Clinics in Surgery

Research Article
Published: 27 Sep, 2021

Prediction of Cardiovascular Diseases from Risk Factor: An Application of Machine Learning

Yousaf Ali Khan[1,2]*

[1]Department of Mathematics and Statistics, Hazara University Mansehra, Pakistan

[2]School of Statistics, Jiangxi University of Finance and Economics, China

Abstract

Aims: Heart diseases are the leading cause of high mortality and incapacity across the globe. Research in recent years confirmed that, the fees of heart illnesses-associated deaths have reduced in some medically advanced countries, however still high in less and medium medically advanced countries and this need critical attention. Regardless of the seriousness of heart illnesses in low- and middle-income nations, no interest given to the prevention of Cardiovascular Disease (CVD) associated risk factors in Continent of Asia, especially in my home land. Similarly, financial and political variability is hastening the costs of heart sicknesses within these countries. On the other hand, the domain of information mining (DM), which aim at extracting excessive-stage knowledge from raw statistics, provide exciting automatic tools in lots of subject of research.

Methods: This paper addressed the prediction of heart diseases from hazard elements through decision-making tree. This paper introduces data mining technique in public fitness with the aim to extract high-degree knowledge from raw data, which facilitates in prediction of heart diseases from risk factors and its prevention. The existing work intends to introduce new technique of risk elements in heart diseases using novel data mining strategies. Latest actual-international affected person's information (e.g. smoking, area of resident, age, weight, blood stress, chest pain, Low-Density Lipoproteins (LDL), High-Density Lipoproteins (HDL), blocked arteries became accrued by way of the use of questionnaire through direct interview technique from patients. Variable decision trees are constructed for cardiovascular disease records primarily based on chance factors and ranking of risk elements.

Results: The results show that there is correct prediction of Cardiovascular Disease (CVD) from risk factor, if records on chance factors are available. As direct results of this study, the use of tobacco, loss of physical exercise, and weight-reduction plan are the main factors playing vital role in the prediction of heart diseases, which is the most important reason of mortality in developing countries, especially in my country.

Conclusion: We gain ranking of endangerment factors through variable decision tree, which allows improving public safety, as properly in selection regarding heart diseases remedy and prevention. It additionally facilitates in guidelines making related to prevention of cardiovascular diseases risk factors in low- and middle-income nations.

Keywords: Machine learning; Heart diseases; Prevention; Decision tree; Risk factors; Prediction; Hybrid technique; Low-Density Lipoproteins (LDL); High-Density Lipoproteins (HDL)

OPEN ACCESS

Correspondence:
Yousaf Ali Khan, Department of Mathematics and Statistics, Hazara University Mansehra 23010, Pakistan, E-mail: yousaf_hu@yahoo.com

Received Date: *26 Jul 2021*
Accepted Date: *20 Sep 2021*
Published Date: *27 Sep 2021*

Citation:
Khan YA. Prediction of Cardiovascular Diseases from Risk Factor: An Application of Machine Learning. Clin Surg. 2021; 6: 3315.

Introduction

Evidence in this modern day shows that, globally, Cardiovascular Disease (CVD or heart diseases is the main purpose of life loss, and about eighty to eighty-six percent of these expiries arise in below average-earnings nations [1-4].

As of around 16 million expiries that occur due to Non-Communicable Diseases (NCDs), eighty-two percent are in developing countries and thirty-seven percent of those losses are associated with CVD [2-4]. Though, there may be a large version inside the humanity fees, consistent with intercourse, years live, culture, socio-economic status, and environmental vicinity. The universal heart diseases-associated expiry costs for males (age ≤ seventy years) is three times advanced than for females and twice in low income comparatively cheap regions than in prosperous zones [3]. Maximum South Asian countries, such as Nepal, Sri Lanka, Pakistan, Bangladesh, and India contain extra than 1/4

of the emerging realms and are diagnosed to have a better threat of Cardiovascular Disease (CVD) in comparison to different part of the globe [3]. A large population-based totally cohort observe diagnosed that the occurrence of CVD in South Asian patients become high in comparison to Chinese and Canadian patients [4,5]. There are several behaviors, such as social, organic, and psychological risk factors, that raise the CVD burden in below average earnings nations. The inter heart case-manipulate looked at it, and conducted a survey in fifty-two international locations all over Australia, Africa, America, Middle-East and Asia discovered nine modifiable participants for Acute Myocardial Infarction (AMI) that results in cardiovascular diseases. These modifiable chance factors encompass high blood pressure, diabetes, smoking, belly weight problems, mental index, lack of workout, loss of end result and greens, Apo lipoprotein B/ Apo lipoprotein A1 ratio, and the use of liquor [2,6]. Several research concerning the danger elements of cardiovascular diseases have been carried out, nevertheless, there is shortage of fiction in Pakistan situation. The Pakistan fitness examination body in 2016 concluded that in Pakistan, the threat element for non-communicable diseases is swelling [3].

This research, predict heart diseases from risk factor and rank risk elements in line with their importance using novel two variable decision tree machine learning approach that is a beneficial data mining method for classification, prediction with extra function of factors ranking.

The contributions of this paper are as follows

- In this paper, an extraordinary Data Mining (DM) method became introduce in the discipline of public health with the purpose of extracting excessive-level knowledge from raw data, to make efficient prediction of the future behavior.

- Efficient classification based on correlated variables of cardiovascular disease became accomplished which helps in predicting heart diseases from risk elements as well as importance of risk elements, which facilitates in polices making associated with prevention of coronary heart diseases hazard factors specially in low-income regions.

The remaining of the paper is organized as follows: Section 2 offers materials and methods, comparison of the method with existing ones and simulation studies. Section 3 provides growing decision trees primarily based on correlated variables, and discussion on outcomes. In the end, Section 4 concludes this research with suggestions and policy implementations.

Motivation

The advance progresses in facts technology bring about a big quantity of facts that desires to be analyzed and controlled to advantage beneficial statistics expertise to predict future conduct. Then again, the sector of Data Mining (DM), which aim at extracting high-stage know-how from uncooked facts, offer interesting automatic equipment in many studies fields one want to aid it to the fitness care area. A decision tree signifies a tree-established organization that plays a break up test in its inner knot and guesses an aim magnificence of a specimen in its child node. With their easiness and clearness, decision trees are extensively use in data reduction [2,3,7]. Two variable decision trees is a nonparametric copula based machine learning approach with a further characteristic of factor ranking, based on highly correlated variables using Maximal Information Coefficient (MIC) as classification index.

Comparisons of the proposed tree with traditional decision tree

In modern era a large amount of classification practices from both machines learning and statistical societies have been suggested by [8-10]. A renowned technique of organization is the orientation of decision trees (e.g., CART: [11]; ID3: [8] C4.5: [9]). A choice tree is a tree structure diagram involving of root nodes, child nodes, and branches. Every root-node characterizes a decision on a data feature or a function of data features, and correspondingly outgoing branch resembles to a potential conclusion of the occasion. Every single child-node signifies a class. In directive to categorize raw data, the classifier investigates the feature values of the sample beside the decision tree. A route is outlined from the root-node to a child-node, which grasps the class prediction for that model. Verdict trees can definitely be transformed into IF-THEN procedures and then used for policymaking [7]. The effectiveness of prevailing choice tree procedures (e.g., CART: [11]; C4.5: [8]), has been healthy recognized for comparatively trivial data sets [11,12].

Recently, several considerations on the generation of choice tree how to create it effective, consistent, precise and appreciated. Numerous researches on decision trees have been accomplished to hypothesis progressive arrangements of trees in direction to progress extra accuracy. Initially, it has been deliberate for the construction of root-node to be either univariate or multivariate. Second, there are numerous strategies how to advance their flexibility and consistency such as multiple decision trees. They assembled a large number of decision trees (up to 100,000) by altering sampling data. Then, they strained to discover finest calculation of ordering by balloting outcomes of multiple trees (random forest).

Usually, decision trees are constructed for single variable. More than one variable decision trees can classify more than one variable at a time at each root-node. Univariate decision trees are not unique and have accuracy problems; they are large in size and time consuming. Traditional decision tree use entropy for classification. Some time we need more than 100,000 trees called random forest to achieve the desired results.

In ordered to address all these issues related to decision trees, this research propose a novel nonparametric copula based decision tree for two random variables using mutual information coefficient as classification index. The dependence structure among variables at root-node and each child-node was explored through nonparametric copula density and the value of Mutual Information Coefficient (MIC) was determined at each test node. Higher difference of MIC value among factor levels from pre-specified value allow us to further classify the data and obtained the child-node. Continue this procedure until the last node and stop growing the branches of the tree, when the difference of MIC value is less than or equal to the pre-specified value of MIC. It is to be noted that in our proposed method the stopping criteria (the value of MIC) will be pre-determined and which is different for each data set depending upon the size, factor levels and dependence structure among variables/attributes. One of the many other essential features of two variable decision trees is, raking of factors according to their importance.

Copula-based decision tree for two random variables is proven useful data mining tool, which benefits over other existing decision tree due to its simplicity, uniformity, accuracy and efficiency. The proposed method of two variables copula-based decision tree assisted as a beneficial data-mining tool for classification, prediction and

ordering of factors by way of their importance in various fields of life sciences such as machine learning, banking and finance, health sciences, experimental and applied sciences.

