Predictive Model on Cardiovascular-Disease using Ensemble Techniques - Supervised Classification Algorithms in Health Care Industries

[1]Dr.Bechoo Lal
Department of Information Technology,
Western College, University of Mumbai,
Mumbai, India,
E-mail: drblalpersonal@gmail.com

[2]Shivanand Pokhriyal
PGP- Data Science
Purdue University - USA
E-mail: shivanand.pokhriyal@outlook.com

ABSTRACT

Background: Cardiovascular Diseases are rapidly growing in all kinds of aged people beyond the gender specification because of unhealthy diet, stress, lack of regular exercise, busy work schedule, weight, alcoholic and some other factors which are creating cardiovascular diseasesproblems. It is typical chronic illnesses with a high recurrence rate in health-related industries. In some of the cases, a heart attack occurs suddenly without any omens. Patients typically live in their homes rather than in hospitals and are often unable to access medical care in an emergency. Cardiovascular disease leads to a significant difficulty for the doctors to know the patient's status in time, and it becomes one of the significant reasons for death.

Method: The researcher proposed a predicting model using ensemble techniques with different machine learning algorithms and to optimizethe accuracy of predictive model on cardiovascular-diseases problem. It is used and explored the accuracy of decision tree classifiers, random forest, K-neighborhood and support vector machine classifiers to find out which predictive model is more efficient for the accuracy point of view to predict cardiovascular-diseases problem in patients based on their health previous history. The proposed predictive model is more accurate and approved that support vector machine (SVM) gives good result than another predictive model. So the researcher accepted and adopted support vector machine SVM classier to predict whether a person has cardio or not with good accuracy of 73%.

Results: The accuracy of predictive model shows that Decision Tree Classifier 63%, Random Forest Classifier 70%, K-Neighbors Classifier 72%, and finally Support Vector Machine (SVM) classifier produced the 73%. As per the data analysis of accuracy level of algorithms, we can see that the SVM and KNN are performing better than other models. The researcher found thatSVM gives a better result than other models,in terms of accuracy score, Auc score and F1_score, and SVMgives good result. So the researcher decided to accept and adopt support vector machine (SVM) classifier to predict whether a person has cardio or not with good accuracy of 73%.

Conclusion:Finally the researcher concluded that the patient's age, weight,stress, cholesterol, smoking habits, alcoholic behaviors, irregular exercise,and unbalanced diet are the significant factors for a cardiovascular-disease problem in the real world. The proposed predictive model is more accurate and approved at Auc score and F1_score SVM gives good result. So the researcher assured that support vector machine (SVM) classifier to predict whether a person/patient has cardio or not with good accuracy of 73%.

Keywords: CVD, KNN, SVM,cardiovascular,

I.INTRODUCTION

Cardiovascular diseases (CVDs) are the top causes of death globally. An estimated 17.9 million lives are lost each year (31% of all deaths worldwide) due to CVDs. The term 'cardiovascular disease' covers a range of diseases which affect the heart or the blood vessels, such as coronary heart disease.Machine learning and

Artificial Intelligence models can be implemented to assist healthcare systems in the accurate detection and diagnosis of CVD in patients. This could both relive pressure on healthcare systems, as well as allowing for more specialized and early treatment in those who are predicted to suffer with CVDs (World Health Organization, 2021).

This prediction can be made using a machine learning algorithm, developed through using the already existing medical data produced by healthcare. Trends in the data, both with those suffering and not suffering with CVDs, can be analyzed and identified, thus creating an accurate and sustainable machine learning model to predict the likelihood of someone suffering from CVDs (World Health Organization, 2021).The researcher described three key factors in the model that causes cardiovascular responses to stress: the need for laboratory-life generalizability, the role of the link between natural exposure and individual response conditions, and the importance of both time exposure and cardiovascular response in response. On the assumption that laboratory-based cardiovascular reactivity predicts responses to the natural environment (Schwartz, Amy R et al, 2003).

Heart disease leads to great difficulty for doctors to know a patient's condition in time, and it becomes one of the most important causes of death. To overcome these problems, the solution needs to design, implement, and adequately verify the relevant basic information. To overcome these challenges, real-time patient health data can be viewed (Patro, S.P. et.al., 2020). The researcher examined the related functionality of the models (discrimination, measurement, and reorganization) and the potential for outcome selection and optimism for liking newly introduced models and models developed by the authors. The use of alternative bias, numerical calculations, and qualification classification statistics was consistent (George C M Siontis et.al., (2012).

Although traditional risk factors pose a high risk of CVD, predictable models such as the Framingham risk score (FRS) are naturally limited in their ability to discriminate between people who will or may not experience severe CVD cases (2,3). Clearly, the greatest risk left for CVD exists in individuals who are independent of these risk factors (4,5). The aim of this study was to test whether the inclusion of homocysteine (Hcy) in a model based on coronary heart disease (CVD) improves risk classification. (Vikas Veeranna et al, 2010).

II. LITERATURE REVIEW

Manpreet Singh et.al, (2016) proposed a more accurate model predictive model forcardiovascular disease (CVD) could be a leading cause of death for older men and women. Various research projects have used machine learning / data algorithms to predict CVD, but these methods are plagued by a) ambiguity of predictable model structure, b) inability to present human intelligence, and c) inadequate data [1].Y et.al., (2016) examined the relationship between coronary heart disease (CVD) and 10-year risk of coronary heart disease (CVD) in China and developed statistical models for predicting CVD risk in Chinese with profiles of various features risks [2]. Schwartz, Amy R et. Al., (2003) stated that cardiovascular regeneration is thought to mediate the relationship between stress and heart disease. The researcher found only limited production and suggested that cardiac responses to stress could be better understood when tested in nature. Examination of anticipation and retrieval measures, by the magnitude of the response, may lead to a more effective model of depressive disorder [3].

