Journal of Computer and Knowledge Engineering, Vol. 3, No. 2, 2020.(21-30)
DOI: 10.22067/CKE.2021.63297.0

Assessment of the Metabolic Syndrome in Children and Adolescents in Birjand, Iran: A Data Mining Approach*

Research Article

Fatemeh Taheri[1] Vahide Babaiyan[2] Mohsen Saffarian[3] Seyed Mahmood Kazemi[4] Kokab Namakin[5] Toba Kazemi[6] Azra Ramazankhani[7]

Abstract: Metabolic Syndrome (MS) is the collection of risk factors for coronary artery disease (CAD). It is responsible for cardiovascular disease (CVD), type 2 diabetes, cancer, renal and mental diseases with the transition from childhood to adulthood. The MS is increasing in recent decades in different societies, especially in Iran. This study was conducted to determine the predictive factors of MS in children and adolescents in Birjand (Capital of South Khorasan Province, Iran) using a data mining approach. The four different analyses for females and males (6-11 and 12-18 year-old groups) were carried out using three popular decision tree models. The most important prognostic factors for MS were high level of TG and low level of DBP and HDL in 6-11-year-old group, and high level of waist circumference (WC) and low level of TG for 12-18-year-old group. The most important factors were TG and HDL in females and WC and TG in males. Raising teens and families' awareness of the risk factors, screening children and teens, monitoring and controlling the risk factors through life style correction such as more physical activity and healthy eating are recommended.

Keywords: Index Terms, Metabolic Syndrome, Data Mining, Decision Tree, School-Age Population

1. Introduction

The Metabolic Syndrome (MS) is the collection of risk factors for coronary artery disease (CAD) such as abdominal obesity, dyslipidemia, hypertension, and insulin resistance, which was firstly described by Reaven in 1988. Different definitions are presented for MS. Its prevalence differs from both areas and definitions [1]. It is responsible for cardiovascular disease (CVD), type 2 diabetes, cancer, renal and mental diseases with the transition from childhood to adulthood [2-4]. In recent decades, MS in many countries, including Iran has been on the rise due to a progressive increase in the prevalence of obesity in children and adolescents and changing the lifestyles [1, 2, 4, 5]. It is predicted that CVD will be the most important cause of mortality in the world and also 30% of all deaths in developing countries by 2020 [2]. Due to the increase in death from CVD in the world including Iran, and the matter that the risk factors for these diseases begin from childhood and adolescence, the best way to reduce the risk factors in children and adolescents is the identification and control of them. Given the racial and geographic factors for each region, risk factors should be identified and appropriate interventions for preventing their deadly effects should be designed [2, 4, 6].

Various studies in this field have been conducted in recent years. For example, Lee et al. [7] developed and validated a risk prediction model for metabolic syndrome in 2 years based on an individual's baseline health status and body weight. Cheng-Sheng et al. [8] used various statistical machine learning techniques to visualize and investigate the pattern and the relationship between metabolic syndrome and several risk varieties. In another study by Yanchao Tang et al. [9], a data mining method was used to identify significant physiological indexes and traditional Chinese medicine constitutions. In another study [10], authors performed a systematic review to investigate the various statistical and machine learning techniques used to support the clinical diagnoses of metabolic syndrome from the earliest studies to January 2020. They concluded that the decision tree and machine learning techniques have the highest predictive performance for metabolic syndrome.

In recent years, attention has been paid to the risk factors in children and adolescents, and several studies have been conducted in various countries as such, including Iran. According to [4], the prevalence of MS in Iran in the age group of 3-21 years old is 3% -16% based on ATP III criterion [4]. According to Caspian National Study [2], MS is seen among 14.1% of the children and adolescents, 2.5% of adolescents in 23 provinces of Iran [11], and 9.5% of the adolescents in the age group of 10-19 in Tehran [12]. Prevalence of MS in adolescents between 12-18 years old in United Arab Emirates is 12%, including 21% males and 4% females [13]. According to ATP III definition, prevalence of MS in males is 9.8% in Kuwait, 4.2% in China, 5.1% in America, 5.8% in India, 9% in Korea, 10.3% in Iran, 20.2% in Mexico, and 30.9% in Brazil [14]. The prevalence of MS in the 13-16-year-old adolescents in Vietnam is 3.9%-12.5% based on different definitions [15]. In China, the prevalence of MS is 23.9% in the age group of 7-17 year-old based on

* Manuscript received October, 25; 2020 accepted. February, 22, 2021.

[1] Professor, Cardiovascular Diseases Research Centre, Department of Pediatric, Birjand University of Medical Sciences, Birjand, Iran.

[2] Corresponding Author. Instructor, Department of Computer Engineering, Birjand University of Technology, Birjand, Iran: Email: babaiyan@birjandut.ac.ir

[3] Assistant Professor, Department of Industrial Engineering, Birjand University of Technology, Birjand, Iran.

[4] Assistant Professor, Department of Industrial Engineering, Birjand University of Technology, Birjand, Iran.

[5] Professor, Cardiovascular Diseases Research Centre, Department of Pediatric, Birjand University of Medical Sciences, Birjand, Iran.

[6] Professor, Cardiovascular Diseases Research Centre, Department of Cardiology, Birjand University of Medical Sciences, Birjand, Iran.

[7] Professor, Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Science, Shahid Beheshti University of Medical Sciences,Tehran, Iran.

Cook's definition [16]. The prevalence of MS is also 14% in the 6-14 year- old- group in Brazil [17] and 3.5% in Canadian adolescents [18]. Findings suggests that several factors, such as high density lipoproteins (HDL), low density lipoprotein (LDL), body mass index (BMI) can predict the risk of MS. Therefore, using a multi-variate model to identify populations at higher risk for MS incidence has been a field of active research. Several prediction models for MS have been derived using classical statistical models such as logistic and Cox regressions. Chandola et al investigated the association between stress at work and MS using logistic regression and the data of a cohort study in London civil service departments [19]. Statistical models have some disadvantageous and usually require some assumptions for fitting model, and the results of the analysis can be misleading by violating these assumptions [20].

