Journal of System Management (JSM)
Online ISSN: 2538-1571, Print ISSN: 2322-2301
10(4), 2024, pp. 1-13
DOI: 10.30495/SJSM.2024.918345

Received: 27/11/2023
Accepted: 11/03/2024

RESEARCH ARTICLE **Open Access**

A Comparative Study of Data Science Techniques based on Ensemble Classification Algorithms in Healthcare Systems (Case study: Diabetic patients)

Farnoosh Bagheri[1], Reza Kamran Rad[2*], Morteza Darvishi[3]

Abstract
The adoption of unhealthy lifestyles by individuals can lead to the development of various health conditions, including hypertension, high blood fat levels, and diabetes, posing significant risks to their well-being. This study focuses on examining the lifestyle of patients with diabetes and high blood fat levels in the city of "bordekhoon," conducted at Health Care Centers. Diabetes is a global health concern that is rapidly increasing and is associated with substantial costs. By applying data mining techniques, early detection of diabetes can be achieved, which can help prevent the progression of the disease and mitigate complications such as cardiovascular issues, vision problems, and kidney diseases. Nowadays, data mining-based approaches are employed to predict diabetes and hypertension, aiming to enhance early diagnosis accuracy and obtain valuable insights. In this paper, a combination of classification techniques (Ensemble Method) is used to predict and identify two types of diabetes. Factors such as gender, diet, fasting plasma glucose (FPG), physical activity, cigarette consumption, age, genetic predisposition, and body mass index (BMI) are modeled and analyzed using IBM SPSS Modeler 18 software. The accuracy of the employed techniques is ultimately presented.

Keywords: *Data Mining, Mixed Classification Techniques, Healthcare System, Diabetes*

Introduction

Diabetes is the most frequent cause of death in most developed countries (Nazarzadeh M et al., 2015). According to the World Federation, the number of people infected with diabetes is more than 400 million in 2019, which is forecast to reach 650 million in 2035. So that more than half of the people are unaware of their illness. The disease is the death factor of 67 million in the year. According to the statistics of the Diabetes Association in Iran, 11 % of the population has diabetes in the country over the age of 25, which means the number of people infected with Type 2 diabetes is more significant than that of Type 1 diabetes. Those statistics accounted for 10 % of men and 11.5 % of women. In the past, the incidence of diabetes in villages and large cities was very different. Still, in the countryside, 8 % of the population has a high population of 25, an increase of 12 % in diabetes in villages in recent years.

Meanwhile, the capital " s share of the incidence of diabetes is 12.8 % over 25 years. Except for pregnancy diabetes, which is born during pregnancy, there are two main types: Type 1 and Type 2. In Type 1 diabetes, beta cells are pancreatic beta-producing hormones, the Blood glucose hormone, mistakenly attacked and destroyed by the immune system. Type 2 diabetes is the most common type, accounting for 75 to 85 % of the total. This happens when body cells stop

[1] Department of Quantitative Studies, University of Canada West, Canada.
[2*] Industrial Engineering Department, Faculty of Engineering, University of Semnan, Semnan, Iran. (Corresponding Author: r.kamranrad@semnan.ac.ir)
[3] Department of Industrial Engineering, Faculty of Engineering, Semnan University, Semnan, Iran

their insulin response, or beta cells cannot produce sufficient hormones. This type of clustering, especially in Type 2 diabetes, is highly heterogeneous and restricts medical instructions to the fact that they only respond to poor metabolic control and cannot be appropriate to predict that patients need severe treatments. Clustering can be a powerful tool to identify individuals at risk, take care of personal therapeutic regimes for these individuals, and lead physicians toward optimal treatment (Sohrabi et al., 2015). Data mining is a way to automatically analyze data and hidden patterns that do not happen manually (Elsappagh et al., 2015). Data mining can effectively be used to predict and rapidly and cheaply recognize diseases (Baronepel et al., 2015). The importance of anticipation of diabetes is that the patient can prevent its harmful effects by modifying the diet and exercise. Sou et al. could predict based on artificial neural network methods, decision trees, regression and dependency rules based on 3-D and 2-D body photographs with an accuracy of 89 % whether the person is diagnosed with diabetes or not (Chaoton et al., 2006). Poornami et al., using the support vector machine method, also detected 93 percent of the second type of diabetes. Their data for 768 patients using eight variables, including hypertension and insulin injections, have been used to predict `self–prodiction (Santi WP, 2010). Using the same method, the support vector machine Barakat et al. were able to improve the accuracy of diagnosis and accuracy of 94 percent of Type 2 diabetes. They extracted their research data from the data related to 4682 individuals (Barakat NB et al., 2010). By the same token, many Thai researchers could accurately identify more than 90 % of metabolic syndrome in individuals. They used data from 5683 people (Worachartcheewan, et al., 2015) [8]. Diwani et al. investigate and compare the performance of different data mining methods in identifying type 2 diabetes patients. This study applies supervised machine learning methods such as decision tree algorithms and Bayesian networks to the datasets. Then, the performance and success of each of these algorithms in diagnosing and recognizing diabetes patients were evaluated. The results of this study showed that the Bayesian naive algorithm, with an accuracy of 76.30 %, is better than the decision tree algorithm, with an accuracy of 73.82% (Diwani et al., 2014). In another study, Aljumah et al., with the aim of predictive analysis of diabetes treatment in Saudi Arabia, used the technique of support vector machines to discover patterns that best determine the best cure for diabetic patients at different ages.

