Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2024;10(1): 82-97

Original Article

Mining Hypertension Predictors using Decision Tree: Baseline Data of Kharameh Cohort Study

Abbas Rezaianzadeh[1], Samane Nematolahi[2], Maryam Jalali[1]*, Shayan Rezaeianzadeh[3], Masoumeh Ghoddusi Johari[4], Seyed Vahid Hosseini[1]

[1]Colorectal Research Center, Shiraz University of Medical Sciences. Shiraz, Iran.
[2]Non Communicable Diseases Research Center, Bam University of Medical Sciences, Bam, Kerman, Iran.
[3]Student Research Committee, Shiraz University of Medical Sciences. Shiraz, Iran.
[4]Breast Disease Research Center, Shiraz University of Medical Sciences. Shiraz, Iran.

ARTICLE INFO

Received 03.12.2023
Revised 20.01.2024
Accepted 17.02.2024
Published 15.03.2024

Key words:
Decision tree;
Hypertension;
Kharameh cohort.

ABSTRACT

Introduction: Hypertension is a serious chronic disease and an important risk factor for many health problems. This study aimed to investigate the factors associated with hypertension using a decision-tree algorithm.
Methods: Methods: This cross-sectional study was conducted through the census in Kharameh City between 2014 and 2017. The study included 2510 hypertensive and 7840 non-hypertensive individuals. To create the decision tree, 70% of the cases were randomly allocated to the training dataset. In comparison, the remaining 30% were used as the testing dataset for the performance evaluation of the decision tree. Two models were assessed. In the first model (model I), 15 variables including age, gender, body mass index (BMI), years of education, occupation status, marital status, family history of hypertension, physical activity, total energy, number of meals, salt, oil type, drug use, alcohol use, and smoke entered into the model. in the second model (model II) 16 variables including age, gender, BMI and Blood factors as Hematocrit (HCT), Mean corpuscular hemoglobin concentration (MCHC), Platelet Count (PLT), Fasting blood sugar (FBS), Blood Urea Nitrogen (BUN), creatinine (Cr), triglycerides (TG), cholesterol (CHOL), Alkaline phosphatase (ALP), High-density lipoprotein (HDL), Gamma-glutamyl transpeptidase (GGT), low-density lipoproteins (LDL) and Urinary specific gravity (SG) were considered. A confusion matrix was used to measure the performance of the decision tree. Additionally, accuracy, sensitivity, specificity, and the receiver operating characteristics (ROC) curve were determined to compare the models.
Results: For the model I, the accuracy, sensitivity, specificity and AUC value were 79.2%(77.8-80.6), 82.4%(80.1-84.5), 78.24%(76.4-80), and 0.80%(0.79-0.82), respectively. For model II, the corresponding values were 79.5%(78.2-80.8), 81.0%(78.3-83.6)79.0%(77.5-80.5)and 0.80%(0.79-0.81), respectively. Confusion matrix of model I showed that of the 1188 cases with hypertension in the training data set, 979 cases were classified correctly and, for model II of the 2812 non-hypertension cases, 2222 cases were classified correctly.
Conclusion: We have suggested a decision tree model to identify the risk factors associated with hypertension. This model can be useful for early screening and improving preventive and curative health services in health promotion.

*.Corresponding Author: jalali3944@yahoo.com

Introduction

Hypertension is a serious chronic disease whose occurrence is increasing worldwide [1] and its prevalence increases with age.[2] It is an important risk factor for many health problems such as stroke, heart disease, and retinopathy.[3] Similarly, some risk factors like obesity, glucose intolerance, dyslipidemia, hyperinsulinemia, and hyperuricemia that are related to cardiovascular disease increase the risk of hypertension.[2]

The usual diagnostic method for detecting hypertension is the measurement of blood pressure using a sphygmomanometer. Although this seems simple, the issue is that the blood pressure is not stable and high all the time. Nevertheless, the pathological changes and indicators of disease usually occur sooner than blood pressure increment and the screening of this health problem is usually neglected. therefore, developing a valid diagnosis method based on related biomarkers might be beneficial for early diagnosis and avoiding the progression of diseases, though no early diagnosis approach has been found yet.[1] There are many studies about effective factors on the prevalence of hypertension such as age, BMI, physical activity, Aerobic endurance sport, ratio of waist to hip, total cholesterol, triglyceride, sex, educational level, and family history.[4-8]

Traditional models such as Fisher's Linear Discriminant Analysis and logistic regression have unsuitable performance in the case of big data, many risk factors, and the presence of outliers and missing data.[9-13] Data mining approaches such as decision trees are suggested for handling such conditions and are known as more accurate and have lower error rates than the classic models. Other strengths of the random forest approach include not overfitting; robustness to the noise; owing internal mechanism to estimate error rates, called out-of-the-bag (OOB) error; providing indices of variable importance; working with mixes of continuous and categorical variables; and additionally, it can be used for data imputation and cluster analysis. These properties have made random forests increasingly popular in the last few years.[14-16]

