scientific reports

Check for updates

OPEN

Predicting high sensitivity C-reactive protein levels and their associations in a large population using decision tree and linear regression

Somayeh Ghiasi Hafezi[1,2,8], Toktam Sahranavard[3,8], Alireza Kooshki[3,8], Marzieh Hosseini[4], Amin Mansoori[2]✉, Elham Amir Fakhrian[3], Helia Rezaeifard[3], Mark Ghamsary[5], Habibollah Esmaily[1,6]✉ & Majid Ghayour-Mobarhan[7]

High-sensitivity C-reactive protein (hs-CRP) is a biomarker of inflammation predicting the incidence of different health pathologies. In this study, we aimed to evaluate the association between hematological and demographic factors with hs-CRP levels using decision tree (DT) and linear regression (LR) modeling. This study was conducted on a population of 9704 males and females aged 35 to 65 years recruited from the Mashhad Stroke and Heart Atherosclerotic Disorder (MASHAD) cohort study. We utilized a data mining approach to construct a predictive model of hs-CRP measurements, employing the DT methodology. DT model was used to predict hs-CRP level using biochemical factors and clinical features. A total of 9,704 individuals were included in the analysis, with 57% of them being female. Except for fasting blood glucose (FBG), hypertension (HTN), and Type 2 diabetes mellites (T2DM), all variables showed significant differences between the two groups. The results of the LR models showed that variables such as anxiety score, depression score, Systolic Blood Pressure, Cardiovascular disease, and HTN were significant in predicting hs-CRP levels. In the DT models, depression score, FBG, cholesterol, and anxiety score were identified as the most important factors in predicting hs-CRP levels. DT model was able to predict hs-CRP level with an accuracy of 72.1% in training and 71.4% in testing of both genders. The proposed DT model appears to be able to predict the hs-CRP levels based on anxiety score, depression scores, fasting blood glucose, systolic blood pressure, and history of cardiovascular diseases.

Keywords High sensitivity C-reactive protein, Hematological factors, Demographic parameters, Decision tree

As a role player in every pathogenesis, inflammation is the most common mechanism in almost every disease and situation in living organisms. Inflammatory pathways have been the center of numerous mechanisms from autoimmune diseases, to infections and cancers[1]. One of the well-known and available tests by which we can measure inflammation in a human system is C-reactive protein also known as CRP[2]. As CRP is mostly produced by hepatocytes in response to pro-inflammatory cytokines, the level of this protein can be detected by a relatively newly introduced test called high sensitivity CRP (hs-CRP). This marker is shown to successfully predict various diseases such as cardiovascular diseases, diabetes, cancers as well and autoimmune diseases which are the inseparable part of these diseases' blood profile tests[3,4]. Alongside the diseases that can alter the level of hs-CRP in the human sera, it has been imagined that other serum biomarkers can affect the hs-CRP levels in the serum by different and sometimes unknown mechanisms[5]. Cardiovascular diseases (CVD), depression, anxiety, and

[1]Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran. [2]Department of Applied Mathematics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran. [3]Student Research Committee, Faculty of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran. [4]Department of Biostatistics, College of Health, Isfahan University of Medical Sciences, Isfahan, Iran. [5]School of Public Health, Loma Linda University, Loma Linda, CA, USA. [6]Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. [7]Metabolic Syndrome Research Center, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. [8]Equally contributed to this work: Somayeh Ghiasi Hafezi, Toktam Sahranavard and Alireza Kooshki. ✉email: am.ma7676@yahoo.com; aminmansoori@um.ac.ir; esmailyh@mums.ac.ir

diabetes as the major diseases that hs-CRP level can be altered, are widely studied and reported that they initiate inflammatory responses, stimulate the production of inflammatory cytokines, and increase hypothalamic-pituitary axis activity[6].

The decision tree (DT) classifier is considered one of the most famous methods for data classification. Different researchers from various fields including medicine have considered the problem of expanding DTs from available data, such as machine learning, pattern recognition, and statistics[7]. DT is a nonparametric model which is relatively easy to use and interpret compared to other models[8]. DT is widely used in medicine and is a trustworthy method to predict the outcome using it.

When it comes to sensitive inflammatory factors such as hs-CRP, DT can help organize the vast range of variables that can affect the level of hs-CRP. DT models help prioritize the important variables in branches that can affect the level of hs-CRP more effectively. Therefore, in this study, we tend to use DT modeling in predicting the association between hematological and demographic factors of a large population with the hs-CRP level. These findings may aid physicians and researchers in better understanding the factors interfering with and altering the hs-CRP level as one of the major paraclinical findings in almost every disease and condition.

Method

Study population

The data were obtained from phase I of the Mashhad stroke and heart atherosclerotic disorder (MASHAD) study, a 10-year cohort conducted in northeast of Iran. Several 9704 participants aged 35 to 65 years were recruited by stratified cluster random sampling technique. All participants provided written informed consent and the study protocol was approved by the Ethical Committee of Mashhad University of Medical Sciences. All participants and authors where blind to their results of the laboratory data. More details regarding the study design and methodology have been described before[2,9,10]. All methods were performed in accordance with the Declaration of Helsinki guideline and regulations[11].

Baseline examination

Participants provided blood samples for analysis. Blood samples were collected via venipuncture of an antecubital vein between 8 and 10 a.m. after a 14-h overnight fasting period. The samples were placed in 20 ml vacuum tubes while the individuals were in a sitting position, following a standard protocol. Within 30–45 min of collection, the blood specimens were centrifuged at room temperature to separate the serum and plasma into six 0.5 ml aliquots, which were subsequently sent to the Bu Ali Research Institute in Mashhad. We tried to assess the hs-CRP level of all samples similarly and avoid pre-analytical variations. Due to the delicacy of hs-CRP level we tried to be as much punctual as possible and all samples were measured in almost similar timing. Additionally, aliquots of serum were preserved at -80°C for future analysis. Low-density lipoprotein cholesterol (LDL-C) was determined from serum total cholesterol (TC), triglyceride (TG), and high-density lipoprotein cholesterol (HDL-C) concentrations using the Friedewald formula[12], but only if serum TG concentrations were lower than 400 mg/dL. Dyslipidemia was defined as TC equal to or greater than 200 mg/dL (5.18 mmol/l), LDL-C equal to or greater than 130 mg/dl (3.36 mmol/l), TG equal to or greater than 150 mg/dL (1.69 mmol/l), or HDL-C less than 40 mg/dl (1.03 mmol/l) in men and less than 50 mg/dL (1.30 mmol/l) in women. Type 2 diabetes mellitus (T2DM) is characterized as fasting blood glucose (FBG) equal to or greater than 126 mg/dl or by current treatment with oral hypoglycemic agents or insulin[13].

Hypertension was diagnosed in individuals with systolic blood pressure (SBP) at or above 140 mmHg and/or diastolic blood pressure (DBP) at or above 90 mmHg, as well as those who were taking antihypertensive medication[14].

Psychometric assessments were conducted using the Beck's Anxiety Inventory (BAI) to calculate anxiety scores, with ranges indicating various levels of anxiety. Likewise, the Beck Depression Inventory-II (BDI-II) was utilized to assess depression, with specific score ranges corresponding to different levels of depression[15].

Statistical analysis

All data were analyzed using the R Statistical Software (v4.1.2; R Core Team 2021) and the IBM SPSS Statistics (Version 27). All continuous data are expressed as mean ± SD and frequency (%) for categorical variables. All p-values < 0.05 were regarded as statistically significant. We used the t-test for continuous variables, and the Mann–Whitney test for non-normal data to compare the mean or median of the subjects hs-CRP < 3 mg/dL and hs-CRP $\geq$ 3 mg/dL. Also, we used the ANOVA test for continuous variables, and the Kruskal Wallis test for non-normal data to compare the mean or median of the subjects hs-CRP < 1 mg/dL, 1 mg/L < hs-CRP $\leq$ 3 mg/dL and hs-CRP > 3 mg/dL. The chi-square test was implemented to investigate the association between the categorical variables, the binary and Three-part category outcome, hs-CRPs[1].

To assess the multicollinearity between independent variables the variance inflation factor (VIF) as well as the computation of correlation coefficient was used. Generally, a correlation higher than 0.7 was considered a highly correlated variable that helps to recognize the possibility for multicollinearity[1].

The logistic regression and linear regression were used to compute the odds ratios (OR) and coefficients respectively with their 95% confidence interval based on three models: All Models include the variables CVD, hypertension (HTN), SBP, Anxiety Score, Depression Score, FBG and copper adjusted for Age, physical activity (PAL), LDL, HDL, TG, History of CVD and T2DM. Also, model A adjusted for sex, Model B for male, and Model C for female were presented. The outcome in linear regression was ln (hs-CRP) and logistic regression was hs-CRP < 3 mg/dL and hs-CRP $\geq 3.\mathrm{mg/dL}$ All of the analyses were done separately for males and females.

DT model

We utilized a data mining approach to construct a predictive model of hs-CRP measurements, employing the Decision Tree (DT) methodology. A decision tree is a non-parametric method tailored to the characteristics of the target variable, designed to create a predictive model based on predictor variables [16–19]. Specifically, in this study, we incorporated the CHAID technique within the DT. CHAID serves purposes in prediction, classification, and identifying interactions between variables. Various algorithms can be employed to construct a decision tree, such as CART, ID3, C4.5, and CHAID, each aiming to identify the most influential feature through chi-square tests, also known as CHAID. The Pearson metric is the default correlation measure in most programming libraries, for instance, Pandas in Python. The chi-square formula is used to determine significance,

$$\frac{\sqrt{(y-y')^2}}{y'}$$

With y representing actual values and y' representing expected values, while successive splits indicate the order of importance of the predictor variables. To assess the accuracy, precision, and sensitivity of the decision tree algorithm, we employed the confusion matrix using SPSS software version 27. This allowed for the evaluation of the decision tree's performance.

Results

Characteristics of the study population

Overall, 9704 individuals were eligible for analysis (57% female). The mean age of the participants was 48.87 ± 8.43 and 47.55 ± 8.09 in males and females, respectively. The clinical characteristics of the participants at the baseline have been summarized in Table 1. The biochemical factors and clinical features were compared between both men and women using t-test, Mann–Whitney U test for non-normal data, and chi-square test for categorical data. All variables had significant differences between the two groups ($p < 0.001$) except FBG ($p = 0.345$), HTN ($p = 0.114$), and T2DM ($p = 0.143$).