The datasets used in this study were analyzed with the data collection of the country's nonpidemiological risk factors in the World Health Organization (WHO). For this purpose, the regression analysis technique and ODM 1 tool were used as data mining software for data analysis in this study. In this study, the data set to determine the effectiveness of different treatments was investigated in two young and old age groups. The results of this study show that treatment's effectiveness in two different age groups is different. In fact, in this study, a proper model for an effective treatment plan was provided that drug treatment in the age group of young people should start later to reduce the side effects of the disease.

In contrast, the treatment of drugs in the age group of older adults must begin immediately, along with other treatments, as there is no alternative option. The common mode for treating this disease in both groups was prescribing drugs for patients with Type 2 diabetes to control the effects of diabetes (Aljumah et al., 2013). Another paper used the genetic algorithm and neuro-fuzzy inference system to detect type 1 diabetes (Sreedevi et al., 2013). Research in diabetes diagnosis using neural networks has been developed based on factors affecting the genesis of the disease (Lakshmi KV et al., 2013; Srivastava et al., 2011). For the diagnosis of diabetes, based on risk factors with 16 steps of entry, an article was published by Sumathy et al., which facilitates diagnosis and helps patients predict diabetes by themselves (Sumathy et al., 2010). In another study, Seravana Kumar et al. proposed a system using open-source software and technique reduction techniques for diabetic data analysis. This system predicts the type of diabetes and its associated risks (Saravana Kumar et al., 2015). Aishwarya Iyer used a classification technique to study latent patterns in the kind of diabetes data set.

Bayse and decision trees technique was used in this model. A comparison of the performance and effectiveness of both algorithms is presented as a result (Aishwarya et al., 2015). Rajesh et al. used classification techniques. To improve performance, they used the C4.5 decision-tree algorithm to find hidden patterns among the dataset data (Rajesh et al., 2012). Muhammad et al. used decision-tree algorithms C4.5, neural network, clustering, and visualization to predict diabetes (Dost Muhammad Khan et al., 2011).

In various medical laboratories, there is a large amount of data and information, and sometimes, the value of this information can be counted as human salvation from death. These data and data can help detect and prevent many diseases.

The tests for determining the amount of fat and blood sugar were the most important and most common tests. That can diagnose and prevent many diseases, including heart disease, vascular, brain, and diabetes. If fasting blood sugar (FBS) is found in a larger or equal test of 126 mg per 100 cc, the person is considered diabetic. According to the definition that research is based on, the person with any diabetes can be identified and predicted. In contrast, a study that has already been used to predict diabetes using data mining techniques could only predict a type of diabetes and cause data and research variables from patients with a specific kind of diabetes and a set of healthy individuals. Also, the types of blood fats are synergistic in the amount of blood sugar (Esteghaamati et al., 2004). (Bagheri and Entezarian, 2023) evaluated recent researches in the quest to uncover emerging patterns in the use of business intelligence in marketing based on text mining techniques to extract pertinent terms in the realms of business intelligence and marketing.

A parallel machine scheduling by consideration of failures and energy consumption decline has been presented by (Rabbani et al., 2023). In this paper they minimized early and late delivery penalties, and enhancing tasks and designed a mathematical model for this problem that considered processing times, delivery time, rotation speed and torque, failure time, and machine availability after repair and maintenance.

(Ashrafijoo et al., 2022) applied study which is conducted using description based on testing as method. The discussion is established on analytical-computational methods. In their research, the documents and statistics of the Tehran Stock Exchange have been used to obtain the desired variables.

(Abdi and Abolmakarem, 2021) proposed two-module procedure including module deals with implementation of association rule mining. Using the well-known association rule mining techniques namely FP-Growth, several association rules have been extracted which indicated the effective factors on the waiting time for selling residential units. The main objective of the second module was to develop a fuzzy inference system which could determine the factors influencing the waiting time for selling residential units from historical data, so that the model can be used to estimate the time it to sell the property for a real estate agency.

(Safarzadeh et al., 2023) identified the influencing factors on implementation of smart city plans based on approach of technical and social systems. For this goal, the library study was done and then based on that a research plan is written that include using expert opinions and data mining technique, feature selection, clustering and also Delphi technique to identify and screen factors and then using clustering, the final factors were leveled.

This study uses a new classification method, such as hybrid classification methods (Ensemble), for analysis and prediction of type 1 and 2 diabetes and calculating the accuracy of the way. General variables of age and gender contribute to the incidence of diabetes. According to the literature mentioned above, the current research applies to the factors of age, gender and different amounts of sugar and blood fat, the variable of smoking and heredity. In this study, we first describe the data collection and the research variables. Then, in the second part of the research method and the third section, we study the case study, and in the fourth section, we analyze the data collected. Then, the evaluation of methods is concluded.