Patro, S.P. et.,(2020) emphasized that heart disease is a common chronic disease with high levels of recurrence. This research project aims to develop a framework for predicting heart disease using high-risk substances according to differentiating mechanisms; closest neighbors are K, Na Bayve Bayes, vector support machine, and Lasso and ridge regression algorithms. The vector support mechanism provides 92% accuracy, and the F1 accuracy is 85% [4]. George C M Siontis et.al., (2012) tested the evidence for comparisons of models used to predict cardiovascular risk and collect data for their predictive performance. Data extraction data from the study design, risk models tested, and results. In 32 comparisons, the effect was applied to the initial design of only one type of comparable model, and of the 25 comparisons (78%) results-compliant models had a better position under the receiver using the feature curve [5].

Johanna A. G. Damen et.al., (2016) highlighted models that predict the occurrence of CVD in the general population. The predictive horizon was not specified for 49 models (13%), and in 92 (25%) important details were not available to make the model used to predict each risk. 132 advanced models (36%) were verified

externally and only 70 (19%) were independent investigators. Model performance varied and measures such as discrimination and measurement were reported in only 65% and 58% of external confirmation, respectively [6] .US Kaptoge (2012) studied people without known heart disease, estimating that under current treatment guidelines, CRP level assessment or this fibrinogen in people at moderate risk of cardiac event can help prevent one additional event over a ten-year period in 400 to 500 people tested. The addition of information on high-density lipoprotein cholesterol to a predictor model of heart disease that includes age, smoking status, blood pressure, diabetes history, and total cholesterol level increased C-index, risk discrimination rate, by 0.0050. The researcher estimated that among 100,000 adults 40 years of age or older, 15,025 individuals would initially be at risk of cardiovascular risk if the most common risk factors were used to calculate risk [7].

Farida Meghatriaac, and Omar Belhamiti (2021) study a predictive model of cardiovascular risk (CVD) and type 2 diabetes mellitus (T2DM) in obese people and investigate its impact on lifestyle changes, lifestyle and risk factors are included. to explain the usefulness of the advanced model; show the role of a healthy lifestyle (diet, regular exercise, smoking, and increasing alcohol consumption in moderation) in reducing cardiovascular disease (CVD) and can significantly reduce the risk of developing type 2 diabetes (T2DM) [8]. Duen-Yian Yeha Ching et.al., (2011) stated that cerebrovascular disease has been ranked second or third among the top ten causes of death in Taiwan and has caused an estimated 13,000 deaths annually since 1986. When cerebrovascular disease occurs, it leads not only to high medical care costs, but also to death. But, by looking at preventive medicine, it is necessary to build a predictive model to improve an accurate diagnosis of cerebrovascular disease [9].

Michael J. Pencina et. Al., (2009) analyzed common risk factors that predict the severity of severe CVD in extended follow-up. 30-year-old activities provide additional details of completing those 10-year-old occupations the average risk (male hypertension and anti-hypertensive treatment, total and HDL cholesterol levels, smoking, diabetes) are initially estimated. , was highly correlated with severe CVD episodes and remained significant when updated regularly at follow-up. The body weight index was associated with a 30-year risk of severe CVD only in models who did not review the risk factors. The performance of the Model was very good as it showed the discrimination confirmed by cross c = 0.803 and measuring chi-square = 4.25 (p-value = 0.894). In contrast, the prediction of a 30-year risk based on different uses of 10-year activities showed inadequacy [10].

Farshad Farzadfar (2019) analyzed that heart disease is the leading cause of death worldwide and a major public health concern. Although several models of cardiovascular risk predictors have been produced for a wide range of people over the past decade, the legitimacy of these types is a cause for concern. These types can lead to risk degradation, which can lead to non-existent risk situations. As a result, providing a valid model for the classification of heart disease and general heart disease has become a major priority for scientists and organizations working in this field [11]. Nina P. Paynter et. Al., (2021) emphasized that cardiovascular diseases because of high blood pressure, smoking status, diabetes, blood cholesterol levels, high sensitivity to active protein, and a family history of premature myocardial infarction). The researcher proposed a predictive model based on traditional risk factors, high C-functional sensitivity, and family history of premature myocardial infarction did not contribute to model bias as measured by ic-index (0.807 to 0.809) and did not improve Net Reclassification Improvement points (-0.2%; P = 0.59) or points for Integrated Discrimination Development (0.0; P = 0.18) [12].

Donald M. Lloydjones et.al., (2006) stated that proteins that work with C (CRP) as a diagnostic tool for predicting cardiovascular disease (CVD). The researcher found that there is no clear evidence that, for most people, CRP adds a significant amount of the above assumptions given to measuring risk using traditional CVD risk factors. The use of CRP can add to the risk estimates in a limited set of middle-aged people predicted by Framingham risk score [13]. Nina Friis-Møller et.al., (2010) the researcher emphasized that patients with infected people receiving antiretroviral combination therapy may face metabolic disorders, which may increase their risk of heart disease (CVDs).The models performed well with the area below the receiver curve of 0.783 (range 0.642-0.820) for myocardial infarction, 0.776 (0.670-0.818) for heart disease and 0.769 (0.695-0.824) for CVD. Models with more accurate measurement of outcomes in groups than Framingham scores [14].

Benjamin S. Wessler et. Al., (2015) stated that CPMs are available for comprehensive screening of cardiovascular conditions, with significant reductions in literature. CPM 168 human samples, and 79 models of patients with heart failure. There are 77 different index / result classification of the de novo models in this database, 450 (63%) reported c-statistic and 259 (36%) reported specific data on the scale [15]. Anne B. Newman et. Al., (1999) analyzed Peripheral arterial disease (PAD) in the legs, which is indisputably measured by the ankle-arm (AAI) index associated with cardiovascular disease (CVD) and its severity. The risk of coronary heart failure (related risk [RR] = 1.61) and total mortality (RR = 1.62) in those who do not have CVD initially but with low AAI remained significantly higher after correction of cardiac risk factors. In-hospital PAD events occur months to years after AAI ratings, with a modified RR of 5.55 (95% CI, 3.08 to 9.98) for those at risk of event events. Significant statistical decline in survival was observed at 0.1 [16] each.