Data Mining (DM) techniques such as decision trees (DT) and artificial neural networks (ANNs) are popular techniques to estimate complex relationships in different areas, especially in medical sciences [21-23]. The DT provides a very flexible structure and do not require pre-specifying the interactions between variables. This method has fewer assumptions in modeling, which can be used as an explorative approach to classify samples into groups with similar characteristics. DT is a nonparametric classification model which has attracted great attention in medical sciences. Ramezankhani et al presented a DT for predicting the risk of type 2 diabetes (T2D) and for exploring the interactions between risk factors in those models [24]. To the best of our knowledge, there are few studies in the literature concerned with applying DT model for MS. Worachartchee wan et al. [25] employed decision tree as a decision support system to identify the individuals with MS among a Thai population. Results demonstrated the strong prediction power of the DT in classifying the individuals with and without MS, displaying an overall accuracy in excess of 99%.

In this study, three types of DT algorithms (Chi-squared automatic interaction detector [CHAID], classification and regression tree [CART], and quick unbiased efficient statistical tree [QUEST]) are applied for rapid and automated identification of MS and the relative importance of risk factors leading to MS in children and adolescents in Birjand, South Khorasan.

2. Methods

2.1 Study Population

This cross-sectional descriptive study was conducted from 2012 to 2013 on 4,340 students aged 6-18-year-old living in Birjand including, 2,329 females and 2,011 males. The samples were chosen through multi-stage cluster sampling from school-aged students. Weight and height measurements were performed by standard methods. Weight was measured with a digital scale (Seca, German) with 100 grams of error, and height with an accuracy of 0.5 cm. Waist circumference (WC) was measured by meter in standing position at the distance of between the last rib and the iliac in exhale with an accuracy of 0.5. Blood pressure (BP) was measured under standard circumstances. The laboratory blood tests such as fasting blood sugar (FBS), LDH and triglyceride (TG) were conducted on students.

Blood samples were collected in 5 ml vacuum tubes, (with separator gel, clot activator, Bacton Dickinson U.K). Blood samples were separated at a distance of fewer than fifteen minutes for example in 10 minutes' by sigma centrifugation in 3000 RPM. FBS, total serum cholesterol (TC), and TG, HDL-C and LDL-C concentrations were measured within less than an hour by commercially available enzymatic reagents (ROCHE Kits, Germany with a closed full automated analyzer system of ROCHE COBAS INTEGRA). ATP III was used as the diagnostic criteria for MS in this study. Having at least three of the following are characterized as having MS according to ATP III: abdominal obesity and WC $\geq$ 90th percentile for age and gender, TG $\geq$ 110 mg/dl, HDL < 40, systolic or diastolic blood pressure $\geq$ 90th percentile and FBS $\geq$ 110 mg/dl. Exclusion criteria were genetic syndrome, endocrine disorders, physical impairments and using drugs which somehow affect MS components.

In the population, we excluded individuals with missing data on important variables (n=521). The final population with the size of 3819 students (242 abnormal and 3577 normal) was used in this study. Baseline characteristics of the study population are presented in Table 1 below.

Table 1. Baseline characteristics of MS study in Birjand, Iran

Field	Abnormal (242)	Normal (3577)	P-Value*
WEIGHT	62.022**	41.035	1
LENTH	155.017	146.791	1
WC	81.153	63.586	1
SBP	121.591	101.779	1
DBP	68.967	57.897	1
FBS	92.033	88.384	1
CHOL	171.905	157.067	1
TG	157.269	82.642	1
HDL	36.64	49.934	1
LDL	101.9	86.609	1
BMI	25.289	18.44	1

*P-Value of the significant test for difference between mean of variables for Abnormal and Normal Groups. P-Values greater than 0.95 indicates important and significant difference
** Mean of variable

There are many methods for feature selection and extraction to determine the most important related attributes for the labels of each class. The goal of both methods is to reduce the feature space to improve data analysis. This is especially important when the used dataset contains many features.

The main difference between feature selection and extraction is that the first one is reduced by selecting a subset of features without changing them, while feature extraction reduces their dimensions by deforming the main features to create other features that should be more significant. Feature

selection improves the knowledge of the process under study because it shows the features that most affect the intended phenomenon. In addition, the learning time of the learning machine is adopted and its accuracy must be considered because it is very important in data mining and machine learning applications.

However, in this study, we used decision tree models, which applies an embedded feature selection algorithm to construct trees. In section 3, you will see five most relevance features in the construction of decision trees.

The distribution of the study population by gender and age is shown in Table 2.

Table 2. Distribution of the study population by gender and age

Distribution of dataset		Abnormal	Normal
Gender	Male	139	1600
	Female	103	1977
Age-group	6-11	92	1564
	12-18	150	2013

2.2 Statistical analysis

The research dataset is partitioned into two subsets using stratified random sampling. The training dataset consists of 80% of the dataset used for model building and the remaining 20% for testing model.

2.3 Data balancing

Most prevalent classification models perform well when classes of target variable are distributed identically and model challenges increase when the distribution of observations is imbalanced [26]. Often medical data are mainly combined by the majority of normal class with only a minority of abnormal class [27]. In these cases, standard and popular models tend to generate high precision for the majority class and low precision for the minority class. Oversampling techniques are used to dominate this problem by adjusting the prior probability of the majority and minority class in training dataset to acquire more balanced data in each class. Some studies in medical sciences showed the effectiveness of SMOTE (Synthetic Minority Oversampling Technique) for handling imbalanced datasets [27,28]. In the present study, we balanced training datasets using SMOTE.

2.4 Methods for DT modeling

CART method is a popular DT model [29] used for predicting continuous and categorical target variables. CART split data subsets use all variables for creating two child nodes. The best variable is found using an impurity measure to produce as homogenous as possible data subsets with respect to the target variable. CART procedure is terminated by cost-complexity tree pruning. CHAID method [30] is made frequently by breaking subsets of dataset into two or more child nodes. This method begins with the whole dataset. If a variable is continuous, it is transformed into ordinal type before using the algorithm. This method merges non-significant categories for every variable by using P-value and Chis-square statistics. Any final category of a variable will eventuate in one child node if it is used to split the node. QUEST method [31] is used for univariate and linear combination splits. The QUEST process includes selecting a variable for each node and finding its point split by using ANOVA F-test or Leven's test for continuous and ordinal variables, and Pearson's chi-square for nominal variables. The variable with the greatest dependency on the target is chosen for splitting. The algorithm uses quadratic discriminant analysis (QDA) for searching optimal splitting point for the independent variable.