The clinical characteristics of the participants in three different hs-CRP levels at the baseline have been summarized in Table 2. All variables had significant differences between the three groups ($P < 0.001$) except the family history of CVD ($p = 0.62$).

Two data mining techniques were used to investigate the relationship between biochemical factors and confiner variables predictors and binary response variables (hs-CRP < 3 mg/dL, and hs-CRP ≥ 3 mg/dL) and

P value	Female N = 5819	Male N = 3885	All N = 9704	Variables	
< 0.001	47.55 ± 8.09	48.87 ± 8.43	48.08 ± 8.26[a]	Age	
0.008	121.37 ± 19.37	122.37 ± 17.10	121.78 ± 18.53	SBP	
< 0.001	78.46 ± 11.60	80.04 ± 10.62	79.09 ± 11.24	DBP	
< 0.001	1.70 ± 0.23	1.45 ± 0.30	1.60 ± 0.29	PAL	
< 0.001	118.55 ± 35.61	113.59 ± 34.4	116.57 ± 35.24	LDL	
< 0.001	44.87 ± 9.87	39.79 ± 9.25	42.85 ± 9.94	HDL	
< 0.001	2.73 ± 0.91	2.95 ± 0.97	2.82 ± 0.95	LDL / HDL	
< 0.001	194.34 ± 39.71	186.79 ± 37.8	191.33 ± 39.15	Cholesterol (mg/dL)	
< 0.001	1.81 (1.06,3.88)	1.41 (0.89,2.86)	1.62 (0.99,3.50)[b]	hs-CRP (mg/dL)	
< 0.001	10 (4,18)	6 (2,12)	8 (3,15)	Anxiety score	
< 0.001	12 (6,19)	9 (4,15)	11 (5,18)	Depression score	
< 0.001	117 (83,197)	125 (86,180)	120 (84.25,172)	TG	
0.345	82 (74,94)	82 (74,93)	82 (74,93)	FBG	
0.114	1343 (23.1%) 4457 (76.6%)	951 (24.5%) 2919 (75.1%)	2294 (19.8%)[c] 7376 (63.8%)	Yes No	HTN
0.143	848 (14.6%) 4909 (84.2%)	521 (13.4%) 3294 (84.8%)	1369 (11.8%) 8195 (70.9%)	Yes No	T2DM
< 0.001	93 (1.6%) 5726 (98.4%)	116 (3.0%) 3769 (97.0%)	209 (1.8%) 9495 (82.1%)	Yes No	CVD
< 0.001	1663 (28.6%) 4096 (70.4%)	980 (25.2%) 2847 (73.3%)	2643 (22.9%) 6943 (60.0%)	Yes No	Family history of T2DM
< 0.001	2139 (36.8%) 3642 (62.6%)	1230 (98.7%) 2606 (67.1%)	6248 (54.0%) 3369 (29.1%)	Yes No	Family history of CVD

Table 1. Baseline characteristics of male and female. *Abbreviations:* SBP systolic blood pressure, DBP diastolic blood pressure, PAL physical activity, LDL low-density lipoprotein, HDL high-density lipoprotein, hs-CRP high sensitivity C—reactive protein, TG triglyceride, FBG fasting blood glucose, HTN hypertension, T2DM type 2 diabetes, CVD cardiovascular disease. a. Mean ± sd for continuous and normal variables and p value of two sample t test. b. Median (Q1, Q3) for continuous and abnormal variables and p value of Mann Whitney U test. c. Count (percentage) for categorical variables and p value of chi square test.

P value	hs-CRP ≥ 3 mg/dL N = 2795	1 mg/dL ≤ hs-CRP < 3 mg/dL N = 4353	hs-CRP < 1 mg/dL N = 2486	All N = 9634	Variables	
< 0.001	48.16 ± 8.26	48.98 ± 8.21	48.08 ± 8.05a	48.08 ± 8.05a	Age	
< 0.001	121.81 ± 18.04	124.5 ± 20.17	121.77 ± 18.52	121.77 ± 18.52	SBP	
< 0.001	79.32 ± 11.19	80.22 ± 11.39	79.09 ± 11.24	79.09 ± 11.24	DBP	
0.036	1.59 ± 0.28	1.59 ± 0.29	1.59 ± 0.28	1.59 ± 0.28	PAL	
< 0.001	117.17 ± 38.60	120.31 ± 37.97	116.57 ± 32.24	116.57 ± 32.24	LDL (mg/dL)	
0.001	43 ± 9.90	43.14 ± 9.98	42.84 ± 9.94	42.84 ± 9.94	HDL (mg/dL)	
< 0.001	2.82 ± 0.93	2.90 ± 1.02	2.82 ± 0.94	2.82 ± 0.94	LDL / HDL	
< 0.001	191.94 ± 38.60	199.39 ± 41.23	191.33 ± 39.15	191.33 ± 39.15	Cholesterol (mg/dL)	
< 0.001	7 (3,14)	8 (3,15)	9 (4,17)	8 (3,15)[b]	Anxiety score	
< 0.001	10 (5,17)	10 (5,17)	12 (6,19)	11 (5,18)	Depression score	
< 0.001	131(94,179)	122(86,175)	106(74.25,152)	120 (84.25,172)	TG (mg/dL)	
< 0.001	86(77,102)	83(75,93)	79(72,87)	82(74,93)	FBG (mg/dL)	
< 0.001	1159(46%) 1327(53%)	1774(40%) 2579(59%)	912(32%) 1883(67%)	3885(40%)[c] 5819(59%)	Male Female	Sex
< 0.001	435(17%) 2045(82%)	1044(24%) 3296(75%)	796(28%) 1989(71%)	2294(23%) 7376(76%)	Yes No	HTN
< 0.001	207(8%) 2260(91%)	597(14%) 3726(86%)	552(20%) 2204(79%)	1369(14%) 8195(85%)	Yes No	T2DM
0.001	50(2%) 2436(97%)	74(1%) 4279(98%)	84(3%) 2711(96%)	209(2%) 9495(97%)	Yes No	CVD
< 0.001	617(25%) 1843(74%)	1171(27%) 3130(73%)	834(30%) 1922(69%)	2643(27%) 6943(72%)	Yes No	Family history of T2DM
0.620	859(34%) 1605(65%)	1497(34%) 2819(65%)	991(36%) 1778(64%)	3369(35%) 6248(64%)	Yes No	Family history of CVD

Table 2. Baseline characteristics with tertial divided hs-CRP. *Abbreviations:* SBP systolic blood pressure, DBP diastolic blood pressure, PAL physical activity, LDL low-density lipoprotein, HDL high-density lipoprotein, hs-CRP high sensitivity C—reactive protein, TG triglyceride, FBG fasting blood glucose, HTN hypertension, T2DM type 2 diabetes, CVD cardiovascular disease. a. Mean ± Sd for continuous and normal variables and p value of ANOVA test. b. Median (Q1, Q3) for continuous and abnormal variables and p value of Kruskal Wallis. c. Count (percentage) for categorical variables and p value of chi square test.

Ln (hs-CRP). So, the main objective of this study was to anticipate hs-CRP using the LRs and DT models (binary and tertiles hs-CRP) and to determine their associated factors, especially biochemical factors markers. For this purpose, the dataset in the DT model was randomly split into two parts: training data, and test data (25%-75%). The training dataset was utilized to develop the DT model, which was then validated using test data (25%) that hadn't been used during training.

The association between biochemical factors, clinical features, and hs-CRP using logistic regression (LR) and linear regression model

Table 3 showed the result of linear regression model with both log and binary hs-CRP. CVD, HTN, SBP, Anxiety Score, Depression Score and FBG (mg/dL) were reported to be included in the analysis and the models adjusted for Sex, Age, PAL, LDL, HDL, TG, History of CVD and T2DM as confounding factors. The results of LR in model are divided into A (both genders), B (male), and C (female) (Table 3A). In this part, all biochemical factors and clinical features entered the model in adjusted status and to examine the significance of each, non-significant variables were excluded from the model until all the biochemical factors and clinical features in the model become significant.

Table 3A shows the ORs and their 95% confidence intervals (CIs) for incident hs-CRP of linear regression in models A, B, and C. The results in Model A showed that the anxiety score, depression score, and SBP were significant ($P < 0.05$), while CVD and HTN were not significant. The most important variable with a high effect was the depression score with OR = 1.009 with 95% CI = (1.005, 1.013) and CVD with OR = 1.357 with 95% CI = (1.099, 1.677) respectively in models B and C. Also, the model C risk of incident hs-CRP for each unit increase in CVD = Positive was linearly increased by 0.357. In other words, for each unit increasing in CVD = Positive, the chance of incident hs-CRP ≥ 3 increases by 1.357 times but in the models A and B were not significant. The results of Models A and B showed that the depression score (OR = 1.003, 95%CI = (1.000, 1.006) and OR = 1.009, 95%CI = (1.005, 1.013)) respectively were increased hs-CRP ≥ 3 but model C was not significant. The results of all models showed the anxiety score was significant. Other biochemical factors and clinical features indices were excluded because of multicollinearity.

The method of the models in Table 3B is similar to the method of the models in Table 3A. Table 3B shows the anxiety score, depression score, SBP, and FBG were significant ($P < 0.05$) in model A, while the SBP of model B, and depression score in model C were not significant. The most important variable with a high effect was the anxiety score (OR = 1.009) in models A, and C and depression score (OR = 1.024) in model B.