Problem Statement and Proposed Methods

Classification techniques are commonly used in data mining to predict data based on a model. These techniques involve using the values of certain attributes to predict the value of a specific attribute. Classification is a process that involves finding a model to determine the class of objects based on their characteristics (Hische et al., 2012). In classification algorithms, the initial dataset is divided into two sets: the training dataset and the experimental dataset. The model is built using the training dataset, and the experimental dataset is used to validate and assess the accuracy of the model. This paper provides an explanation of several classification algorithms, including decision tree-based methods, support vector machines, neural networks, compound techniques (Ensemble), and Bagging and boosting techniques.

Decision tree-based methods (DT)

One of the most well-known and valuable classification methods is the use of decision trees. In a decision tree, the internal nodes contain rules about the attributes, and the leaves are labeled with the corresponding classes. To predict the class of an object, the values of its attributes are evaluated against the constraints of the internal nodes, and a path is followed from the root to a leaf node, where the label represents the class of the object (Meng et al., 2013).

Support vector machine (SVM)

The underlying principle of the support vector machine is the linear classification of data, where it aims to find a line that maximizes the margin of separation. To find the optimal line for the data, nonlinear optimization techniques are employed, which are widely recognized and constrained methods (Karaolis et al., 2010).

Artificial neural network.(ANN)

An artificial neural network is a data processing system that is inspired by the human brain's ability to process and analyze data using a large number of interconnected and parallel processors. In these networks, a data structure is designed, with the help of programming knowledge, to mimic the behavior of neurons. By establishing connections between these neurons and employing a learning algorithm, the network is trained to establish relationships between attributes and class labels (Toussi et al., 2009).

Bagging and boosting techniques

The Concept Boosting in Data Mining

This concept is utilized to generate multiple models for prediction or classification. The boosting algorithm was first introduced by Schapire (Schapire et al., 1999). It has been demonstrated that a weak classifier can be transformed into a robust classification set in the PAC (probably approximately correct) form. Among the top 10 data mining algorithms, 'Adaboost' is one of the most well-known algorithms within this family. In this method, the bias is reduced by the variance and it increases similarly to the margins in support vector machines. The algorithm utilizes the entire dataset to train each class, but it places more emphasis on complex data after each training iteration to achieve more accurate classification. This iterative and adaptive approach depends on the data distribution, with a focus on samples that have been misclassified. Initially, all records are assigned the same weight, and unlike bagging, the weights increase with each iteration. The weights of samples that are misclassified are increased, while the weights of samples that are correctly classified are reduced. Additionally, a separate weight is assigned to each classifier based on its overall accuracy, which is then used during the testing phase. Reliable classifiers will have a higher weight factor. Finally, when presented with a new sample, each classifier assigns a weight, and the class label is determined through majority voting.

The concept of Bagging in data mining

This concept is utilized to aggregate the predicted ratings from multiple models. Let's consider a scenario where you intend to build a prediction classification model, but the dataset is small. In this case, you can select subsets (or replacements) from the dataset and employ "CHAID" and "RT&C" decision trees to generate predictions for the selected subsets. Typically, different decision trees will be generated for different subsets. Subsequently, a simple voting mechanism is employed to make predictions by considering the predictions obtained from the various decision trees. The final prediction will be based on the majority vote among the different decision trees.

For a more comprehensive understanding of the current problem and the proposed approaches, it is important to highlight the following points. This article focuses on the development of four classification methods aimed at grouping diabetic patients and predicting the type of their disease based on their conditions. To achieve this objective, a questionnaire was designed to assess the patients' conditions. The resulting data was then utilized to create the proposed models for the specified goals. One notable advantage of this research compared to previous studies is the novel application of these approaches in the field of healthcare, particularly for diabetic patients.

Case Study Based on Healthcare Systems

In this research, data were prepared through a questionnaire at the health center, which the research questionnaire has been applied in Appendix 1. The questionnaire's questions were collected through medical health professionals.

Examples of Sample Data

Before describing the data, medical definitions must be expressed in different types of blood fats and blood sugar. Blood fat is divided into two types: cholesterol and triglyceride. Cholesterol is divided into two types: harmful fatty (LDL) and beneficial fats (HDL). Harmful fat or Lipoprotein with low density has a large amount of cholesterol and a small amount of protein. The function of unhealthy fats is to carry cholesterol and other fats in the blood. The increase in harmful fat in the blood can lead to the aggravation of the arteries of the heart and the brain and the emergence of heart and heart disease. Beneficial fats have a large amount of protein and a small amount of cholesterol, which prevents the cholesterol from the blood that is leaving the veins and taking it to the liver .Helpful fats protect the heart, and the small amount in the blood can be a factor in the causes of heart disease.

The triglyceride is the same fat that is in food. If a lot of calories go into the body, the body converts extra calories into triglyceride and stores the cells in fat. Increasing the amount of triglyceride in the blood will clog the arteries and then damage the pancreas. So, insulin is not produced enough by the pancreas. Because insulin causes a drop in blood sugar, its lack is causing a rise in blood sugar, thus causing diabetes (Asadollahi et al., 2015). FBS is the amount of fasting blood sugar. Everyone has a dose of sugar in their blood, which naturally falls between 70 and 110 milligrams per 100 cc annually. The data used in this study is related to the Bordekhoon City Health Centre data in Dayyer County, Bushehr Province, which was collected in 2020. Data related to 935 clients through questionnaires have been systematically derived, including the general variables of age and gender, the amount of blood sugar and blood fat, and other variables in the form of a table.