2.5 Model performances

The performance evaluation of models is usually done using testing set. By considering a classifier and a pattern, there are four possible outcomes. If the pattern is positive and it is classified as positive, it is counted as a true positive (TP); however, if it is classified as negative, it is counted as a false negative (FN). If it is negative and is classified as negative, it is counted as a true negative (TN), whereas if it is classified as positive, it is counted as a false positive (FP). These measures are used as a basis for calculating many common performance metrics such as classification accuracy, sensitivity, specificity and G-measure. Sensitivity measures the proportion of positives which are correctly classified, while specificity measures the proportion of negatives which are correctly classified.

The classification accuracy is the proportion of true results (both true positives and true negatives) in the patterns. The geometric mean (G-Mean) measures the balance between model performance on the negative and positive class and avoids overfitting to the negative class [32].

Receiver operating characteristics (ROC) graphs and area under ROC (AUC) are used to evaluate the models. ROC graphs are useful for organizing classifiers and visualizing their performance. AUC measures the discrimination performance of a model, that is, the extent to which a model successfully separates the positive and negative observations and ranks them correctly [33].

3. Results

3.1 Performance of DT models

The DT models were constructed using the balanced training subsets of four different analyses (females, males, 6-11 and 12-18 year-old group). The performance measurements for the three types of DT models in testing data are shown in Table 3.

Also, the ROC plots for the models on four testing datasets are shown in Figure 1. The accuracy of CART model in Table 3 and its AUC for all datasets are higher than other models. The comparison between these DT models shows that for all four analyses (Male, Female, 6-11 year-old and 12-16 year-old), CART is the best model; therefore, it was chosen as the final model.

Table 3. Performances of the DT models for four subset analyses in testing data

DT models	Performance measures	CART	CHAID	QUEST
Males	Sensitivity	93.846	88.846	87.692
	Specificity	90.759	91.749	86.755
	Accuracy	92.180	90.410	87.030
	G-mean	0.923	0.903	0.872
	AUC	0.976	0.955	0.949
Females	Sensitivity	98.701	91.775	92.208
	Specificity	93.684	92.895	91.316
	Accuracy	95.580	92.470	91.650
	G-mean	0.962	0.923	0.918
	AUC	0.982	0.975	0.953
age-group [6-11]	Sensitivity	96.914	95.370	95.988
	Specificity	95.380	91.419	87.789
	Accuracy	96.170	93.460	92.030
	G-mean	0.961	0.934	0.918
	AUC	0.975	0.974	0.929
age-group [12-18]	Sensitivity	97.396	97.396	94.531
	Specificity	97.354	91.534	92.593
	Accuracy	97.380	94.490	93.570
	G-mean	0.974	0.944	0.936
	AUC	0.987	0.982	0.957

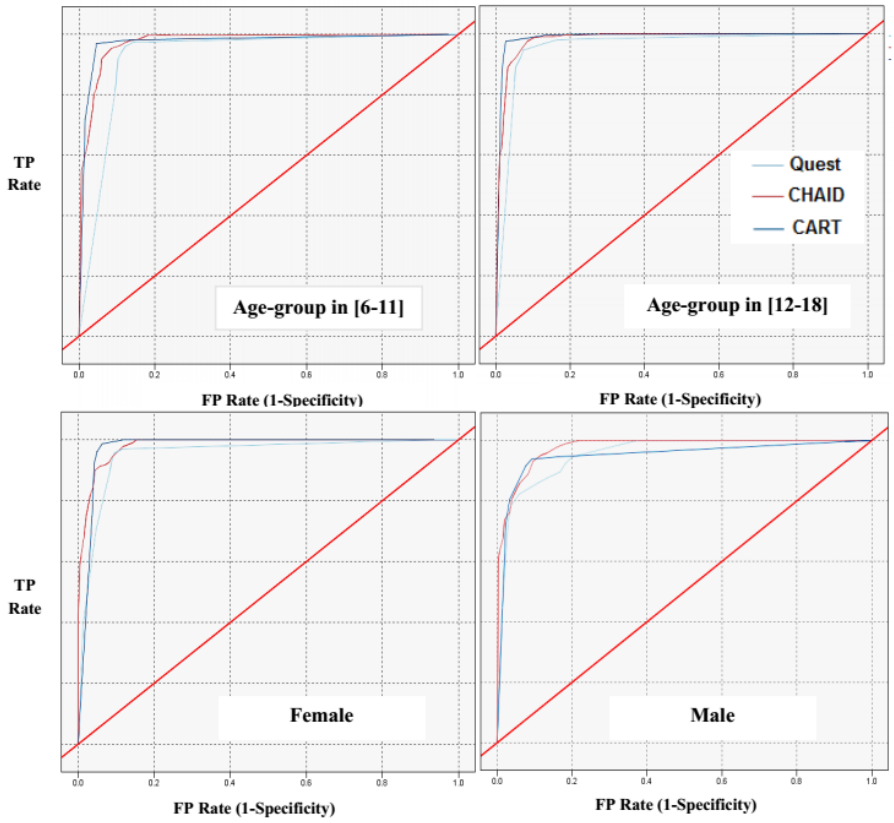

Figure 1. ROC plots for the 3 classifier models on 4 subsets of testing data

3.2 Results of the best Classification tree models

In this section, the results of CART models obtained separately from both sexes in the two age groups (6-11 and 12-18 year-old) are interpreted and explained. Figure 2 depicts the DT for females in the dataset, including the predictor variables and the cut points for each predictor. It uses three variables (TG, SBP and HDL) to classify six decision rules. Each rule identifies a special subgroup with a certain probability of outcome (normal or abnormal) for each person belonging to that subgroup. Table 4 shows the six predictive rules obtained by DT in Figure 2.

Table 4 shows that in the group of the females that TG is ≤ 111.5, If SBP is ≤120.56, the individual would be 100% normal (rule 1). With an increase in SBP to > 120.56, the risk of MS increases; hence, to reduce the risk of MS, the level of HDL should be more than 39.46. In this way the individual would be normal with the probability of 93.75% (rule 3), and if HDL is ≤ 39.46, the individual would be abnormal with the probability of 83.7% (rule 2). If TG level becomes > 111.5, HDL plays an important role in the individual's status. So if HDL is ≤ 45.9, then individual would be abnormal with the probability of 85.3% (rule 4). If HDL becomes greater than 45.9, SBP is an important factor in determining the individual's status. Therefore, if SBP is more than 121.44, the individual is abnormal with the possibility of 100% (rule 6), and finally if SBP is less than 121.44, the individual is normal with the possibility of 100% (rule 5).