Comparison of decision tree with other classification techniques

Copula-based decision tree:

- Use MIC for Classification
- Unique Decision Tree
- Based on association among variables.
- Applicable for two random variables only.
- Simple and efficient.
- Accurate and valid in both large and small data set.
- Ranking of factors according to their importance.
- Graphical presentation of hidden dependence is valid.
- Used for prediction, classification and ranking etc.

Traditional trees:

- Use Entropy for classification
- Use for one variable
- Complicated when the data is sufficient large
- Accuracy Problem
- Structure of tree is very complicated when data is large.
- Difficult to interpret.
- Not Unique

Random forest

- Use Cross Validation for Construction of Forest.
- Use Gini Index for classification.
- Need 50,000 to 100,000 trees for accurate result.
- Decision based on balloting.
- Very complicated and difficult to interpret.

Support vector machine:

- A supervised learning technique.
- Useful in case of large data set only.
- Use for prediction.
- A regression technique.
- Tuned parameter on testing data set.
- Applicable when the data matrix is non orthogonal.

Numerous alternatives to decision trees for data exploration are available in the literature, such as neural networks, nearest neighbor methods, support vector machine, naive Bayes, and logistic regression. Quinlan empirically compared decision trees to neural networks [9] and to genetic classifiers [13]. Author in [14] compared CART with multilayered perceptions and observed no difference in accuracy. Author in [15] compared neural networks and decision trees for analyzing Electrocardiograms (ECG) and outlined that no procedure is superior to the others. Author in [16] compared decision trees with back propagation neural networks on high dimensional data problems. They found that there was not much difference between both techniques.

Research issue

Owing to high variety of losses from heart sicknesses in south Asia in the last decades. There is a severe need, to evolved efficient prediction equipment which classify diseases information, extract high-degree knowledge from raw statistics and give the chance to accurately detecting sicknesses from different factors; which allows in enhancing the high quality of public health as well as exposure to prevention of coronary heart diseases.

Methodology and Data

Heart diseases data

Pakistan is dealing with a double load of both transmissible and non-infectious diseases. The 2013 international consignment of disorder file expected that the thirty out of hundreds of the global expiries are linked to cardiovascular diseases [17]. This unique move in disorder connected pattern will have extra inferences for aptness care carrier transport abilities and aid distribution. A few of the estimates about the commonplace infection amongst Pakistani mature residents consists of forty-one out of 100 high blood pressure, twenty-two percent tobacco use, eighteen percent excessive fat, twenty-one out of hundred weight problems, eight to ten percent diabetes mellitus, and dyslipidemia (men, thirty-four percent; ladies, forty-nine percent), 16 and 2.8 percent knock [1]. These guesses are rising within the state, and the price of non-communicable and transmittable diseases is nearly identical. This epidemiologic change has an effect at the spreading of the adaptable chance elements of heart diseases in mother land, such as extended pressure stages, unnatural consuming behaviors, inactive routine and rise in smoking charges.

A fact for heart disease in Pakistan is incomplete; populace research was led in 1965 and 1973 that confirmed that the superiority of coronary heart disorder is among zero percent and four out of hundred in countryside and concrete zones. A result of 1994 countrywide fitness survey of Pakistan on health complications suggests an excessive occurrence of threat elements for CVD in both countryside and metropolitan populations [6,18]. But, direct data on (e.g. age, sex, area of residence, Blood Stress (BP), Body Mass Index (BMI)/weight, smoking, low and high-density lipoprotein, chest pain, blocked arteries associated features) of heart sufferers aren't available. Consequently, statistics from 321 sufferers who visited Ayyub Medical Complex, Abbottabad Pakistan in daytime for test-up at some point of the month of September 2019 by questionnaires through direct interview approach are been gathered for this research. The questionnaire was reviewed via expert heart illnesses doctors (e.g. heart specialist) and helped in amassing a few precise data like Low-Density Lipoproteins (LDL), High-Density Lipoproteins (HDL) and blocked arteries etc. The very last model contained 12 questions in a single A4 sheet.

Copula decision tree

Information mining is the abstraction of understood, earlier unidentified and rotationally valuable facts from figures. Also, it is extraction of big database into beneficial records or facts. Data Mining is continually inserted in techniques for finding and describing structural styles in information as a tool for assisting and makes prediction. Among many different type of techniques, decision tree is

one of the most popular classification technique used for classification and prediction problem. A decision tree characterizes a tree-shape organization that accomplishes a cut up, look at in its inside knob and forecasts an objective class of a sample in its child knob. With their straightforwardness and limpidity, decision trees are mostly used in lots of decision-making fields on a daily life [7,19,20].

Novel bivariate nonparametric copula based decision tree is a useful facts mining approach construct for two random variables, which are extraordinarily correlated. It dependence structure is explored by means of copula using Maximal Information Coefficient (MIC) as classification index [7]. Copula based decision tree works on the principle of correlation, therefore once we have a pair observation, first, we need to have scrutiny of whether or not there may be a correlation between the variables or not. If sure, then initiate nonparametric copula density estimation; estimate the density of paired observation, and obtain the contour plot (a graphical exploration of correlation between two variables). The graphical exploration will explain how to cut out the pair variables in an awful way; by taking association among them in account, one subgroup could have a very high correlation while the alternative has a very low correlation. To determine the uniqueness between the two measures of dependence, you will get the factors of classification. Repeat this process at every child node of the tree and cross down. Pre-specified the minimal value of MIC as stopping criteria for tree branches' that rely upon the dataset you have. In this way, one may acquire a decision tree for highly correlated random variables. The main idea of generating a copula-based decision tree using the MIC as the classification index can be illustrated as below in Figure 1:

Unique aspect of copula-based decision tree over traditional decision tree

Following are the advantages of copula-based decision tree over a traditional decision trees such as CART: [11]; ID3: [8] C4.5: [9].

1. It is for two random variables.

2. Nonparametric copula-based exploration of data under consideration.

3. Factor based classification. Where each child-node represent important factor.

4. Rank the factors according to their importance.

5. Use Maximal Information Coefficient (MIC) as classification index instead of entropy.

6. Pre-specification of MIC value for stop growing tree branches.

7. Obtain unique decision tree.

8. Time efficient and small in size with high accuracy.

The concept of copula: Copula is a Latin word which means a link, tie or bound. Copulas are multivariate distributions function whose marginal's are uniform at the interval (0.1). Copulas are important amongst statistician for two reasons.

Firstly, it provides a scale-free measure of dependence and secondly, a starting point for assembling relations of bivariate distribution. Copula concept is primarily based on famous Skylar's theorem dated back [4], which stated that multivariate distribution function may be disintegrated into the marginal's and a copula, which detentions the dependence among variables. Because of the wider variety of copula applications, copula attracted the attention of researchers in the world and become establishing device in many fields such as applied statistics and machine learning [5,6].

Assume that random variables $X_1, X_2, \ldots, X_p$ with joint cumulative distribution function

$F(x_1, x_2, \ldots, x_p) = P(X_1 \leq x_1, \ldots X_p \leq x_p)$ and marginal cumulative distribution function

$F_j(x)=P(X_j \leq x)$ for all j=1,2, … .p

Then copula can be defined as:

$$F(x_1,x_2,..,x_p) = C[F_1(x_1),F_2(x_2), \ldots, F_p(x_p)] \quad (1)$$

A Copula $\mathcal{C}$ is unique joint distribution of these marginal distributions i.e. $F_1(x_1),F_2(x_2), \ldots, F_p(x_p)$, if and only if $F_j(x)$ are continuous. One among many others important features of copula based estimation is that extrapolation of marginal distributions can be separated from the dependence structure. As copulas are not directly observable, estimation of the copula density "c" can be done in two steps. First, we need to estimates the marginal $F_1(x_1),F_2(x_2), \ldots,$

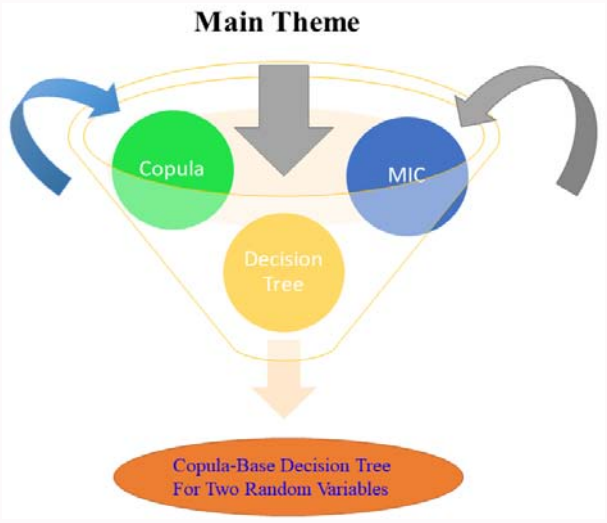

Figure 1: Conceptual Framework of the Copula based decision tree.