Keaven M. Anderson et. Al., (1991) analyzed an estimate of the endpoints of cardiovascular endpoints, based on estimates of several known risk factors. Subjects (n = 5573) were actual and offspring studies at the Framingham Heart Study, aged 30 to 74, and initially had no heart disease. Statistics predicted risk factors for the following: myocardial infarction, heart disease (CHD), death from CHD, heart disease, and death from heart disease. The parametric model used has been found to have several advantages over existing standard retrieval models. In contrast to asset disposal, it can provide forecasts for varying lengths of time, and opportunities can be more accurately expressed than the equivalent risk model [17].

Sanne A E Peters et. Al., (2011) develops a predictable number of additional tests that can be obtained mainly in people with no symptoms of cardiovascular risk. The function of predicting image markers with discrimination, measurement and re-editing is removed. Additional predictive value, measured by the difference of c-index, FMD, CIMT, carotid plates or CAC from 0.00 to 0.01 for FMD, from 0.00 to 0.03 for CIMT, ranges from 0.01 to 0.05 for carotid plaque and from 0.05 to 0.13 for CAC. Although the definition of moderate cardiovascular risk varies across all studies, NRI was significantly higher in those with moderate cardiovascular risk [18]. Milne, RJ. et. al., (1997) developed 5-year models of cardiovascular risk reduction and the effectiveness of monotherapy costs from a public perspective. The model shows that in the lowest case (60-year-old non-smokers and non-smokers with a SBP of 160mm Hg and a five-year risk of cardiac risk of 12%), celiprolol (271 mg / day) is 2 - effective much more than atenolol (77.4 mg / day) in reducing the risk of coronary event and is equally effective in reducing the risk of cerebrovascular event. Both drugs are internationally priced in the treatment of 5-year-old patients with complete cardiovascular risk greater than 10% and are more expensive for those patients at higher risk levels for complete heart disease [19].

P Brindle et. Al., (2006) analyzed the efficacy of the framingham risk level varies widely between individuals and the evidence supporting the use of risk factors for primary blood vessels is scarce. Predicted 10-year risk of CVD and CHD (review A), and fatal or non-lethal cardiovascular or coronary events, risk levels, complete or cardiovascular risk, prescription dose reduction and lifestyle changes - Related behavior (review B). In Review B, four randomized controlled trials in people with high blood pressure or diabetes did not find strong evidence that cardiovascular risk assessment by a physician improves health outcomes [20]. Vikas Veeranna et.al., (2010) stated that cardiovascular disease (CVD) is the leading cause of death in the United States, accounting for more than a third of all deaths (1). What we see is novel and has shown that high Hcy concentration predicts future CVD and CHD events in contrast, people representing U.S. adults. The researcher found that plasma Hcy levels improve risk prediction when added to FRS, enabling reorganization of the population at “moderate risk” of CHD events [21].

III. PROBLEM STATEMENT AND RESARCH OBJECTIVES

This research study is based on “predictive model on cardiovascular-disease using ensemble techniques - supervised classification algorithms in health care industries”, the researcher formulated the research problems and current research issues on cardiovascular disease in real world. It is one of the significant research issues in Health Industries. The researcher stated that some of the research objectives such as:

1. To study the factors which are causes for cardiovascular disease problem in the real world.

2. To develop a predictive model based on Cardiovascular-Disease using Ensemble Techniques - Supervised Classification Machine learning algorithms.
3. Analyze and evaluate the accuracy of predictive model.

IV. NATURE OF DATASET AND VARIABLES / FEATURES DESCRIPTION

In this research study the researcher used the cardiovascular-disease datasets which is accessed from the kaggle.com. In cardiovascular-diseasedatasets having7000 recordsof different patient'scases historywith 13 attributes/features are there in. The features are showing the patients id, age, gender specification, weight, height, alcoholic, smoking etc, whichplays a significant role for cardiovascular-disease problem in current health industries.

Data RangeIndex: 70000 entries, 0 to 69999
Data columns: (total thirteen columns)/ features
Data Source: Kaggle.com

Table 1.1: cardiovascular-disease datasets

S. No.	Features/Variables	Non-Null Count	Datatype
0	id	70000 non-null	int64
1	age	70000 non-null	int64
2	gender	70000 non-null	int64
3	height	70000 non-null	int64
4	weight	70000 non-null	float64
5	ap_hi	70000 non-null	int64
6	ap_lo	70000 non-null	int64
7	cholesterol	70000 non-null	int64
8	gluc	70000 non-null	int64
9	smoke	70000 non-null	int64
10	alco	70000 non-null	int64
11	active	70000 non-null	int64
12	cardio	70000 non-null	int64

During the thorough analysis of datasets / data pre-processing the researcher found that cardiovascular-disease datasets are having some outlier's data which would create a misleading statistic for the predictive model. The researcher identifies the cardiovascular-disease datasets and analyzed that 1451entries are having outliers out of 7000, which are represented 2% of overall data representation, the researcher decided to remove the outliers. Now we have 7000-1451= 5549 data considering for the training and testing to the predictive model on cardiovascular-disease classifications.

The researcher used the data set with 5549 entries where 3884 for training dataset and entries for testing dataset for classifier to classify the data in terms of 70:30 ratio. The entire datasets are represented as follows:

1. Let the complete datasets be represented D={D1 ,D2, D3, D4…….D5549},
2. Let the training datasets be presented as Train={D1 ,D2, D3, D4……D3884},
3. Let the test data be represented as Test={D326 ,D327, D328, D328…….D1665},

The splitting of dataset is based on the random manner, system automatically divided the two different datasets in terms of ration 70:30 manner which is one of the standard mapping parameters to train and test the dataset in machine learning model.

V. RESEARCH DESIGN AND METHODOLOGY

In this research study the researcher used the different machine learning algorithms to build the optimum accuracy of predictive model on cardiovascular-disease using ensemble techniques - supervised classification algorithms in health care industries. The researcher used and explored the decision tree classifiers, random forest, K-neighborhood and support vector machine classifiers to find out which predictive model is more efficient for the accuracy point of view to predict cardiovascular-diseases problem in patients based on their patient's previous cases history.

