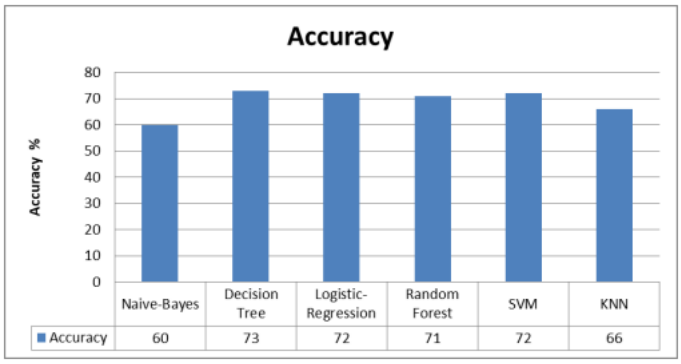


Fig. 3: Accuracy of various learning techniques

B. Precision

Real positive cases from all the positive forecasts are represented here. On the other hand, Fig. 4 shows the precisions of several methods.

Precision = True Positive/ (True Positive + False Positive)

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Fig. 4: Precision of learning techniques

C. Recall

It identifies the positive classes with the most accurate predictions. Fig. 5 shows the recall levels for each tested method in comparison to each other.

Recall = True Positive/ (True Positive + False Negative)

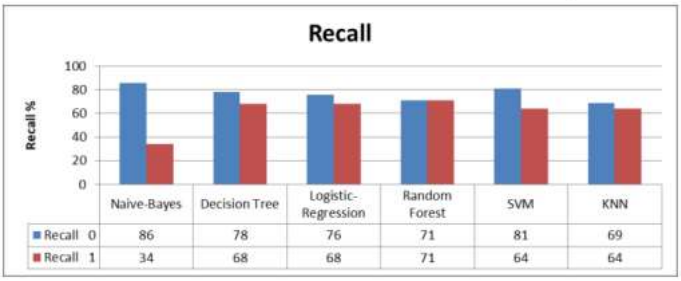


Fig. 5: Recall of learning techniques

D. F1 score

Test accuracy is calculated using Harmonic Mean, which evaluates Recall and Precision. F1 scores are shown in Fig. 6 based on several algorithms.

F1 score = (2 * Precision * Recall))/ (Precision + Recall

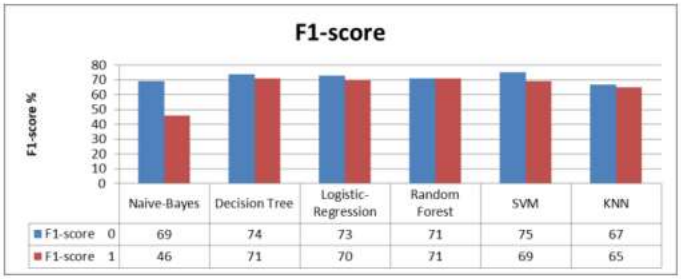


Fig. 6: F1 score of learning techniques

E. Dimensionality

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

Reduction Figure 2 shows a negative relationship between things like height, smoking, drinking, and physical activity. When testing these models, these entities were taken out of their dataset and tested against them. Reduced dimensions were used to forecast the CVD.

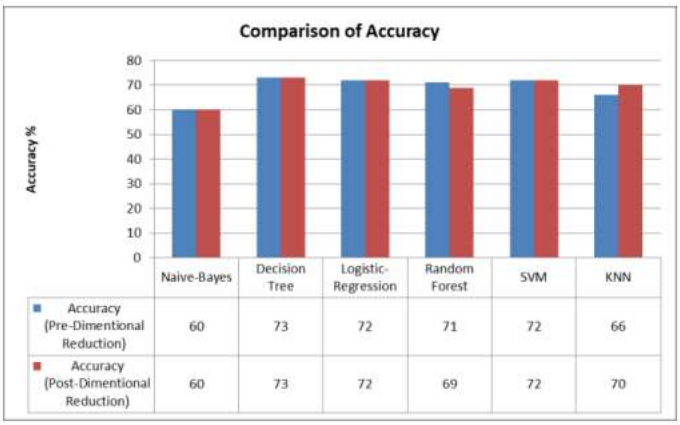


Fig. 7: Comparison of Accuracy

Figure 7 compares the accuracy of algorithms before and after they have been reduced in dimension. The results of measures were influenced when the number of attributes was lowered to eight. The Random Forest algorithm's accuracy was reduced from 71% to 69%. However, the KNN algorithm's accuracy increased from 66% to 70%. Algorithm precision values before and after the dimensionality reduction are depicted in Figure 8. Logistic-precision Regression's value has changed somewhat, whereas Random Forest's and KNN's precision values have changed significantly. Figures 9 and 10 show how Random Forest and KNN's memory and F1 scores have changed significantly.