Research Variables

The following table describes the variables of the research variables and their characteristics. Gender, a person's cigarette, and genetics (family history) are binary variables, and the age variable, blood sugar, is the amount of body activity per minute per week. The BMI variable is considered a weighted average relative to the person's height, and the nominal variable BFBS, if taken, represents type 1 diabetes. If it gives 2, type 2 diabetes; if zero is available, the person is healthy.

Table 1.
Description of the research variables

Variables	Description	Variable type
SEX	Gender	Binary
AGE	Age	Numerical
FBS	Blood Sugar	Numerical
Body activity level		Numerical
The mood of being a smoker		Binary
Food Program		Binary
BMI	Body mass	Numerical
Genetics	Heredity	Binary
BFBS	Diabetes	Tree value (Type1, Type2, No diabetes)

The table presented above illustrates the second step of the data mining process, which involves identifying the variables and data pertinent to the problem. This table provides a detailed examination of the description and type of each significant variable in the classification models.

Data Analysis and Assessment of the Proposed Methods

First, we examine the reliability of the research questionnaire. Cronbach's alpha coefficient is used to obtain the reliability of the questionnaire. To compute the SPSS 25 software. Cronbach's alpha coefficient was calculated by 0.835, showing the questionnaire's stability.

Table 2.
Cronbach's alpha coefficient

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha
0/835	0/835	0/835

The demographic population of the statistical population

In this stage, using statistical analysis and illustration data, we come to a preliminary knowledge of the data that this knowledge will contribute to explaining and interpreting the modelling results. In the descriptive analysis, we determine the number and percentage of each decision variable in the dataset.

Table 3.
Demographic information of the research population

Age variable						
Age range	20-36	39-36	41-39	62-41	65-62	>65
Count	77	40	147	244	17	299
Percentage (%)	8.24	4.28	15.72	26.1	1.82	31.95
Body mass variable (BMI)						
Body mass interval	23-21	25-23	26-25	28-26	29-28	>29
Count	245	151	33	23	320	97
Percentage (%)	26.2	16.1	3.5	2.45	34.2	10.37
The variable of body activity						

Exercise Time (minutes)	72-58	72-100	100-123	>123		
Count	129	556	189	13		
Percentage (%)	13.7	59.46	20.21	1.4		
Gender variable				Diabetes variable		
Gender	men	Women	Type of diabetes	0	1	2
Count	174	761	The number of infected people	245	121	568
Percentage (%)	18.61	81.39	Percentage (%)	26.2	12.94	60.75

In Table 4, we investigated the descriptive statistics of the research variables.

Table 4.
Descriptive statistics of the variables

Variables	Min	Max	Mean	Std.Dev	Skewness
SEX	-	-	-	-	-
AGE	18	75	47.268	16.608	0.019
BMI (kg/cm)	19.840	35	25.913	3.197	0.045
Food Program	0	1	-	-	-
Heredity	0	1	-	-	-
Body activity (min)	18	200	78.161	18.668	1.721
Blood sugar (mg / dL)	36	190	127.632	41.106	0.229
stressful activity	0	1	-	-	-
Cigarettes	0	1	-	-	-
BFBS	0	2	-	-	-

According to the data prepared from analysis with the SPSS modeller, the distribution function of the continuous variables of age is gamma, the body activity variable is lognormal, and the blood sugar variable and the fat mass variable (BMI) is Triangular. Furthermore, this table also includes the values of descriptive indices, such as central tendencies and dispersions, for each of the variables. These indices have been determined and reported to provide transparency regarding the patients' conditions. The correlation of the research variables is presented in Table 5.

Table 5
Correlation of variables

Variables	Blood Sugar	BMI	Age	Body activity
Blood Sugar	1	0.708	0.686	-0.091
BMI	0.708	1	0.595	0.241
Age	0.686	0.595	1	0.057
Body activity	-0.091	0.241	0.057	1

By choosing the Anderson-Darling technique to fit the variables, the high correlation of blood sugar variables with age and BMI presents the interpretation that increases the amount of blood sugar by increasing age and body fat mass. Due to this, the incidence of diabetes in older ages is higher than that of older people. Also, the higher the BMI is, the higher the probability that an individual has diabetes.

Forecast method Type 1 diabetes and Type 2 diabetes

Logistic regression (LR)

Using the Minitab16 software, we describe the nominal logistic regression to obtain the NLR probabilities with respect to the reference variable of the absence of diabetes (that is, zero):

$$\frac{\boldsymbol{\pi_1}}{\boldsymbol{\pi_0}} = \mathbf{exp}(221.44 + 8.9580(\mathrm{SEX}) - 1.7208(\mathrm{AGE}) - 4.5376(\mathrm{BMI}) - 2.7286(\mathrm{DIET}) - 2.9849(\mathrm{INHERITANCE}) - 1.2344(\mathrm{EXERCISE}) + 0.14731(\mathrm{FBS}) - 39.1326(\mathrm{STREES}) - 6.3137(\mathrm{SIGARET})$$

$$\frac{\boldsymbol{\pi_2}}{\boldsymbol{\pi_0}} = \mathbf{exp}(2.15070 + 3.0478(\mathrm{SEX}) + 0.0598(\mathrm{AGE}) - 0.1540(\mathrm{BMI}) - 2.62357(DIET) + 1.05870(\mathrm{INHERITANCE}) + 0.003104(\mathrm{EXERCISE}) - 0.0031045(\mathrm{FBS}) - 7.18319(\mathrm{STREES}) + 0.19406(\mathrm{SIGARET})$$