Figure 3 depicts the DT for males. It used five variables (WC, TG, HDL, SBP, BMI) to categorize seven decision rules. Table 5 shows the seven predictive rules obtained by DT in Figure 3.

Table 5 shows that WC has a significant impact on the risk of MS in males. If it is ≤74.03, then TG is important in determining the individual's status, and if TG is ≤118.5, the individual would be normal with the probability of 97% (The Rule 1). If TG is >118.5, then SBP has more diagnostic power. For more precise diagnosis, BMI should be considered. If SBP is ≤120.17 and BMI is ≤20.71, the individual would be normal with the probability of 94% (rule 2), while if BMI is > 20.71, the individual would be abnormal with the probability of 71% (rule 3). If males' WC is > 74.03, HDL has high diagnostic power. If their HDL is ≤ 40.97, the individual would be abnormal with the possibility of 95% (rule 5). If HDL is > 40.97, then TG is important in the individual's status. If TG is > 109.8, then the individual would be abnormal with the possibility of 82% (rule 7). And finally, if TG is ≤ 109.8, then the individual would be normal with the possibility of 99% (rule 6).

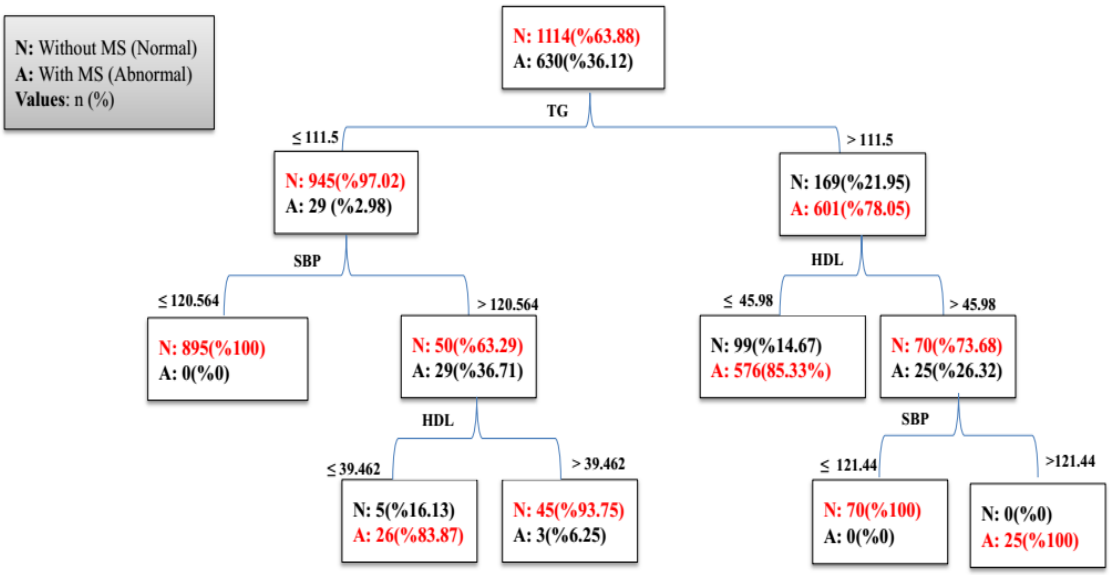

Figure 2. Decision tree for females

Table 4. The six predictive rules obtained through DT for females

Group	Definition (rules)	Probability*	Predicted class†
1	TG≤111.5 and SBP≤120.564	1.00	Normal
2	TG≤111.5 and SBP>120.564 and HDL≤39.462	0.84	Abnormal
3	TG≤111.5 and SBP>120.564 and HDL>39.462	0.937	Normal
4	TG>111.5 and HDL≤45.98	0.85	Abnormal
5	TG>111.5 and HDL>45.98 and SBP≤121.44	1.00	Normal
6	TG>111.5 and HDL>45.98 and SBP>121.44	1.00	Abnormal