In a study exploring hypertension and hyperlipidemia prediction models using common risk factors, six data mining methods were used to find the related common risk factors of the two diseases. The applied methods were logistic regression analysis, C5.0 decision tree, Classification and Regression Tree, Chi-squared Automatic Interaction Detector (CHAID), exhaustive CHAID, and discriminant analysis. Each data mining method provided different significant risk factors. So, each risk factor which was significant in at least three methods was selected. The commonly selected risk factors were age, systolic and diastolic blood pressure, triglyceride, creatinine, uric acid and Glutamate pyruvate transaminase. In the second stage of this study, the multivariate adaptive regression splines (MARS) were applied to construct a predictive model for hypertension and hyperlipidemia based on the common risk factors of these two diseases. This method aimed to overcome some above methods. The proposed method had an accuracy of 93.1%.[17]

In another similar study with the purpose of essential hypertension prediction, the performance of data mining approaches including three decision trees, four statistical algorithms, and two neural networks were compared and the effective predictors were age, gender, body mass index, smoking, family history of hypertension, lipoprotein(a), triglyceride, uric acid, total cholesterol. neural network performance was better than others in predicting hypertension and the quick unbiased efficient statistical tree had the lowest performance.[18]

In this study we aimed to identify the associated predictors of hypertension in Kharameh cohort study (KHCS) in a subgroup with an age range of 40-70 years old between 2014-2017 years, using decision tree modelling.

Material and Methods

Participants

population

This cross-sectional study considered the population of Kharameh city between 2014 and 2017 through census. The city is located in the south of Iran and has a population of 61,580 citizens. Kharameh study is a subset of Prospective Epidemiological Research Studies in Iran (PERSIAN Cohort), and one of its main objectives is to identify and explore non-communicable disease risk factors and their prevalence. the reference number is IR.SUMS.REC.1393.S7421.[19]

Hypertension diagnosis was confirmed by two internists and followed the hypertension management recommendations outlined in the European guidelines. It was determined by systolic blood pressure ≥140 mmHg or diastolic blood pressure ≥90 mmHg or using antihypertensive medication or hypertension that was previously diagnosed.[19]

After obtaining written consent, data on participants' demographic profiles, including age, gender, marital status, education level, occupation, place of residence, social and economic status, and behavioral factors, such as smoking, alcohol consumption, drug use, and physical activity were collected through interviews, laboratory experiments, and physical examinations and questionnaire. The questionnaire had been previously tested and validated by the Persian cohort national team.

Exclusion criteria and sample size

The exclusion criteria included no cooperation in evaluating the procedure and mental ability.[19, 20] In addition, the participants whose total daily energy intake (Kcal) was out of the range of mean±3SD were excluded.[21, 22] Finally, the study included 2510 hypertensive and 7840 non-hypertensive individuals.

Input parameters

The parameters used in our analysis are as follow:

- Anthropometric data: Age, weight (Kg), height (m), body mass index (BMI (kg/m^2)
- Gender
- Occupation status
- Education year
- Cigarette smoking, Alcohol consumption, drug use.
- Number of meals
- Oil type
- habit of adding salt when taking lunch and dinner. (yes/sometimes/no)
- Glucose levels: fasting blood sauger FBS;
- Kidney function: specific gravity (SG, creatinine Cr and blood urea nitrogen (BUN;
- Liver tests: serum glutamic-oxaloacetic transaminase (SGOT, Serum glutamic pyruvic transaminase (SGPT, Alkaline phosphatase (ALPand gamma-glutamyl transferase (GGT;

- Blood component: WBC, RBC, HGB, HT, HCT, PLT, MCV, MCH, MCHC.
- Lipid profile: cholesterol (Chol), high-density lipoproteins (HDL, low-density lipoproteins LDL and triglycerides (TG;
- Total energy(kcal);
- Physical activity.

Measurement

A valid and reliable standardized questionnaire was used to collect demographic information.[20] The SECA Germany scale was used to measure weight. A validated physical activity questionnaire was used to assess physical activity, including sleep duration, sport, and working for a day. The Metabolic Equivalent Task (MET) index was computed according to units per hour per day.[23] Additionally, the FFQ using NUTRITIONIST-IV software was also applied to derive the energy (dietary data).[20]

Data analysis methods

The data mining method used in this study to explore prediction rules and patterns was a decision tree. The important aim of this method is to find a predictive model using the features.[24] The decision tree algorithm has two parts: the first part is the nodes (including the root node and internal node) and the second part is the leaf (including end node). Each node represents an attribute and each link represents a decision. Each leaf shows the outcome.[24, 25] Splitting criteria are used at the internal nodes for constructing the trees. These criteria aim to minimize the impurity of internal nodes. Node impurity is used to measure the homogeneity in each node and leaf. The process is such that a node will be split when the impurity is reduced, otherwise it will be a leaf. In the state of reducing the impurity, two branches will be formed and consequently, two new nodes will be made. In particular, a rate is also achieved for each predictor variable that is used for selecting the variables for inclusion into the model.[24]

One of the common tree algorithms that can handle classification and regression issues is CART (Classification And Regression Tree).[26] It works based on the division of the nodes based on the threshold of a predictor. The Gini index is a common method used to determine the non-homogeneity in the decision tree method and its value ranges from 0 to 1.[27]

The confusion matrix was used to measure the performance of the decision tree. This matrix is based on True Positives (prediction and actual both are yes), True Negatives (prediction is no and actual is yes), False Positives (prediction is yes and actual is no), and False Negatives (prediction is no and actual is no).[28] Additionally, accuracy, sensitivity, specificity, and the ROC curve were determined to compare the models.[29] The AUC is used to determine the predictive accuracy of a diagnostic test. The higher the AUC the better the predictive accuracy.[30]

To evaluate the performance of data mining methods such as decision trees, the data set is commonly divided into two sets: the training and the test set. The model is made based on the training data set and then tested on the testing data set. In our study, 70% of the data were considered as training and 30% as the testing data set.