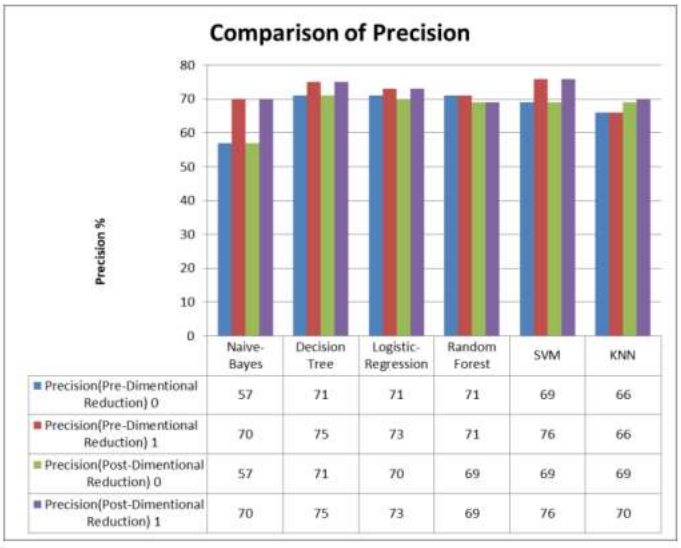
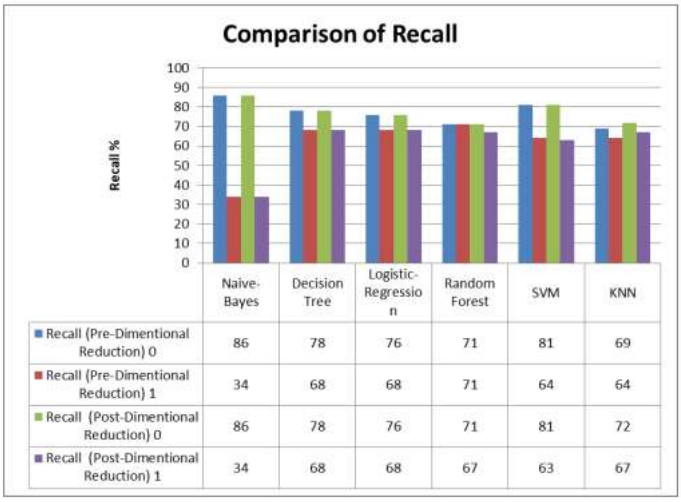


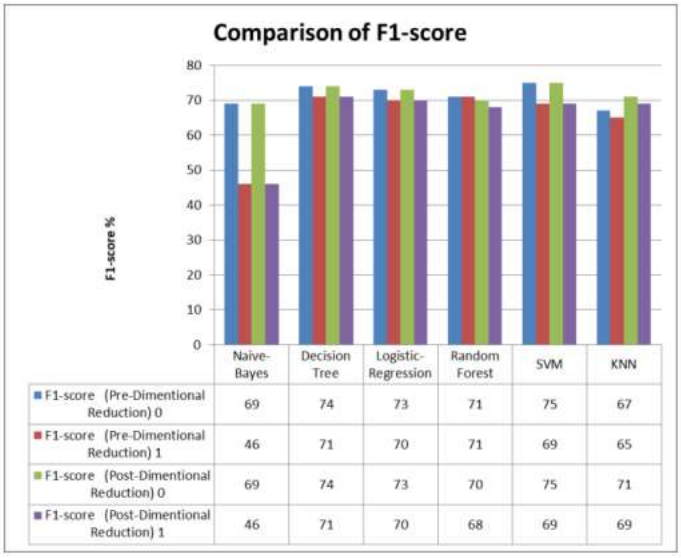
Fig. 8: Comparison of Precision

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022



	Naive-Bayes	Decision Tree	Logistic-Regression	Random Forest	SVM	KNN
Recall (Pre-Dimentional Reduction) 0	86	78	76	71	81	69
Recall (Pre-Dimentional Reduction) 1	34	68	68	71	64	64
Recall (Post-Dimentional Reduction) 0	86	78	76	71	81	72
Recall (Post-Dimentional Reduction) 1	34	68	68	67	63	67

Fig. 9: Comparison of Recall



	Naive-Bayes	Decision Tree	Logistic-Regression	Random Forest	SVM	KNN
F1-score (Pre-Dimentional Reduction) 0	69	74	73	71	75	67
F1-score (Pre-Dimentional Reduction) 1	46	71	70	71	69	65
F1-score (Post-Dimentional Reduction) 0	69	74	73	70	75	71
F1-score (Post-Dimentional Reduction) 1	46	71	70	68	69	69

Fig. 10: Comparison of F1 score

V. CONCLUSION AND FUTURE WORK

Classification algorithms have been used to analyse a cardiovascular dataset in this work. By giving a prediction accuracy of 73%, the Decision Tree algorithm outperformed the competition. Random Forest and KNN algorithms are affected by a reduction in the number of dimensions in a dataset, which impacts their performance. The results show that dataset size has a positive or negative impact on algorithm performance. The Principal Component Analysis (PCA) and High Correlation Filter (HCF) will be used for dimensionality reduction in the next step. The CVD dataset will be used to evaluate and create a better illness prediction model with the help of ensemble machine learning algorithms.