$F_p(x_p)$ and then from these marginal we estimate the copula density. For estimation of copula density, three approaches exist in literature. one may additionally, recall parametric procedure for copula density "*c*" and estimates the parameters by means of Maximum Likelihood Estimation (MLE) -method. Regardless of the fact that a rich literature of parametric models for estimation of copula density is to be had, though as copula is not directly observable and live in a hidden dependence structure, that is why parametric modeling for copula density has high chance of misspecification [17]. Another method of estimating copula density is semiparametric; in which one assume parametric models for copula and non-parametric model for marginal [7].

Alternatively, in non-parametric estimation of copula density one expect non-parametric models for each marginal and copula, nonparametric estimation of copula density resolves the problem of misspecification and consequently gives greatest generality. A number of nonparametric strategies for estimation of copula can be found in literature (e.g. multivariate empirical distribution and marginal empirical distributions method, smother estimation based on kernel estimation method, B-spline technique, reflect-mirrored image method, transformation technique, beta kernel smoothing approach and local linear kernel approach) among one particular class is kernel estimation. They are a regular tool for investigation and extensively used in various disciplines. But crucial problem with density valuation is that a copula and its density are defined on a compact cube $[0,1]^3$ because of this that boundary bias associated with kernel bend valuation may be present and one should have to be carefully address this problem. In bivariate case, it's far important to ensure stable estimator of a copula density over entire region, particularly close to the corners [0,0] and [0,1] [7,17].

Probit transformation and copula density estimation: It's far-flung difficult to estimate the density of kernel copula c of (U, V) directly due to the constrained nature of its comfort I = $[0,1]^2$. Because of this, we define

$$S=\Phi^{-1}(U) \text{ and } T=\Phi^{-1}(V) \quad (2)$$

Here Φ is usual Gaussian cdf and Φ-1is the Probit transformation. Given that each U and V are $U_{[0,1]}$, S and T both are standard normal variables, which, however, does not longer mean that the vector (S, T) is bivariate normal. With the intention to harness the case if the copula of the joint cdf of (S, T), say f_{ST}, is the Gaussian copula, that is, if the copula C of F_{xy} itself is the Gaussian copula, as copulas are invariant to growing changes in their margins (Nelsen, 2006, theorem 2.4.3). The impression is that, if c(u, v) >0 Lebesgue- a.e. over I, (s, t) has unconstrained aid R^2 and estimating its density f_{ST} cannot be afflicted by boundary problems. Further, because of its normal margins, we assumed that f_{ST} to be nicely-behaved, and its estimation turns out to be smooth. Specially, below slight assumptions, f_{ST} and its partial derivatives as much as the second order will be visible to be uniformly bounded on R^2, even within the case of unbounded copula density c. As the copula of FST is C, S~N (0.1) and T~N (0.1), we can write Skylar's theorem for (S, T) as:

$$F_{ST}(s, t) = P(S \le s, T \le t) = C(\Phi(s), \Phi(t)), \forall (s, t) \in R^2 \quad (3)$$

Upon differentiation with reference to s and t, the joint density f_{ST} of (S, T) is discovered to be:

$$f_{ST}(s, t) = c(\Phi(s), \Phi(t))\, \varphi(s)\varphi(t) \quad (4)$$

where φ is the standard normal density. Overturning this appearance, we obtain

$$c(u,v) = \frac{f_{ST}\left(\Phi^{-1}(u)\right),\left(\Phi^{-1}(v)\right)}{\phi\left(\Phi^{-1}(u)\right)\phi\left(\Phi^{-1}(v)\right)} \quad (5)$$

Thus, for any (u, v) ∈ $[0,1]^2$, fST of f_{ST} on R^2 is the estimator of the copula density inside of I, *viz*.

$$\hat{c}^{(\tau)}(u,v) = \frac{\hat{f}_{ST}\left(\Phi^{-1}(u)\right),\left(\Phi^{-1}(v)\right)}{\phi\left(\Phi^{-1}(u)\right)\phi\left(\Phi^{-1}(v)\right)} \quad (6)$$

Symbol (τ) in above expression refers to the indication of Probit transformation. $\hat{c}^{(\tau)}$ has interesting properties and removes all boundary biases that is why we utilized estimation of bivariate kernel copula density through Probit transformation in this research work [18]. For complete assessment of copulas and non-parametric bivariate copula density estimation through Probit transformation see [7,21-27].

Measures of functional dependence: Extract useful information to predict future behavior. To expect destiny behavior, we need to have deep information of relationship among variables in big data set. Relationships among variables are often tested in terms of whether they exchange together or one by one. Correlation co-efficient (dependence degree) serves this cause efficaciously and as a result, emerges as the workhorse of quantitative studies and evaluation. A number of dependence measures are available in literature, used for measuring association among variables. More common of them are Pearson's correlation, spearman's-(rho) and Kendall's-tau. Pearson's correlation coefficient measures linear association among variables, however fail to measure relationship among variables while the relationship is not linear. In addition, Spearman and Kendall's correlation coefficients are effective, when there is monotonic relationship among variables and much less efficient in identifying relationship between variables when the connection is linear. Therefore, there may be an exceptional need of measure of dependence, which serve in all scenarios irrespective of variables functional form and capture a large variety of association amongst variables. Reshef et al. [28] proposed Maximal Information Coefficient (MIC); a statistics of dependence measure mainly based on mutual information which measure relationship among variables and which is effective in all functional form of variables. The mutual information between two random variables *X* and *Y* in expressions of their combined probability distribution P(X,Y) is expressed as:

$$I[X;Y] = \int \mathrm{dx}\, \mathrm{dyp}(x,y) \log_2 \frac{p(x;y)}{p(x)p(y)} \quad (7)$$

I[X; Y] in above Eqn. is always nonnegative and *I[X;Y]*= zero indicates that the two variables are independent . Any positive value of I[X; Y] recommends mutual dependence among variables, the stronger the dependence is, the higher the value of I[X;Y]. Whereas, maximal information of dependency measure is written as:

$$MIC\{x;y\} = I_{MIC}\{x; y\}/Z_{MIC} \quad (8)$$

Here $I_{MIC}\{x; y\}$ is the mutual information among two random variables *x* and *y* computed using a pre define binning scheme and $Z_{MIC} = log^2(\min(m_x, m_y))$ is a penalty in which m_x and m_y, is the number of bins containing observations on the *y* and *x-axes*. The range of MIC is from zero to one. Hence, in this way Reshef et al. [28] reduced down the range of mutual information from zero to infinity to 0 to 1 and hence, *MIC* becomes a novel degree of dependence for large data set and is applicable to any functional form between pair of variables. See [29-31] for a complete evaluation of maximal information coefficient and its estimation for two variables [7].

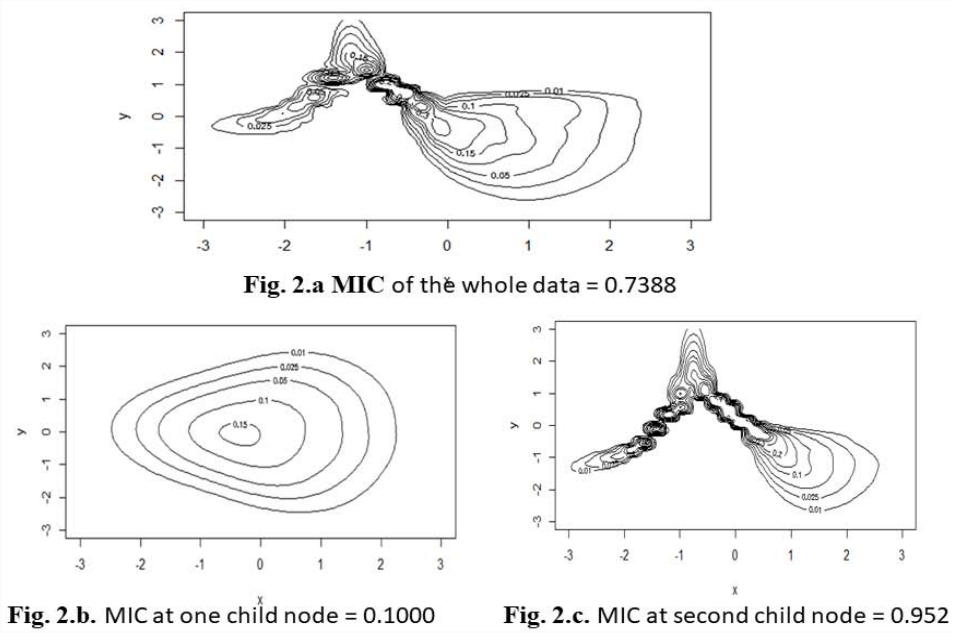

Fig. 2.a MIC of the whole data = 0.7388

Fig. 2.b. MIC at one child node = 0.1000

Fig. 2.c. MIC at second child node = 0.952

Figure 2: Stage one classification using MIC as a classification index. The difference of MIC=0.852.