5.1. DECISION TREE CLASSIFIER

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

$$E(T,X)=\sum_{c\in X} P(c)E(c) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

T=Target Variables whereas X= Features/ Independent Variables

Decision trees are used for handling non-linear data sets effectively. The decision tree tool is used in real life in many areas, such as engineering, civil planning, law, and business. Decision trees can be divided into two types: categorical variable and continuous variable decision trees. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

Algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

5.2. RANDOM FOREST CLASSIFIERS

Random forest is a supervised learning algorithms which is used for both classification as well as regression. It is used for classification problems. As the researcher emphasized that forest is made up of trees and more trees means more robust forest. Random forest algorithms creates decision tree on the data samples and then get the prediction on each of them and finally select the best solutions by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over fitting by averaging the result.

Algorithm:

Step-1: First start the selection of random samples from a given dataset.

Step-2: Next, this algorithm will construct a decision tree for every sample then it will get the prediction result from every decision tree.

Stept-3: In this step, voting will be performed for every predicted results.

Step-4: At last, select the most voted prediction result as the final prediction result.

5.3. K-NEAREST NEIGHBORS CLASSIFIER

It is an algorithm which classifies a new data point based on its proximity to other data point groups. Higher the proximity of new data point from one group, higher is the likelihood of it getting classified into that group. Distance between data points is measured by distance metrics like Euclidean distance, Manhattan distance, Murkowski distance, mahalanobis distance, tangential distance, cosine distance and many more.

For the data points X and Y with n features

X=(x1, x2, x3, x4, ……xn) and Y=(y1, y2, y3, y4,……yn)

$$D(X,Y) = (\sum_{1=1}^{n}(|xi - yi|)^{p})^{1/p} \text{……..(2)}$$

For data points X and Y with n features

Using the distance metrics, it is easy to create neighborhood of n closest neighborhood to the new data point.. to get the class of the new data point, we look at the class groups which have more data points in the created neighborhood and the class groups which are closer to our new data point compared to other groups in the neighborhood, based on these two factors we determine the class of our new data point.

5.4. SUPPORT VECTOR MACHINE (SVM)

Predicting qualitative responses in machine learning is called a classification. Support vector machine (SVM) is the classifier that maximize the margin. The main objective of this classifier is to find a line or (n-1) dimension hyper plain that separate the two classes present in the n- dimensions space.

$$G(x)=w^{T}x+b$$

Maximize k such that

$$- w^{T}x + b>=k \text{ for } di==1\text{……..(3)}$$

$$- w^{T}x + b <=k \text{ for } di==-1\text{……(4)}$$

Value of g(x) dependent of ||w||

1. Keep || w|| =1 and maximize g(x) or ,
2. g(x) >=1 and minimize ||w||

We will use above two approach and formulate the problem as

$$\Phi(w)=1/2\ w^{T}w - \text{minimize}\text{……(5)}$$

Subject to $di(w^{T}x + b) >=1$ for every i

Integrating the constant in lagrangian form

$$\text{Minimize: } J(w, b,a)=1/2\ w^{T}w - (\sum_{1=1}^{n} aidi(|wTx + b|) + (\sum_{1=1}^{n} ai \text{)…(6)}$$

⇨ $ai[di(w^{T}xi + b_0) -1]=0$

⇨ Either $ai=0$ or $di(w^{T}xi + b_0)=0$

It implies that non-zero lagrangian coefficient correspondent to the support vector data points. Using the above equation, we can generalize that:

$$J(w, b,a)= - \sum_{1=1}^{n} ai + \frac{1}{2}\ wTw - wT \sum ai\ dixi - b + \sum_{1=1}^{n} aibi \quad \textbf{.....(7)}$$

J represents the dual form which is only dependent on a as rest are all know scalars. A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. So you're working on a text classification problem.

VI. RESULTS AND DISCUSSION

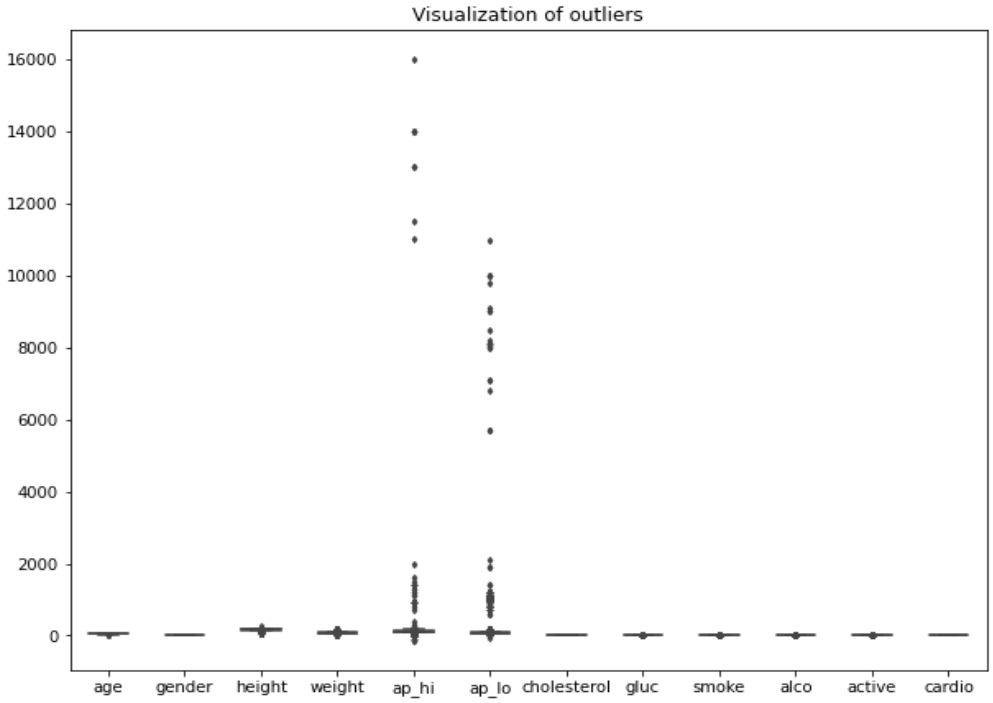

Fig.1.1: Analysis Report of Outliers in Datasets-cardiovascular-disease

The above data statistics shows that Ap_hi, and Ap_lo attributes are having outliers which are too far from the expected data points. The history of outliers are the noisy data which creates a misleading statistic during the data processing. The researcher used the data pro-processing technique to remove the outliers andhaving complete dataset 7000 on cardiovascular disease whereas 1451 are represented as outliers, that is total 2% from overall datasets which is very minimum, so the researcher decided to remove the outlier's datasets. The further data operational data 5549 will be consider for the training and testing data to train the predictive machine learning model (fig.1.1).