A: Linear regression with log (hs-CRP)						
	Model A N = 9704		Model B (Male) N = 3885		Model C (Female) N = 5819	
Variables	Exp* (β (CI 95%))	P value	Exp (β (CI 95%))	P value	Exp (β (CI 95%))	P value
CVD	1.127 (0.978,1.298)	0.097	1.003 (0.830,1.211)	0.973	1.357 (1.099,1.677)	0.005
HTN	1.049 (0.981,1.122)	0.161	1.028 (0.930,1.137)	0.583	1.078 (0.985,1.179)	0.100
SBP	1.002 (1.001,1.004)	0.001	1.001 (0.999,1.004)	0.191	1.002 (1.000,1.004)	0.009
Anxiety Score	1.005 (1.003,1.008)	0.000	1.006 (1.002,1.010)	0.004	1.005 (1.002,1.008)	0.001
Depression Score	1.003 (1.000,1.006)	0.007	1.009 (1.005,1.013)	0.000	1.000 (0.997,1.003)	0.817
FBG (mg/dL)	1.003 (1.002,1.003)	0.000	1.002 (1.001,1.003)	0.000	1.003 (1.003,1.004)	0.000
B: Logistic regression with binary hs-CRP (hs-CRP < 3 vs hs-CRP > 3)						
	Model A N = 9704		Model B (Male) N = 3885		Model C (Female) N = 5819	
Variables	OR (CI 95%)	P value	OR (CI 95%)	P value	OR (CI 95%)	P value
CVD	1.289 (0.953,1.743)	0.100	1.257 (0.826,1.912)	0.286	1.419(0.911,2.208)	0.122
HTN	1.029 (0.886,1.195)	0.710	1.102 (0.867,1.402)	0.428	1.004(0.828,1.218)	0.966
SBP	1.006 (1.002,1.010)	0.001	1.004 (0.997,1.010)	0.243	1.007(1.002,1.011)	0.003
Anxiety Score	1.009 (1.004,1.015)	0.001	1.012 (1.002,1.022)	0.020	1.009(1.002,1.015)	0.011
Depression Score	1.007 (1.001,1.012)	0.020	1.024 (1.014,1.034)	0.000	0.998(0.991,1.005)	0.609
FBG (mg/dL)	1.006 (1.005,1.007)	0.000	1.004 (1.002,1.006)	0.000	1.007(1.006,1.008)	0.000

Table 3. Regression with log (hs-CRP) and with binary hs-CRP (hs-CRP < 3 vs hs-CRP > 3) response variable. The models adjusted for Sex, Age, PAL, LDL, HDL, TG, History of CVD and T2DM as confounding factors. *Exponential. *Abbreviations:* CVD cardiovascular disease, HTN hypertension, SBP systolic blood pressure, FBG fasting blood glucose.

The results of the LR training and testing confusion matrix for biochemical factors, and clinical features in sexual factors are shown in Table 7. The LR algorithm evaluated the various (hs-CRP < 3 mg/dL and hs-CRP ≥ 3mg/dL) risk factors and categorized them into two groups. Training and testing specificity ranged from 94 to 99.8% with highest in male's group. Precision in training and testing in LR model ranged from 48 to 80% with highest in testing the model for males. Lastly, accuracy ranged from 67.3 to 76.7% with highest in male's testing model. Overall, the results were more favorable in male subgroup followed by both genders and lastly, female's subgroup. The Area under the curve (AUC) and F1 score are also reported in the Table 7. The highest AUC belongs to female subgroup training (63.4%) and highest F1-score belongs to test for both genders (12.5%). Generally, the F1-score in LR model was lower compared to both DT models.

The association between biochemical factors, clinical features, and hs-CRP using DT models

Binary hs-CRP using DT models

Figure 1 illustrates the outcomes of the DT training for biochemical factors, and clinical features in male factors. The DT algorithm determined the various binary factor (hs-CRP < 3 mg/dL vs hs-CRP ≥ 3mg/dL) risk factors and categorized them into 3 layers. According to the DT model, the first variable (root) has the highest significance for classifying data, while the subsequent variables have lower significance. As shown in Fig. 1, Depression Score has the most crucial effect on hs-CRP development risk, followed by FBG, Cholesterol, and Anxiety Score. In the subgroup with 10 < depression score ≤ 22 and FBG > 97 mg/dL, 38.3% of participants were hs-CRP (highest risk of hs-CRP ≥ 3mg/dL). Meanwhile, among those with a Depression Score ≤ 10 and Cholesterol ≤ 217 mg/dL, 82.1% of subjects were identified as hs-CRP (lowest risk of hs-CRP < 3 mg/dL). Detailed rules for hs-CRP for males created by the DT model are demonstrated in Table 4.

The results of the DT training for biochemical factors, and clinical features in female factors are shown in Fig. 2. The DT algorithm evaluated the various hs-CRP < 3 mg/dL vs hs-CRP ≥ 3 mg/dL risk factors and categorized them into three layers. In the DT model, the first variable (root) is of the highest importance, with the following variables in the next levels of significance, accordingly. As shown in Fig. 2, FBG has the most crucial effect on hs-CRP development risk, followed by Anxiety Score, Cholesterol, TG, and Depression Score. In the subgroup with 97 < FBG ≤ 120, Anxiety Score ≤ 25, and Cholesterol > 241, 62.5% of participants also had FBG > 120 mg/dL, TG > 120 mg/dL, and Depression Score > 25, 57.7% of participants were hs-CRP (highest risk of hs-CRP ≥ 3 mg/dL). Meanwhile, among those with FBG ≤ 82 mg/dL, Cholesterol ≤ 158, and FBG ≤ 76, 84% of subjects were identified as hs-CRP (lowest risk of hs-CRP < 3 mg/dL). In layer 82 < FBG ≤ 97 and Anxiety Score ≤ 10 by increasing cholesterol, it increases the risk of certain hs-CRP ≥ 3 mg/dL. Also, with FBG > 120 mg/dL and TG > 120 mg/dL by increasing Depression Score, 12.6% of participants were hs-CRP (highest risk of hs-CRP ≥ 3 mg/dL). Detailed rules for hs-CRP for females created by the DT model are demonstrated in Table 4.

Figure 3 illustrates the outcomes of the DT training for biochemical factors, and clinical features in sexual factors. The DT algorithm determined the various binary factor (hs-CRP < 3 mg/dL vs hs-CRP ≥ 3mg/dL) risk factors and categorized them into 3 layers. According to the DT model, the first variable (root) has the highest significance for classifying data, while the subsequent variables have lower significance. Figure 3 illustrates that

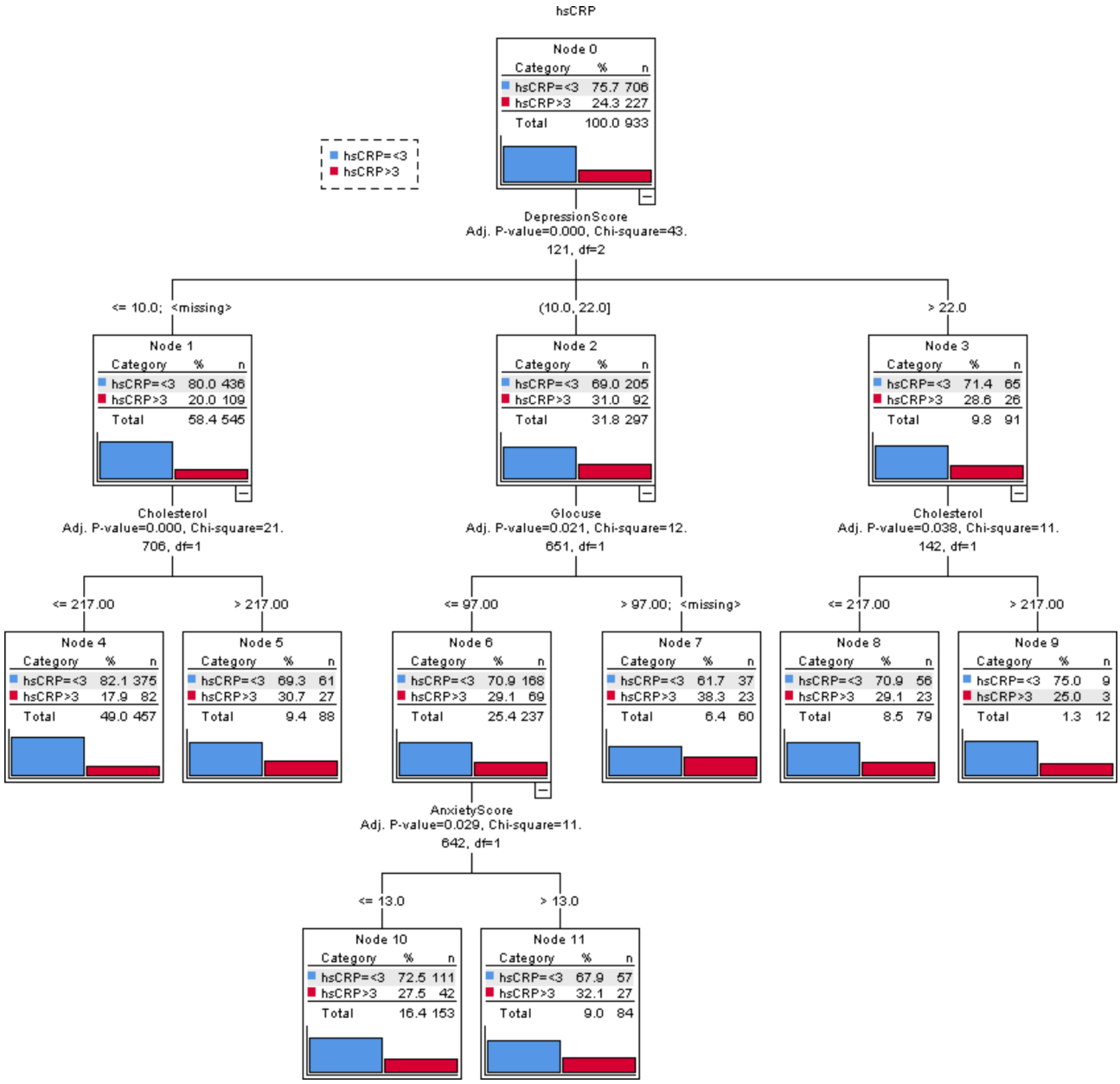

Fig. 1. Decision tree for hsCRP (binary hsCRP) event in male.