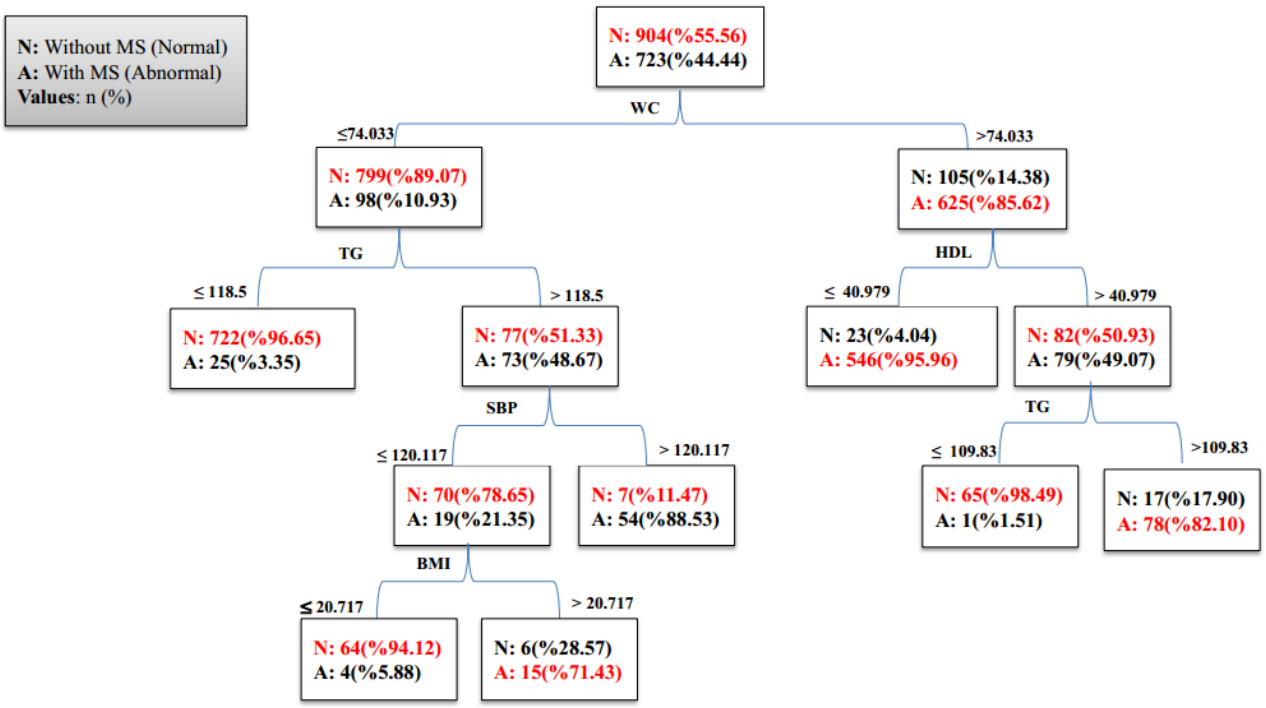

Figure 3. Decision tree for males

Table 5. The seven predictive rules obtained through DT for males

Group	Definition (rules)	Probability*	Predicted class†
1	WC≤74.033 and TG≤118.5	0.97	Normal
2	WC≤74.033 and TG>118.5 and SBP≤120.117 and BMI≤20.717	0.94	Normal
3	WC≤74.033 and TG>118.5 and SBP≤120.117 and BMI>20.717	0.71	Abnormal
4	WC≤74.033 and TG>118.5 and SBP>120.117	0.89	Abnormal
5	WC>74.033 and HDL≤40.979	0.96	Abnormal
6	WC>74.033 and HDL>40.979 and TG≤109.830	0.99	Normal
7	WC>74.033 and HDL>40.979 and TG>109.830	0.82	Abnormal

Figure 4 depicts the DT for the 6-11 year-old individuals in the dataset. It uses four variables (TG, HDL, DBP, SBP) for generating six decision rules. Table 6 shows the six predictive rules obtained using DT (see Figure 4).

Table 6 shows the importance of TG in the individuals' status. As can be seen, if TG is less than 110.46, HDL has high diagnostic power. If HDL is less than 39.98, SBP is important. If SBP is less than 111.15, the individual would be normal with the possibility of 98% (rule 1), while if SBP is more than 111.15, the individual would be abnormal with the possibility of 94% (rule 2). As for the individuals with TG less than 110.46 and HDL more than 39.98, the possible normality rate is 98%. However, if TG is more than 110.46, diastolic pressure has an important role in the individual's status. In addition, if DBP is more than 60.02, the individual would be abnormal with the possibility of 98% (rule 6), while if DBP is less than 60.02, HDL has high diagnostic power. If HDL is less than 40.5, the individual would be abnormal with the possibility of 89%, while if HDL is more than 40.5, the individual would be normal with the possibility of 92%.

Figure 5 depicts the DT for the 12-18 year-old group. It uses four variables (WC, TG, HDL and SBP) for generating six decision rules.

Table 7 shows the six predictive rules obtained using DT (see Figure 5).

Table 7 shows the importance of WC in the individual's status. If WC is ≤75.05, TG has an important role in diagnosis. If TG is ≤123.22, the individual would be normal with the possibility of 99% (rule 1). If TG is >123.22 then

SBP is important. Moreover, if SBP is ≤ 121.15, the individual would be normal with the possibility of 92% (rule 2), while if SBP is >121.15, the individual would be abnormal with the possibility of 91% (rule 3). If WC is > 75.05, HDL has high diagnostic power. If HDL is ≤ 40.96, the individual would be abnormal with the possibility of 82% (rule 4). If HDL is > 40.96, SBP has an important role in diagnosis. If it is ≤121.63, the individual would be normal with the possibility of 98% (rule 5), while if it is > 121.63, the individual would be abnormal with the possibility of 68%.

Thus, it is concluded that the most important prognostic factors for MS are the high level of TG and the low level of HDL for the 6-11-year-old group, and also the high level of TG, low level of HDL and WC for the 18-12-year-old group. These factors include TG and HDL in females and HDL, WC, and TG in males.

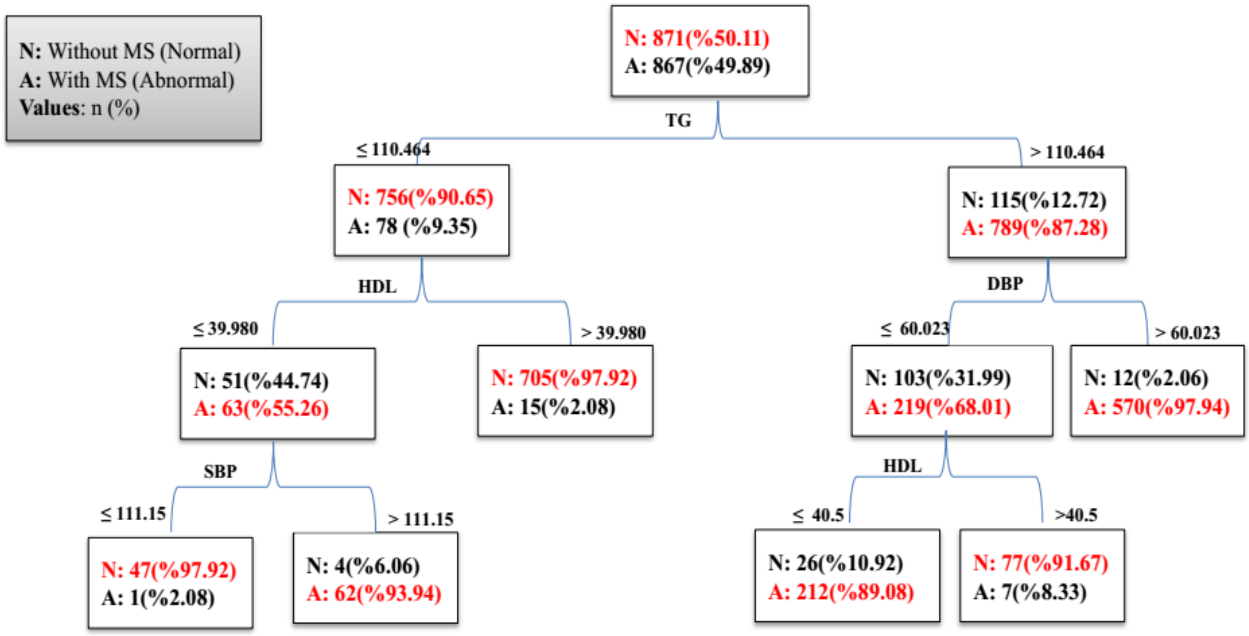

Figure 4. Decision tree for the 6-11 year-old group

Table 6. The six predictive rules obtained through DT for the 6-11 year-old group