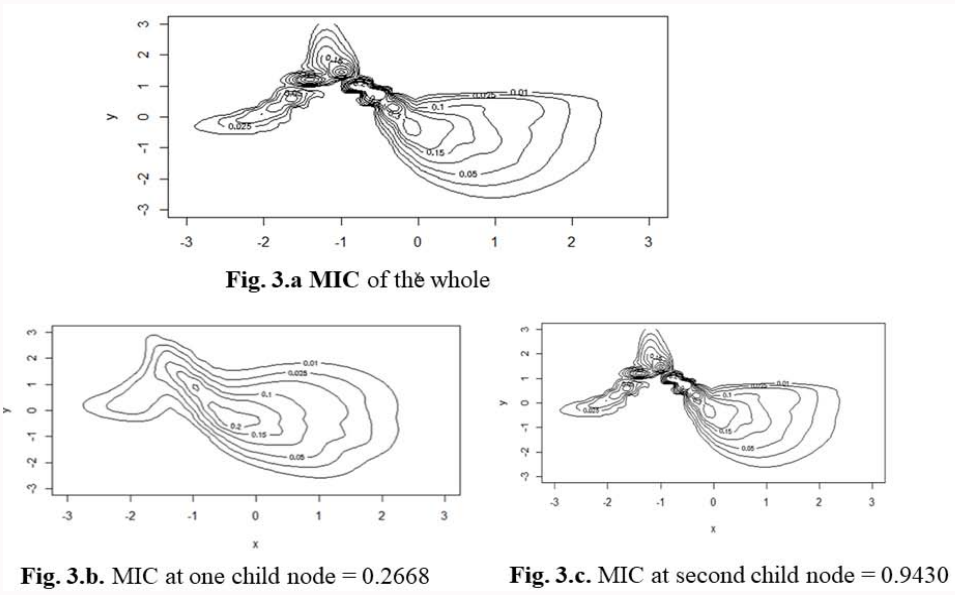

Fig. 3.a MIC of the whole

Fig. 3.b. MIC at one child node = 0.2668

Fig. 3.c. MIC at second child node = 0.9430

Figure 3: Stage two classification using MIC as a classification index. The Difference of MIC=0.6762.

We applied copula kernel based nonparametric density estimation of two random variables and combined it with measure of dependence to establish new classifier used for classification in different field of applied statistics especially in machine learning, say novel measure of dependence based classification in machine learning.

Algorithmic representation of the proposed idea of machine based decision tree classifier for prediction of heart diseases from risk factor is well explained as;

Machine Based Heart Diseases Algorithm:

Initiate

Step-1: Calculate the correlation among variables of interest for input data.

Step-2: Search the factors having high MIC differences at their level.

Step-3: Make classification and obtained child-nodes.

Step-4: Repeat step 1-3 at each child node and obtain the value of MIC.

Step-5: Stop growing branches of decision tree when MIC difference at factor level is less then pre-specified value.

Step-6: Construct the variable decision tree.

Illustration from simulation

For simulation study, we assumed that x follows a uniform distribution with parameters a=-.5 and b=5, and y follows skewed

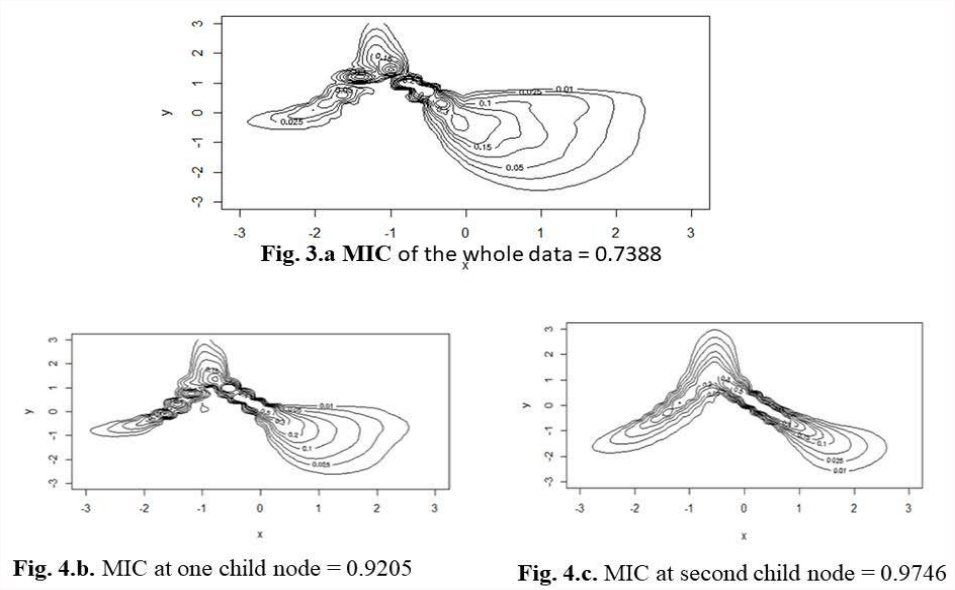

Figure 4: Stage three classification using MIC as a classification index. The difference MIC=0.0541.

normal with parameters omega =1.30 and alpha =5. We then draw a random sample of 13000 observations from the stated distribution with specified parameters and plot them on *xy*-plan. There is a strong correlation between the variables, which is detected by using MIC. We estimate the density with the help of nonparametric copula kernel density estimation using Probit transformation and graphically explore the hidden dependence through contour diagram for the whole dataset, as shown in Figure 2a. It is to be noted that the contour diagram is the graphical exploration of the association among variables, whereas MIC is the quantitative measure of dependence.

After taking a look at the contour diagram for the full dataset, we are aware that it could be classified into subgroups. The correlation between each group is sufficiently different, as in Figure 2b, 2c. We classify the data into two groups in such a way that the dependence between the variables under consideration is different. MIC accounts the correlation among a pair of variables, no matter their functional form. We account MIC to measure the dependence in each subgroup, higher the difference of MIC between two subgroups, approve the classification of data into further two subgroups. In this way, we can have obtained the classification of the data in many additional subgroups. Every time we account MIC difference, if the difference of two subgroups MIC is sufficiently large, we can approve the classification, as shown in Figures 2a-2c. Classification of data into any subgroup wherein the individuality of MIC is adequately large. In addition, we need to evaluate the differences of MIC on every factor level, which helps us in the ranking of factors according to their importance.

Figures 3a-3c represents stage two classifications of data into two subgroups, where the difference of MIC is very large. Likewise, Figures 4a-4c presents stage three classification of the whole data into two subgroups by means of a third factor. When we attain a pre-specified level of MIC at each factor levels, then we stop further classification. These simulation studies validate the proposed method of two variable classifications by means of maximal information coefficient.

Growing Copula Based Decision Trees and Discussion

A set of two decision trees have grown for cardiovascular disease records whenever through taking distinct correlated variables to become aware of the elements significance and to expecting cardiovascular disease greater correctly. We start exploring the data by locating correlation between weight and High-Density Lipoprotein (HDL) there is an extraordinary negative association between variables; means that because the weight will increase the HDL goes down which causes blocking of arteries and eventually, result in coronary heart attack. We evaluate the density of the notably correlated variables with the aid of nonparametric bivariate copula density estimation through Probit transformation to find the hidden dependence structure of related variables as shown in Figure 5a then we estimate Maximum Information Coefficient (MIC) which is represented in Table 1 and Table 2.

We discover that location of residence is the maximum influencing component for these two variables (e.g. weight and HDL) amongst all, so we classify our information through "*residence*" in subgroups and gain two baby-nodes of our tree "*mountain*" and "*plain*".

We repeat the whole procedure on both infant-nodes of the tree. At toddler node "mountain" factor "smoking" is playing important position inside the dating of two variables so we classify aspect "*smoking*" at child node "*mountain*" and attain two sub-infant nodes "*smoking*" and "*non-smoking*" and on child-node "*plain*" component "blocked arteries" has massive MIC distinction at their degree. At child-node "*plain*", we classify factor "*blocked arteries*" and acquire sub-child nodes for this child-node "*blocked arteries yes*" and "*blocked arteries no*". Subsequent for toddler-node "Plain" after blocked arteries aspect "smoking" is crucial and has big MIC distinction at stages as shown in Table 3.

We forestall growing our tree branches on all other sub-child nodes of child-node "*plain*" and continue at sub-sub-child node "*smoking yes*" and in addition classify component "*gender*" into "*male*" and "*female*" as proven in Figure 5a and prevent growing