We considered two models in this study. In the first model (model I), 15 variables including age, gender, body mass index, years of education, Occupation status, marital status, family history of hypertension, physical activity, total energy, number of meals, salt, oil type, drug use, alcohol use and smoke entered into the model. We used the identical training and test set for the second model. 16 variables including age, gender, BMI, and Blood factors such as HCT, MCHC, PLT, FBS, BUN, CERAT, TG, CHOL, ALP, HDL, GGT, LDL, and SG were considered in model II. The rationale for considering these models stems from the nature of the factors. An effort has been made to thoughtfully organize the available factors in a logical manner. The concept of using two models is based on a previous study.[31]

Statistical analysis was done in R software version 4.1.3 and rpart package. Independent sample t-test, Mann-Whitney, and Chi-square test were used to compare demographic characteristics and clinical variables between hypertension and non-hypertension groups. Mean ± SD and frequency (percentage) are reported for quantitative and qualitative data respectively.

Result

Demographic characteristics and clinical variables and their comparison between two groups of hypertension and non-hypertension are demonstrated in Table 1. There were 2510 and 7864 cases with and without hypertension in our study.

Table 1. Comparison of baseline characteristics between hypertension and non-hypertension groups

Variable	Hypertension (n=2510)	Non-hypertension (n=7864)	P-value
Age (year)	56.4±7.85	50.7±7.95	<0.0001*
BMI (Kg/m^2)	27.5±4.37	25.7±4.38	<0.0001*
MET	37.0±4.62	38.9±6.40	<0.0001*
Energy	2184.4±677.6	2458.2±712.0	<0.0001*
Education years	3.21±4.08	4.81±4.51	<0.0001*
WBC	6.67±1.81	6.44±1.75	<0.0001*
RBC	5.11±0.62	5.19±0.64	<0.0001*
HGB	14.05±1.54	14.33±1.63	<0.0001*
HCT	41.4±4.07	42.08±4.21	<0.0001*
MCV	81.8±8.63	81.8±9.09	0.924
MCH	27.8±3.50	27.9±3.72	0.151
MCHC	33.9±1.22	34.03±1.35	<0.0001*
PLT	248.9±63.6	241.3±60.9	<0.0001*
FBS	109.5±43.02	96.4±29.7	<0.0001*
BUN	14.6±4.82	13.83±3.87	<0.0001*
Cr	0.98±0.25	0.96±0.15	0.001*
TG	138.4±76.1	127.8±81.5	<0.0001*
CHOL	185.3±43	187.5±41.3	0.024*
SGOT	22±10.3	22.3±9.39	0.198
SGPT	25.5±17.5	25.1±15.3	0.279
ALP	226 ±75.5	206.2±62.5	<0.0001*
Gender			
Male	663 (26.4%)	3799 (48.3%)	<0.0001*
Female	1847 (73.6%)	4065 (51.7%)	
Marriage status			

Continue table 1.

Variable	Hypertension (n=2510)	Non-hypertension (n=7864)	P-value
Single	454 (18.1%)	707 (9.0%)	<0.0001*
Married	2056 (81.9%)	7157 (91%)	
Occupation status			
Yes	790 (31.5%)	4479 (57%)	<0.0001*
No	1720 (68.5%)	3385 (43%)	
Smoke Cigarette			
Yes	374 (14.9%)	2165(27.5%)	<0.0001*
No	2136 (85.1%)	5699 (72.5%)	
Drug use			
Yes	216 (8.6%)	1368 (17.4%)	<0.0001*
No	2294 (91.4%)	6496 (82.6%)	
Alcohol Use			
Yes	34 (1.40%)	275 (3.50%)	<0.0001*
No	2476 (98.6%)	7589 (96.5%)	
Number of meals			
Less than 3 meals	210 (2.70%)	57 (2.30%)	0.003*
3 meals	2200 (87.6%)	6734 (85.6%)	
4 meals	200 (8.0%)	802 (10.2%)	
5-6 meals	52 (2.10%)	115 (1.50%)	
More than 6 meals	3 (0.0%)	1 (0.0%)	
Adding Salt to the food during eating			
Yes	178 (7.1%)	1301 (16.5%)	<0.0001*
Sometimes	386 (15.4%)	1944 (24.7%)	
No	1946 (77.5%)	4619 (58.7%)	
Oil type			
Margarine	645 (25.7%)	2499 (31.8%)	<0.0001*
Butter, cream	2 (0.10%)	18 (0.20%)	
Solid vegetable oil	328 (13.1%)	903 (11.5%)	
Solid animal oil	1489 (59.3%)	4382 (55.7%)	
Liquid soybean oil	9 (0.40%)	22 (0.30%)	
Liquid sunflower oil	37 (1.50%)	40 (0.50%)	

