Journal of Advanced Zoology

ISSN: 0253-7214

Volume 44 Issue Special Issue-5 Year 2023 Page 1215:1226

Prediction of Cardiovascular disease using machine learning algorithms on healthcare data

Salmoli Chandra*, J. Chanda **, Sukumar Chandra**

**JIO Institute, Sector-4, ULWE, NAVI MUMBAI, MAHARASHTRA- 410206, India*

*** B.P. P. I.T & M., Maulana Abul Kalam Azad University of Technology, Kolkata, India*

**** Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, West Bengal, India*

**** Corresponding Author :*

Mail address: sukumarchandra14@gmail.com

Article History

Received: 08July2023

Revised: 29 Aug 2023

Accepted: 02 Oct 2023

ABSTRACT:

Cardiovascular Disease (CVD) is a leading cause of death worldwide, with the potential to cause serious conditions such as heart attacks and strokes. Early assessment of CVD can significantly reduce mortality rates. In recent studies, machine learning algorithms have been applied to Electronic Health Records (EHR) to estimate risk factors for myocardial infarction. This article explores the use of various machine learning techniques on a healthcare dataset to predict a 10-year risk of future coronary heart disease (CHD). The dataset used in this study was obtained from the Framingham and Massachusetts cardiovascular study. We found that our models achieved varying levels of accuracy: 64% for logistic regression, 83% for Naïve Bayes classifier, 42% for Support Vector Machine (SVM), 65% for Random Forest, 78% for KNN classifier, and 70% for XGBOOST classifier. It is revealed that a patient with no history of heart disease may benefit from an algorithm such as Naive Bayes Classifier, while an older patient with a history of heart disease may require an algorithm such as Support Vector Machine. These factors can help guide the physician in selecting the most appropriate algorithm for each individual patient, ensuring that the diagnosis is as accurate as possible and that the treatment plan is tailored to meet the patient's unique needs.

Keywords: *Support Vector Machine, Cardiovascular diseases, Machine learning*

Introduction:

Cardiovascular Disease (CVD) is a major health concern worldwide, causing millions of deaths each year with increasing rates [1, 2, 3]. Cardiologists and surgeons often struggle with estimating the risk of heart failure. It is crucial to accurately predict the risk of heart failure in order to identify and treat complex cardiovascular diseases at an early stage. Machine learning models from medical databases can be used for that purpose. Ishaq et al. [4] has made an attempt to predict heart failure disease using SMOTE and data mining techniques.

Machine learning techniques can be more useful than traditional modeling techniques in some cases. Turkmenoglu and Yildiz [5], as well as Chicco and Jurmen [6], have used machine learning models in their data analysis. Different machine learning models can lead to different conclusions, and it is important to choose the appropriate model for a specific use case. Therefore, it is necessary to carefully evaluate the performance of each model to determine which one is best suited for the task at hand.

To further explore the use of machine learning in predicting cardiovascular disease, this study utilizes the Framingham Heart Study data set. By applying multiple machine learning techniques and using the Python framework, we aim to gain valuable insights into CVD prediction. The Framingham data set contains 3,390 records and 17 attributes related to patient information, making it an extremely useful resource for predicting CVD.

METHODOLOGY

Data collection was made by many researchers from the Framingham dataset [7, 8, 9]. Here we have also used the Framingham dataset. Demographic features such as Sex (male or female), Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) , behavioural features like smoking (whether or not the patient is a current smoker), Cigs Per Day (the number of cigarettes that the person smoked on average in one day and can be considered continuous as one can have any number of cigarettes, even half a cigarette, medical parameters like BP, Prevalent Stroke, Prevalent Hypertension, Diabetes, Total Cholesterol level, Systolic blood pressure, diastolic blood pressure , Body Mass Index, heart rate, glucose level, 10-year risk of coronary heart disease (CHD) are taken into consideration.

Data cleaning has been made to make the data free from error. Then missing value treatment has been treated as medians. Inconsistencies have been detected by graphical methods. Data imbalance has been resolved by equalising data for both the classes getting rid of oversampling. Outliers have been handled and out of range or false data have been eliminated removing duplicate values. Same values in multiple places have also been removed. Label encoding has been used for few categorical features. Multicollinearity has been removed using the VIF factor and after that exploratory data analysis method has been adopted. made. Then the ML algorithm has been applied to find the accuracy, precision score, F1 score and the recall. Missing value (null value) treatment has been carried out through missing observations in several columns which has further been treated with its median value that corresponds to that column.

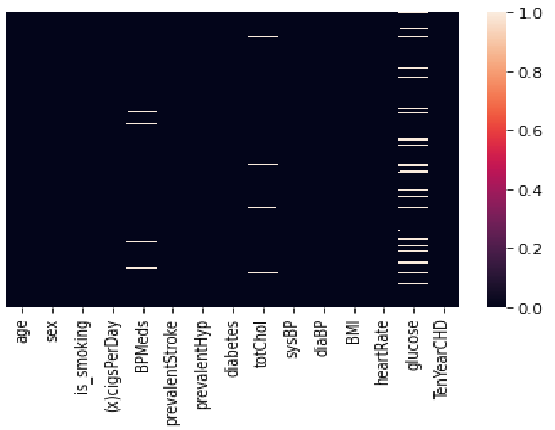


Figure-1. Position of null values

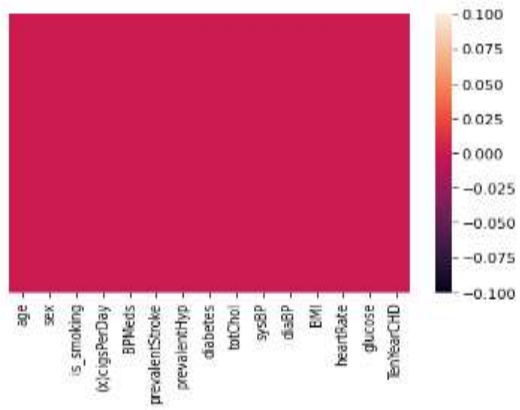


Figure-2. After treating null values

Exploratory data analysis that we performed on our train dataset helped us to realize how different features do influence the target variable.