FBG followed by Cholesterol and Anxiety Score, Sex, and TG has the greatest impact on the hs-CRP presence risk. In the subgroup with FBG $>$ 120 mg/dL, Sex = Female, and TG $>$ 91mg/dL, 59.3% of participants were hs-CRP (highest risk of hs-CRP $\geq$ 3mg/dL). Participants with FBG $\leq$ 79 mg/dL, Cholesterol $\leq$ 159 mg/dL, and SBP $\leq$ 106.66 mmHg had lower hs-CRP $\geq$ 3mg/dL, according to the DT model, than those with higher TG and FBG levels (0.30 vs. 0.593 incident rate). Table 4 illustrates the specific hs-CRP rules developed by the DT model. Therefore, Anxiety Score, FBG, Depression Score, and Cholesterol were thus determined to be the most crucial variables in the DT model's sexual factor. Also, the confusion matrixes were presented in Table 5 for DT models binary hs-CRP. The highest AUC belongs to female training with 67.8% and the highest F1-score belongs to training in male subgroup with 86.7%. generally, the F1-score showed more favorable results in DT models compared to LR model (Table 5).

Tertial hs-CRP using DT model

Figure 4 illustrates the outcomes of the DT training for biochemical factors, and clinical features in male factors. The DT algorithm determined the various tertial factors (hs-CRP $\leq$ 1 mg/dL, 1 mg/dL $<$ hs-CRP $\leq$ 3 mg/dL, and hs-CRP $>$ 3) risk factors and categorized them into 3 layers. According to the DT model, the first variable (root) has the highest significance for classifying data, while the subsequent variables have lower significance. As shown in Fig. 4, FBG has the most crucial effect on hs-CRP

Male			
NO	**Rules**	**hs-CRP < 3 mg/dL (%)**	**hs-CRP ≥ 3 mg/dL (%)**
1	Depression Score ≤ 10 & Cholesterol ≤ 217 mg/dL	82.1	17.9
2	Depression Score ≤ 10 & Cholesterol > 217 mg/dL	69.3	30.7
3	10 < Depression Score ≤ 22 & FBG ≤ 97 mg/dL & Anxiety Score ≤ 13	72.5	27.5
4	10 < Depression Score ≤ 22 & FBG > 97 mg/dL & Anxiety Score > 13	67.9	32.1
5	10 < Depression Score ≤ 22 & FBG > 97	61.7	38.3
6	Depression Score > 22 & Cholesterol ≤ 217 mg/dL	70.9	29.1
7	Depression Score > 22 & Cholesterol > 217 mg/dL	75	25
Female			
NO	**Rules**	**hs-CRP < 3 mg/dL (%)**	**hs-CRP ≥ 3 mg/dL (%)**
1	FBG ≤ 82 mg/dL & Cholesterol ≤ 158 mg/dL & FBG ≤ 76 mg/dL	84	16
2	FBG ≤ 82 mg/dL & Cholesterol ≤ 158 mg/dL & FBG > 76 mg/dL	75.9	24.1
3	FBG ≤ 82 mg/dL & 158 mg/dL < Cholesterol ≤ 208 mg/dL & TG ≤ 63 mg/dL	77.1	22.9
4	FBG ≤ 82**mg/dL** & 158 mg/dL < Cholesterol > 208 mg/dL & TG > 63 mg/dL	76.4	23.6
5	FBG ≤ 82 mg/dL & Cholesterol > 208 mg/dL & LDL ≤ 123.1 mg/dL	76.9	32.1
6	FBG ≤ 82 mg/dL & Cholesterol > 208 mg/dL & LDL > 123.1 mg/dL	70.8	29.2
7	82 mg/dL < FBG ≤ 97 mg/dL & Anxiety Score ≤ 10 & Cholesterol ≤ 208 mg/dL	75.5	24.8
8	82 mg/dL < FBG ≤ 97**mg/dL** & Anxiety Score ≤ 10 & 208 < Cholesterol ≤ 221	66	34
9	82 mg/dL < FBG ≤ 97 mg/dL & Anxiety Score ≤ 10 & 221 mg/dL < Cholesterol ≤ 241 mg/dL	62.8	37.2
10	82 mg/dL < FBG ≤ 97 mg/dL & Anxiety Score ≤ 10 & Cholesterol > 241 mg/dL	62.8	37.2
11	82 mg/dL < FBG ≤ 97 mg/dL & Anxiety Score > 10 & Cholesterol ≤ 187 mg/dL	72.2	28.7
12	82 mg/dL < FBG ≤ 97 mg/dL & Anxiety Score > 10 mg/dL & Cholesterol > 187 mg/dL	63.9	36.1
13	97 mg/dL < FBG ≤ 120 mg/dL & Anxiety Score ≤ 25 & Cholesterol ≤ 158 mg/dL	51.7	48.3
14	97 mg/dL < FBG ≤ 120**mg/dL** & Anxiety Score ≤ 25 & 158 mg/dL < Cholesterol ≤ 241 mg/dL	62	38
15	97 mg/dL < FBG ≤ 120 mg/dL & Anxiety Score ≤ 25 & Cholesterol > 241 mg/dL	37.5	62.5
16	97 mg/dL < FBG ≤ 120 mg/dL & Anxiety Score > 25	51.7	48.3
17	FBG > 120**mg/dL** & TG ≤ 120 mg/dL	61.3	38.7
18	FBG > 120 mg/dL & TG > 120 mg/dL & Depression Score ≤ 10	54.9	45.1
19	FBG > 120 mg/dL & TG > 120**mg/dL** & 10 < Depression Score ≤ 25	53.8	46.2
20	FBG > 120 mg/dL & TG > 120 mg/dL & Depression Score > 25	42.3	57.7
Male & Female			
NO	**Rules**	**hs-CRP < 3 mg/dL (%)**	**hs-CRP ≥ 3 mg/dL (%)**
1	FBG ≤ 79 mg/dL & Cholesterol ≤ 159 mg/dL & SBP ≤ 106.66	81.2	18.8
2	FBG ≤ 79 mg/dL & Cholesterol ≤ 159 mg/dL & SBP > 106.66	85.1	14.9
3	FBG ≤ 79 mg/dL & 159 mg/dL < Cholesterol ≤ 208 mg/dL & Sex = Male	79.9	20.1
4	FBG ≤ 79 mg/dL & 159 mg/dL < Cholesterol ≤ 208 mg/dL & Sex = Female	76.1	23.9
6	FBG ≤ 79 mg/dL & Cholesterol > 208 mg/dL & HTN^-	75.6	24.7
7	FBG ≤ 79 mg/dL & Cholesterol > 208 mg/dL & HTN^+	68.8	31.2
8	79 mg/dL < FBG ≤ 98 mg/dL & Anxiety Score ≤ 10 & Sex = Male	76.2	23.8
9	79 mg/dL < FBG ≤ 98 mg/dL & Anxiety Score ≤ 10 & Sex = Female	69.1	30.9
10	79 mg/dL < FBG ≤ 98 mg/dL & Anxiety Score > 10 & Cholesterol ≤ 159 mg/dL	62.3	37.7
11	79 mg/dL < FBG ≤ 98**mg/dL** & Anxiety Score > 10 & 159 mg/dL < Cholesterol > 187 mg/dL	79.5	20.5
12	79 mg/dL < FBG ≤ 98 mg/dL & Anxiety Score > 10 & Cholesterol > 187 mg/dL	64.1	35.9

Continued

Male & Female			
NO	Rules	hs-CRP $<$ 3 mg/dL (%)	hs-CRP $\geq$ 3 mg/dL (%)
13	98 mg/dL $<$ FBG $\leq$ 120 mg/dL & Sex = Male	66	34
14	98 mg/dL $<$ FBG $\leq$ 120 mg/dL & Sex = Female & PAL $\leq$ 1.593	62.5	37.5
15	98 $<$ FBG $\leq$ 120 mg/dL & Sex = Female & PAL $>$ 1.593	48.5	53.5
16	FBG $>$ 120 mg/dL & Sex = Male & Depression Score $\leq$ 15	77.6	22.4
17	FBG $>$ 120 mg/dL & Sex = Male & Depression Score $>$ 15	69	31
18	FBG $>$ 120 mg/dL & Sex = Female & TG $\leq$ 91	70	30
19	FBG $>$ 120 mg/dL & Sex = Female & TG $>$ 91	40.7	59.3

Table 4. Detailed rules based on DT models with hs-CRP binary.

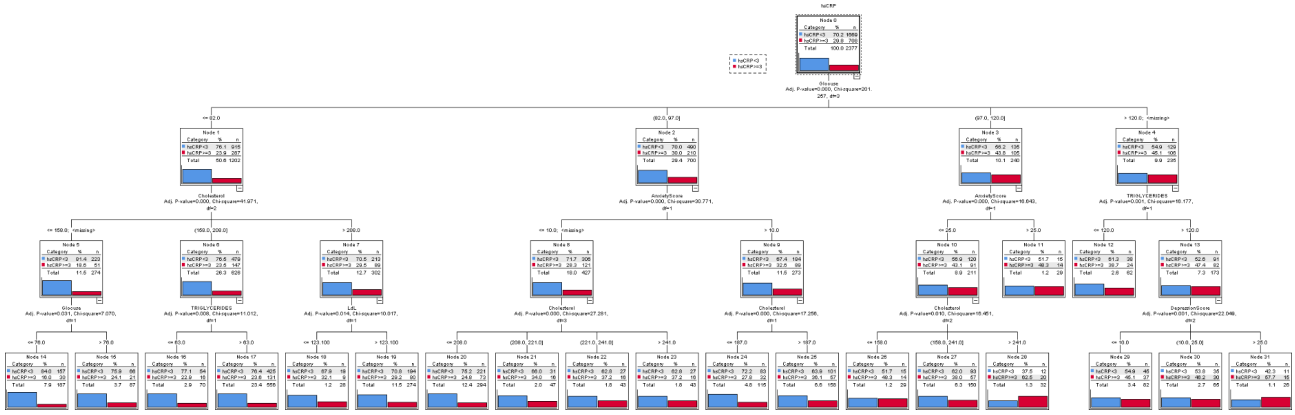

Fig. 2. Decision tree for hsCRP (binary hsCRP) event in female.