MET, Physical activity; White blood cells; RBC, Red blood cell; HGB, Hemoglobin; HCT, Hematocrit; MCHC, Mean corpuscular hemoglobin concentration; MCV, Mean corpuscular volume; MCH mean corpuscular hemoglobin; PLT, Platelet count; FBS, Fasting blood sugar; BUN, Blood urea nitrogen; Cr, Creatinine; (SG, Urinary specific gravity; TG, Triglycerides; CHOL, Cholesterol; HDL, High-density lipoprotein; LDL, Low-density lipoproteins; ALP, Alkaline phosphatase; GGT, Gamma-glutamyl transpeptidase;

The decision tree for model I is shown in Figure 1. The variables that remained in the model were age, physical activity, adding salt, energy, and BMI. The result of the confusion matrix on the testing dataset for model evaluation is shown in Table 2. The accuracy of this model was 79.2 (77.8-80.6) (Table3). The Sensitivity was 82.4% (of the 1188 cases with hypertension in the training data set, 979 cases were classified correctly). Also, the Specificity was 78.2% (of the 2077 non-hypertension cases, 1625 cases were classified correctly). In addition, a ROC curve based on the testing data set for this model is shown in Figure 2. The AUC was 0.80% for model I (Table 3). The if-then rules for this model are demonstrated in

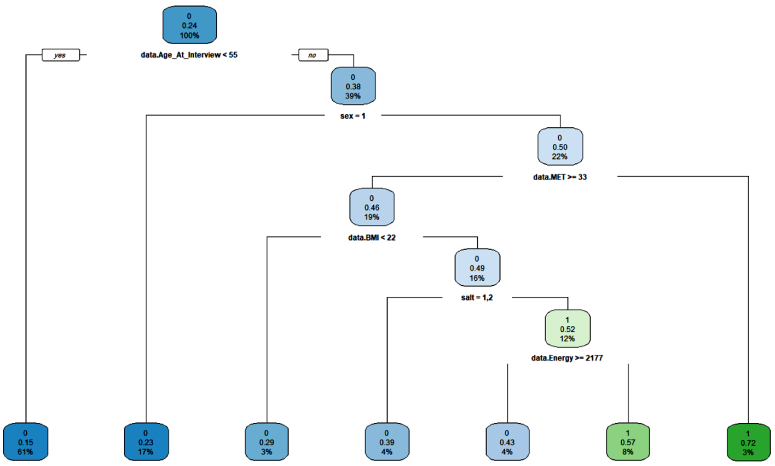


Figure 1. Decision tree with training dataset in model I

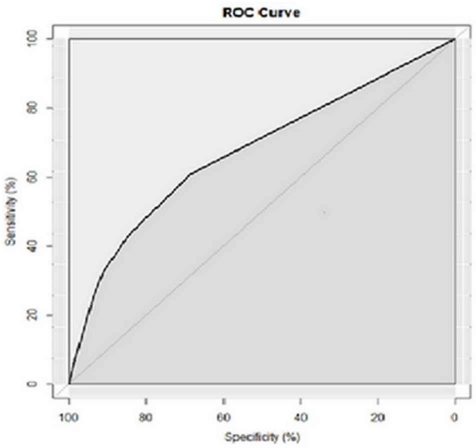
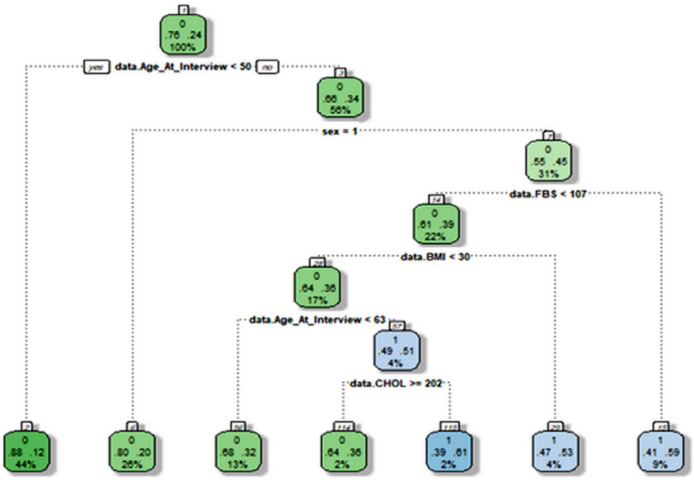


Figure 2. Roc curve of the decision tree with test data in model I



Rattle 2023-Dec-31 11:49:13 LaptopCity

Figure 3. Decision tree with training dataset in model II

Table 4. According to this table, the probability of being hypertension when Age >= 55, sex is female, physical activity >= 33, BMI >= 22, salt is 3, and Energy < 2177 was 57% and the probability was 72% when Age >= 55 & sex is 2 & physical activity < 33.

Table 2. Confusion Matrix with test data in Model I

	Actual outcome	
Predicted outcome	Hypertension	No-hypertension
Hypertension	979	452
No-Hypertension	209	1625

Table 3. Criteria of decision tree with test data in model I

Measure	% (Confidence Interval)
Accuracy	79.2 (77.8-80.6)
Sensitivity	82.4 (80.1-84.5)
Specificity	78.2 (76.4-80.0)
AUC	0.80 (0.79-0.82)

Age, sex, FBS, BMI, and cholesterol remained in the model. The decision tree for this model is shown in Figure 3 and the ROC curve based on the testing data set for this model is shown in Figure 4.