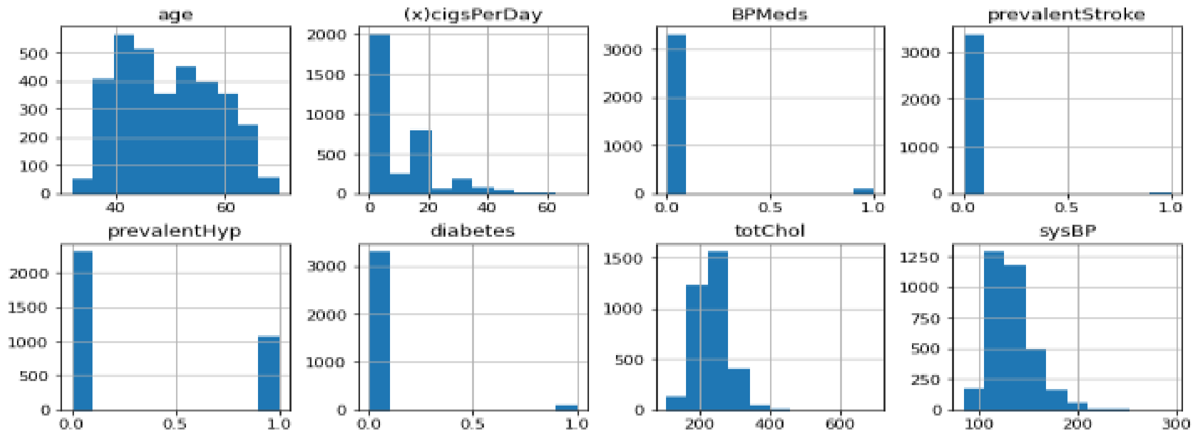
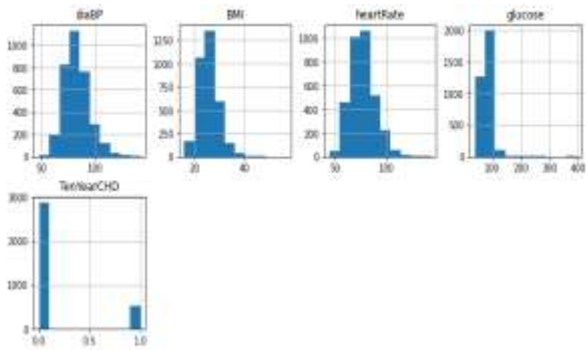


Figure-3. Data distribution



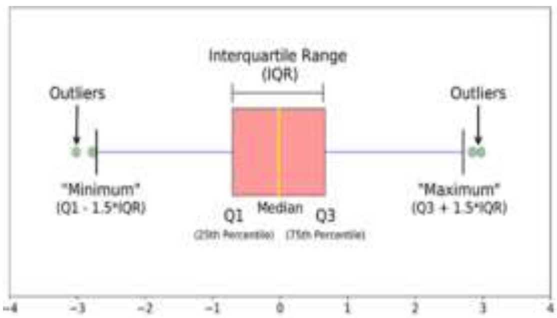


Figure 4 – outlier treatment

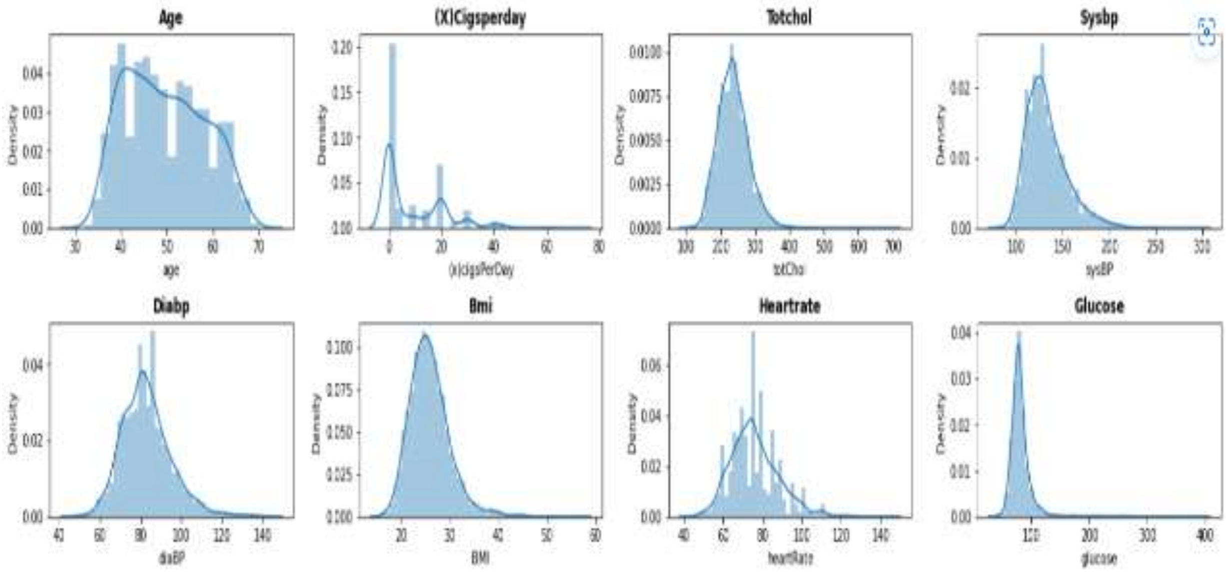
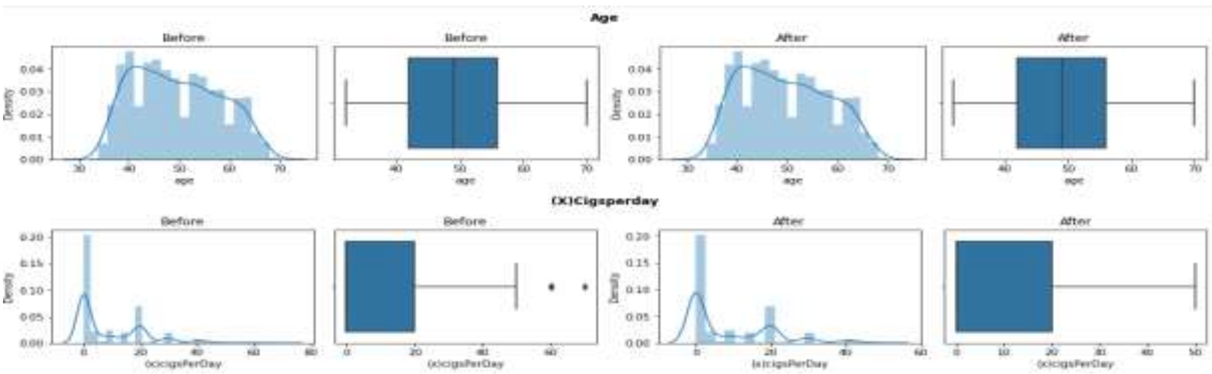