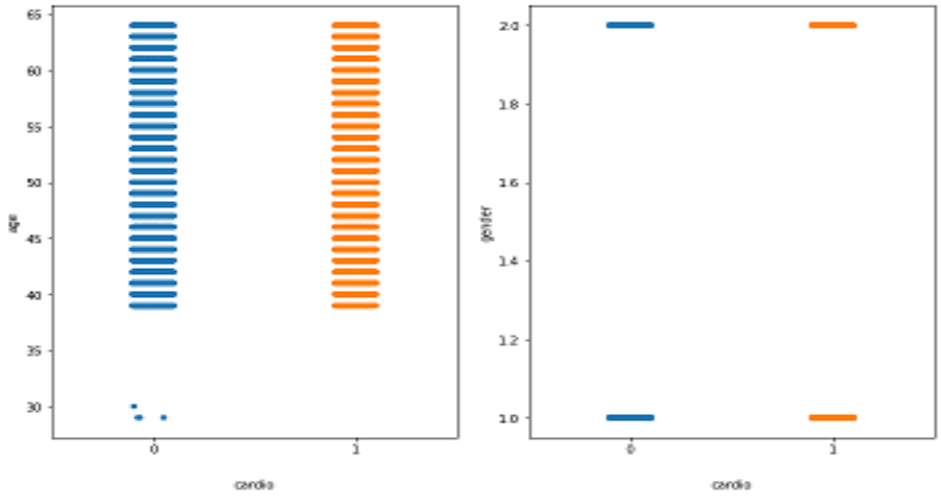

Fig.1.2: Statistics of Age and Gender Specification- cardiovascular-disease Datasets

The above data analysis report shows that age and gender specification with respect to cardiovascular-disease problems in patients. As per the data specification the patients age is one of the significant attributes for cardiovascular-disease problems, the fig.1.2 represented that as the age is growing the cardiovascular problem will be chances of increased in patients whereas gender specification does not affect or associated with any kinds of cardiovascular-disease problems in patients (fig.1.2).

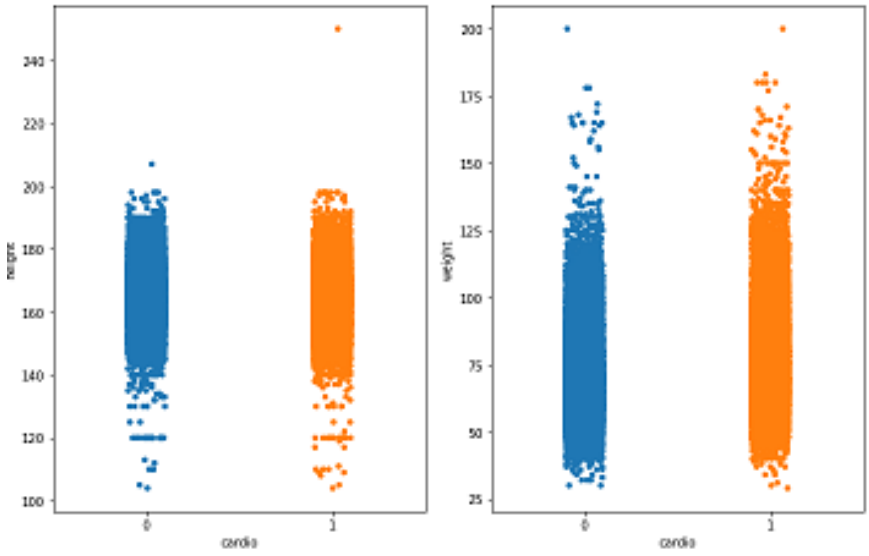

Fig.1.3: Statistics of Weight and Height- cardiovascular-disease Datasets

The above data analysis report shows that weight and height with respect to cardiovascular-disease problems in patients. As per the data specification the patient's weight is one of the significant attributes for cardiovascular-disease problems, the fig.1.3 represented that as the weight is growing the cardiovascular problem will be chances of increased in patients whereas height attributes do not affect or associated with any kinds of cardiovascular-disease problems in patients (fig.1.3).

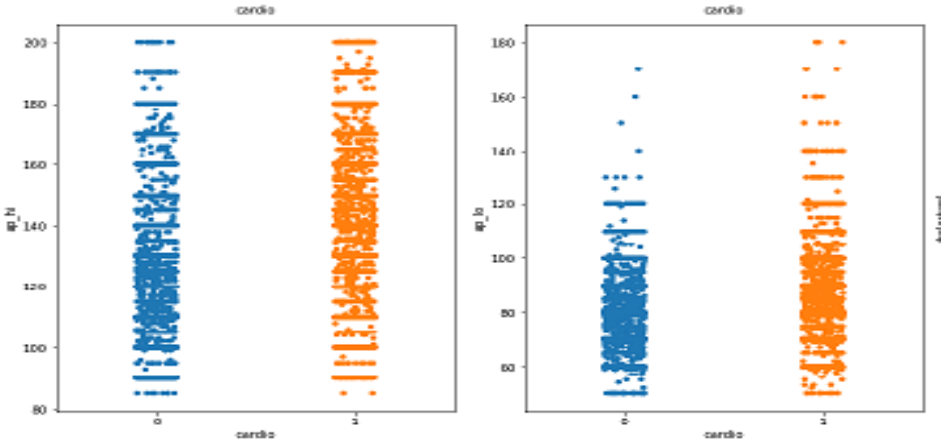

Fig 1.4: Statistics of Ap_lo and Ap_hi- cardiovascular-disease Datasets

The above data analysis report shows that Ap_lo and Ap_hi with respect to cardiovascular-disease problems in patients. As per the data specification the patients Ap_lo is one of the significant attributes for cardiovascular-disease problems, the fig.1.4 represented that as the Ap_hi is growing the cardiovascular problem will be chances of increased in patients where as Ap_lo attributes also have a significant affect or associated with any kinds of cardiovascular-disease problems in patients.In above graph we can see that if ap_lo is more than 120

there is high chance of cardio. Ifthe age is less 38 there is very less chance or no chance of cardio, if the weight is more than 175 there is achance of cardio(fig.1.4).