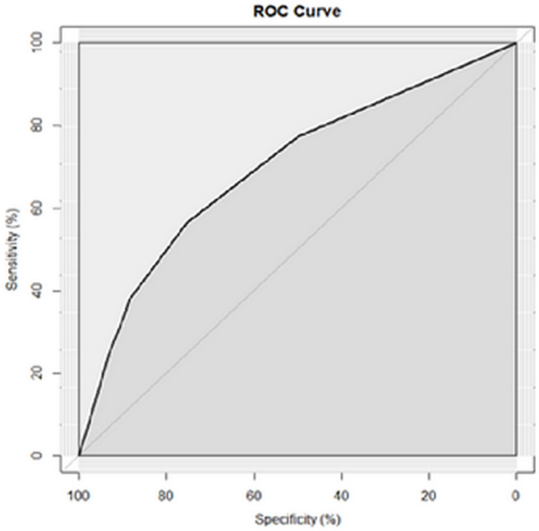


Figure 4. Roc curve of the decision tree with test data in model II

The result of the confusion matrix on the testing dataset for model evaluation is shown in Table 5. The AUC was 0.80% and the accuracy of this model was 79.5%(78.2-80.8)

Table 4. Decision Rules for model I

R1	When Age < 55 then class: non- hypertension (propability:0.15)
R2	When Age >= 55 & sex is Male then class: non- hypertension (probability: 0.23)
R3	When Age >= 55 & sex is Female & MET >= 33 & BMI (kg/m2) < 22 then class: non- hypertension (probability: 0.29)
R4	When Age >= 55 & sex is Female & MET >= 33 & BMI (kg/m2) >= 22 & salt is yes or sometimes then class: non- hypertension (probability: 0.39)
R5	When Age >= 55 & sex is Female & MET >= 33 & BMI(kg/m2) >= 22 & salt is no & Energy >= 2177 then class: non- hypertension (probability: 0.43)
R6	When Age >= 55 & sex is Female & MET >= 33 & BMI(kg/m2) >= 22 & salt is no & Energy < 2177 then class: hypertension (probability: 0.57)
R7	When Age >= 55 & sex is Female & MET < 33 then class: hypertension (probability: 0.72)

Table 5. Confusion Matrix with test data in Model II

	Actual outcome	
Predicted outcome	Hypertension	No-hypertension
Hypertension	709	590
No-hypertension	166	2222

Table 6. Criteria of decision tree with test data in model II

Measure	% (Confidence Interval)
Accuracy	79.5 (78.2-80.8)
AUC	0.80 (0.79-0.81)
Sensitivity	81.0 (78.3-83.6)
Specificity	79.0 (77.5-80.5)

Table7. Decision Rules extracted from decision tree model II.

R1	When Age < 50 then class: non- hypertension (probability: 0.12)
R2	When Age >= 50 & sex is Male then class: non- hypertension (probability: 0.20)
R3	When Age is 50 to 63 & sex is Female & FBS < 107 & BMI(kg/m^2) < 30 then class: non- hypertension (probability: 0.32)
R4	Ehen Age >= 63 & sex is Female & FBS < 107 & BMI(kg/m^2) < 30 & CHOL >= 202 then class: non- hypertension (probability: 0.36)
R5	Ehen Age >=50 & sex is Female & FBS <107 & BMI(kg/m^2) >= 30 then class: hypertension (probability: 0.53)
R6	When Age >= 50 & sex is Female & FBS >= 107 then class: hypertension (probability: 0.59)
R7	When Age >= 63 & sex is Female & FBS < 107 & BMI (kg/m^2) < 30 & CHOL < 202 then class: hypertension (probability: 0.61)

(Table 6). The Sensitivity was 81.0% (of the 875 cases with hypertension in the training data set, 706 cases were classified correctly. Also, the Specificity was 79.0% (of the 2812 non-hypertension cases, 2222 cases were classified correctly). The if-then rules for this model are demonstrated in Table 7. Based on the tree model in the subgroup with Age >=50 and female cases, FBS <107 and BMI >= 30, the probability of being hypertensive was 53% and for a subgroup of Age >= 50, female groups, FBS >= 107 this probability was 59% while it increased to 61% when cholesterol appeared and the subgroup was Age >= 63 & sex was female, FBS < 107, BMI < 30 and cholesterol < 202.

Discussion

In this paper, we identified the associated predictors of hypertension in KHCS using two decision tree models. The results showed that the accuracy of our proposed models was high enough.

The mechanisms behind the development of hypertension are not fully understood, and the causes remain partially unexplained.[32] However, several factors have been associated with hypertension, including BMI, gender, age, family history of hypertension, smoking status, lipoprotein (a), triglyceride, uric acid, and total cholesterol.[33]

The performance of classifying methods and the best performance may vary from one dataset to another.[33] For example, in a study of hypertension prediction, Chi-squared Automatic Interaction Detector (CHAID) which is a decision tree algorithm, had better performance than logistic regression and C5.0 had the worst predictive power.[34] In a study of survival prediction of breast cancer, the performance of the decision tree (C5) was better than logistic regression and neural networks.[35] In another investigation, it was found that classification trees (ID3, C4.5, CHAID, and CART) were performed conveniently in data sets related to veterinary epidemiology.[36] Moreover, multivariate additive regression splines had better performance than artificial neural networks, CART and linear models in predicting forest characteristics. Also, in predicting cardiovascular risk, the performance of CART was slightly more than logistic regression and artificial neural networks.[37]