Figure-5- visualisation of data distribution before outlier treatment



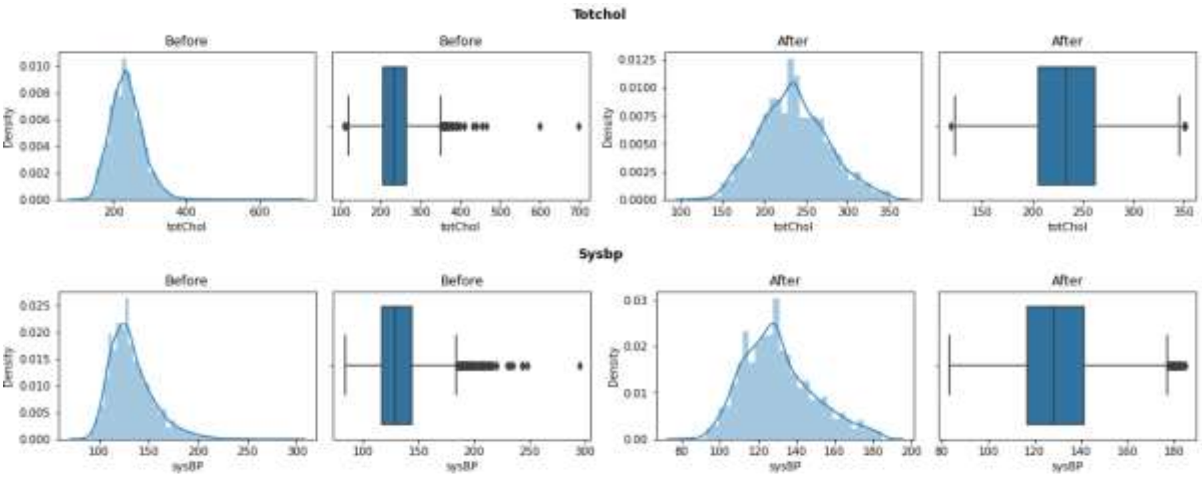


Figure-6- Visualisation of data distribution with and without outlier treatment

Checking and dropping the duplicate values and cleaning the dataset, label encoding has been made by converting the categorical variables into numeric form. Univariate analysis method has been used on one variable with the aim of finding out and identifying the characteristics of the variable.