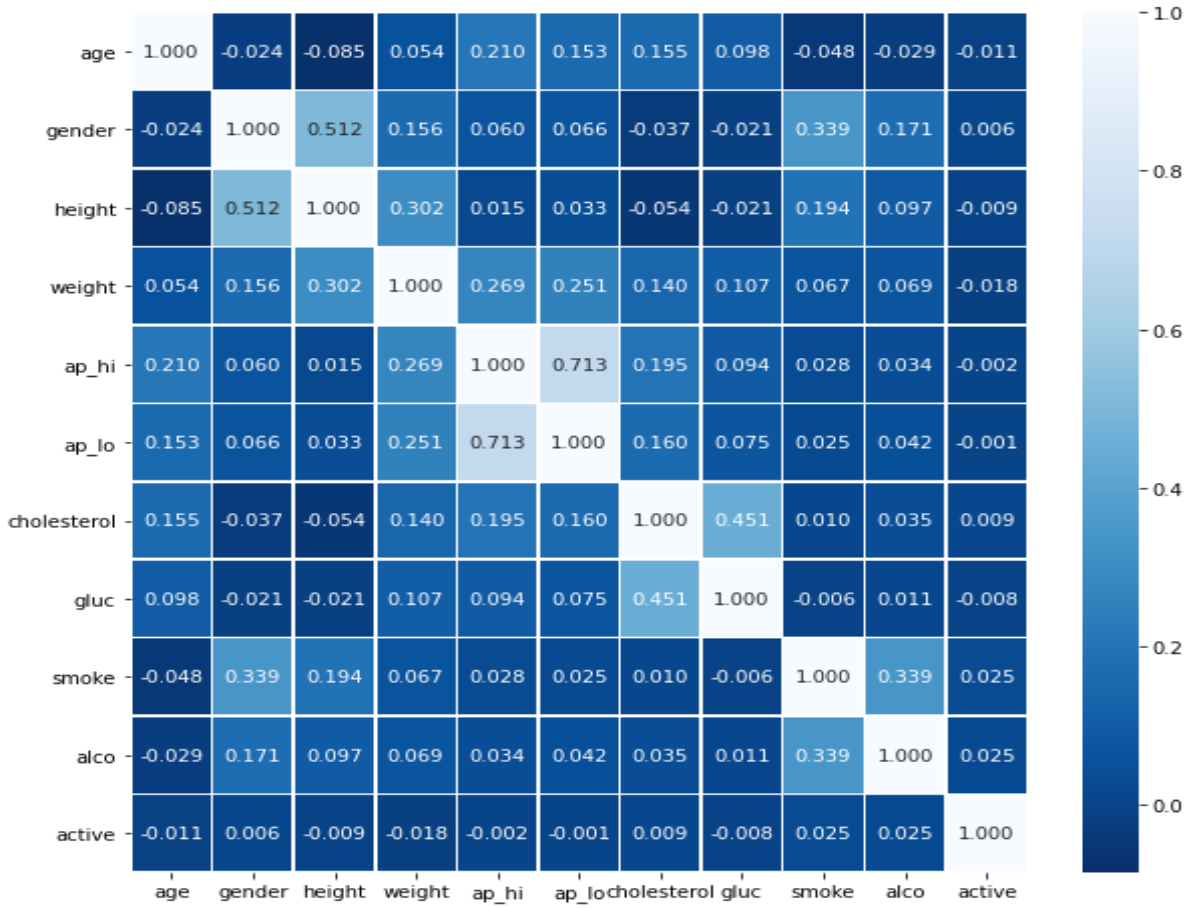

Fig. 1.5: Correlation Matrix between Data Attributes/ Features

The above heat map shows the correlation matrix and represented the relationship between attributes/ features in patients' datasets. As we can see that there is not much collinearity between any data, but the association of weight, smoke is 0.067, cholesterol, and glucose 0.075, AP_hi and glucose 0.094 are showing a strong association and have a significant impact on cardiovascular-disease problems in patients (Fig. 1.5).

Table 1.2: Accuracy Score of Predictive Model

S. No	Supervised Machine Learning Classifiers	Accuracy
1	Support Vector Machine (SVM) Classifier	0.732215
2	Decision Tree Classifier	0.630489
3	K-Neighborhood Classifier(KNN)	0.723705
4	Random Forest Classifier	0.703282