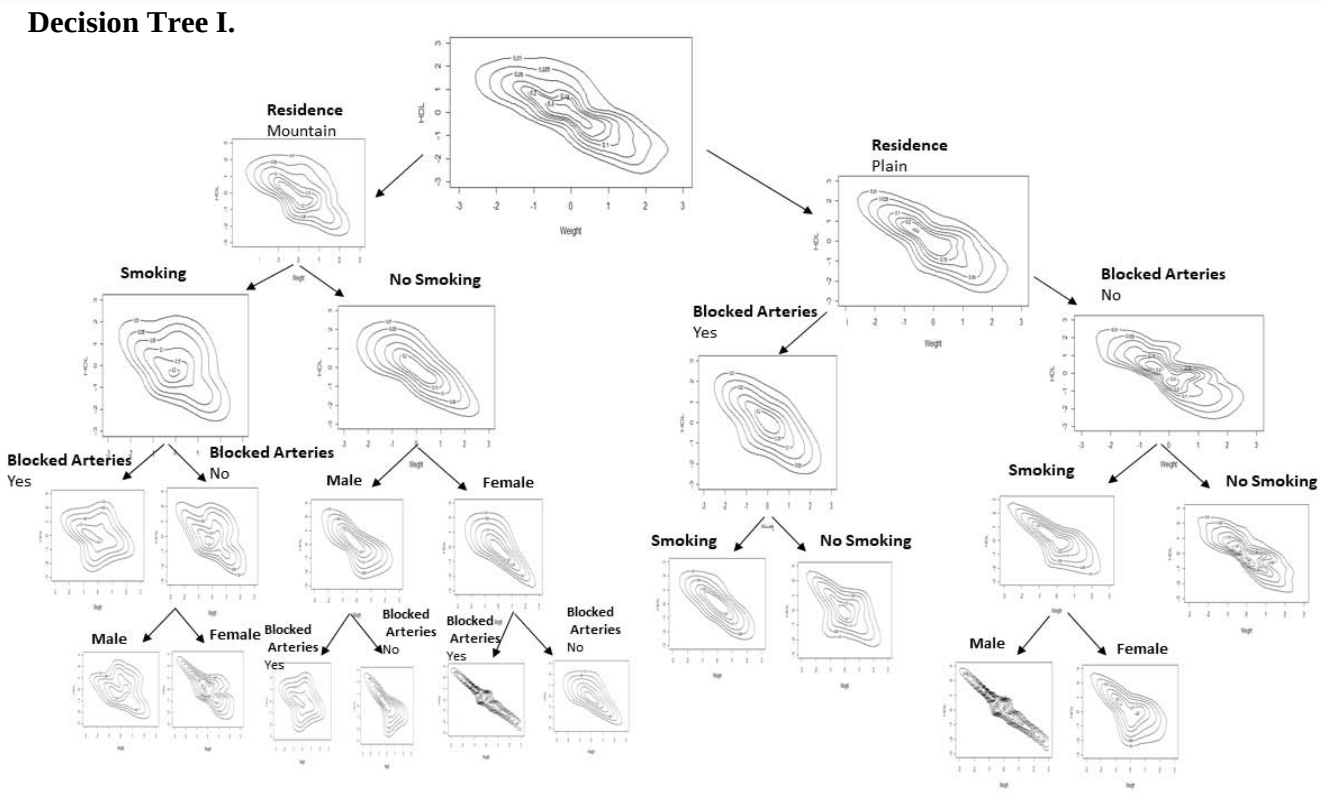

Figure 5a: Contour representation of two variable decision tree one.

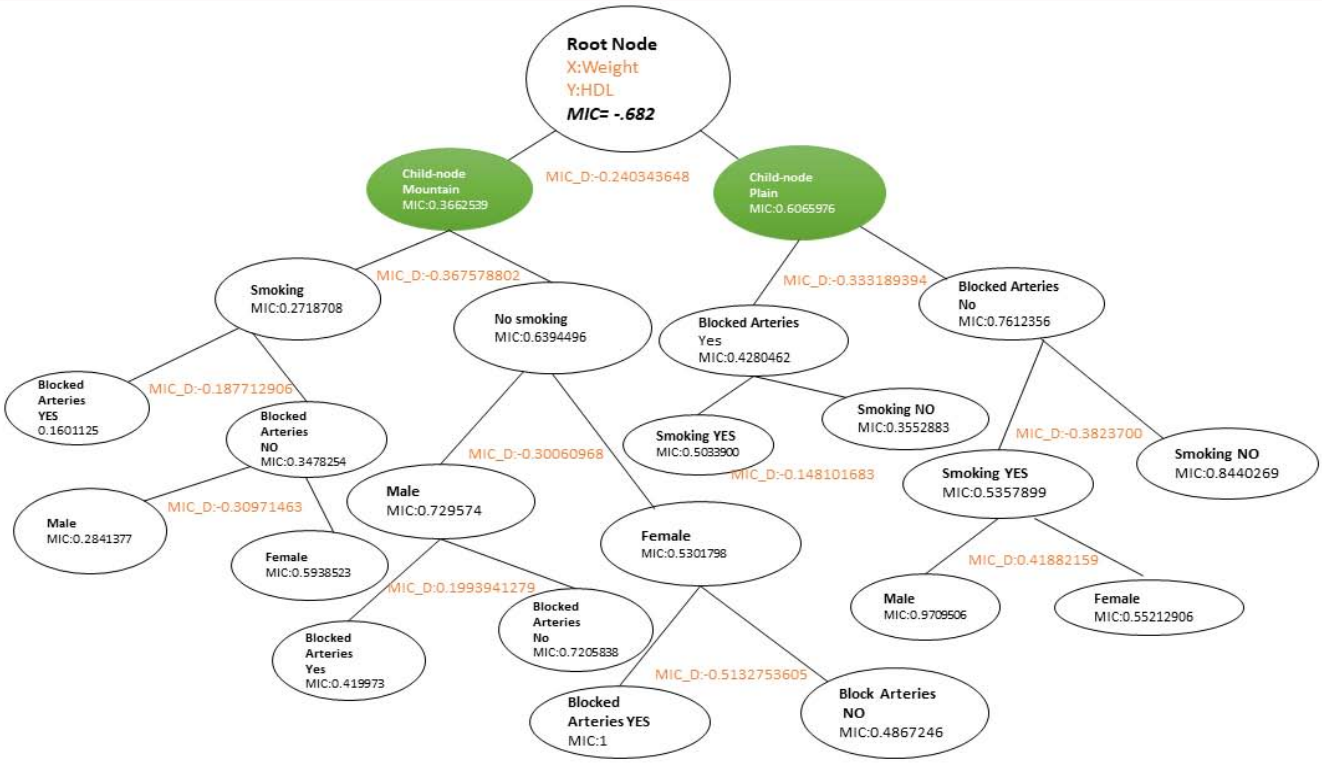

Figure 5b: Classical representation of two variable decision tree for heart Diseases based on (Weight, HDL).

tree branches whilst the distinction of MIC at factor tiers is ≤ 0.14. Then again, for child-node "*mountain*" at sub-infant node "*smoking*" as shown in the Figure 5a we classify "*blocked arteries*" to gain in addition two sub-sub-toddler nodes as shown in Figure 5b.

We further classify factor "*gender*" at sub-sub-child node "*blocked arteries no*" and stop growing branches at sub-sub child node "*blocked*

Table 1: Factor selection for classification of heart diseases decision tree for (Weight, HDL).

Stage One	Factors	*MIC* of Factors Levels	Difference of *MIC* between levels	Classification Factor
	Gender	0.7400879 (0.001)**	0.043146976	
		0.6969409 (0.000)***		
	Smoking	0.640047 (0.0011)**	-0.153852886	
		0.7938999 (0.001)**		Residence
	Blocked Arteries	0.5966124 (0.003)**	-0.07626301	
		0.6728754 (0.000)***		
		0.3662539 (0.0010)**	-0.240343648	
	Residence	0.6065976 (0.001)***		

Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **, *** correspond to Significance at 5% and 1% respectively

Table 2: Factor selection at child nodes for classification of decision tree for (Weight, HDL).

Stage Two	Factors at Child Node one	MIC of Factors Level	Difference of MIC between levels	Classification Factor
	Gender	0.3237797 (0.001)**	-0.2598772	
		0.5836570 (0.000)***	-0.367578802	Smoking
	Smoking	0.2718708 (0.0021)**	0.04431659	
		0.6394496 (0.0011)**		
	Blocked Arteries	0.4179988 (0.000)**		
		0.3736822 (0.0013)**		
	Factors at Child Node two			
	Gender	0.6677318 (0.000)***	0.108967889	
		0.5587639 (0.000)***	-0.191909131	Blocked Arteries
	Smoking	0.5059405 (0.000)**	-0.333189394	
		0.6978496 (0.0011)***		
	Blocked Arteries	0.4280462 (0.001)***		
		0.7612356 (0.001)**		

Note: Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **' *** correspond to Significance at 5% and 1% respectively

Table 3: Factor selection at sub-child nodes for classification of decision tree for (Weight, HDL).

Stage Three	Factors at First Sub Child Node One	*MIC* of Factors Level	Difference of MIC between levels	Classification Factor
Smoking	Gender	0.2436628 (0.0011)**	-0.1227145	Blocked Arteries
		0.3663773 (0.0013)**	-0.187712906	
	Blocked Arteries	0.1601125 (0.000)***		
		0.3478254 (0.000)***		
	Factors at First Sub Child Node two			
Nonsmoking	Gender	0.729574 (0.000)***	0.199394127	*Gender*
		0.5301798 (0.001)***	-0.1008583	
	Blocked Arteries	0.5216406 (0.001)***		
		0.6225265 (0.0013)***		
	Factors at Second Sub Child Node One			
Blocked Arteries	Gender	0.4290541 (0.000)***	0.0541511	
		0.3749030 (0.000)***	0.148101683	*Smoking*
	Smoking	0.5033900 (0.000)***		
		0.3552883 (0.0014)**		
	Factors at Second Sub Child Node Two			
Blocked Arteries	Gender	0.8844106 (0.000)***	0.1263281	
		0.7580825 (0.000)***	-0.38237	*Smoking*
	Smoking	0.5357899 (0.0013)**		
		0.8440269 (0.0001)***		

Note: Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **, *** correspond to Significance at 5% and 1% respectively

Table 4: Factor selection at sub-sub-child nodes for classification of decision tree for (Weight, HDL).