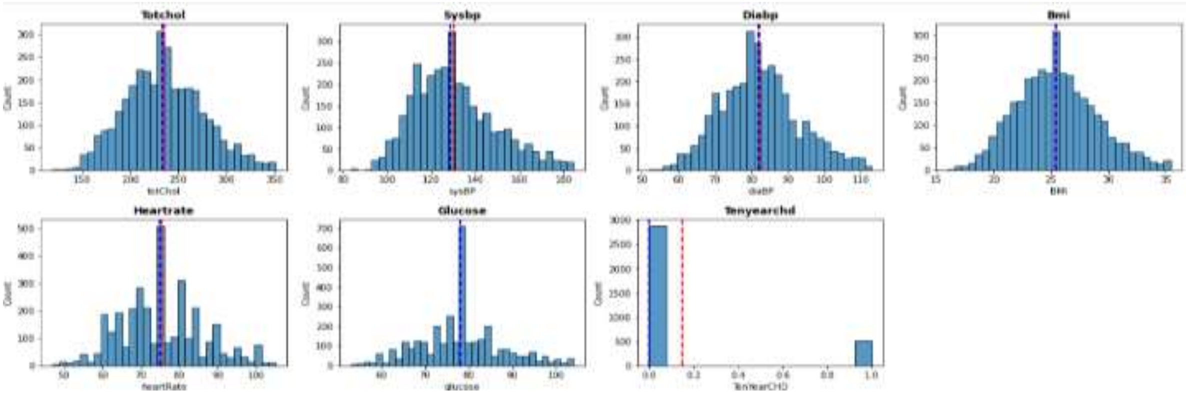
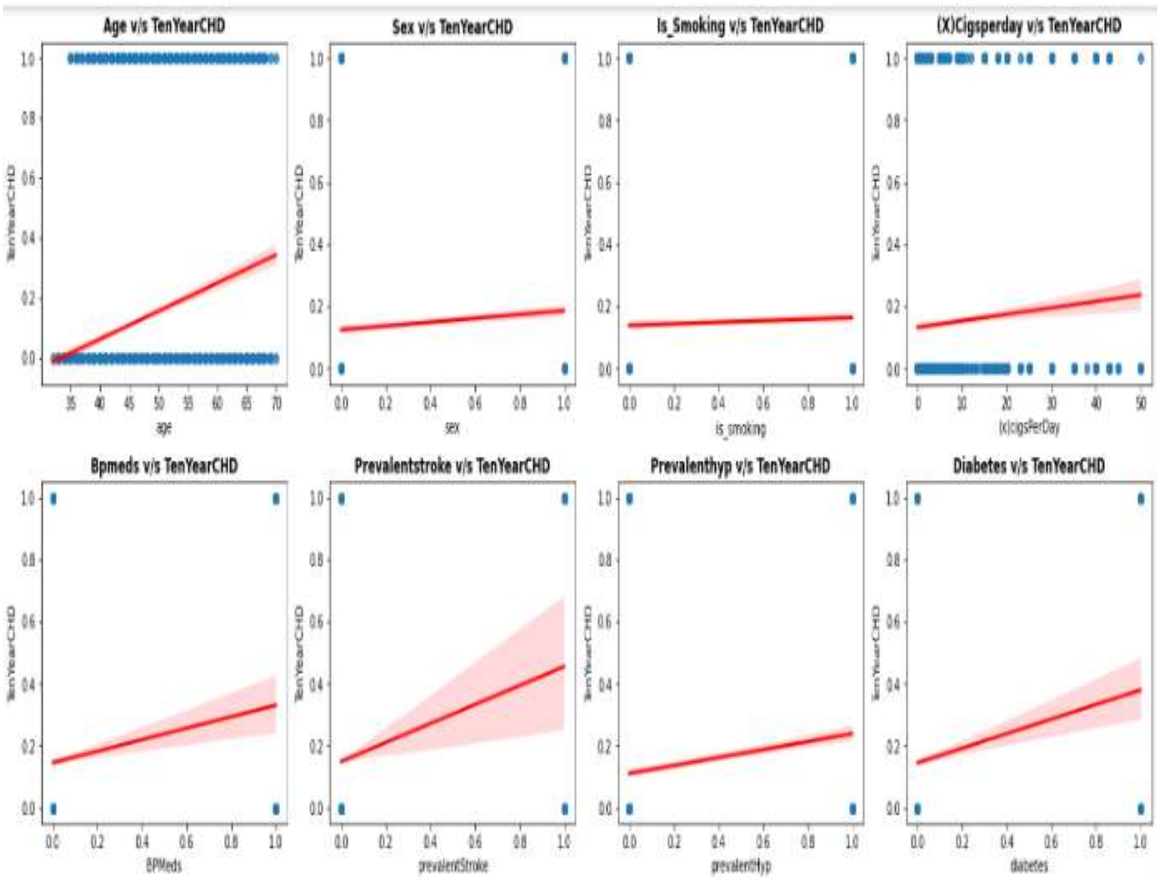


Figure-7- Univariate analysis

Bivariate analysis for any combination of categorical and continuous variables has also made using the python framework.



.

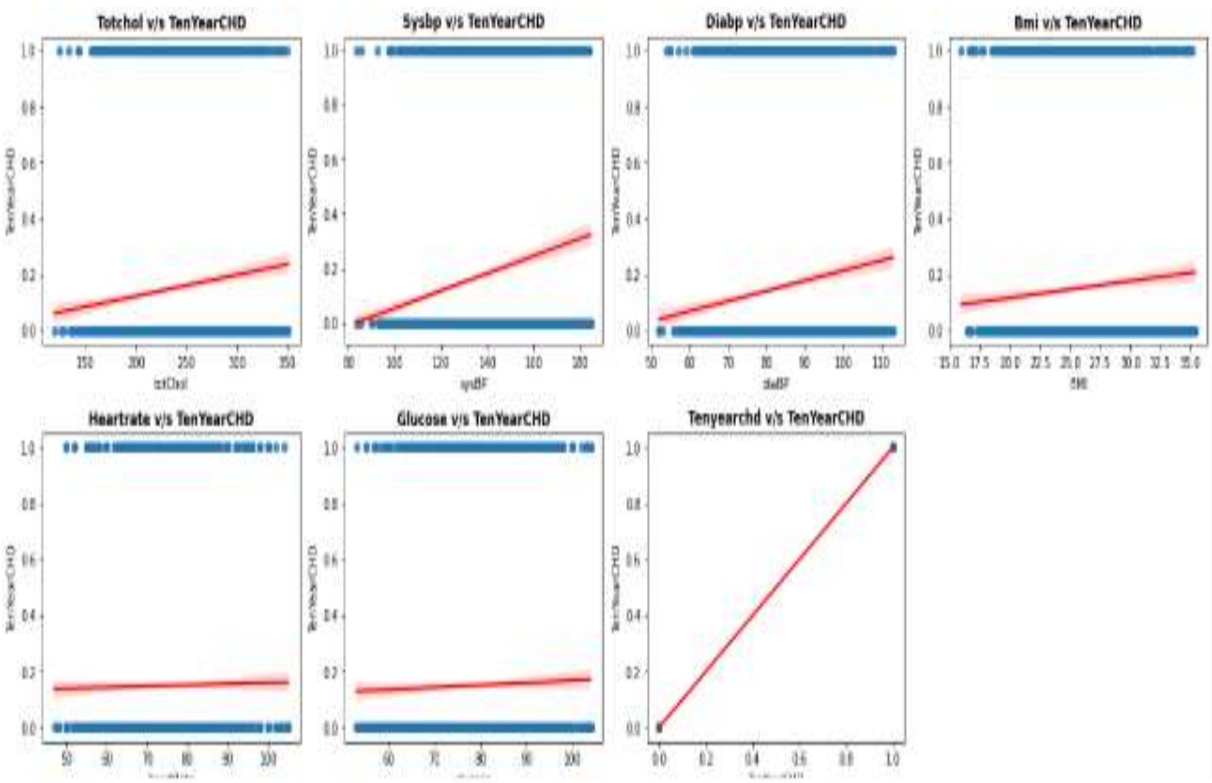


Figure-8- Bivariate analysis

Then, multicollinearity treatment has been made to check correlation among various independent variables in order to draw a reliable statistical inference.