Using the SPSS modeler 18 software, 922 recorded true records and there are 13 wrong records. The accuracy of the model prediction is about 99 percent. In addition, the third equation for above nominal logistic regression is $\boldsymbol{\pi_0 + \pi_2 + \pi_3 = 1}$. It is important to note that by employing the three equations mentioned above, the probability of a diabetic patient having any type of the disease can be predicted with a remarkable accuracy of 99**%**.

Review of non-combinatorial classification methods

Using separate classifier techniques, we analyzed dataset analysis. We used clustering classifiers, support vector machines, neural networks and decision trees.

Clustering

All techniques are applied to preprocessing data. The selected variables were selected by numerical variables (four variables), and clustering quality was classified as very good 0.8 and was classified into 5 clusters. The highest importance in this analysis is related to the blood sugar variable, and the lowest importance is associated with the variable of body activity. Table 6 is the supplementary information of the cluster.

Table 6.
Clusters Information

Cluster	Quantity	Percent (%)	Variables importance			
			BMI	**AGE**	**Blood sugar**	**Body activity**
Cluster-1	310	33.2	Importance:1.00 Mean:22.36	Importance:1.00 Mean:40.6	Importance:1.00 Mean:86.40	Importance:0.34 Mean:75.12
Cluster-2	316	33.8	Importance:1.00 Mean:28.90	Importance:1.00 Mean:67.66	Importance:1.00 Mean:179.33	Importance:0.34 Mean:75.43
Cluster-3	158	16.9	Importance:1.00 Mean:28.18	Importance:1.00 Mean:38.7	Importance:1.00 Mean:107.64	Importance:0.34 Mean:101.59
Cluster-4	25	2.7	Importance:1.00 Mean:26.74	Importance:1.00 Mean:61.80	Importance:1.00 Mean107.33	Importance:0.34 Mean:90.6
Cluster-5	126	13.5	Importance:1.00 Mean:24.15	Importance:1.00 Mean:20.40	Importance:1.00 Mean:128.36	Importance:0.34 Mean:60.71

The results show that the variables in each cluster have characteristics, each of the most important in their cluster, and the most appropriate variable is the importance of value 1, and the lowest is the amount of 0.34. Here, the age variable, blood sugar and BMI are of the highest importance, and the variable of body activity has the lowest importance in this technique.

Support vector machine analysis

The results of this technique are presented in accordance with Table 7 by SPSS modeller 18. The accuracy of the forecast report is about 99 %.

Table 7.
SVM Technique Accuracy

Correct	924	98.82
wrong	11	1.18
total	935	

The results displayed in Table 7 demonstrate the commendable performance of the support vector machine method in accurately classifying diabetic patients based on the influential variables.

Neural network algorithm (NN)

Artificial neural networks, also known as neural networks, are advanced computational systems and modern methods used for machine learning, knowledge representation, and ultimately generating predictions of output responses from complex systems. In this study, neural networks were applied to a dataset consisting of 935 samples, and the obtained results were analyzed using SPSS Modeler 18.