Group	Definition (rules)	Probability*	Predicted class†
1	TG≤110.464 and HDL≤39.98 and SBP≤111.15	0.98	Normal
2	TG≤110.464 and HDL≤39.98 and SBP>111.15	0.94	Abnormal
3	TG≤110.464 and HDL>39.98	0.98	Normal
4	TG>110.464 and DBP≤60.023 and HDL≤40.5	0.89	Abnormal
5	TG>110.464 and DBP≤60.023 and HDL>40.5	0.92	Normal
6	TG>110.464 and DBP>60.023	0.98	Abnormal

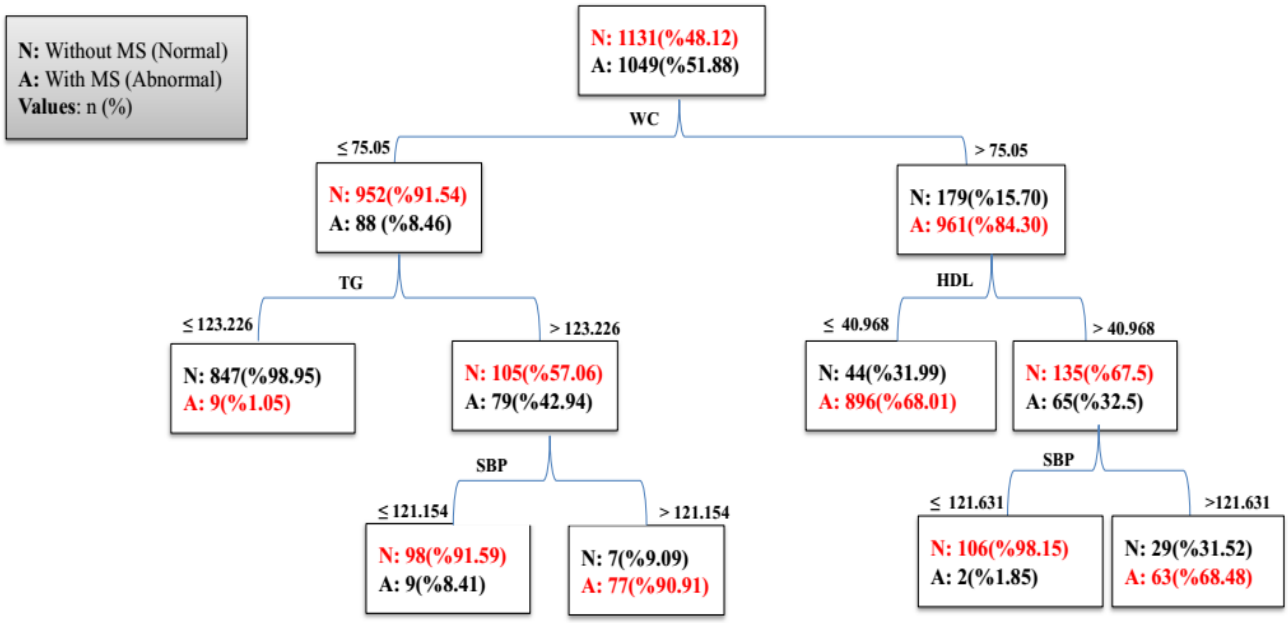

Figure 5. Decision tree for the 12-18 year-old group

Table 7. The six predictive rules obtained through DT for the 12-18 year-old group

Group	Definition (rules)	Probability*	Predicted class†
1	WC≤75.05 and TG≤123.226	0.99	Normal
2	WC≤75.05 and TG>123.226 and SBP≤121.154	0.92	Normal
3	WC≤75.05 and TG>123.226 and SBP>121.154	0.91	Abnormal
4	WC>75.05 and HDL≤40.968	0.82	Abnormal
5	WC>75.05 and HDL>40.96 and SBP≤121.631	0.98	Normal
6	WC>75.05 and HDL>40.968 and SBP>121.631	0.68	Abnormal

3. Discussion

According to the results, the most important prognostic factors for MS were the high level of TG and the low level of DBP and HDL in the 6-11-year-old group, and the high level of WC and the low level of TG for the 12-18-year-old group. The most important factors were TG and HDL in females and WC and TG in males. These varieties could be due to the significant differences of the under-studied risk factors in both sexes and age groups.

The importance of WC in predicting MS in males rather than females and 12-18 age groups rather than 6-11 years old could be due to the higher prevalence of abdominal obesity in males and the 12-18 age groups. The importance of abnormal TG and HDL in predicting MS, which can be seen in both sexes and age groups, is due to their high prevalence reported in several studies.

In a review article, the prevalence of dyslipidemia, hypercholesterolemia, the high level of LDL and the low level of HDL among Iranian children and adolescents were 3-48%, 3-50%, 5-20% and 5-88%, respectively. The most common dyslipidemia included the low level of HDL and the high level of TG [34]. MS was reported among Iranian children and adolescents as 14.1% in National Caspian Study, and the most common risk factor was dyslipidemia with the prevalence of 45.7%. In this study the prevalence of dyslipidemia, hypertension and MS were higher in overweight children. It has also reported that the most common factor for MS in Iranian and Turkish children and adolescents were the high level of TG and the low level of HDL, which can be due to racial differences [2]. Taheri et al. (2012) reported the prevalence of dyslipidemia in 11-18 year- old -adolescents in Birjand as 34.3%. The low level of HDL in 24.7% and the high level of TG in 14% were the most common dyslipidemia. The high level of total cholesterol was 6.1% and the high level of LDL was 3.5% [35]. Similar studies have reported that the prevalence of dyslipidemia in children living in Birjand is often seen in obese children compared with non-obese ones [36]. Rashidi et al. (2014) studied 2246 adolescents from 10 to 19 years old in Ahvaz and reported that the prevalence of MS was 9% (11% in males and 7% in females). In this study, the most common components were the high level of TG (23.5%) and the low level of HDL (24.1%). It was also found that the prevalence of MS was higher in obese or overweight individuals [37]. The most common components of MS in In another study, adolescents aged 10-18 years old in Korea were reported to have a high level of TG (21.2%) and a low level of HDL (12.6%) [38]. Among the 10-19 year-old teenagers in India, the prevalence of MS was 9.9% and the most common component was the low level of HDL in

58.3% of the cases [39]. In a study conducted in Ahvaz on adults with diabetes type II, the prevalence of MS was 73.1%, whereas TG and HDL had a high value in predicting of MS in females [40]. Also, WC had a high value in predicting MS. With 1 cm increase in WC in the 13-year-old individuals, the chance of MS increases 5% at the age of 22 [41]. In another study, overweight Italian children and adolescents aged 4-17 years old showed that the ratio of WC to height is the most predicting factor for MS. In 56% of the individuals, the ration of WC to height was more than 0.6. This increase was associated with the high risk of dyslipidemia and hypertension. Dyslipidemia was the most component of MS (43.6% and 24.15%),which was associated with hypertension [42]. Other studies on 6-13 school-aged students have showed that WC and SBP are independent predicting factors for insulin resistance [43].