The decision tree algorithm is a practical and advantageous classification method. Some of the superiorities of this method include simple interpretation, providing rules, and handling linear and non-linear associations.[38-40]

In this study, we investigated two tree models to find hypertension risk factors. In the first model (model I), 15 variables were entered into the model and the remaining ones were BMI, age, physical activity, adding salt during mealtime, and energy intake. Age, gender, and BMI were associated with hypertension in other studies.[8] Physical inactivity is the reason for many diseases worldwide[41] and regular physical activity can improve the risk factors for noncommunicable diseases such as hypertension.[42, 43] Salt consumption is the main risk factor for blood pressure increment and a high amount of salt in the diet can increase the risk of stroke and other diseases, in addition to affecting blood pressure.[44] Several studies have confirmed a higher risk of hypertension in men than women at similar ages.[45-47] However, in our study, as presented in the rules of the model I, both subgroups that identified hypertension (the subgroups with a probability of more than 50%) showed that being female is a risk factor for hypertension. Even though hypertension is reported more in men than in women, it is stated that blood pressure increases in women after menopause. However, other factors may affect this mechanism such as obesity and type II diabetes.[48] The age in both rules was more than 55. Around this age, most women have entered their post-menopausal years.

Energy intake was a new component that we entered into model I and it remained in the tree model. Less emphasis has been placed on whether energy intake affects blood pressure significantly. In a study of the association between hypertension and energy intake, it is found that more energy intake at breakfast is related to a lower prevalence of hypertension and more energy intake late in the evening was related to a higher prevalence of hypertension and an increase in blood pressure.[49] In this study, the information about time-of-day of energy intake were not available. Future studies are needed to investigate the association between macronutrients, time-of-day energy intake, and blood pressure.

In the second model, Age, sex, FBS, BMI and cholesterol remained in the tree model among 16 variables. High blood pressure is observed in more than two-thirds of diabetics and usually occurs simultaneously with hyperglycaemia development. There are many pathophysiological explanations for this association, including insulin resistance in the nitric-oxide pathway and the stimulative impact of hyperinsulinaemia on sympathetic drive.[50]

High cholesterol and high blood pressure have a complicated relationship. More than 60% of people with high blood pressure also have high cholesterol.[51] The link between high blood pressure and high cholesterol goes in both directions. When the body can’t clear cholesterol from the bloodstream, that excess cholesterol can deposit along artery walls. When arteries become stiff and narrow from deposits, the heart has to work overtime to pump blood through them. This causes blood pressure to rise. Over time, high blood pressure can damage arteries in its way by making tears in artery walls where excess cholesterol can collect.[52] Whether cholesterol is a risk factor for hypertension remains controversial. Several studies that have investigated the relationship between cholesterol and blood pressure have reported inconsistent results. Some had found a positive correlation[53] while others found a negative correlation,[54-56] or no association at all.[57, 58]

This study had some limitations. It was a cross-sectional design; Therefore, the data does not provide sufficient evidence to establish definitive causal relationships. so, cohort studies are required in addition to finding more complete results of the long-term influence of factors on hypertension. furthermore, some behaviors such as alcohol or drug use are stigmatized in Iran and people usually do not reveal information about them. This may cause biased association. finally, genetics and race were not available, despite they have been mentioned as effective factors in previous studies. Nevertheless, the large sample size as well as the highly accurate data of the KHCS are strengths of this study. In addition, evaluating a large number of variables is another point of our study.

Conclusion

Our findings focus on the crucial need to establish and develop strategies for early prognosis, and to take preventive actions to raise awareness among people about the impact of high blood pressure. In addition, our results can be useful for early screening and improving preventive and curative health services in health promotion.

Acknowledgements

The authors are grateful to the officers and data management staff of Kharameh cohort. This article is part of the Persian cohort study in Kharameh.

Availability of data and materials

The datasets used and analyzed during the current study are available by sending an email to the owner of data (Abbas Rezaianzadeh).

Consent for publication

Not Applicable.

Competing interests

The authors report no conflict of interest.

Funding

None

Authors’ contributions

Study concept and design: MJ; Acquisition of data: MGJ, AR, SVH, and SR; Analysis and interpretation of data: SN, MJ; Drafting of the manuscript: MJ ; Critical revision of the manuscript for important intellectual content: MJ, SN, AR and MGJ; Statistical analysis: SN; Administrative, technical, and material support: MGJ, AR, SVH, and MSR; Study supervision: MGJ, AR, SVH. All authors have read and approved the final manuscript.

Ethical issues

The study was approved by ethics committee and confirmation were taken from Shiraz University of Medical Sciences (ethical code:IR. SUMS.REC.1393.S7421). Confidentiality of their personal data was emphasized.

References

1. Han Z, Wen LJAoTM. Development and validation of a decision tree classification model for the essential hypertension based on serum protein biomarkers. J Annals of Translational Medicine. 2022;10(18).

2. Staessen JA, Wang J, Bianchi G, Birkenhäger WHJTL. Essential hypertension. J The Lancet. 2003;361(9369):1629-41.

3. Liu L-S, Wu Z, Wang J, Wang W, Bao Y, Cai J, et al. 2018 Chinese guidelines for prevention and treatment of hypertension-A report of the revision committee of Chinese guidelines for prevention and treatment of hypertension. J Journal of Geriatric Cardiology. 2019;16(3):182-245.