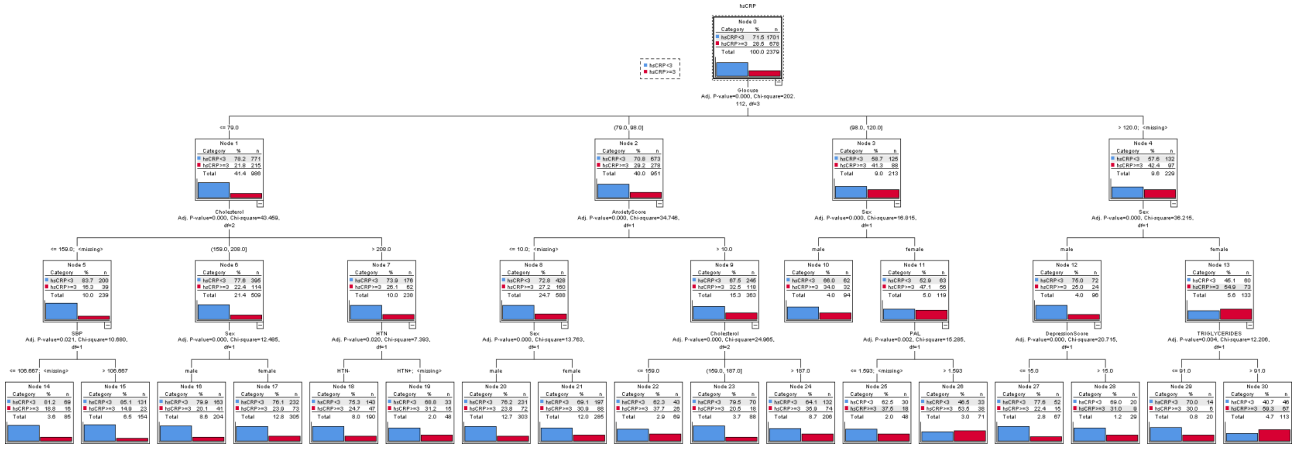

Fig. 3. Decision tree for hsCRP (binary hsCRP) event in sexual.

development risk, followed by Depression Score, Cholesterol, and SBP. In the subgroup with FBG $>$ 97 and Depression Score $\leq$ 16 and increasing SBP increased hs-CRP levels (0.25 vs. 0.26 incident rate in hs-CRP $\leq$ 1mg/dL and0.50 vs.0.56 incident rate in 1 mg/dL $<$ hs $-$ CRP $\leq$ 3mg/dL). Meanwhile, among those with FBG$\leq$82, Cholesterol$\leq$193 and History.CVD+, 17% of subjects were identified as hs-CRP (lowest risk of hs-CRP$<$3 mg/dL), and 48.1% of subjects were identified as hs-CRP $\leq$ 1 mg/dL. Detailed rules for hs-CRP for males created by the DT model are demonstrated in Table 6.

The results of the DT training for biochemical factors, and clinical features in female factors are shown in Fig. 5. The DT algorithm evaluated the various (hs-CRP $\leq$ 1 mg/dL, 1 mg/dL$<$ hs-CRP $\leq$ 3 mg/dL, and hs-CRP $>$ 3mg/dL) risk factors and categorized them into three layers. In the DT model, the first variable (root) is of the highest importance, with the following variables in the next levels of significance, accordingly. As shown in Fig. 5, FBG has the most crucial effect on hs-CRP development risk, followed by Cholesterol, SBP,

DT model binary for sexual

(a) Training (n =7157)

Actual	Predicted Count	
	hsCRP<3	hsCRP>=3
hsCRP<3	4877	262
hsCRP>=3	1761	356

Sensitivity =94.9% | Precision =73 % | Accuracy =72.1%

(b) Testing (n =2379)

Actual	Predicted Count	
	hsCRP<3	hsCRP>=3
hsCRP<3	1605	96
hsCRP>=3	584	94

Sensitivity =94.3 % | Precision =73.3 % | Accuracy = 71.4%

DT model binary for Male

(a) Training (n =2913)

Actual	Predicted Count	
	hsCRP<3	hsCRP<3
hsCRP<3	2203	25
hsCRP>=3	654	31

Sensitivity =95.5 % | Precision =77 % | Accuracy= 76.7%

(b) Testing (n =933)

Actual	Predicted Count	
	hsCRP<3	hsCRP<3
hsCRP<3	697	9
hsCRP>=3	224	3

Sensitivity =98.7 % | Precision =75 % | Accuracy = 75%

DT model binary for Female

(c) Training (n =7258)

Actual	Predicted Count	
	hsCRP<3	hsCRP<3
hsCRP<3	4974	197
hsCRP>=3	1827	260

Sensitivity =75.7 % | Precision =73.1 % | Accuracy =72.1%

(d) Testing (n =2377)

Actual	Predicted Count	
	hsCRP<3	hsCRP<3
hsCRP<3	1596	73
hsCRP>=3	629	79

Sensitivity =95.6% | Precision =71.7 % | Accuracy = 70.5%

Table 5. Performance indices of the DT models with binary hsCRP.
[a]General structure of compounds **10a-n**[b]IC_{50} for positive control (Acarbose): 750.1 ± 1.3 μM[c]Values are the mean ± SD. All experiments were performed at least three independent assays

TG, and Depression Score. In the subgroup with FBG $>$ 98, Cholesterol $>$ 241 and Depression Score $\leq$ 16, 54.7% of participants also had FBG $>$ 98 mg/dL, Cholesterol $>$ 241 mg/dL, and Depression Score $>$ 16, 65.4% of participants were hs-CRP (highest risk of hs-CRP $\geq$ 3 mg/dL). Meanwhile, among those with FBG $>$ 98 mg/dL, 187 mg/dL$<$ Cholesterol $\leq$ 241 mg/dL, and HTN^{-}, 6.3% of subjects were identified as hs-CRP (lowest risk of hs-CRP $<$ 1 mg/dL). In layer 82 $<$ FBG $\leq$ 86 and SBP $>$ 128.667 by increasing TG, it increases the risk of certain hs-CRP $\geq$ 3 mg/dL. Detailed rules for hs-CRP for females created by the DT model are demonstrated in Table 6.

The results of the DT training for biochemical factors, and clinical features in sexual factors are shown in Fig. 6. The DT algorithm evaluated the various (hs-CRP $\leq$ 1 mg/dL, 1 mg/dL$<$ hs-CRP $\leq$ 3 mg/dL, and hs-CRP $>$ 3mg/dL) risk factors and categorized them into three layers. In the DT model, the first variable (root) is of the highest importance, with the following variables in the next levels of significance, accordingly. As shown in Fig. 6, FBG has the most crucial effect on hs-CRP development risk, followed by Cholesterol, sexual, Anxiety Score, and Depression Score. In the subgroup with 79 $<$ FBG $\leq$ 85, 145 $<$ Cholesterol $\leq$ 222 increasing the Anxiety Score increases the risk of certain hs-CRP $\geq$ 3 mg/dL and hs-CRP $\leq$ 1 mg/dL also decreasing the risk of certain 1 $<$ hs-CRP $\leq$ 3. Meanwhile, among those with 85 mg/dL$<$ FBG $\leq$ 97 mg/dL, Cholesterol $>$ 241 mg/dL, and Sex = Female, 61.1% of subjects were identified as hs-CRP (highest risk of 1 mg/dL$<$ hs-CRP $\leq$ 3 mg/dL). In layer 85 $<$ FBG $\leq$ 97, 145 $<$ Cholesterol $\leq$ 188 by increasing LDL, it increases the risk of certain 1 mg/dL$<$hs-CRP $\leq$ 3 mg/dL but decreases another various hs-CRP. Detailed rules for hs-CRP for females

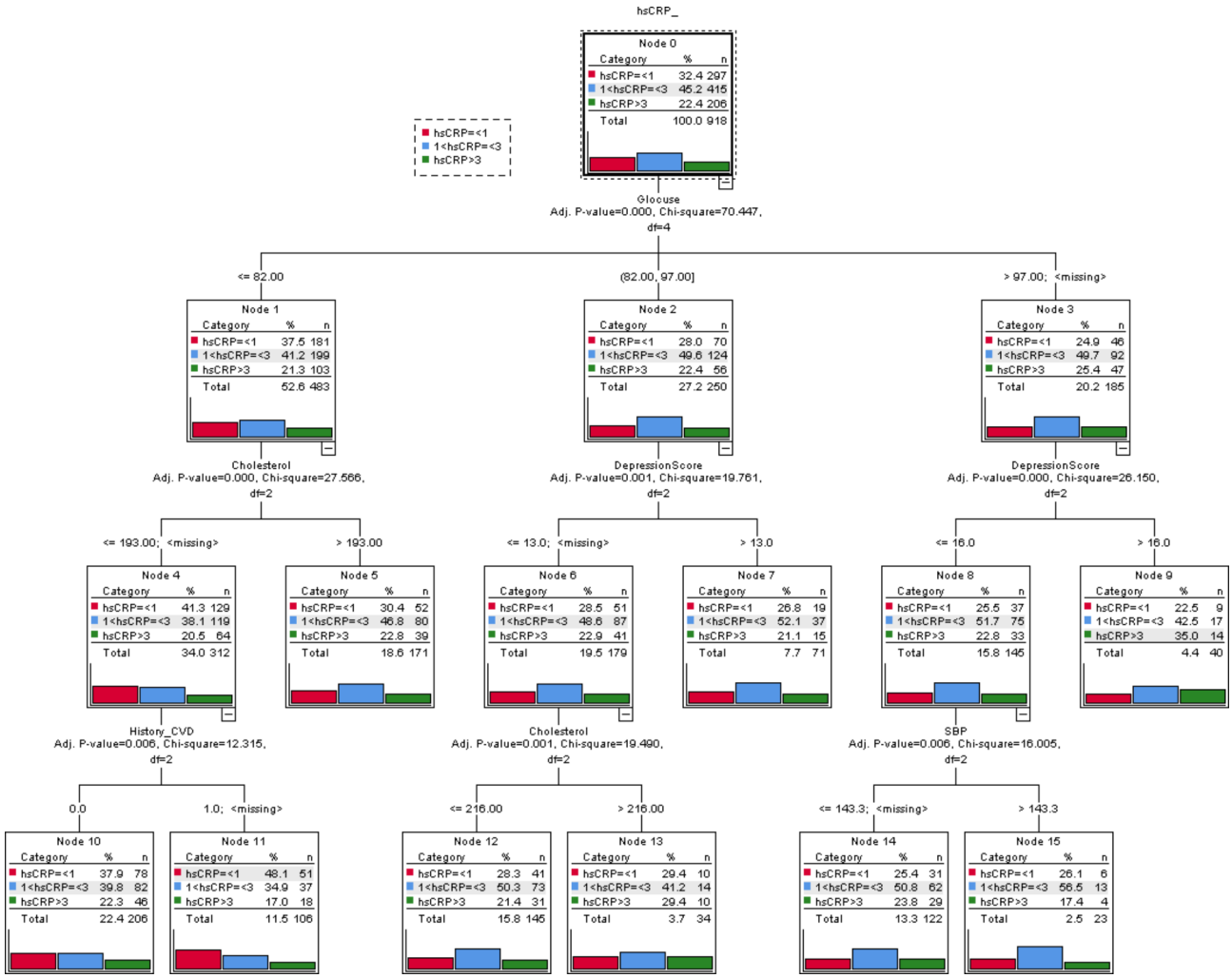

Fig. 4. Decision tree for hsCRP (tertials hsCRP) event in male.

created by the DT model are demonstrated in Table 6. Also, the confusion matrixes were presented in Table 7 for DT models tertial hs-CRP. Table 7 illustrate also the F1-score and AUC of this model. The highest F1-score belongs to female subgroup training with 29.59%. the highest AUC belongs to testing model for both genders with 74.7%. generally, the F1-score in this model were not as favorable as binary DT model but was superior to LR model.