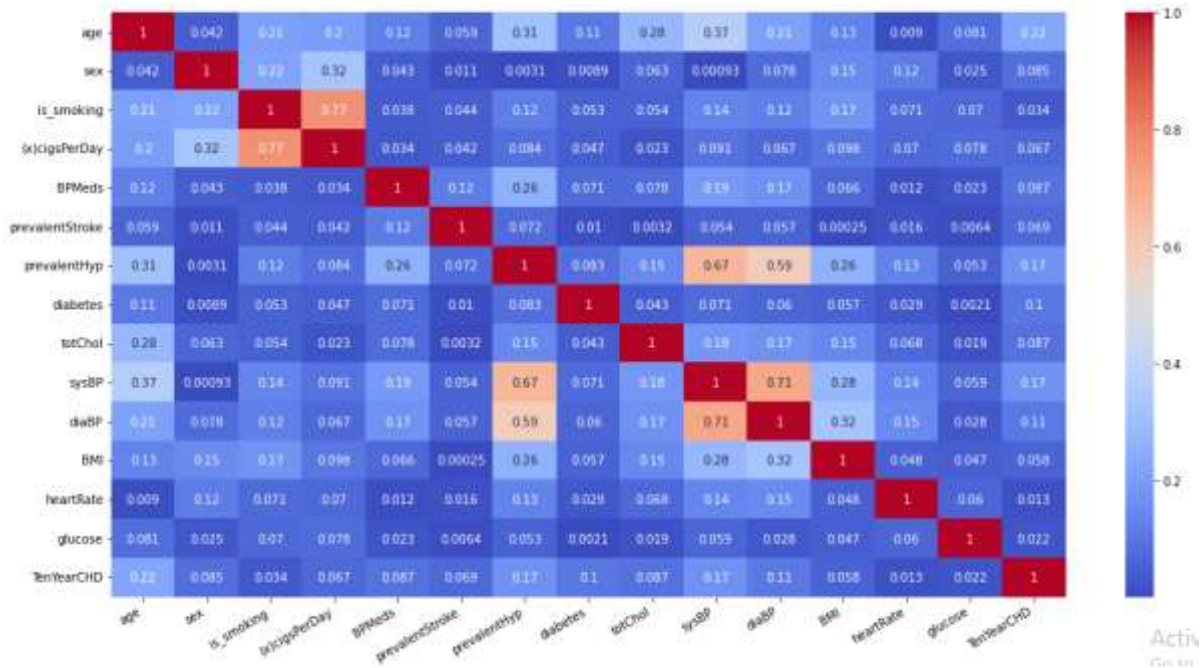


Figure-9- Multicollinearity graph

Variance Inflation Factor (VIF) has been used usually with a VIF score less than 5 (V.I.F = 1/1-R^2) and then considering the features with VIF score less than 10, the data left with the following features in the above, have been with the features like age, education, sex, cigs per day, prevalent HYP, BP meds, diabetes and prevalent stroke.

	variables	VIF
0	age	2.555058
1	sex	1.958771
2	(x)cigsPerDay	1.722460
3	prevalentHyp	1.669085
4	BPMeds	1.120401
5	diabetes	1.041588
6	prevalentStroke	1.023992

Fig-10-VIF score

In order to build the model, class imbalance issue has been solved using SMOTE and TOMEK links ; Confusion matrix is plotted for comparison of actual target values with predicted values.

		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative -	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive +	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Figure-11- Confusion matrix

Then Performance measurement has been made through AUC-ROC curve at various threshold settings.

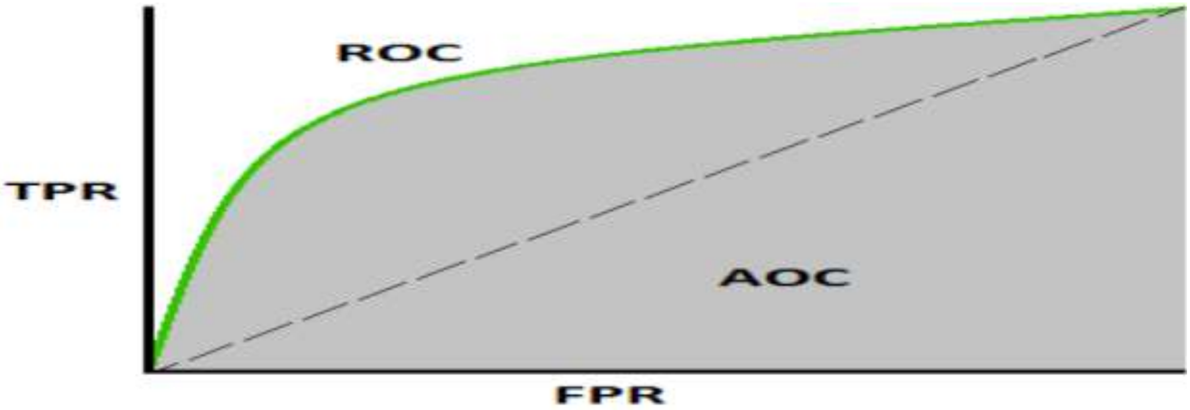


Figure-12- AUC-ROC curve

Classification report is made to get the accuracy score, precision, recall and F1 score in the form of Metrics. Feature importance is also assigned for both classification and regression problem