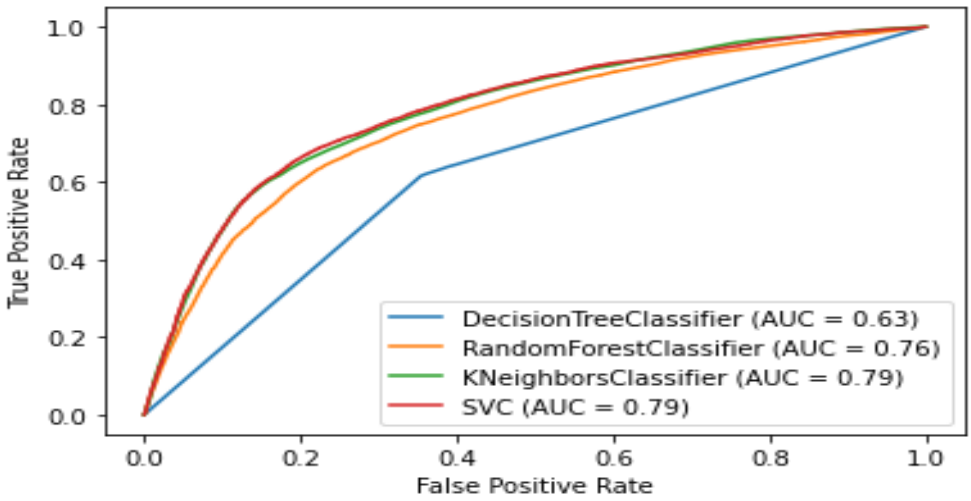

Fig. 1.6: ROC Curve of Different Predictive Model Accuracy

The above data analysis report shows that the accuracy of "predictive model on cardiovascular-disease using ensemble techniques - supervised classification algorithms in health care industries" using different supervised machine learning algorithms such as Decision Tree Classifier, Random Forest Classifier, K-Neighbors Classifier, and Support Vector Machine (SVM) classifier.The different accuracy of predictive model shows that Decision Tree Classifier 63%, Random Forest Classifier 70%, K-Neighbors Classifier 72%, and Support Vector Machine (SVM) classifier produced the 73% (Table 1.2). As per the data analysis and interpretation we can see that the SVM and KNN are performing better than other models. The researcher found thatSVM gives a better result than other models,in terms of Accuracy score, Auc score and F1_score SVM gives good result. So the researcher accepted and adopted the Support Vector Machine (SVM) to predict whether a person has cardio or not with good accuracy of 73% (Fig.1.6).

DISCUSSION

In the research study the researcher focused oncardiovascular-disease problems and predictive model on using ensemble techniques - supervised classification algorithms in health care industries. The researcher stated that the perception of other researchers on cardiovascular-disease problems are such as: Manpreet Singh et.al,(2016) emphasized a novel approach to tackle these issues and design a very robust and reasonably accurate model on Structural Equation Modeling (SEM) and Fuzzy Cognitive Map (FCM). The researcher used Canadian Community Health Survey, 2012 data set to test our approach. The designed model has 79% area under the ROC curve and 74% accuracy. The researcher believing the adding more attributes and having an expert heart specialist panel would further improve the accuracy of the system. Wang Y et.al.,(2016) analyzed by univariate and multivariate methods. (1) The 10-year accumulated coronary event rates were 1.41% for men and 0.62% for women and stroke event rates, 2.02% for men and 1.37% for women. (2) The predicated absolute risk of CVD increased with the coexistence of risk factors. There were synergic effects among different CVD risk factors. Different combination of major risk factors had different impact on CVD risk.

The causal model of cardiovascular responses to stress should generalize to the real world, assess interactions between individual predispositions and environmental exposures, and focus on sustained pathogenic exposures and responses (Schwartz, Amy R et.al., 2003).An analysis of the stability of stationary solutions is also obtained to theoretically confirm the mathematical validity. Numerical simulations presented to explain the usefulness of the developed model, they show the role of a healthy lifestyle (diet, regular exercise, smoking, and increasing alcohol consumption above moderate levels) in alleviating the burden of cardiovascular disease (CVD) and can significantly reduce the risk to develop type 2 diabetes mellitus (T2DM (Farida Meghatriaac, and Omar Belhamiti, 2021).

Finally the researcher emphasized thatthe patient's age, weight, smoking habits, alcoholic behaviors, unbalanced diet are the significant attributes/features for acardiovascular-disease problem in real world. The researcher

giving assurance the proposed predictive model is more accurate and approved at Auc score and F1-score, confusion matrix to predict whether a person has cardio or not with good accuracy of 73%.

VI. CONCLUSION

In this research study the researcher enhanced and optimized the accuracy of machine learning algorithms to predict cardiovascular-disease problems using ensemble techniques -supervised classification algorithms in health care industries. The researcher used 7000 different patients'cases history records associated to cardiovasculardisease with12 factors / features which are significant to draw a conclusion on patient's cases history. The researcher used the different supervised machine learning algorithms such as Decision Tree Classifier, Random Forest Classifier, K-Neighbors Classifier, and Support Vector Machine (SVM) to classify the patients whether they are having cardiovascular-disease problems or not. The accuracy of the different predictivemodelverifiesthat the Decision Tree Classifier 63%, Random Forest Classifier 70%, K-Neighbors Classifier 72%, whereas Support Vector Machine (SVM) classifier produced 73%. The data analysis and interpretation shown that SVM and KNN work better than other models. The proposed predictive model is more accurate and approved at Auc score and F1-score, where Support Vector Machine (SVM) gives good result. So the researcher stated that the proposed predictive model based on support vector machine (SVM) classifier is one of significant and capable to predict whether patientsare having cardiovascular-disease problems or not with good accuracy of 73%and it can be further deployed in Health Care Industries.

REFERENCES

1. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, Vijay K. Mago (2016). Creating a predictive model for heart disease using Structural Equation Model & Fuzzy Cognitive Map, 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE Xplore: 10 November 2016, DOI: 10.1109 / FUZZ-IEEE.2016.7737850.
2. Y, Liu J, W,M, Qi Y, Xie W, Li Y, Sun J, Liu J, Zhao D (2016). Life risk of for young Chinese and middle-aged people: Chinese Regional Collection Study., Europe PMC, J Hypertens, 34 (12): 2434-2440, Dec 2016.
3. Schwartz, Amy R, A, Gerin, William, Davidson, Karina W, Pickering, Thomas G. (2003). From the Causal Model of Cardiovascular Responses to Stress and the Development of Cardiovascular Disease, copyright © 2003 by American Psychosomatic Society, Psychosomatic Medicine: January 2003 - Volume 65 - Issue 1 - p 22-35, doi: 10.1097 / 01.PSY.0000046075.79922.61.
4. Prato, S.P., Padhy, N. & Chiranjevi, D. (2020). An effective living speculation model for the ambient for predicting heart disease using supervised learning. Evol. Intel. (2020). https://doi.org/10.1007/s12065-020-00484-8.
5. George C M Siontis, Ioanna Tzoulaki, Konstantinos C Siontis, John P An Ioannidis (2012). Comparison of established risk assessment models for heart disease: a systematic review, BMJ 2012; 344 doi: https://doi.org/10.1136/bmj.e3318.
6. Johanna A. G. Damen, Lotty Hooft, Ewoud Schuit, Thomas P A Debray (2016). Types of predictors of heart disease risk for many people: a systematic review, BMJ 2016; Opportunity 353: https://doi.org/10.1136/bmj.i2416.
7. S Kaptoge (2012). C - reactive protein, Fibrinogen, and Cardiovascular Disease Prediction, New England Journal of Medicene, N Engl J Med 2012; 367: 1310-1320, DOI: 10.1056 / NEJMoa1107477.
8. Farida Meghatriaac, and Omar Belhamiti (2021). Predictability model for cardiovascular disease and type 2 diabetes in obese people, Volume 146, May 2021, 110834, Elsevier Pvt. Ltd.
9. Duen-Yian Yeha Ching, Hsue Chengb, Yen-WenChenb (2011). An example of cerebrovascular hypothesis using a data mine, Expert Systems with Applications, Volume 38, Issue 7, July 2011, Pages 8970-8977, Elsevier, https: //doi.org/10.1016/j.eswa.2011.01.114.
10. Michael J. Pencina, Ralph B. D'Agostino, Martin G. Larson, Joseph M. Massaro, and Ramachandran S. Vasan, M (2009). Predicting Thirty Years of Risk of Heart Disease: Framingham Heart Study, COVID-19,Informaltion, and 2009 Jun 23; 119 (24): 3078–3084, doi: 10.1161 / CIRCULATIONAHA.108.816694.