Stage Four Sub-sub child nodes	Last Factor left	MIC of Factors Level	Difference of MIC between levels	Classification Factor
	Male	0.1908745 (0.000)***	0.05294912	Stop growing
	Female	0.1379254 (0.0019)**		
Plain	Male	0.2841377 (0.0017)**	-0.30971463	Gender
	Female	0.5938523 (0.000)***		
Gender Male	Blocked	0.4199731 (0.003)**	-0.30060968	*Blocked Arteries*
	Arteries	0.7205838 (0.000)***		
Gender Female		1 (0.000)***	-0.51327536	*Blocked Arteries*
		0.4867246 (0.0001)***		
Smoking YES	Male	0.6488821 (0.0011)**	0.12503474	Stop growing
	Female	0.5238473 (0.001)**		
Smoking NO	Male	0.3246061 (0.000)***	-0.052261462	Stop growing
	Female	0.3768676 (0.000)***		
Smoking YES	Male	0.9709506 (0.000)***	0.41882159	Gender
	Female	0.5521290 (0.001)***		
Smoking NO	Male	0.9454190 (0.003)**	0.127930522	Stop growing
	Female	0.8174884 (0.001)***		

Note: Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **, *** correspond to significance at 5% and 1% respectively

arteries yes" as the *MIC* difference is ≤ 0.14 shown in Table 4.

Whereas, at sub-child node "*non-smoking*" gender is playing influencing role in the relationship of weight and HDL. So we further classify factor "*gender*" at this sub-child node and obtain two sub-sub-child nodes as shown in Figure 5b and then further classify blocked arteries. At these sub-sub-child nodes as "*blocked arteries*" is the least influencing factor on the relationship between weight and HDL at this child node as shown in Figure 5b. Figure 5b is the classical representation of decision tree with MIC at each factor level and the difference of MIC between each factor.

Post decision tree analysis

Weight and High-Density Lipoproteins (HDL) are anticipated to be inversely related variables in literature. Taking weight and HDL as base variables, we explore the correlation between the two variables and look at a high terrible correlation between two variables. Which means that as weight will increase the best cholesterol level goes down and inversely the coolest cholesterol level might be appropriate if the patient is not always overweight. We construct contour diagram the use of bivariate copula density estimation to discover the hidden dependence shape between the variables. Figure 5a constitute the contour diagram of the two correlated variables with appreciate to every aspect level of child-nodes and sub-child nodes, as shown in Figure 5b aspect "*place of residence*" play great role inside the dating among weight and HDL. And stale- course, if one has some bodily activities and taking natural objects of meals he/she have sound health and correct cholesterol level as well. We can more correctly predict from place of house "plain" as compare to "*mountain*" because the *MIC* at factor level is higher than mountain. A component blocked artery is the second essential factor after region of residence "*plain*". As proven in Figure 5b those sufferers whose arteries are not block and they are from undeniable residence we predict extra exactly that they have no heart diseases. Smoking is the third important factor in undeniable place of house. We can efficiently predict that the ones males who smokes despite the fact that their arteries are not blocked having place of house "*plain*" there good cholesterol stage could

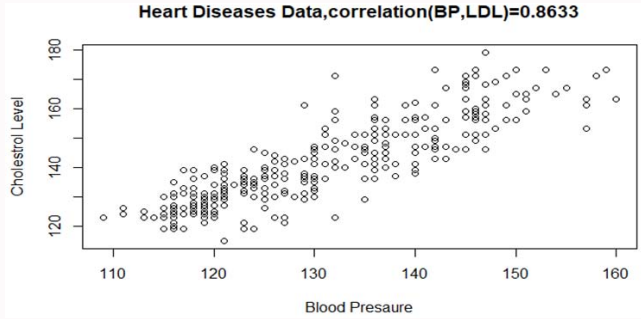

Figure 6: Scatter plot representing correlation between (Blood pressure and Cholesterol level).

be low. Then again, at child-node "mountain" smoking play huge position in predicting cardiovascular diseases.

We will easily expect that if the affected person is male and not smoking whose vicinity of residence is mountain, his HDL level is very good and has no heart illnesses as proven in Figure 5b alongside *MIC* value at each child-node and sub-baby nodes. And the difference of MIC among factor ranges subsequent, we do not forget blood pressure and Low-Density Lipoproteins (LDL) as our base variables, we discover the correlation between these two variables. As proven in Figure 6.

There is effective upward robust correlation among blood pressure and LDL. This means that one is efficaciously anticipated from the alternative. Then we built copula contour diagram by way of using nonparametric bivariate copula kernel density estimation as shown in the Figure 7a under, to discover the hidden dependence among variables. We then discover all factors separately and locate the MIC difference amongst every factor level as shown in Table 5.

We found that component "*smoking*" appreciably impact the connection between the two correlated variables (BP, LDL) as MIC distinction at component stage for smoking is higher amongst all factors as shown in Table 5. We classify things "*smoking*" in two

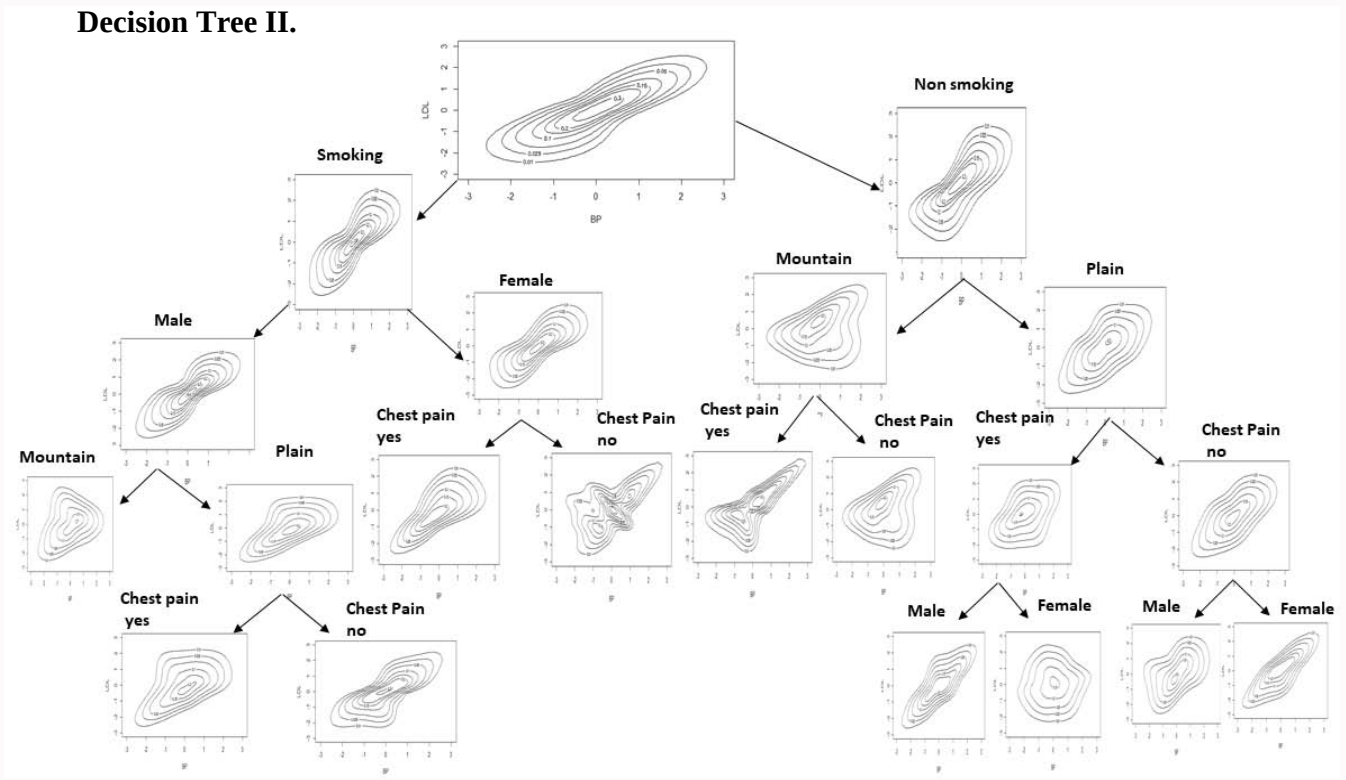

Figure 7a: Contour representation of two variable decision tree three.