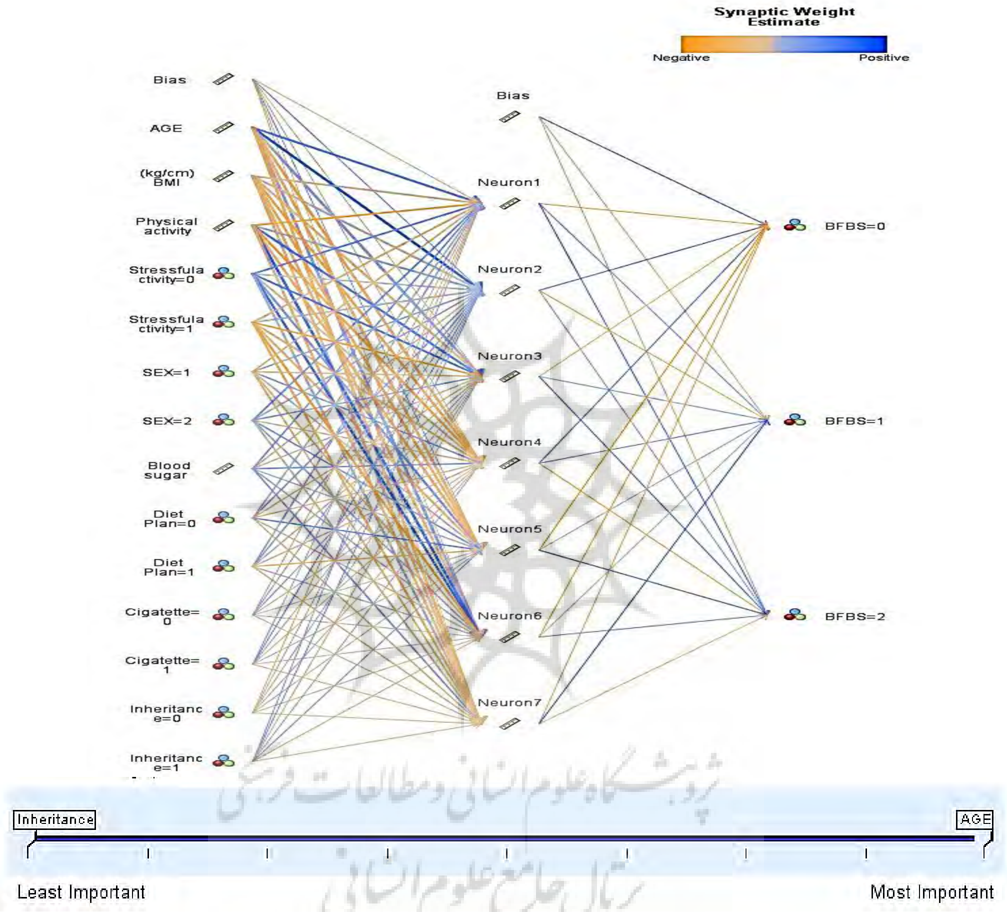

Figure 1. *The neural network trained for diabetes diagnosis*

Figure 1 shows the conceptual scheme of the neural network trained to diagnose diabetes in this study. The correlation of each variable on the neurons and then each neuron (as the latent layers of the neural network) on the diabetes variable is defined in the colour of the synapses. The positive effect is shown in the colour blue line and the negative effect of the orange line. The greatest importance of the neural network of research affects the age variable, and the most minor importance in heredity. The foresight accuracy reports about 98.50 % and gives the highest importance to the age variable at about 25 %. In this respect, 921 set the record correctly and wrongly predicted 14 records.

Decision trees technique (C 5.0)

This algorithm is developed by the ID3 algorithm. It is based on the hunt algorithm. The C5.0 algorithm decides both classes of class and continuously builds a tree. C5.0 Eliminations observed for eliminating non-essential branches to improve classification accuracy. In this technique, the greatest impact of the research variables is related to stress and age, whose importance is estimated at 0.4 and 0.6, respectively. The accuracy of prediction in this technique is estimated at around 99.04. The tree graph is shown in Fig. 2.