11. Farshad Farzadfar (2019). Types of cardiovascular risk predictors: challenges and ideas, Lancet Global Health, September 02, 2019DOI: https: //doi.org/10.1016/S2214-109X (19) 30365-1.
12. Nina P. Paynter, Daniel I. Chasman, Julie E. Buring, Dov Shiffman (2021). /10.7326/0003-4819-150-2-200901200-0000.
13. Donald M. Lloyd-Jones, ScM, Kiang Liu, Lu Tian, Philip Greenland (2006). Subsequent Review: C - reactive protein Tests in Risk Prediction for Cardiovascular Disease, The Annals of Internet Medicene, July 4, 2006, https://doi.org/10.7326/0003-4819-145-1-200607040-00129.
14. Nina Friis-Møller, Rodolphe Thiébaut, Peter Reiss, Rainer Weber (2010). Predicting cardiovascular risk in infected patients: data collection on adverse effects of Anti -Drug Study, European journal Prevention and Cardiac Rehabilitation, Volume 17, Issue 5, October 1, 2010, Pages 491-501, https://doi.org/10.1097/HJR.0b013e328336a150.
15. Benjamin S. Wessler, Lana Lai YH, Whitney Kramer, Michael Cangelosi, Gowri Raman, Jennifer S. Lutz, and David M. Kent (2015). Diseases, July, 2015, doi.org/10.1161/CIRCOUTCOMES.115.
16. Anne B. Newman, Lynn Shemanski, Teri A. Manolio, Mary Cushman, Maurice Mittelmark, Joseph F. Polak, Neil R. Powe, David Siscovick (1999). Weapons Index As Predictor of Cardiovascular Disease and Mortal in Cardiovascular Health Study, Cardiovascular Health Research Group, Mar 1999, https://doi.org/10.1161/01.ATV.19.3.
17. Keaven M. Anderson, Patricia M. Odel, Peter WFWilson, William B. Channel (1991). Cardiovascular Diseases, American Heart Journal, Volume 121, Issue 1, Part 2, January 1991, Pages 293-298, Elsevier Pvt. Ltd.
18. Sanne AE Peters, Hester M den Ruijter, Michiel L Bots, Karel GM Moons (2011). Development of the risk of coronary heart disease by subclinical atherosclerosis: a systematic review, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands, http://dx.doi.org/10.1136/heartjnl-2011-300747.
19. Milne, R.J., Vander Hoorn, S. & Jackson, R.T.(1997). Predictable Example of Health Benefits and Success Costs of Celiprolol and Atenolol in Major Prevention of Cardiovascular Disease in Hepatitis Patients. Pharmacoeconomics 12, 384-408 (1997). https://doi.org/10.2165/00019053-199712030-00010.
20. P Brindle, A Beswick, T Fahey, S Ebrahim (2006). Accuracy and impact of risk assessments on major cardiovascular prevention: systematic reviews, cardiac medicine, 2006, http://dx.doi.org/10.1136/hrt.2006.087932.
21. Vikas Veeranna, Sandip K. Zalawadiya, Ashutosh Niraj, Jyotiranjan Pradhan, Brian Ference, Robert C. Burack, Sony Jacob, and Luis Afonso (2010). Homocysteine and Reclassification of Cardiovascular Disease Risk, Journal of the American College of Cardiology, Vol. 58, Issue 10, PP. 1025-1033, 2011 Aug.
22. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

Abbreviations Used

SVM	:	Support Vector Machine
KNN	:	K-Neighbors Classifier
CVDs	:	Cardiovascular Diseases
WHO	:	World Health Organization
T2DM	:	Type 2 Diabetes Mellitus
CRP		Clear Evidence That, For Most People,
PAD	:	Peripheral Arterial Disease
AAI	:	Ankle-Arm Index
CHD	:	Child Heart Disease

Authors Profiles

Bechoo Lal, PhD. became a Member (M) of IAENG: International Association of Engineers, USA with membership (108820) in 2010, a Senior Member (SM) in 2019. I am doctorate PhD in Computer Science, PhD-Information System from University of Mumbai, Master of Computer Application from Banaras Hindu University (BHU), PGP- Data Science from Purdue University, USA and Graduation in Statistics (Hons). Currently I am working as a Coordinator- IT department, Western College, University of Mumbai, Maharashtra, India. My research areas are data science, big data analytics and data mining. In addition to my research currently working on predictive model using different machine learning algorithms with high accuracy and efficient manner to reduce the computational and system complexity.

Shivanand Pokhriyal has more than 17 years of experience in Information Technology in various roles. Most of this work experience is around data-related fields.He has a postgraduate in Data science from Purdue University, USA. He is also a postgraduate in Information technology from Symbiosis University, Pune.His meticulous nature and love for data analysis keep him motivated for his research in the field.

…