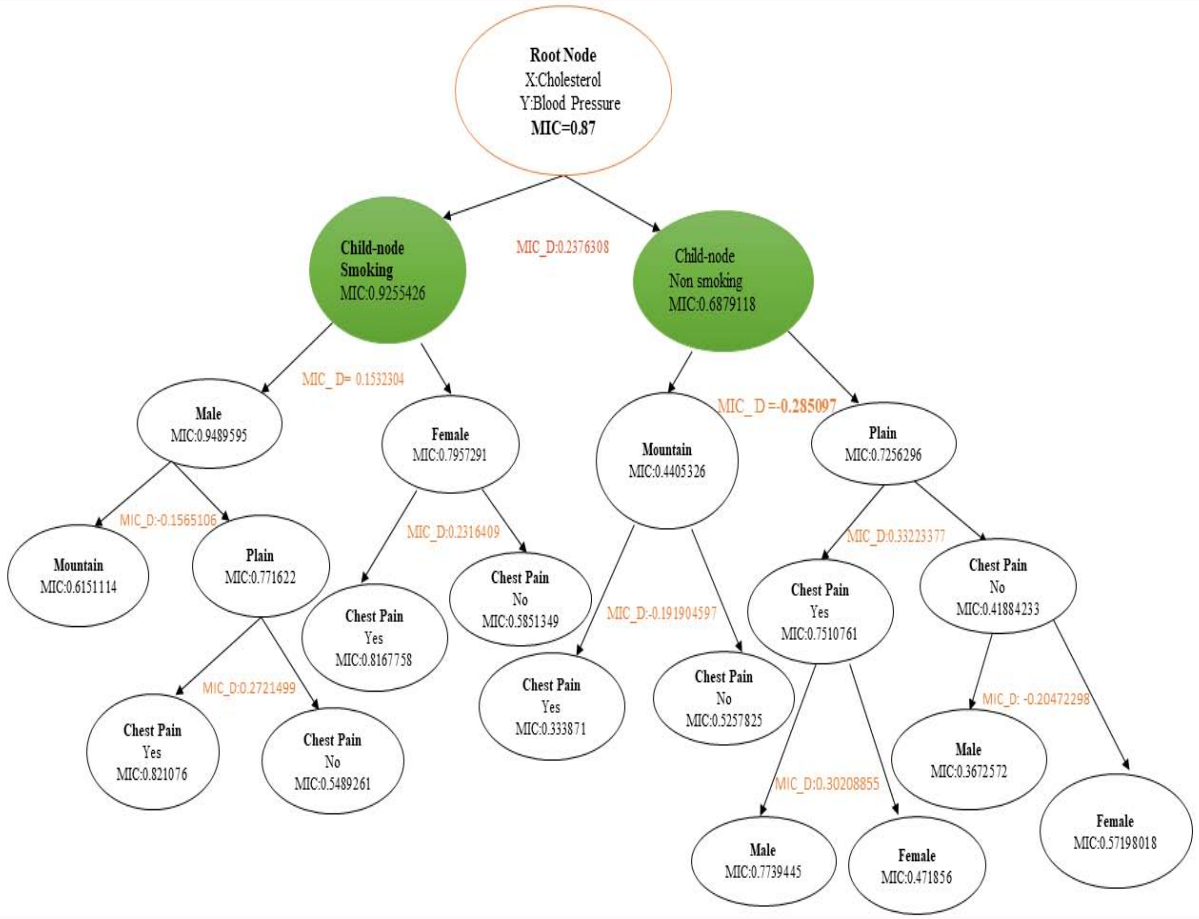

Figure 7b: Classical representation of two variable decision tree for heart diseases based on Blood pressure and Cholesterol Level.

Table 5: Factor selection for classification of heart diseases decision tree for (BP, LDL).

Stage One	Factors	MIC of Factors Levels	Difference of MIC between levels	Classification Factor
	Gender	0.7420809 (0.0001)***	0.02.300532	
		0.7650862 (0.0000)***		
	Smoking	0.9255436 (0.0000)***	0.2376318	Smoking
		0.6879118 (0.00013)**		
	Chest Pain	0.5529996 (0.0000)***	0.070840982	
		0.4821586 (0.0001)***		
		0.3014501 (0.0011)**	0.196856018	
	Residence	0.4983061 (0.0000)***		

Note: Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **, *** correspond to significance at 5% and 1% respectively

Table 6: Factor selection at child nodes for classification of decision tree for (BP, LDL).

Stage Two	Factors at Child Node one	MIC of Factors Level	Difference of MIC between levels	Classification Factor
	Gender	0.9489595 (0.0000)**	0.153230403	
		0.7957291 (0.0000)***		Gender
	Residence	0.2567465 (0.0003)**	-0.044200364	
		0.3009469 (0.0001)**		
	Chest pain	0.6299574 (0.000)***	0.036684593	
		0.5932728 (0.0010)**		
	Factors at Child Node two			
	Gender	0.6170139 (0.0000)***	-0.140690193	
		0.7577041 (0.0001)**		
	Residence	0.4405326 (0.00011)**	-0.285096942	Residence
		0.7256296 (0.0001)**		
	Chest pain	0.5546234 (0.0000)**	-0.038649415	
		0.5932728 (0.0000)**		

Note: Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **'*** correspond to significance at 5% and 1% respectively

subgroups and alternative two infant-nodes "*smoking*" and "*non-smoking*". Now we repeat the whole system on both baby-nodes, assemble the contour diagram the use of copula density at every baby node as shown in Figure 7a similarly classify, the factor on the idea of *MIC* variations at tiers on each infant-nodes as proven in Table 6.

At child-node "smoking" gender is the second maximum influencing factor after smoking which change the connection between blood strain and Low-Density Lipoproteins (LDL) considerably whereas, on child node "*non-smoking*" region of house plays extensive position within the relationship between blood pressure and LDL. Again, we repeat the entire system at every sub-baby node "*male*", "*female*", "*mountain*" and "*plain residence*" construct contour diagram the use of bivariate copula at every sub-child node and then evaluate the MIC difference at component level at each sub-baby node and classify the factor which has maximum *MIC* difference at level. At sub-child node "*male*", this distinction is maximum for vicinity of residence, for sub-child node "*female*", "mountain" and "plain residence" chest pain has highest *MIC* difference at levels, as proven in Table 7 and Table 8 respectively.

We hold our tree growing until we classify the facts with respect to all element; we prevent developing tree branches when MIC distinction is ≤ 0.14. From Figure 7a above, the whole technique is clear which goes at the lower back of constructing two variable decision tree primarily based on copula. Figure 7b underneath is the classical representation of decision tree along with MIC value at each infant and sub-child nodes and the distinction of MIC at aspect ranges.

Post decision tree analysis: Blood strain (BP) and Low-Density Lipoprotein (LDL) are the two fundamental variables which give bases to all heart illnesses and those variables are noticeably correlated. There may be a superb upward sturdy correlation between the two variables. Which means that after LDL increases there is a growth in the BP; in different phrases, blood strain is excessive because cholesterol level is high, the decrease the cholesterol level the decrease the blood stress. From two variable decision trees as supplied in Figure 7b we see that aspect "*smoking*" has substantial have an impact on inside the relationship of blood strain and cholesterol level this means that if we analyze the connection of blood pressure and cholesterol level with-recognize to component smoking its alternate the relationship significantly. This relationship is robust for factor level "*smoking*" and week for component degree "*non-smoking*". In this way from factor level "smoking" we can extra accurately predict heart diseases compared to "*non-smoking*" as the two variables are perfectly correlated with admire to component stage "*smoking*". From the tree, it is far clear that factor level "*male*" of gender allows extra in prediction of heart diseases compared to issue degree "*female*". Area of residence is any other vital aspect after gender and alternates the relationship among two variables blood stress and cholesterol degree. This data is more focus for plain place for adult males who are smoking. Because of this that we successfully expect that if the affected person belongs to plain vicinity and is male who is smoking and sense chest ache he has excessive blood pressure due

Table 7: Factor selection at sub-child nodes for classification of decision tree for (BP, LDL).

Stage Three	Factors at First Sub Child Node One	MIC of Factors Level	Difference of *MIC* between levels	Classification Factor
	Residence	0.6151114 (0.0000)***		Residence
		0.7716225 (0.0000)***	-0.1565106	
	Chest Pain			
	Factors at First Sub Child Node two			
	Residence	0.1886812 (0.00011)***		
		0.3878706 (0.00013)***	-0.1991894	
	Chest Pain	0.8167758 (0.00010)**		Chest Pain
		0.5851349 (0.00011)***	0.231640889	
	Factors at Second Sub Child Node One			
	Gender	0.4613051 (0.00011)**		
		0.4766799 (0.00013)***	-0.015374815	
	Chest Pain	0.3338779 (0.00011)***		Chest Pain
		0.5257825 (0.00010)**	-0.191904597	
	Factors at Second Sub Child Node Two			
	Gender	0.3729094 (0.0000)***		
		0.3040193 (0.0000)***	0.068890129	
	Chest Pain	0.7510761 (0.0000)***		
		0.41884233 (0.0001)***	0.332333377	Chest Pain

Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **' *** correspond to significance at 5% and 1% respectively

Table 8: Factor selection at sub-sub child nodes for classification of decision tree.