LOGISTIC REGRESSION

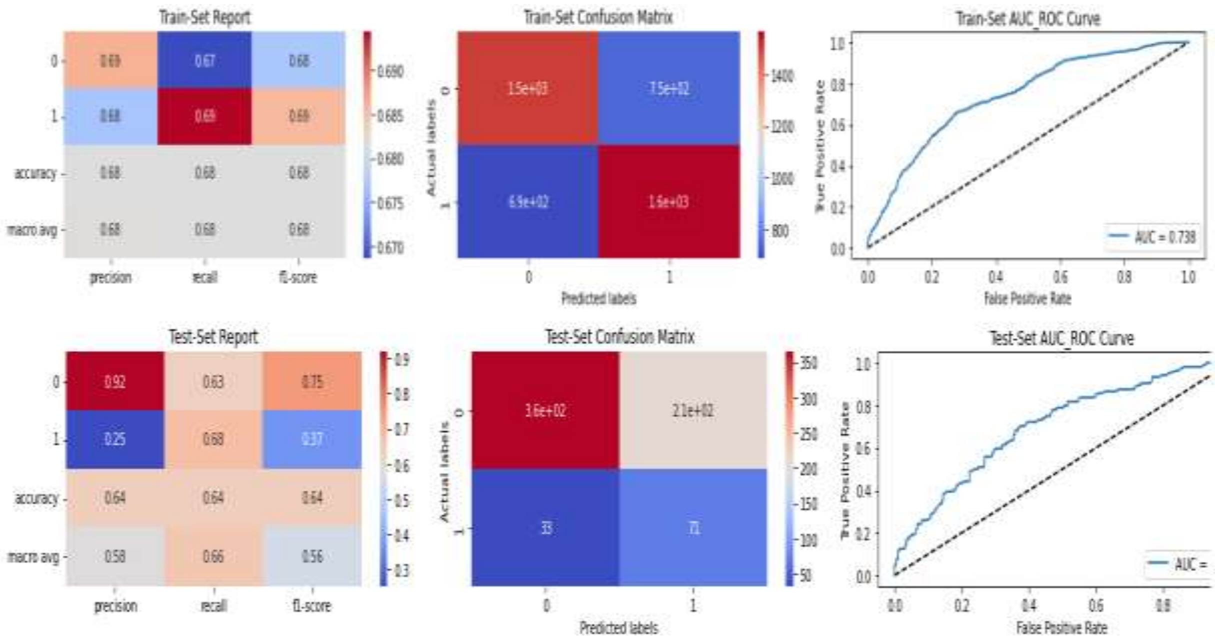


Figure13- Here it is showing that accuracy for logistic regression is 0.64, Precision is 0.25, recall is 0.66, F 1 score is 0.3

NAÏVE BAYES CLASSIFIER:

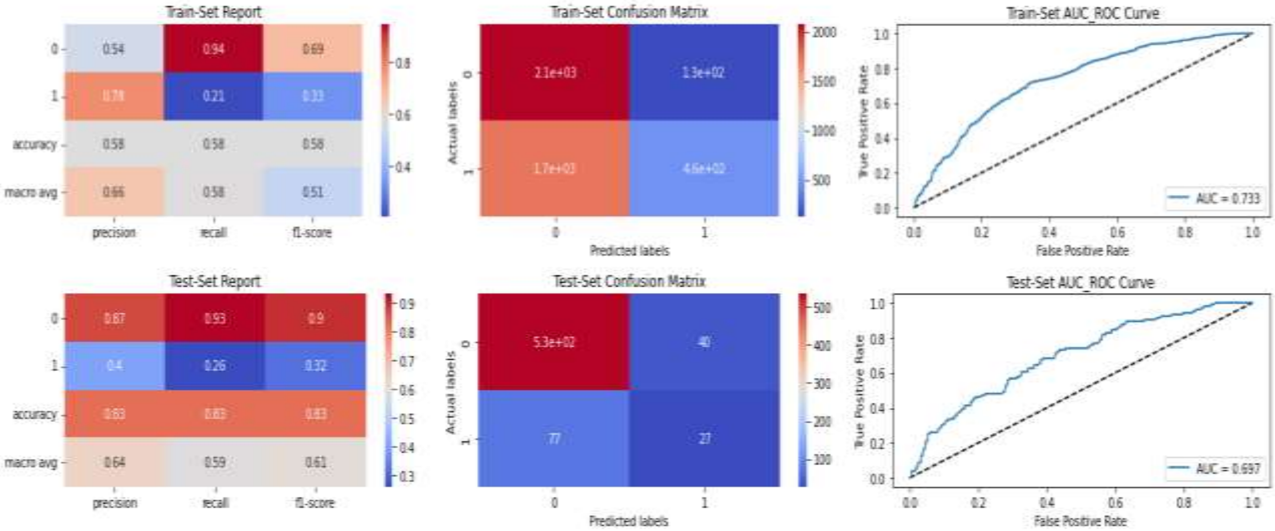


Figure14- Here it is showing that accuracy for Naïve Bayes Classifier is 0.83, Precision is 0.4, recall is 0.26, F 1 score is 0.32

Support vector classifier:

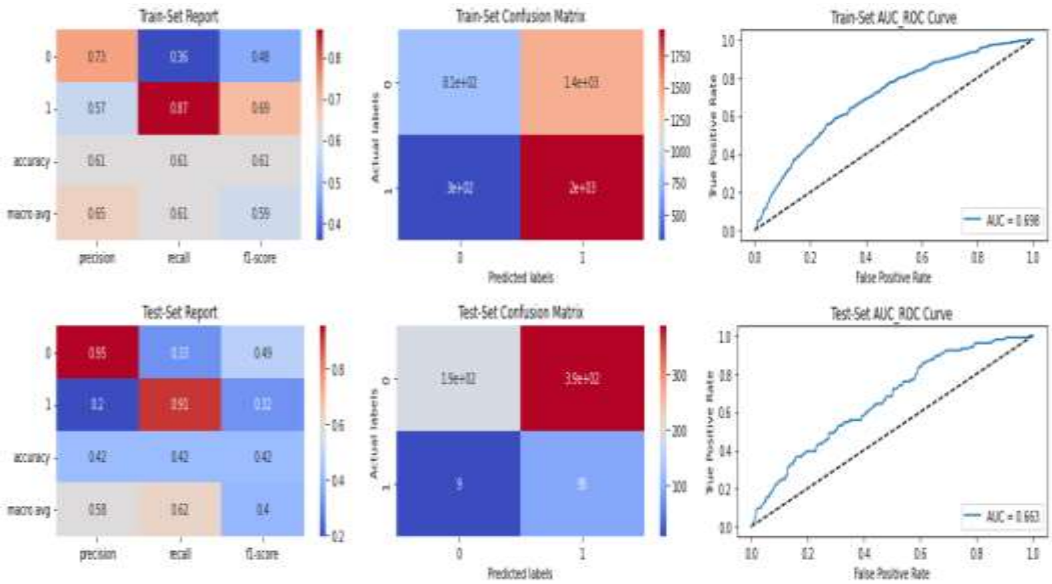


Figure15- Here it is showing that accuracy for support vector classifier is 0.42, Precision is 0.2, recall is 0.91, F 1 score is 0.32