4. Conclusion

In this study, assessing MS data in children and adolescents living in Birjand, Iran was performed using data mining based approach. In fact, this is an important multidisciplinary study conducted on the risk factors of metabolic syndrome and the obtained models are used as predictive tools to empower the Birjand's health care system. We successfully used a decision tree model to build 4 predictive models separately for 6-11 year-old and 12-18 year-old male and female individuals. According to the results, the most important prognostic factors for MS are the high level of TG and the low level of DBP and HDL in the 6-11-year-old group, and the high level of WC and the low level of TG in the 12-18-year-old group. The most important factors are TG and HDL in females and WC and TG in males. In conclusion, raising teens and families' awareness of the risk factors, screening children and teens, monitoring and controlling the risk factors through life style correction such as more physical activity and healthy eating are recommended.

References

[1] L. Silveira, C. Buonani , P. Monteiro, A. Mello, B. M. Antunes, and I. F. Freitas Júnior, "Metabolic Syndrome: Criteria for Diagnosing in Children and Adolescents", *Endocrinol Metab Synd,* Vol. 2, no.3, pp. 118, 2013.

[2] R. Kelishadi, S. Hovsepian, M. Qorbani, F. Jamshidi, Z. Fallah, S. Djalalinia, "National and sub-national prevalence, trend, and burden of cardiometabolic risk factors in Iranian children and adolescents 1990 – 2013", *Arch Iran Med,* Vol. 17, pp. 71-80, 2014.

[3] M. Jari, M. Qorbani, , M. E. Motlagh, R. Heshmat, G. Ardalan, and R. elishadi, "Association of overweight and obesity with mental distress in Iranian adolescents: The CASPIAN-III study", *Int J Prev Med,* Vol. 5, no.3, pp. 256-261, 2014.

[4] R. Kelishadi, S. Hovsepian, S. Djalalinia, F. Jamshidi, and M. Qorbani, "A systematic review on the prevalence of metabolic syndrome in Iranian children and adolescents", *J Res Med Sci,* Vol. 21, pp. 88, 2016.

[5] F. Taheri, T. Chahkandi, T. Kazemi, and B. Bijari, "Prevalence of Obesity and Overweight among Adolescents of Birjand, East of Iran", *Iranian Journal of Diabetes and Obesity,* Vol. 6, pp. 176-181, 2014.

[6] G. W.-H. Tan, K.-B. Ooi, L.-Y. Leong, and B. Lin, "Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-Neural Networks approach", *Computers in Human Behavior,* Vol. 36, pp. 198-213, 2014.

[7] S. Lee, H. Lee, J. R. Choi, and S. B. Koh, Scientific Reports, "Development and Validation of Prediction Model for Risk Reduction of Metabolic Syndrome by Body Weight Control: A Prospective Population-based Study", *Scientific Reports,* Vol. 10, no. 1, pp. 1-9, 2020.

[8] C.-S. Yu, Y.-J. Lin, C.-H. Lin, S.-T. Wang, S.-Y. Lin, S. H. Lin, et al., "Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study", *JMIR Med Inform,* Vol. 8, no. 3, pp. e17110, 2020.

[9] Y. Tang, T. Zhao, N. Huang, W. Lin, Z. Luo, and C. Ling, "Identification of Traditional Chinese Medicine Constitutions and Physiological Indexes Risk Factors in Metabolic Syndrome: A Data Mining Approach", *Evidence-Based Complementary and Alternative Medicine,* Vol. 2019, 2019.

[10] H. A. Kakudi, C. K. Loo, F. M. Moy, L. C. Kau, and K. Pasupa, "Diagnosis of Metabolic Syndrome using Machine Learning, Statistical and Risk Quantification Techniques: A Systematic Literature Review", *Malaysian Journal of Computer Science,* Vol. 34, No. 3, pp. 221-241, 2021.

[11] P. Khashayar, R. Heshmat, M. Qorbani, M. Motlagh, T. Aminaee, G. Ardalan, "Metabolic Syndrome and cardiovascular Risk Factors in a National Sample of Adolescent Population in the Middle East and North Africa:The CASPIAN III study", *International Journal of Endocrinology,* pp. ID702095,8 pages, 2013.

[12] H. Chiti, F. Hoseinpanah, Y. Mehrabi, and F. Azizi, "The Prevalence of MS in Adolescents with Varying Degreesof Body Weight: Tehran Lipid and Glucose Study TLGS", *Iranian Journal of endocrinology Metabolism,* Vol. 11, no. 6, pp. 625-637, 2010.

[13] A. E. Mehairi, A. A. Khouri, M. M. Naqb, S. J. Muhairi, F. A. Maskari, N. Nagelkerke, "Emirati adolescents: A school-based study", *PLoS One,* Vol. 8, no. 2, pp. e56159, 2013.

[14] A. L. Abdulwahab, Prevalence of Metabolic Syndrom among male Kuwaiti Adolescents aged 10-19 years, Health, Vol. 5, pp. 938-942, 2013.

[15] T. K. Hong, N. H. Trang, and M. J. Dibley, "Prevalence of metabolic syndrome and factor analysis of cardiovascular risk clustering among adolescents in Ho Chi Minh City", *Vietnam, Prev Med,* Vol. 55, no. 5, pp. 409-411, 2012.

[16] D. Yu, L. Zhao, J. Ma, J. Piao, J. Zhang, X. Hu, "Prevalence of MS among 7- 17year- old overweight and obese children and adolescents, Wei Sheng Yan Jiu", *Journal of Hygiene Research,* Vol. 41, no. 3, pp. 410-413, 2012

[17] N. Rosini, S. A. Z. Oppermann Moura, R. D. Rosini, M. J. Machado, and da Silva, E. L, "Metabolic Syndrome and Importance of Associated Variables in Children and Adolescents in Guabiruba - SC, Brazil", *Arq Bras Cardiol,* Vol. 105, pp. 37-44, 2015.