4. Pickering TG, Hall JE, Appel LJ, Falkner BE, Graves J, Hill MN, et al. Recommendations for blood pressure measurement in humans and experimental animals: part 1: blood pressure measurement in humans: a statement for professionals from the Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research. J Circulation. 2005;111(5):697-716.

5. Colin Bell A, Adair LS, Popkin BMJAjoe. Ethnic differences in the association between body mass index and hypertension. J American journal of 2002;155(4):346-53.

6. Pescatello L, Franklin B, Fagard R, Farquhar W, Kelley G, Ray CJMSSE. Exercise and hypertension: American College of Sports Medicine position stand. J Med Sci Sports Exerc 2004;36(3):533-53.

7. Cornelissen VA, Fagard RHJH. Effects of endurance training on blood pressure, blood pressure–regulating mechanisms, and cardiovascular risk factors. J Hypertension 2005;46(4):667-75.

8. Akdag B, Fenkci S, Degirmencioglu S, Rota S, Sermez Y, Camdeviren HJAit. Determination of risk factors for hypertension through the classification tree method. Advances in therapy 2006;23:885-92.

9. Beaty TH, Neel JV, Fajans SSJAjoe. Identifying risk factors for diabetes in first degree relatives of non-insulin dependent diabetic patients. J American journal of epidemiology 1982;115(3):380-97.

10. Pan X-R, Yang W-Y, Li G-W, Liu J, Prevention ND, care CCGJD. Prevalence of diabetes and its risk factors in China, 1994. J Diabetes care. 1997;20(11):1664-9.

11. Goss EP, Ramchandani HJJoE, Finance. Comparing classification accuracy of neural networks, binary logit regression and discriminant analysis for insolvency prediction of life insurers. Journal of Economics Finance. 1995;19(3):1-18.

12. Efron BJJotASA. The efficiency of logistic regression compared to normal discriminant analysis. J Journal of the American Statistical Association. 1975;70(352):892-8.

13. Fan X, Wang LJTJoee. Comparing linear discriminant function with logistic regression for the two-group classification problem. J The Journal of experimental education 1999;67(3):265-86.

14. Somvanshi M, Chavan P, Tambade S, Shinde S, editors. A review of machine learning techniques using decision tree and support vector machine. 2016 international conference on computing communication control and automation (ICCUBEA); 2016: IEEE.

15. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça AJBrn. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. J BMC research. 2011;4(1):1-14.

16. Kurt I, Ture M, Kurum ATJEswa. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. J Expert systems with applications. 2008;34(1):366-74.

17. Chang C-D, Wang C-C, Jiang BCJEswa. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. J Expert systems with applications 2011;38(5):5507-13.

18. Ture M, Kurt I, Kurum AT, Ozdamar KJESwA. Comparing classification techniques for predicting essential hypertension. J Expert Systems with Applications. 2005;29(3):583-8.

19. Keshani P, Jalali M, Johari MG, Rezaeianzadeh R, Hosseini SV, Rezaianzadeh AJJoB, et al. The Association between Dietary Antioxidant Indices and Cardiac Disease: Baseline Data of Kharameh Cohort Study. ournal of Biostatistics Epidemiology 2022;8(4):458-70.

20. Poustchi H, Eghtesad S, Kamangar F, Etemadi A, Keshtkar A-A, Hekmatdoost A, et al. Prospective epidemiological research studies in Iran (the PERSIAN Cohort Study): rationale, objectives, and design. American journal of epidemiology. 2018;187(4):647-55.

21. Rezazadeh A, Rashidkhani BJJons, vitaminology. The association of general and central obesity with major dietary patterns of adult women living in Tehran, Iran. Journal of nutritional science vitaminology 2010;56(2):132-8.

22. Jalali M, Keshani P, Ghoddusi Johari M, Rezaeianzadeh R, Hosseini SV, Rezaeianzadeh AJBRI. The Association between Index of Nutritional Quality (INQ) and Obesity: Baseline Data of Kharameh Cohort. BioMed Research International. 2022;2022.

23. Mirmiran P, Esfahani FH, Mehrabi Y, Hedayati M, Azizi FJPhn. Reliability and relative validity of an FFQ for nutrients in the Tehran lipid and glucose study. Public health

nutrition. 2010;13(5):654-62.

24. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh AJJorihs. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. Journal of research in health sciences 2018;18(2):412.

25. Jadhav SD, Channe HJIRJET. Efficient recommendation system using decision tree classifier and collaborative filtering. Int Res J Eng Technol 2016;3(8):2113-8.

26. Ghiasi MM, Zendehboudi S, Mohsenipour AAJCm, biomedicine pi. Decision tree-based diagnosis of coronary artery disease: CART model. 2020;192:105400.

27. Khalilia M, Chakraborty S, Popescu MJBmi, making d. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics decision making 2011;11:1-13.

28. Maria Navin J, Pankaja RJIJoE, Research T. Performance analysis of text classification algorithms using confusion matrix. 2016;6(4):75-8.

29. Nhu V-H, Shirzadi A, Shahabi H, Singh SK, Al-Ansari N, Clague JJ, et al. Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. International journal of environmental research public health 2020;17(8):2749.

30. Lee J-SJIA. AUC4. 5: AUC-based C4. 5 decision tree algorithm for imbalanced data classification. IEEE Access. 2019;7:106034-42.

