MATHEMATICAL MODELING AND COMPUTING, Vol. 8, No. 4, pp. 638–645 (2021)

Use of Data Mining in the prediction of risk factors of Type 2 diabetes mellitus in Gulf countries

Boutayeb W.[1], Badaoui M.[2], Al Ali H.[3], Boutayeb A.[1], Lamlili M.[1]

[1] *University Mohammed First, Faculty of Sciences, Oujda, Morocco*
[2] *University Mohammed First, Highest School of Technologies, Oujda, Morocco*
[3] *Emirates Aviation University, Dubai, United Arab Emirates*

(Received 23 May 2021; Accepted 7 June 2021)

Prevalence of diabetes in Gulf countries is knowing a significant increase because of various risk factors, such as: obesity, unhealthy diet, physical inactivity and smoking. The aim of our proposed study is to use Data Mining and Data Analysis tools in order to determine different risk factors of the development of Type 2 diabetes mellitus (T2DM) in Gulf countries, from Gulf COAST dataset.

Keywords: *Gulf COAST dataset, Data Mining, decision tree, principal component analysis, Type 2 Diabetes.*

2010 MSC: 62-07, 68T99, 62H25 **DOI:** 10.23939/mmc2021.04.638

1. Introduction

The World Health Organization (WHO) defines type 2 diabetes mellitus (T2DM) as a chronic disease characterized by a high concentration of glucose in the blood (hyperglycemia) resulting from defects in insulin secretion, insulin action, or both [1]. Unhealthy eating habits characterized by high consumption of carbohydrates and free fats in addition to physical inactivity are the main risk factors of this type of diabetes [2–4].

Type 2 diabetes mellitus that used to affect middle aged persons is now seeing a significant rise among obese children [5]. Due to the fact that patients don't necessary feel the signs of TD2M, the disease can be unnoticed. A cross sectional study by Donaghue et al. in Iran [6], shows that 30˘50% cases of type 2 diabetic patients remain undiagnosed. Consequently, the risk of microvascular complications (retinopathy, neuropathy and nephropathy) is increased [6, 7].

In the last decade the prevalence of T2DM has known an alarming increase in Gulf countries with highest prevalence in Kingdom of Saudi Arabia 31.6%, Oman 29%, Kuwait 25.4%, Bahrain 25.0% and United Arab Emirates 25.0% [2].

The application of data mining tools in health-care in general [8–11] and detection of diabetes risk factors in particular simplifies the understanding of huge biomedical datasets [12]. In health-care data mining can be used to better identify chronic diseases and extract risk factors [8, 9]. Several studies based on data mining tools have been carried out in order to extract information and interpret data to predict diabetes [13–15]. Tabak L. et al. tested six classifiers in the prediction of diabetes by comparing their performances in terms of sensitivity, specificity and total accuracy [13]. Mukesh et al. applied a bayesian network in order to classify 226 persons in three classes: Not diabetic, Pre-diabetic or Diabetic [14]. Azrar A. et al. established a comparative study of different data mining algorithms (Naive bays, KNN and Decision Tree) for early prediction of diabetes [15].

In this work a case-control epidemiological study is performed on a dataset of 4061 patients. The population of this study is characterized by cardiac patients living in the four Gulf countries (UAE, Oman, Kuwait, Bahrain) followed in their corresponding hospitals:

— A Principal Component Analysis is carried out on cleaned data.

— The different risk factors of T2DM are detected and analysed by data mining technic: Decision Tree.

2. Dataset

A case-control epidemiological study was performed on the "Gulf COAST" database. The dataset was collected for the Gulf COAST registry, it was retrieved between the years 2012 and 2013 and was obtained by surveys of patients admitted in this time period with Acute Coronary Syndrome (ACS). A total of 29 hospitals were involved in the four gulf countries (UAE, Bahrain, Kuwait and Qatar). The 4061 patients registered in the database, where signed consent was given to use their information for this database.

An accidental sampling is considered since patients were chosen as long as they were checked in their corresponding hospitals. The aim of our study is to predict risk factors of T2DM. Therefore, 3372 is the number of patients who fulfilled the following inclusion criteria:

— Non diabetic and T2DM patients.
— Patients with at least 90% of the input information.

2.1. Chosen variables

For our target patients, the following variables are studied:

— Sociodemographic: age, sex, socio-economic level, level of education and marital status.
— Clinical variables: weight, height and waist and body mass index (BMI).
— Biological variables: glycated hemoglobin (HbA1C), fasting blood glucose and lipid parameters.
— Hygiene and Dietary Measures (HDM) and physical activity.
— Hypertension: diastolic pressure, systolic pressure.

2.2. Characteristics of target population

The studied population is composed of 3372 patients living in Bahrain, Kuwait, Oman or UAE, from which 1907 are type 2 diabetics, ranging from the age of 18 to 99 years old, with an average age $= 60 \pm 12.41$.

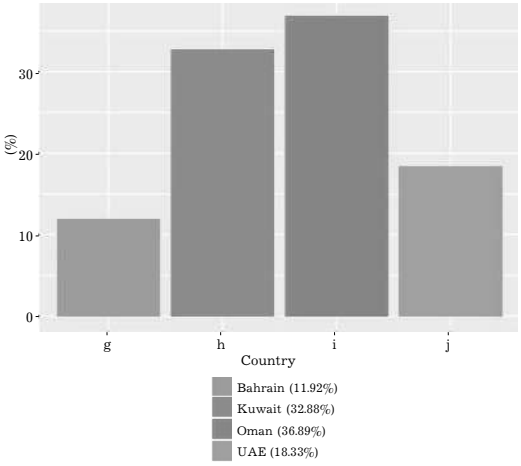

Fig. 1. Distribution of the population by countries.