Random Forest classifier:

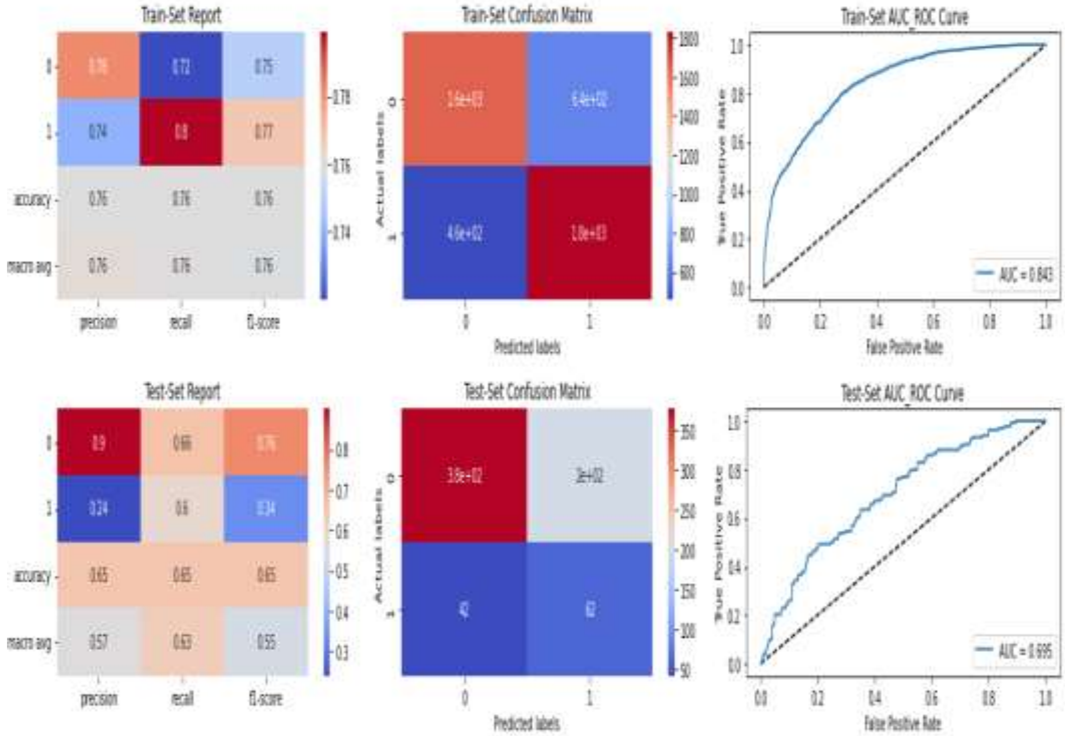


Figure16- Here it is showing that accuracy for random forest classifier is 0.65, Precision is 0.24, recall is 0.55, F 1 score is 0.34

KNN Classifier:

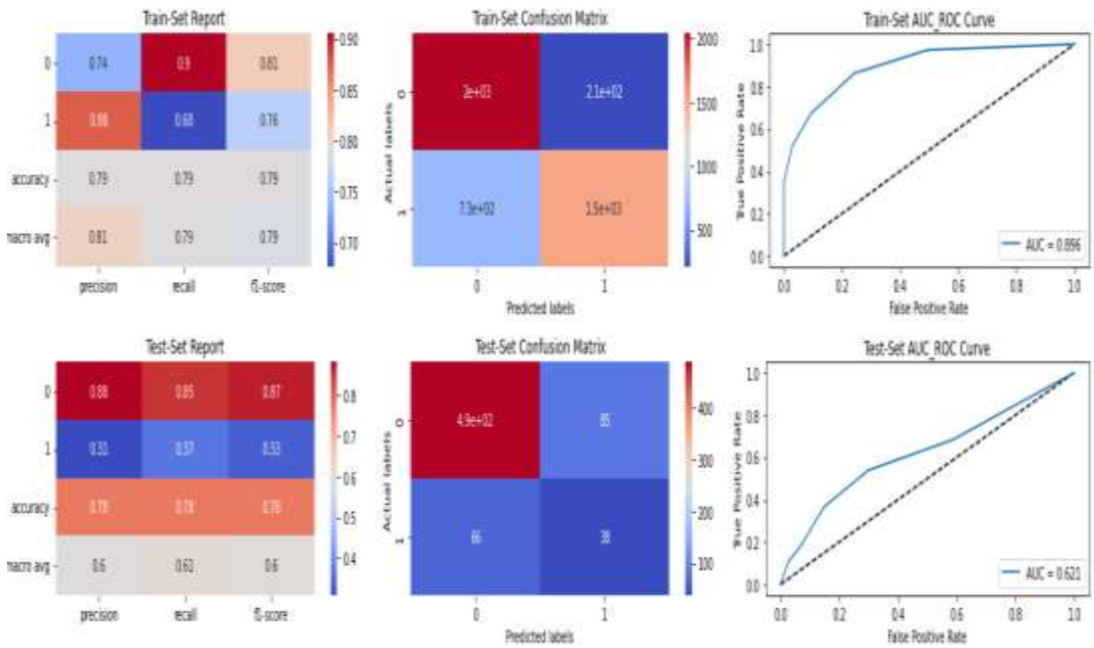


Figure17- Here it is showing that accuracy for KNN classifier is 0.78, Precision is 0.17, recall is 0.38, F 1 score is 0.23