31. Tayefi M, Esmaeili H, Karimian MS, Zadeh AA, Ebrahimi M, Safarian M, et al. The application of a decision tree to establish the parameters associated with hypertension. Computer methods programs in biomedicine 2017;139:83-91.

32. Rivas AM, Pena C, Kopel J, Dennis JA, Nugent K. Hypertension and hyperthyroidism: association and pathogenesis. The American Journal of the Medical Sciences. 2021;361(1):3-7.

33. Ture M, Kurt I, Kurum AT, Ozdamar K. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications 2005;29(3):583-8.

34. Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. International journal of medical informatics. 2001;62(2-3):103-11.

35. Delen D, Walker G, Kadam AJAiim. Predicting breast cancer survivability: a comparison of three data mining methods. J Artificial intelligence in medicine. 2005;34(2):113-27.

36. Stärk KD, Pfeiffer DUJIDA. The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology–An example. Intelligent Data Analysis. 1999;3(1):23-35.

37. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent M-C, editors. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. Proceedings of the AMIA Symposium; 2000: American Medical Informatics Association.

38. Kammerer JS, McNabb SJ, Becerra JE, Rosenblum L, Shang N, Iademarco MF, et al. Tuberculosis transmission in nontraditional settings: a decision-tree approach. American journal of preventive medicine. 2005;28(2):201-7.

39. Wang C-J, Li Y-Q, Wang L, Li L-L, Guo Y-R, Zhang L-Y, et al. Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents. PloS one. 2012;7(8):e43834.

40. Podgorelec V, Kokol P, Stiglic B, Rozman IJJoms. Decision trees: an overview and their use in medicine. Journal of medical systems. 2002;26:445-63.

41. Zhu Z, Feng T, Huang Y, Liu X, Lei H, Li G, et al. Excessive physical activity duration may be a risk factor for hypertension in young and middle-aged populations. Medicine. 2019;98(18).

42. Haskell WL, Lee I-M, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Circulation. 2007;116(9):1081.

43. Park S, Rink LD, Wallace JPJJoh. Accumulation of physical activity leads to a greater blood pressure reduction than a single continuous session, in prehypertension. Journal of hypertension. 2006;24(9):1761-70.

44. He FJ, MacGregor GAJJohh. A comprehensive review on salt and health and current experience of worldwide salt reduction programmes. Journal of human hypertension. 2009;23(6):363-84.

45. Wiinberg N, Høegholm A, Christensen HR, Bang LE, Mikkelsen KL, Nielsen PE, et al. 24-h ambulatory blood pressure in 352 normal Danish subjects, related to age and gender. American journal of hypertension. 1995;8(10):978-86.

46. Khoury S, Yavows SA, O'Brien TK, Sowers JR. Ambulatory blood pressure monitoring in a nonacademic setting: effects of age and sex. American journal of hypertension. 1992;5(9):616-23.

47. Staessen J, Fagard R, Lijnen P, Thijs L, Van Hoof R, Amery. Reference values for ambulatory blood pressure: a meta-analysis. Journal of hypertension Supplement: official journal of the International Society of Hypertension 1990;8(6):S57-64.

48. Burt VL, Whelton P, Roccella EJ, Brown C, Cutler JA, Higgins M, et al. Prevalence of hypertension in the US adult population: results from the Third National Health and Nutrition Examination Survey, 1988-1991. Hypertension. 1995;25(3):305-13.

49. Almoosawi S, Prynne CJ, Hardy R,

Stephen AM. Time-of-day of energy intake: association with hypertension and blood pressure 10 years later in the 1946 British Birth Cohort. Journal of hypertension. 2013;31(5):882-92.

50. Ferrannini E, Cushman WC. Diabetes and hypertension: the bad companions. The Lancet. 2012;380(9841):601-10.

51. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, et al. 2020 International Society of Hypertension global hypertension practice guidelines. J Hypertension. 2020;75(6):1334-57.

52. Egan BM, Li J, Qanungo S, Wolfman TE. Blood pressure and cholesterol control in hypertensive hypercholesterolemic patients: national health and nutrition examination surveys 1988–2010. Circulation. 2013;128(1):29-41.

53. Bonaa K. Association between blood pressure and serum lipids in a population. The Tromso Study. Circulation. 1991;83:1305-14.

54. Li L, Wang Y, Cao W, Xu F, Cao J. Longitudinal studies of blood pressure in children. Asia Pacific Journal of Public Health. 1995;8(2):130-3.

55. Wakabayashi IJMS, Disorders R. Associations of blood lipid-related indices with blood pressure and pulse pressure in middle-aged men. Metabolic Syndrome. 2015;13(1):22-8.

56. Cho K-H, Park H-J, Kim J-R, health p. Decrease in serum HDL-C level is associated with elevation of blood pressure: correlation analysis from the Korean National Health and nutrition examination survey 2017. International journal of environmental research. 2020;17(3):1101.

57. HUGHES K, Leong W, Sothy S, Lun K, Yeo P. Relationships between cigarette smoking, blood pressure and serum lipids in the Singapore general population. international journal of epidemiology. 1993;22(4):637-43.

58. Yan Z, Bi-Rong D, Hui W, Chang-Quan H. Serum lipid/lipoprotein and arterial blood pressure among Chinese nonagenarians/centenarians. Blood Pressure. 2011;20(5):296-302.