Discussion

CRP is the most-known and available serum factor that can indicate inflammation and inflammatory-related processes in the human body. Hs-CRP as a low-grade inflammatory marker presented as a key marker in epidemiological and cardiovascular studies can be altered by a wide range of variables in the human body which all somehow initiate the inflammatory response[20]. Similar studies had assessed the effect of hs-CRP in chronic coronary syndrome patients. but to our knowledge there are not similar studies that asses and predict hs-CRP in large population based on Community collection regardless of ACS[21,22]. We did not set ACS or any other disease as inclusion criteria which make the studies more relatable to society population. Additionally, ethnicity and culture play an important role on the findings of similar studies and previous studies had different ethnicity comparing to our study[21,22]. Thus, this cohort analysis study investigated the demographic and hematological features that might have a relation to the level of hs-CRP. The findings can be assed further with different ethnicities and populations to be included in the guidelines and help physicians in better understand, manage and treat related diseases and the role of hs-CRP in prognosis and incidence of such diseases.

On the other hand, Table 3 Baseline characteristics which have been divided by the tertial of the hs-CRP, indicated that almost all of the mentioned characteristics had been significantly associated with the level of hs-CRP. For instance, age, HDL, LDL, triglyceride anxiety score, depression score, HTN, T2DM, and even family history of T2DM were all significant. Cardiovascular family history and PAL were not associated with the level of hs-CRP in the baseline characteristics. As known and justified before HTN and T2DM both can cause and affect some level of inflammatory response in the human body by various mechanisms. There are several studies on cell–cell interactions from intracellular to extracellular mechanisms that indicate inflammation pathways related

Male				
NO	**Rules**	**hsCRP ≤ 1 (%)**	**1 < hsCRP ≤ 3 (%)**	**hsCRP > 3 (%)**
1	Glucose ≤ 82& Cholesterol ≤ 193 & History.CVD$^-$	37.9	39.8	22.3
2	Glucose ≤ 82& Cholesterol ≤ 193 & History.CVD$^+$	48.1	34.9	17
3	Glucose ≤ 82& Cholesterol > 193	30.4	46.8	22.8
4	82 < Glucose ≤ 97& Depression Score ≤ 13 & Cholesterol ≤ 216	28.3	50.3	21.4
5	82 < Glucose ≤ 97& Depression Score ≤ 13 & Cholesterol > 216	29.4	41.2	29.4
6	82 < Glucose ≤ 97& Depression Score > 13	26.8	52.1	21.1
7	Glucose > 97 & Depression Score ≤ 16 & SBP ≤ 143.3	25.4	50.8	23.8
8	Glucose > 97 & Depression Score ≤ 16 & SBP > 143.3	26.1	56.5	17.4
9	Glucose > 97 & Depression Score > 16	22.5	42.5	35
Female				
NO	**Rules**	**hsCRP ≤ 1 (%)**	**1 < hsCRP ≤ 3 (%)**	**hsCRP > 3 (%)**
1	Glucose ≤ 82& Cholesterol ≤ 150	42.9	42.5	14.6
2	Glucose ≤ 82& 150 < Cholesterol ≤ 208 & Triglycerides ≤ 91	35.6	45.8	18.5
3	Glucose ≤ 82& 150 < Cholesterol ≤ 208 & Triglycerides > 91	28.2	46.8	25
4	Glucose ≤ 82& Cholesterol > 208 & HTN$^-$	21.6	49.8	28.6
5	Glucose ≤ 82& Cholesterol > 208 & HTN$^+$	19.4	45.2	35.5
6	82 < Glucose ≤ 86 & SBP ≤ 128.667 & Cholesterol ≤ 178	22	50	28
7	82 < Glucose ≤ 86 & SBP ≤ 128.667 & Cholesterol > 178	27.2	45.6	27.2
8	82 < Glucose ≤ 86 & SBP > 128.667 & Triglycerides ≤ 91	29	41.9	29
9	82 < Glucose ≤ 86 & SBP > 128.667 & Triglycerides > 91	11.3	47.2	41.5
10	86 < Glucose ≤ 98& Cholesterol ≤ 187 & LDL ≤ 97.81	30.5	44.8	24.8
11	86 < Glucose ≤ 98& Cholesterol ≤ 187 & 97.81 < LDL ≤ 114.72	21.4	53.6	25
12	86 < Glucose ≤ 98& Cholesterol ≤ 187 & LDL > 114.72	31	43.8	20.7
13	86 < Glucose ≤ 98& 187 < Cholesterol ≤ 241 & SBP ≤ 106.66	18.5	59.3	22.2
14	86 < Glucose ≤ 98& 187 < Cholesterol ≤ 241 & 106.66 < SBP ≤ 135	25	35.3	39.7
15	86 < Glucose ≤ 98 & 187 < Cholesterol ≤ 241 & SBP > 135	16.3	44.9	38.8
16	86 < Glucose ≤ 98 & Cholesterol > 241	7.7	51.9	40.4
17	Glucose > 98 & Cholesterol ≤ 145	29	41.9	29
18	Glucose > 98 & 145 < Cholesterol ≤ 187 & Glucose ≤ 119	14.3	50	35.7
19	Glucose > 98 & 145 < Cholesterol ≤ 187 & Glucose > 119	8.1	39.2	52.7
20	Glucose > 98 & 187 < Cholesterol ≤ 241 & HTN$^-$	11.3	45.4	43.3
21	Glucose > 98 & 187 < Cholesterol ≤ 241 & HTN$^-$	6.3	45.2	47.9
22	Glucose > 98 & Cholesterol > 241 & Depression Score ≤ 16	7.5	37.7	54.7
23	Glucose > 98 & Cholesterol > 241 & Depression Score > 16	3.8	30.8	65.4
Female & Male				
NO	**Rules**	**hsCRP ≤ 1 (%)**	**1 < hsCRP ≤ 3 (%)**	**hsCRP > 3 (%)**
1	Glucose ≤ 79& Cholesterol ≤ 159	41.1	43.5	15.4
2	Glucose ≤ 79& 159 < Cholesterol ≤ 188 & Sex = Male	35.5	40	23.9
3	Glucose ≤ 79& 159 < Cholesterol ≤ 188 & Sex = Female	32.4	44.9	22.7
4	Glucose ≤ 79& 188 < Cholesterol ≤ 198 & Sex = Male	42.9	45.7	11.4
5	Glucose ≤ 79& 159 < Cholesterol ≤ 198 & Sex = Female	18.6	34.9	45.5
6	Glucose ≤ 79& Cholesterol > 1988	23.3	49.3	27.4
7	79 < Glucose ≤ 85 & Cholesterol ≤ 145	32.6	45.5	20.9
8	79 < Glucose ≤ 85 & 145 < Cholesterol ≤ 222 & Anxiety Score ≤ 2	22.2	55.6	22.2
9	79 < Glucose ≤ 85 & 145 < Cholesterol ≤ 222 & Anxiety Score > 2	31.1	40.7	28.2
10	79 < Glucose ≤ 85 & Cholesterol > 222	21.3	49.3	29.3
11	85 < Glucose ≤ 97 & Cholesterol ≤ 145	22	51.2	26.8
12	85 < Glucose ≤ 97 & 145 < Cholesterol ≤ 188 & LDL ≤ 98	23.3	47.7	29.1
13	85 < Glucose ≤ 97 & 145 < Cholesterol ≤ 188 & LDL > 98	22.2	57.3	20.5
14	85 < Glucose ≤ 97 & 188 < Cholesterol ≤ 241 & Sex = Male	28.2	44.7	27.1
15	85 < Glucose ≤ 97 & 188 < Cholesterol ≤ 241 & Sex = Female	16.2	41.9	41.9
16	85 < Glucose ≤ 97 & Cholesterol > 241 & Sex = Male	7.7	53.8	38.5
Continued				

Female & Male				
NO	**Rules**	**hsCRP ≤ 1 (%)**	**1 < hsCRP ≤ 3 (%)**	**hsCRP > 3 (%)**
17	85 < Glucose ≤ 97 & Cholesterol > 241 & Sex = Female	8.3	61.1	30.6
18	85 < Glucose ≤ 97 & Cholesterol ≤ 159	12.1	42.4	45.5
19	85 < Glucose ≤ 97 & 159 < Cholesterol ≤ 222 & Sex = Male	28.1	42.1	29.8
20	85 < Glucose ≤ 97 & 159 < Cholesterol ≤ 222 & Sex = Female	15.7	34.9	49.4
21	85 < Glucose ≤ 97 & 222 < Cholesterol ≤ 241	11.5	26.9	61.5
22	85 < Glucose ≤ 97 & Cholesterol > 241	5.9	38.2	55.9
23	Glucose > 119 & Cholesterol ≤ 145 & Sex = Male & Depression Score ≤ 15	25.9	50.6	23.5
24	Glucose > 119 & Cholesterol ≤ 145 & Sex = Male & Depression Score > 15	9.5	38.1	52.4
25	Glucose > 119 & Cholesterol ≤ 145 & Sex = Female & HTN^-	12	34.8	53.3
26	Glucose > 119 & Cholesterol ≤ 145 & Sex = Female & HTN^+	5.9	47.1	47.1

Table 6. Detailed rules based on DT models with tertile divided hsCRP.