XGBoost classifier:

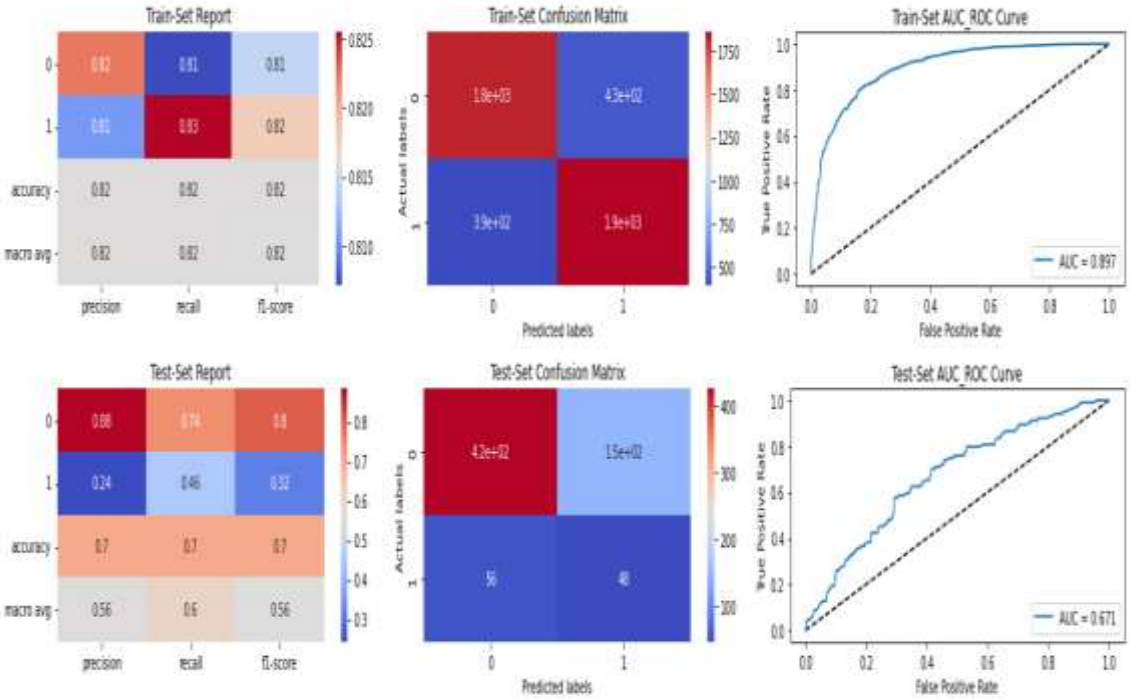


Figure18- Here it is showing that accuracy for XG Boost classifier is 0.7, Precision is 0.28, recall is 0.5, F 1 score is 0.36

CONCLUSION:

In conclusion, we have addressed the issue of class imbalance in the training data set by adding synthetic data points. Our study shows that the high performance of the models on the training set is not due to overfitting but rather a mismatch in the data distribution between the training and test sets.

We also found that the choice of algorithm depends on the specific needs of the patient. For patients without heart disease, a high precision is desired, making the Naive Bayes Classifier (NBC) a suitable algorithm. For patients with heart disease undergoing treatment for other conditions, an algorithm with high recall, such as Support Vector Machine (SVC), is recommended to avoid overlooking the presence of heart disease.

In cases where the patient's correct diagnosis of heart disease is not critical and other diseases are equally important, F1 score can be used to identify other ailments. Our study shows that Logistic Regression and Extreme Gradient Boosting (XGBoost) algorithms have high F1 scores, making them suitable for such cases.

Overall, the choice of algorithm depends on the specific needs of the patient and the priorities of the treating physician. By selecting the appropriate algorithm, doctors can improve the accuracy of their diagnoses and provide better treatment options for their patients.

===

REFERENCES

[1] C. D. Mathers, A. Lopez, C. Stein, D. Ma Fat, C. Rao, M. Inoue, and others. 2001. "Deaths and Disease Burden by Cause: Global Burden of Disease Estimates for 2001 by World Bank Country Groups." In *Disease Control Priorities Project Working Paper 18*. Bethesda, MD.

[2] M. McClellan, D. Kessler, A Global Analysis of Technological Change in Health Care: The Case of Heart Attacks—The TECH Investigators. Health Affairs. 1999;18(3):250–255.

[3] S. Yusuf, R. Peto, J. Lewis, R. Collins, P. Sleight, Beta Blockade during and after Myocardial Infarction: An Overview of the Randomized Trials. Progress in Cardiovascular Diseases. 1985; 27(5): 335–371.

[4] Ishaq, S. Sadiq, M. Umer et al., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021

[5] B. K. Turkmenoglu and O. Yildiz, "Predicting the survival of heart failure patients in unbalanced data sets," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, Istanbul, Turkey, 2021.

[6] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 2020.

[7] R. B. Sr D'agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel, "General cardiovascular risk profile for use in primary care: The Framingham Heart Study," Circulation, vol.117, no. 6, pp. 743–753, 2008

[8] L. Ma, J. Yang, H. B. Runesha, T. Tanaka, L. Ferrucci, S. Bandinelli and Y. Da, "Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heartstudy data," BMC Med. Genet., vol. 11, Art. no. 55, 2010.

[9] D. Levy, M. G. Larson, E. J. Benjamin, C. Newton-Cheh, T. J. Wang,S. J. Hwang, R. S. Vasan, and G. F. Mitchell, "Framingham Heart Study 100K Project: Genome-wide associations for blood pressure and arterial stiffness," BMC Med. Genet., vol. 8, no. Suppl 1, Art. no. S3, 2007