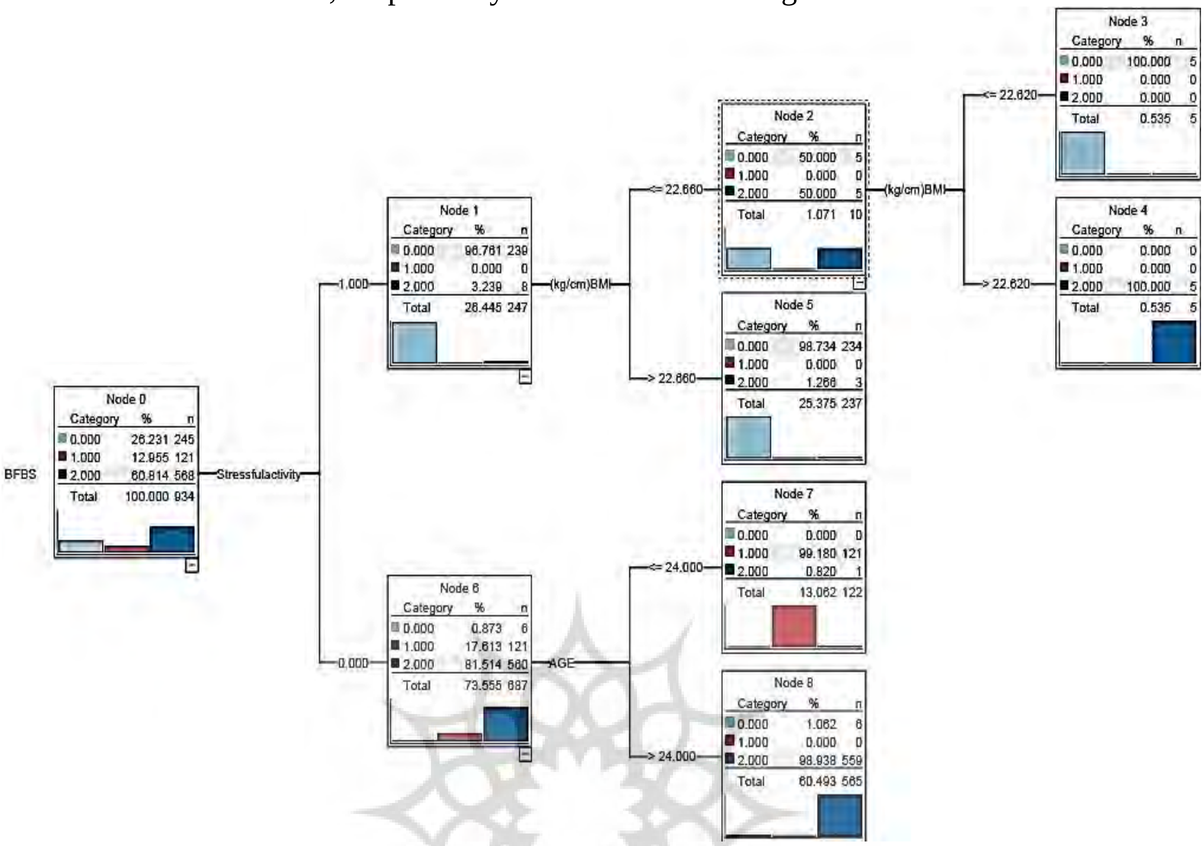

Figure 2. *The decision tree of the research variables*

The decision tree of the present study, at node 0, states that people who did not have diabetes (BFBS = 0) account for about 27 %, the number of people with Type 1 diabetes, around 13 %, and those with Type 2 diabetes, around 61 %. At the second level, where the "stressful activity" variable has achieved the most importance, nodes 1 and 6 are split. In node 1, people who have stressful activities and do not have any type of diabetes are around 97 %. And those who did not have stressful activities are about 81 % of Type 2 diabetes.

At the next level, the BMI variable and the age variable are more important than other variables. It shows that people who have been involved in stressful activities and BMI <= 22, equal probabilities have no diabetes, or type 2 diabetes, and people who have stressful activities, BMI >= 22 ,They are more likely to have no diabetes at all. Or in other words, people who do not have stressful activities and whose age is less than 24 with the possibility of approaching 1 have Type 1 diabetes and people whose age is larger than 24 with the possibility of close to one Type 2 diabetes. Level 4 shows the clusters of the BMI and the diet variable, indicating that the BMI variable contains a node of node 3 and node 4. In node 3, about 100 % of the statistical sample lacks any type of diabetes. Also in node 4 are all samples with Type 2 diabetes.

Review of combinatorial classification methods (Ensemble)

After studying non-combinatorial methods and presentation of results, we want to examine the combined methods of data mining techniques to determine the accuracy of the technique. It is predictable that integrated techniques have better accuracy than individual methods. Because these methods are sequential in sequence, the output of a technique is applied as input to another technique. Finally, after decreasing non-compliance cases at each stage, the final accuracy will be improved.

The combination of neural network techniques (NN) and logistic regression (LR)

The composition results show that the combinatorial prediction accuracy of the two techniques is about 99.22 %. This accuracy is better than forecasting individual techniques. and has reduced about 42 records.

Combined Bagging and Boosting techniques with a decision–tree.

≠ Combining boosting techniques with C5: The accuracy of the forecast rate hit 100 %, reducing the number of nine records.

≠ Combining bagging techniques with C5: The accuracy of the forecast rate hit 99.14 %, reducing the number of one record.

Comparison of classification methods

In the following, the performance of these six sorting methods is compared adaptively in Table 8. As is clear, the accuracy of discrete techniques is 99 %, but the accuracy of combining techniques is equal to 100 %. Hence, the subsequent results demonstrate that the combined classification approaches including C5 and Boosting developed in this paper has outperformed other combined and individual methods significantly.

Table 8.
Comparison of data mining techniques accuracy

Hybrid classification methods		Individual classification methods	
Accuracy (%)	Techniques	Accuracy (%)	Technique
99.22	NN+LR	98.61	SVM
100	C5+Boosting	98.50	NN
99.14	C5+Bagging	98.61	LR

Conclusion and Future Research

Data mining methods in recent years in medicine and health care in the fields of diagnosis and prevention, treatment selection, mortality prediction and prediction of therapeutic costs have been widely used. In this research, using different classification techniques in data mining software SPSS modeler 18, single-application and combination models are made to identify types of diabetes and predict and classify patients with diabetes and non-diabetic. The best accuracy of the assessment model for combining the decision - tree (C5.0) and Boosting technique was presented by 100 % accuracy. In this study, a statistical model based on logistic regression was proposed to predict type 1 and 2 diabetes, which was achieved with precision of 99 percent. The results of this study can be helpful for other researchers in the future, as well as hospitals and health care centers, by creating an electronic record of the medical condition of each patient and the patient's physical condition, including blood pressure, weight, and fat content. Furthermore, it is feasible to extend the proposed methods for categorizing and predicting the status of other diseases, such as infertility in patients. In this context, models can be developed to identify patients who seek infertility treatment and assess their likelihood of successful pregnancy based on physiological characteristics of the couple. This approach can help prevent significant treatment costs being incurred by couples who have a low chance of achieving pregnancy.