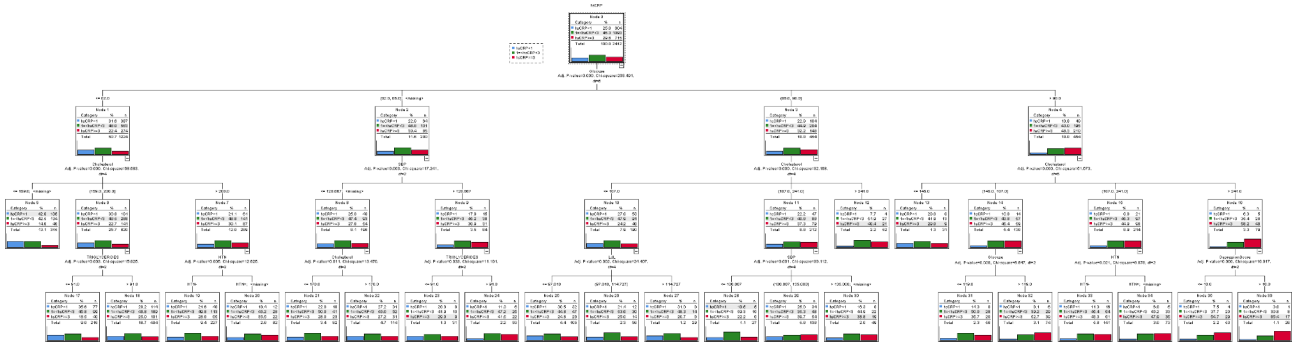

Fig. 5. Decision tree for hsCRP (tertials hsCRP) event in female.

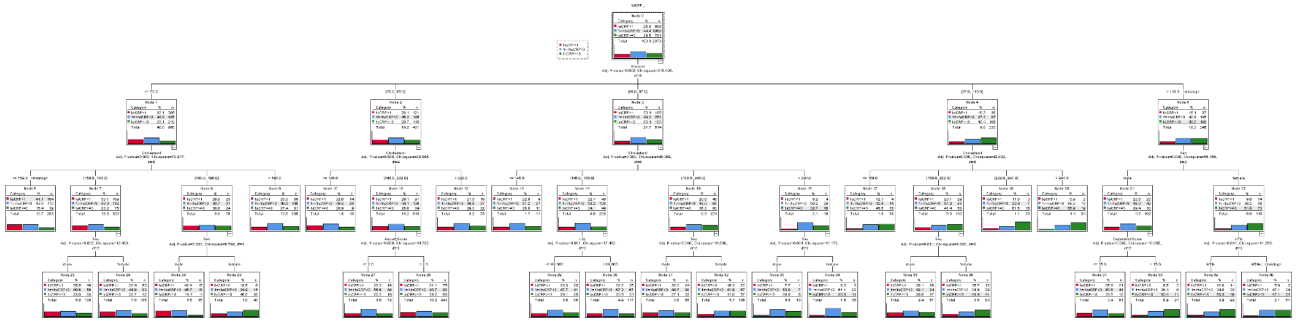

Fig. 6. Decision tree for hsCRP (tertials hsCRP) event in sexual.

to increase of hs-CRP levels[23–28]. Some meta-analyses indicate that a higher level of hs-CRP is directly associated with a higher risk of both T2DM and HTN[29–31]. Additionally, studies have found that a family history of T2DM is associated with the occurrence of T2DM in one self's which indicates that genetics are one of the vital role players in the occurrence of T2DM[10,32].

In both linear regression with log and logistic regression with binary hs-CRP, the results in males and females are pretty similar (Table 3). Studies with similar analysis indicated the importance and priority of LR model than other models in predicting hs-CRP level and their correlation with other factors[33]. In both modes, males' FBG levels as well as depression and anxiety scores were significantly related to the level of hs-CRP whereas in females SBP, CVD, FBG level, and anxiety score were significant. We tried to include variables which are easy to asceses and asses and can be evaluated in regions with no access to laboratory settings. As previously resulted inflammatory markers especially hs-CRP are capable of being a predictor of CVD in women. Previous studies have also resulted in the fact that changes in the cardiovascular system such as an increase in blood pressure and cardiovascular events have a more intense effect on the level of hs-CRP in women compared to men[34]. As

DT model binary for sexual

(c) Training (n =7157)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	322	1458	100
hsCRP>=3	305	2599	383
	120	1448	526
Sensitivity =94.9%		Precision =73 %	Accuracy =72.1%

(d) Testing (n =2379)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	104	468	34
hsCRP>=3	110	828	128
	39	507	155
Sensitivity =94.3 %		Precision =73.3 %	Accuracy = 71.4%

DT model binary for Male

(e) Training (n =2913)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	138	702	22
hsCRP>=3	122	1194	43
	37	601	68
Sensitivity =95.5 %		Precision =77 %	Accuracy= 76.7%

(f) Testing (n =933)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	51	237	9
hsCRP>=3	37	361	17
	18	174	14
Sensitivity =98.7 %		Precision =75 %	Accuracy = 75%

DT model binary for Female

(g) Training (n =7258)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	406	1355	121
hsCRP>=3	396	2482	382
	185	1441	454
Sensitivity =75.7 %		Precision =73.1 %	Accuracy =72.1%

(h) Testing (n =2377)

Actual		Predicted Count	
	hsCRP<3		hsCRP>=3
hsCRP<3	140	424	40
hsCRP>=3	150	807	136
	52	524	139
Sensitivity =95.6%		Precision =71.7 %	Accuracy = 70.5%

Table 7. Performance indices of the DT models with tertial hsCRP.
[a]General structure of compounds **10a-n**[b]IC_{50} for positive control (Acarbose): 750.1 ± 1.3 μM[c]Values are the mean ± SD. All experiments were performed at least three independent assays

consistent in previous studies the mean level of CRP is higher in females than males and this might cause the threshold level of CRP to reveal as significant in women rather than men[35]. Another thought-provoking finding is the alteration of hs-CRP level by anxiety and depression. This fact has been argued before by several studies as direct and direct causes such as obesity, and smoking[36,37]. The exact mechanism of the direct effect of anxiety on CRP levels is not clear but studies suggest that the endocrine and noninflammatory dysfunction are involved[38]. Also, as expected following the previous findings in this matter, FBG level has been related to the level of hs-CRP as an inflammatory factor with several proposed mechanisms[5]. In model A consisting of both genders all previously four discussed variables (anxiety and depression score, FBG level, and SBP) have been significantly related to the level of hs-CRP in both linear regression and logistic regression. Our findings in Table 3 are aligned with previous findings and suggest more coherent information in this area.

Leading to the next table, our findings indicate that in the first node depression score and cholesterol level are determinative variables as in female's FBG level and cholesterol are the ones in charge. As the third section indicates FBG, SBP, and cholesterol are the three major variables in the first node for both genders suggesting the significance of cholesterol in shaping the decision tree in the first place. Cholesterol is also the key variable in the first node in Table 6 which are rules based on DT models with tertial divided hs-CRP. In this table cholesterol among history of CVD and FBG in males and cholesterol and FBG in females and both genders are present in the first node.

Lastly bringing all the data together on the DT model with binary and tertial hs-CRP, findings show the results of these DT models in predicting the hs-CRP based on other variables. Generally, the accuracy, sensitivity, and precession numbers were acceptable considering that hs-CRP is a highly sensitive factor and most human

conditions and statuses can significantly alter the level of this nonspecific inflammatory factor. In DT models with binary hs-CRP, the range of accuracy was between 71.4% to 76.7% with the highest for training DT for males and the lowest for testing DT model for both genders. In the matter of sensitivity, the results were mainly in a higher percentage than the precision with the highest of 98.7 in the testing of males' DT model and the lowest of 75.7% in the training of females' DT model. These results indicate that the DT model can distinguish the true positive and false negative quite successfully but there are some false positive cases which lowered the precision percentage. This might be because hs-CRP can be altered by other variables that were not included in this cohort study. Considering Table 7, the results are nearly the same with the highest accuracy of 76.7% in training of DT model for males and the lowest of 70.5% in testing the DT model for females. Here again, the precision number (highest of 77% in training DT for males and lowest of 71.7% in testing DT for females) is lower than sensitivity (highest of 98.7% in testing DT for males and lowest of 75.7% in training DT for females) in total. By predicting hs-CRP level based on demographic data clinicians can better evaluate the patients on similar population by their demographic data and assess the variables and their role in diseases that are strongly linked to the hs-CRP level. Factors such as depression (based on questioner), cholesterol level and FBG showed to have a significant effect on hs-CRP prediction. Other study had indicated that triglyceride-glucose index can be significantly related to the hs-CRP levels and could help physicians in early prediction and management of ACS patients[22]. Similar data on hs-CRP level and their association with coronary heart disease (CHD) also indicated that DT model portrait favorable findings and results for prediction. The presence of was strongly associated with hs-CRP in DT model[39].

Limitations

There are some limitations on this study that should be asses. Firstly, the DT and LR model used on this study was not further assessed in other datasets. Future studies can use these findings and similar analysis in some other data with various ethnicities to validate these findings. It is crucial for the future of ML and disease prediction that the models can be implied in as much population as possible.

Secondly, follow-ups data would benefit the study by following the patients throw diseases that can be related to hs-CRP level and assess the consequences and prediction of disease. Similar study designs in future can consider follow-ups for better monitoring and evaluating the results on this. By following the patient's morbidity and mortality data analysis could significantly help the data for better understanding of the significant of hs-CRP.

Lastly, the study would benefit from other models to enhance the prediction of hs-CRP level. We recommend that further studies focus on these issues to better clear out the missing part of this findings. Other similar artificial learning models specially ML and neural network analysis could be helpful in comparing various data analysis and their results.