[18] S. Setayeshgar, S. J. Whiting, and H. A. Vatanparast, "Metabolic Syndrome in Canadian Adults and Adolescents: Prevalence and Associated Dietary Intake", *ISRN Obesity*, Vol. 2012, pp. ID 816846, 8 pages, 2012

[19] T. Chandola, E. Brunner, and M. Marmot, "Chronic stress at work and the metabolic syndrome: prospective study", Bmj, Vol. 332, no. 7540, pp. 521-525, 2006.

[20] A. Ramezankhani, A. Kabir, O. Pournik, F. Azizi, and F. Hadaegh, "Classification-based data mining for identification of risk patrns associated with hypertension in Middle Eastern population: A 12-year longitudinal study", *Medicine*, Vol. 95, no. 35, 2016.

[21] E. Hadavandi, J. Shahrabi, and Y. Hayashi, "SPMoE: a novel subspace-projected mixture of experts model for multi-target regression problems", *Soft Computing*, Vol. 20, no. 5, pp. 2047-2065, 2016.

[22] E. Hadavandi, J. Shahrabi, and S. Shamshirband, "A novel Boosted-neural network ensemble for modeling multi-target regression problems", *Engineering Applications of Artificial Intelligence*, Vol. 45, no. 2015, pp. 204-219, 2015.

[23] S. Kazemi, E. Hadavandi, S. Shamshirband, and S. Asadi, "A novel evolutionary-negative correlated mixture of experts model in tourism demand estimation", *Computers in Human Behavior*, Vol. 64, no. 2016, pp. 641-655, 2016.

[24] A. Ramezankhani, E. Hadavandi, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, "Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study", *BMJ open*, Vol 6, No. 12, pp. e013336, 2016.

[25] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, P. Pidetcha, and V. Prachayasittikul, "Identification of metabolic syndrome using decision tree analysis", *Diabetes Research and Clinical Practice*, Vol. 90. no. 1, pp. e15-e18, 2010.

[26] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," *European Conference on Principles of Data Mining and Knowledge Discovery in Knowledge Discovery in Databases: PKDD 2003*, ed: Springer, pp. 107-119, 2003.

[27] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh,and Khalili, D, "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes", *Medical Decision Making*, Vol. 36, no. 1, pp. 137-144, 2014.

[28] N. Chawla, K. Bowyer, L. Hall, and K. WP, "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357, 2002.

[29] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees: CRC press", ISBN 9780412048418 - CAT# C4841, 1984.

[30] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data", *Applied statistics*, Vol. 29, no. 2, pp.119-127, 1980.

[31] W.-Y. Loh, and Y.-S. Shih, "Split selection methods for classification trees", *Statistica sinica*, Vol. 7, no. 4, pp. 815-840, 1997.

[32] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, 2013. "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, Vol. 3, no. 10, pp. 27-38, 2013.

[33] S. Rosset, "Model selection via the AUC", *in Proceedings of the twenty-first international conference on Machine learning*, pp. 89, 2004.

[34] S. Hovsepian, R. Kelishadi, S. Djalalinia, F. Farzadfar, S. h. Naderimagham, and M. Qorbani, "Prevalence of dyslipidemia in Iranian children and adolescents: A systematic review", *J Res Med Sci*, Vol. 20, no. 5, pp. 503-521, 2015.

[35] F. Taheri, T. Chahkandi, T. Kazemi, B. Bijari, M. Zardast, and K. Namakin, "Lipid Profiles and Prevalence of Dyslipidemia in Eastern Iranian Adolescents, Birjand, Iran", *J Med Sci*, Vol. 40, no. 4, pp. 341-348, 2015.

[36] B. Bijari, F. Taheri, T.Chahkandi, T. Kazemi, K. Namakin, and M. Zardast, "The Relationship between Serum Lipids and Obesity among Elementary School", *Journal of Research in Health Sciences*, Vol. 15, no. 2, pp. 83-87, 2015.

[37] H. Rashidi, S. P. Payami, S. M. Latifi, M. Karandish, A. Armaghan Moravej, M. Aminzadeh, "Prevalence of metabolic syndrome and its correlated factors among children and adolescents of Ahvaz aged 10 – 19", *Journal of Diabetes & Metabolic Disorders*, Vol. 13, no. 1, pp. 1-6, 2014.

[38] S. Kim, and W. So, "Prevalence of Metabolic Syndrome among Korean Adolescents According to the National Cholesterol Education Program", *Adult Treatment Panel III and International Diabetes Federation*, *Nutrients*, Vol. 8, no. 10, 2016.

[39] V. Bhalavi, P. Deshmukh, M. Atram, K. Goswami, and N. Garg, "Prevalence of metabolic syndrome and cardio-metabolic risk factors in the adolescents of Rural Wardha", *International Journal of Biomedical Research*, Vol. 5, no. 12, pp. 754-757, 2014.

[40] H. Rashidi, F. farzad, B. Ghaderian, H. B. Shahbazian, M. Latifi, M. Karandish, "Prevalence of Metabolic Syndrome and Its Predicting Factors in Type 2 Diabetic Patients in Ahvaz", *JUNDISHAPUR SCIENTIFIC MEDICAL JOURNAL*. Vol. 11, no. 2, pp. 163-175, 2012.

[41] S. Aaron, J. Kelly, D. R. Steinberge, J. R. Jacobs, H. ChingPing, and A. Moran, "Predicting Cardiovascular Risk in Young Adulthood from Metabolic Syndrome, its Component Risk Factors, and a Cluster Score in Childhood", Int J Pediatr Obes, Vol. 6, no. 0, pp. 283-289, 2011.

[42] N. Santoro, A. Amato, A. Grandone, C. Brienza, V. varese, N. Tartaglione, "Predicting Metabolic in Obese Children and Adolescents Look, Measure and Ask", Obese Facts, Vol. 6, no. 1, pp. 48-56, 2013.

[43] V. Hirschler, C. Aranda, M. D. L. Calcagno, G. Maccalini, and M. Jadzinsky, "Can Waist Circumference Identify Children With the Metabolic Syndrome?", *Arch Pediatr Adolesc Med*, Vol. 159, no. 8, pp. 740-744, 2005.