Stage Four	Last Factor left	MIC of Factors Level	Difference of MIC between levels	Classification Factor
Mountain	Chest Pain Yes	0.31127810 (0.0000)***	-0.0006174	Stop growing
	Chest Pain No	0.31189550 (0.0001)***		
Plain	Chest Pain Yes	0.8210761 (0.00012)**	0.27215	Chest Pain
	Chest Pain No	0.5489261 (0.00011)**		
Chest pain	Mountain Plain	0.3958156 (0.00001)***	0.0468895	Stop growing
		0.3489261 (0.00007)***		
Chest pain	Mountain Plain	0.2686650 (0.0000)***	-0.0426131	Stop growing
		0.3112781 (0.0001)***		
Chest pain	Male	1 (0.0000)***	0.0817042	Stop growing
	Female	0.9182958 (0.0001)***		
Chest pain	Male	0.2686665 (0.0003)**	-0.0167914	Stop growing
	Female	0.2854579 (0.0001)**		
Chest pain	Male	0.7739445 (0.0000)***	0.3639445	Gender
	Female	0.4718445 (0.0001)***		
Chest pain	Male	0.36725472 (0.0000)***	-0.15254467	Gender
	Female	0.57198018 (0.0000)***		

Values in bracket represent corresponding p-value of the MIC statistics. Furthermore, **' *** correspond to significance at 5% and 1% respectively. The pre-specified value of MIC ≤ 0.14 to stop growing tree branches

to high-density lipoproteins and could have cardiovascular diseases. Further, as proven inside the choice tree supplied in Figure 7b above chest pain is gambling important position in the relationship of the two correlated variables *via* factor level *"female"* of gender; manner that if the affected person is female and she has chest ache alongside smoking we can accurately expect that she has cardiovascular diseases. Likewise, for "nonsmoking" child-node location of house "*plain*" is vital and chest pain plays his position after area of house; and we accurately are expecting heart illnesses from aspect degree "*male*" of gender. Which means that if the affected person is male and feature chest pain and belongs to plain place of residence despite the fact that, he is not smoking we expect that he has high blood strain and cholesterol level and has heart diseases.

Conclusion and Suggestions

Cardiovascular diseases are the predominant causes of high mortality and disability rate all over the world. Sturdy from researchers in the previous year's holds that, the quotes of heart diseases-related expiries have declined in a number of developed

nations however nevertheless excessive in low-and middle-income countries and need severe attention [4]. Although the importance of heart diseases in low-and middle-income countries [1-4], no care is given to the anticipation of cardiovascular diseases danger elements in South Asia, particular in my home land.

Further, monetary and administrative insecurity is hastening the rates of cardiovascular diseases. In this paper, we addressed the prediction of heart diseases from hazard factors via decision tree. We efficiently introduce naively develop machine learning approach in public health with the goal to extract high-level information from raw data which facilitates in prediction of coronary heart diseases from risk factors and its prevention. We assemble novel nonparametric copula based decision tree for affected person's facts, which facilitates in prediction as well as rating of risk factors in keeping with their significance. The obtain results show that it is possible to expect heart diseases from risk factors with accuracy, if statistics on those risk aspects are to be had. As direct outcome of this research, use of tobacco, physical activity, and weight loss program are the primary risk factors gambling enormous position within the prediction of cardiovascular disease, which is the essential purpose of deaths in low-and middle-earnings countries [4], particular in Pakistan. We achieve ranking of risk factors via two variable decision trees, which enables in improving public health as nicely in selection regarding cardiovascular diseases remedy and prevention. It additionally allows in strategies making, related to prevention of heart diseases threat elements in low- and middle-income nations.

Implementation and suggestion

We have robust evidences from affected person's facts that the attainment and increase of vascular threats originated early in lifestyles. Unnatural practices in formative years and teen-age rise the chance which incorporates tobacco use, high fat and excessive-calorie consumption, and absence of somatic interest; as an outcome, the rate of cardiovascular disease deaths in below average-income countries are growing daily. The existing scenario requires proper planning and monitoring to overcome this trouble through putting in right database on those risk factors at national level from which we are able to effortlessly expect tendencies of Cardiovascular Diseases (CVD) and make regulations for its prevention in low-and middle-earnings countries especially in my home land. We successfully present a newly developed data mining technique; one may additionally get benefit from it to conquer this trouble through replicating the equal nature of research for his use.

Supporting packages and computational environment

Nonparametric kernel approximation of two variables copula density was estimated using the "kdecopula" platform [32] and *MIC* was projected using the "MINE" suite of [28] as described. Investigations were achieved in R (Studio). Questionnaire, collected data and R-codes will be available publically once the project is completed.

Funding

This project is part of National Health Program jointly funded by Pakistan Medical Research Council and Khyber Pkhutan Khawa Government via Notification no.SO(H-I)/NHP/12879321.

Acknowledgement

I greatly acknowledged the support of my colleagues Dr. Basharat Hussain who helped me a lot in timely completion of this research. I am also thankful to Pakistan as well as Khyber Pkhutan Khawa Government for providing Funds for this project.

Ethics Approval

Consequent upon the approval from Pakistan Medical Research Council, Provincial Government of Khyber Pkhutan Khawa and Admiration of Ayyub Medical Hospital under letter No. SO(U-II)/H. E/12-8/ dated 21/08/2019 data for this project was collected from Heart patient who visited Ayyub Medical Complex Hospital for checkup in day time from September 9th, 2019 to October 9th, 2019.

References

1. Turin TC, Shahana N, Wangchuk LZ, Specogna AV, Al Mamun M, Khan MA, et al. The burden of cardiovascular and cerebrovascular diseases and the conventional risk factors in the South Asian population. Glob Heart. 2013;8(2):121-30.
2. Filion KB, Luepker RV. Cigarette smoking and cardiovascular disease: Lessons from Framingham. Glob Heart. 2013;8(1):35-41.
3. WHO. Global status report on non-communicable diseases 2014. Media Centre. 2017.
4. Yusuf S, Rangarajan S, Teo K, Islam S, Li W, Liu L, et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. N Engl J Med. 2014;371:818-27.
5. Quan H, Chen G, Walker R, Wielgosz A, Dai S, Tu K, et al. Incidence, Cardiovascular complications and mortality of hypertension by sex and ethnicity. Heart. 2013;99(10):715-21.
6. Government of Pakistan. Population Census Organization (2012).
7. Khan YA, Shan QS, Liu Q, Abbas SZ. A nonparametric copula-based decision tree for two random variables using MIC as a classification index. Soft Computing. 2021;25:9677-92.
8. Weiss SM, Kulikowski CA, Computer systems that learn: Classification and prediction methods from statistics. Neural Nets, Machine Learning, and Expert Systems, Morgan Kaufman 1991.
9. Quinlan JR. C4.5: Programs for machine learning, 1993;29:5-44.
10. Weiss SM, Kapouleas I. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods, In Proc. 11th Intl. Joint Conf. on Artificial Intelligence, p.781.787., Detroit, MI, 1989.
11. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Wadsworth International Group. 1984.
12. Mooney R, Shavlik J, Towell G, Grove A. An experimental comparison of symbolic and connectionist learning algorithms, In Proc. 11th Intl. Joint Conf. on Artificial Intelligence, p.775.787., Detroit, MI, 1989.
13. Quinlan JR. Induction of decision trees. Machine learning. 1986;1:81-106.
14. Atlas L, Connor J, Park D, El-Sharkawi M, Marks R, Lippman A, et al. "A performance comparison of trained multilayer perceptrons and trained classification trees," Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics, 1989.v
15. Talmon JL. A multiclass nonparametric partitioning algorithm. Pattern Recognition Letters. 1986;4(1):31-8.
16. Brown, Donald E, Pittard CL. "Classification trees with optimal multi-variate splits." Proceedings of IEEE Systems Man and Cybernetics Conference - SMC 3 (1993): 475-477 vol.3.
17. Hatnagar p, wickramasinghe k, williams j, rayner m, townsend n. the epidemiology of cardiovascular disease in the uk 2014. heart. 2015;101:1182-9.
18. Pakistan Medical Research Council. National Health Survey of Pakistan 1990-1994. 1998.

19. Wang LM, Li XL, Cao CH, Yuan SM. Combining decision tree and naïve Bayes for classification. Knowledge- Based Systems. 2006;19(7):511-5.

20. Aitkenhead MJ. A co-evolving decision tree classification method. Expert Systems with Applications. 2008;34(1):18-25.

21. Cherubini U, Luciano E, Vecchiato W. Copula methods in finance. Wiley Finance series. John Wiley & Sons. 2004.

22. Elidan G. Copulas in Machine Learning. In: Jaworski P, Durante F, Hardle WK, editors. Copulae in Mathematical and Quantitative Finance. Springer-Verlag Berlin Heidelberg. 2013.

23. Sklar A. Functions de R´epartition `a n Dimensions et Leurs Marges. Publications de l'Institut Statistique de Universit´de Paris. 1959;8:229-231.

24. Gijbels I, Mielniczuk J. Estimating the density of a copula function. Communications in Statistics-Theory and Methods.1990;9(2):445-64.

25. Nelsen RB. An Introduction to Copulas, Springer, New York. 1997.

26. Geenens G, Charpentier A, Paindaveine D. Probit transformation for nonparametric kernel estimation of the copula density. Bernoulli. 2017;23(3):1848-73.

27. Filose M. Minerva: Maximal information-based nonparametric exploration r package for variable analysis version 1.3. 2013.

28. Reshef DN, Reshef YA, Finucane, HK Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334(6062):1518-24.

29. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69(6 Pt 2):066138.

30. Simon N, Tibshirani R. Comment on detecting novel associations in large data sets by Reshef et al., Science Dec 16, 2011.

31. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference and prediction Springer Verlag, New York. 2009.

32. Nagler T. Kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities. 2017.