References

Abdi, F., & Abolmakarem, S. (2021). Mining a Set of Rules for Determining the Waiting Time for Selling Residential Units. *Journal of System Management*, *7*(2), 171-203. doi: 10.30495/jsm.2021.1931499.1480

Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, *25*(2), 127-136. https://doi.org/10.1016/j.jksuci.2012.10.003

Asadollahi, K., Delpisheh, A., Asadollahi, P., & Abangah, G. (2015). Hyperglycaemia and its related risk factors in Ilam province, west of Iran-a population-based study. *Journal of Diabetes & Metabolic Disorders*, *14*, 1-14. DOI 10.1186/s40200-015-0203-9

Ashrafijoo, B., Fegh-hi Farahmand, N., Alavi Matin, Y., & rahmani, K. (2022). Designing an Optimal Model Using Artificial Neural Networks to Predict Non-Linear Time Series (case study: Tehran Stock Exchange Index). *Journal of System Management*, *8*(4), 65-80. doi: 10.30495/jsm.2022.1965914.1679

Bagheri, R., & Entezarian, N. (2023). Topic Modeling Emerging Trends for Business Intelligence in Marketing: With Text Mining and Latent Dirichlet Allocation. *Journal of System Management*, (),-. doi: 10.30495/jsm.2023.2000760.1901

Baron-Epel, O., Heymann, A. D., Friedman, N., & Kaplan, G. (2015). Development of an unsupportive social interaction scale for patients with diabetes. *Patient preference and adherence*, 1033-1041. https://doi.org/10.2147/PPA.S83403

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, *14*(4), 1114-1120. DOI: 10.1109/TITB.2009.2039485

Chaoton S, Chienhsin Y, Kuanghung H, Wenko C. Data mining for diagnosing type II diabetes from three-dimensional body surface anthropometrical scanning data. Comput Matemat Appl 2006.

Diwani, S. A., & Sam, A. (2014). Diabetes forecasting using supervised learning techniques. *Adv Comput Sci an Int J*, *3*, 10-18. https://doi.org/10.1016/j.procs.2022.12.107

Esteghamati A. Comprehensive guide diagnosis and treatment diabetes. Spe Soc Diabetes America 2004; 5:16-30.

Eswari, T., Sampath, P., & Lavanya, S. J. P. C. S. (2015). Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, *50*, 203-208. https://doi.org/10.1016/j.procs.2015.04.069

Hische M, Larhlimi A, Schwarz F, Fischerrosinsky A, Bobbert T, Assmann A, et al. A distinct metabolic signature predicts the development of fasting plasma glucose. J Clin Bioinfo2012; 2:2-3.

Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on information technology in biomedicine*, *14*(3), 559-566. DOI: 10.1109/TITB.2009.2038906

Khan, D. M., & Mohamudally, N. (2011). An integration of K-means and decision tree (ID3) towards a more efficient data mining algorithm. *Journal of Computing*, *3*(12), 76-82.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*. https://doi.org/10.48550/arXiv.1502.03774

Lakshmi, K. V., & Padmavathamma, M. (2013). Modelling an expert system for diagnosis of gestational diabetes mellitus based on risk factors. IOSR-JCE, 8(3), 29-32.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99. https://doi.org/10.1016/j.kjms.2012.08.016

Nazarzadeh, M., Bidel, Z., & Sanjari Moghaddam, A. (2016). Meta-analysis of diabetes mellitus and risk of hip fractures: small-study effect. *Osteoporosis International*, *27*, 229-230. DOI 10.1007/s00198-015-3358-9

Purnami, S. W., Embong, A., Zain, J. M., & Rahayu, S. P. (2009). A new smooth support vector machine and its applications in diabetes disease diagnosis. *Journal of Computer Science*, *5*(12), 1003-1008. https://doi.org/10.3844/jcssp.2009.1003.1008

Rajesh, K., & Sangeetha, V. (2012). Application of data mining methods and techniques for diabetes diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(3), 224-229.

Rabbani, Y., Qorbani, A., & Kamran Rad, R. (2023). Parallel Machine Scheduling with Controllable Processing Time Considering Energy Cost and Machine Failure Prediction. *Journal of System Management*, *9*(1), 79-96. doi: 10.30495/jsm.2022.1967931.1689

Srivastava, S., & Tripathi, K. C. (2012). Artificial neural network and non-linear regression: A comparative study. *International*

Journal of Scientific and Research Publications, *2*(12), 740-744. ISSN: 2250-3135

Sreedevi, E., & Padmavathamma, M. (2012, June). Modelling effective diagnosis of risk complications in gestational diabetes mellitus: an e-diabetic expert system for pregnant women. In *Fourth International Conference on Digital Image Processing (ICDIP 2012)* (Vol. 8334, pp. 632-635). SPIE. https://doi.org/10.1117/12.956489

Sumathy, M., Kumar, P., Jishnujit, T. M., & Kumar, K. R. (2010). Diagnosis of diabetes mellitus based on risk ISSN: 0975-8887

Sohrabi, B., & Hamideh, I. (2015). Macro Data Management in the Private and Public Sectors. (In Persian)

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*, 197-227.

Safarzadeh, A., Bazaei, G., & Faghihi, M. (2023). Identification of influencing factors on implementation of smart city plans based on approach of technical and social system. *Journal of System Management*, *9*(2), 197-212. doi: 10.30495/jsm.2023.1972219.1719

Toussi, M., Lamy, J. B., Le Toumelin, P., & Venot, A. (2009). Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and examples for type 2 diabetes. *BMC medical informatics and decision making*, *9*(1), 1-12. DOI:10.1186/1472-6947-9-28

Worachartcheewan, A., Shoombuatong, W., Pidetcha, P., Nopnithipat, W., Prachayasittikul, V., & Nantasenamat, C. (2015). Predicting metabolic syndrome using the random forest method. *The Scientific World Journal*, *2015*. https://doi.org/10.1155/2015/581501