Conclusion

Considering the relatively impressable nature of hs-CRP as an inflammatory factor, the DT model was acceptably able to predict the hs-CRP level in this great cohort study. The results also indicated that factors such as anxiety and depression scores alongside FBG and systolic blood pressure and history of cardiovascular diseases are some of the main factors that can alter the level of hs-CRP. Our results are aligned with previous findings in this area affecting the hs-CRP level in which our study provided comprehensive and detailed findings about hs-CRP and various variables affecting it.

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Received: 30 June 2024; Accepted: 28 November 2024
Published online: 05 December 2024

References

1. Chen, L. et al. Inflammatory responses and inflammation-associated diseases in organs. *Oncotarget* **9**, 7204 (2018).
2. Shrivastava, A. K., Singh, H. V., Raizada, A. & Singh, S. K. C-reactive protein, inflammation and coronary heart disease, Egypt. *Hear. J.* **67**, 89–97 (2015).
3. Li, Y. et al. Hs-CRP and all-cause, cardiovascular, and cancer mortality risk: a meta-analysis. *Atherosclerosis* **259**, 75–82 (2017).
4. Abou-Raya, S., Abou-Raya, A., Naim, A. & Abuelkheir, H. Chronic inflammatory autoimmune disorders and atherosclerosis. *Ann. N. Y. Acad. Sci.* **1107**, 56–67 (2007).
5. Aronson, D. et al. Association between fasting glucose and C-reactive protein in middle-aged subjects. *Diabet. Med.* **21**, 39–44 (2004).
6. Huang, Y., Su, Y., Chen, H., Liu, H. & Hu, J. Serum levels of CRP are associated with depression in a middle-aged and elderly population with diabetes mellitus: a diabetes mellitus-stratified analysis in a population-based study. *J. Affect. Disord.* **281**, 351–357 (2021).
7. Charbuty, B. & Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2**, 20–28 (2021).
8. Mehrpour, O. et al. Utility of artificial intelligence to identify antihyperglycemic agents poisoning in the USA: introducing a practical web application using National Poison Data System (NPDS). *Environ. Sci. Pollut. Res.* **30**, 57801–57810 (2023).
9. Ghayour-Mobarhan, M. et al. Mashhad stroke and heart atherosclerotic disorder (MASHAD) study: design, baseline characteristics and 10-year cardiovascular risk estimation. *Int. J. Public Health* **60**, 561–572 (2015).
10. Harrison, T. A. et al. Family history of diabetes as a potential public health tool. *Am. J. Prev. Med.* **24**, 152–159 (2003).
11. G.A. of the W.M. Association, *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects*, J. Am. Coll. Dent. 81 (2014), pp. 14–18.

12. Castelli, W. P. et al. Incidence of coronary heart disease and lipoprotein cholesterol levels: the framingham study. *Jama* **256**, 2835–2838 (1986).
13. *Noncommunicable Diseases Global Monitoring Framework: Indicator Definitions and Specifications*. Available at https://www.who.int/publications/m/item/noncommunicable-diseases-global-monitoring-framework-indicator-definitions-and-specifications.
14. Giles, T. D., Materson, B. J., Cohn, J. N. & Kostis, J. B. Definition and classification of hypertension: an update. *J. Clin. Hypertens.* **11**, 611–614 (2009).
15. Inventory-Ii, B. D. Beck depression inventory-II. *Corsini Encycl. Psychol.* **1**(1), 210 (2010).
16. Ghazizadeh, H., Shakour, N., Ghoflchi, S., Mansoori, A., Saberi-Karimiam, M., Rashidmayvan, M., Ferns, G., Esmaily, H., Ghayour-Mobarhan, M. Use of data mining approaches to explore the association between type 2 diabetes mellitus with SARS-CoV-2. *BMC Pulm. Med.* **23**(1) (2023). https://doi.org/10.1186/s12890-023-02495-4
17. Poudineh, M., Mansoori, A., Sadooghi Rad, E., Hosseini, Z. S., Salmani Izadi, F., Hoseinpour, M., Mahmoudi Zo M., Ghoflchi, S., Tanbakuchi, D., Nazar, E., Ferns, G., Effati, S., Esmaily, H., Ghayour-Mobarhan, M. Platelet distribution widths and white blood cell are associated with cardiovascular diseases: data mining approaches. *Acta Cardiologica.* **78**(9) 1033–1044 (2023). https://doi.org/10.1080/00015385.2023.2246199
18. Mansoori, A., Farizani Gohari, N. S., Etemad, L., Poudineh, M., Ahari, R. K., Mohammadyari, F., Azami, M., Rad, E. S., Ferns, G., Esmaily, H., Ghayour Mobarhan, M. White blood cell and platelet distribution widths are associated with hypertension: data mining approaches. *Hypertens. Res.* **47**(2) 515–528 (2024). https://doi.org/10.1038/s41440-023-01472-y
19. Mansoori, A., Hosseini, Z. S., Ahari, R. K., Poudineh, M., Rad, E.S., Zo, M.M., Izadi, F.S., Hoseinpour, M., Miralizadeh, A., Mashhadi, Y.A., Hormozi, M., Firoozeh, M.T., Hajhoseini, O., Ferns, G., Esmaily, H., Mobarhan, M. G. (2023) Development of Data Mining Algorithms for Identifying the Best Anthropometric Predictors for Cardiovascular Disease: MASHAD Cohort Study. *HIgh Blood Press Car.***30**(3) 243–253.
20. Ridker, P. M. & Silvertown, J. D. Inflammation, C-reactive protein, and atherothrombosis. *J. Periodontol.* **79**, 1544–1551 (2008).
21. Liu, H.-H. et al. High-sensitivity C-reactive protein and hypertension: combined effects on coronary severity and cardiovascular outcomes. *Hypertens. Res.* **42**, 1783–1793 (2019).
22. Li, Q. et al. The combined effect of triglyceride-glucose index and high-sensitivity C-reactive protein on cardiovascular outcomes in patients with chronic coronary syndrome: A multicenter cohort study. *J. Diabetes* **16**, e13589 (2024).
23. Li, S. et al. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.* **46**, D106–D112 (2018).
24. Li, Y. et al. EV-origin: Enumerating the tissue-cellular origin of circulating extracellular vesicles using exLR profile. *Comput. Struct Biotechnol. J.* **18**, 2851–2859 (2020).
25. Su, Y. et al. Plasma extracellular vesicle long RNA profiles in the diagnosis and prediction of treatment response for breast cancer. *NPJ Breast Cancer* **7**, 154 (2021).
26. Lai, H. et al. exoRBase 2.0: an atlas of mRNA, lncRNA and circRNA in extracellular vesicles from human biofluids. *Nucleic Acids Res.* **50**, D118–D128 (2022).
27. Guo, T.-A. et al. Plasma extracellular vesicle long RNAs have potential as biomarkers in early detection of colorectal cancer. *Front. Oncol.* **12**, 829230 (2022).
28. C. Liu, J. Chen, J. Liao, Y. Li, H. Yu, X. Zhao et al., *Plasma Extracellular Vesicle Long RNA in Diagnosis and Prediction in Small Cell Lung Cancer.*, Cancers (Basel). 14 (2022).
29. Jayedi, A. et al. Inflammation markers and risk of developing hypertension: a meta-analysis of cohort studies. *Heart* **105**, 686–692 (2019).
30. Lee, C. C. et al. Association of C-reactive protein with type 2 diabetes: prospective analysis and meta-analysis. *Diabetologia* **52**, 1040–1047 (2009).
31. Wang, X. et al. Inflammatory markers and risk of type 2 diabetes: a systematic review and meta-analysis. *Diabetes Care* **36**, 166–175 (2013).
32. Moosazadeh, M. et al. Family history of diabetes and the risk of gestational diabetes mellitus in Iran: A systematic review and meta-analysis. *Diabetes Metab. Syndr. Clin Res. Rev.* **11**, S99–S104 (2017).
33. Garcia-Carretero, R., Vigil-Medina, L. & Barquero-Perez, O. The use of machine learning techniques to determine the predictive value of inflammatory biomarkers in the development of type 2 diabetes mellitus. *Metab. Syndr. Relat. Disord.* **19**, 240–248 (2021).
34. Arena, R., Arrowood, J. A., Fei, D.-Y., Helm, S. & Kraft, K. A. The relationship between C-reactive protein and other cardiovascular risk factors in men and women. *J. Cardiopulm. Rehabil. Prev.* **26**, 323–327 (2006).
35. Rogowski, O. et al. Gender difference in C-reactive protein concentrations in individuals with atherothrombotic risk factors and apparently healthy ones. *Biomarkers* **9**, 85–92 (2004).
36. Tayefi, M. et al. Depression and anxiety both associate with serum level of hs-CRP: a gender-stratified analysis in a population-based study. *Psychoneuroendocrinology* **81**, 63–69 (2017).
37. Toker, S., Shirom, A., Shapira, I., Berliner, S. & Melamed, S. The association between burnout, depression, anxiety, and inflammation biomarkers: C-reactive protein and fibrinogen in men and women. *J. Occup. Health Psychol.* **10**, 344 (2005).
38. Naudé, P. J. W., Roest, A. M., Stein, D. J., de Jonge, P. & Doornbos, B. Anxiety disorders and CRP in a population cohort study with 54,326 participants: The LifeLines study. *World J Biol. Psychiatry* **19**, 461–470 (2018).
39. Tayefi, M. et al. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Comput. Methods Programs Biomed.* **141**, 105–109 (2017).

Acknowledgements

The authors are thankful to all such persons who helped them to construct this piece of work and paper in proper shape.

Author contributions

SG, MGM, HE, AM, and TS contributed to the conception, design, and preparation of the manuscript. SG, MH, AK conducted the data collection analysis, and contributed to acquisition and interpretation. AK, EAF, MH, HR and MG made substantial contributions in drafting the manuscript and revising it critically for important intellectual content. All authors have read and approved the final version of the manuscript.

Funding

This study was supported financially by Mashhad University of Medical Science (MUMS), Mashhad, Iran (Project Number: 951214).

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This study protocol was reviewed and approved by the Ethics Committee of MUMS, approval number IR.MUMS.REC.1386.250.

Consent to publish
The consent for publication has been obtained from all the authors.

Additional information
Correspondence and requests for materials should be addressed to A.M. or H.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.