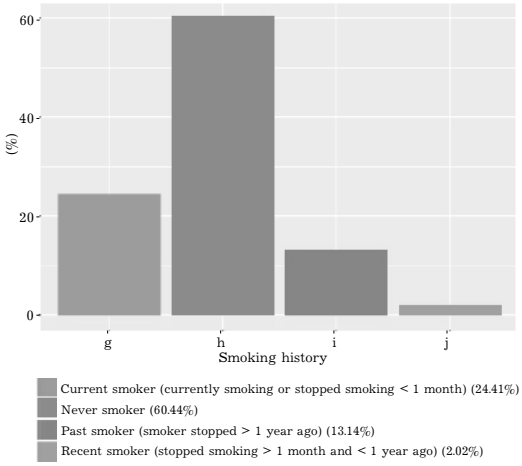

Fig. 2. Smoking habits in the studied population.

International Diabetes Federation, states in a review on sex differences in metabolic regulation and diabetes, that diabetes is more frequent in men than in women [16] which is confirmed in our studied population as diabetes is more common in men (66.4%) than in women (33.6%).

Distribution of our studied population by countries is illustrated in the following figure (see Fig. 1)

It is noticed that the percentage of tobacco consumption in the studied population is 39.57% with a huge gap between men and women (95.5% vs 4.5%). Men are more likely to have smoking history than women (see Fig.2).

3. Data preprocessing

Data cleaning is used to identify erroneous, missing and inaccurate data. Data pre-processing steps are mainly: filtering sparse data, checking content, handling missing values and grouping categories. Observations with important missing values are characterized by sparse data providing insignificant information in the analysis and modeling.

Our dataset, contains missing values in patient profiles. The table (see Table 1) below gives an overview that summarizes the initial state of the data:

Table 1. Initial data overview.

col	index	mod count	NAs	NAp
Country	1	4	0	0
Gender	2	2	0	0
Age	3	-1	0	0
Marital_Status	4	5	0	0
Work	5	3	0	0
Cardiac_Arrest_Admission	6	3	0	0
Non_Cardiac_Condition	7	2	0	0
Hypertension	8	2	0	0
Dyslipidemia	9	2	0	0
DM	10	2	0	0
Year_DM_Diagnosed	11	-1	2051	50.89
DM_Duration	12	-1	1970	48.88
DM_Type	13	1	1867	46.33
DM_Treatment	14	9	1868	46.35
Family_History_DM	15	8	1872	46.45
Smoking_History	16	4	1	0.02
Waist	17	-1	87	2.16
BMI	18	-1	55	1.36
Admission_Blood_Glucose_Value_SI_Units	19	-1	242	6
Fasting_Blood_Glucose_Value_SI_Units	20	-1	1276	31.66
HbA1C_Admission_Value	21	-1	1484	36.82
Lipid_24_Collected	22	2	2	0.05
Cholesterol_Value_SI_Units	23	-1	329	8.16
Triglycerides_Value_SI_Units	24	-1	369	9.16
LDL_Value_SI_Units	25	-1	830	20.6
HDL_Value_SI_Units	26	-1	802	19.9
Stress	27	2	2	0.05
Creatinine_Clearance	28	-1	64	1.59
Education	29	6	0	0
Sleep_Apnea	30	2	0	0
Heart_Rate	31	-1	1	0.02
Systolic_BP	32	-1	1	0.02
Diastolic_BP	33	-1	1	0.02

The inclusion criteria is applied in our recorded data in order to have reliable results.

The following table (see Table 2) shows the state of the database after applying the inclusion criteria.

The existence of conflicting values is noticed from observations. Consequently, they are removed from the analysis. The table (see Table 3) gives an idea of the observations with contradictory values. After checking values of our variables, data normalization is needed in order to respect each variable's standards.

After the extraction of final sample, processing missing data is needed.

Table 2. Database status after filtration.

col	index	mod count	NAs	NAp
Country	1	4	0	0
Gender	2	2	0	0
Age	3	-1	0	0
Marital_Status	4	5	0	0
Work	5	3	0	0
Cardiac_Arrest_Admission	6	3	0	0
Non_Cardiac_Condition	7	2	0	0
Hypertension	8	2	0	0
Dyslipidemia	9	2	0	0
DM	10	2	0	0
Year_DM_Diagnosed	11	44	154	4.57
DM_Duration	12	42	85	2.52
DM_Type	13	2	0	0
DM_Treatment	14	10	0	0
Family_History_DM	15	8	1464	43.42
Smoking_History	16	4	0	0
Waist	17	-1	34	1.01
BMI	18	-1	12	0.36
Admission_Blood_Glucose_Value_SI_Units	19	-1	116	3.44
Fasting_Blood_Glucose_Value_SI_Units	20	-1	862	25.56
HbA1C_Admission_Value	21	-1	973	28.86
Lipid_24_Collected	22	2	0	0
Cholesterol_Value_SI_Units	23	-1	5	0.15
Triglycerides_Value_SI_Units	24	-1	13	0.39
LDL_Value_SI_Units	25	-1	276	8.19
HDL_Value_SI_Units	26	-1	246	7.3
Stress	27	2	0	0
Creatinine_Clearance	28	-1	21	0.62
Education	29	6	0	0
Sleep_Apnea	30	2	0	0
Heart_Rate	31	-1	0	0
Systolic_BP	32	-1	0	0
Diastolic_BP	33	-1	0	0

Table 3. Observations with conflicting values.

DM	Year_DM_Diagnosed	DM_Type	DM_Treatment
No	2008.00	Type 2	Diet
No	1992.00	Type 2	
No	2009.00	Type 2	Diet, Oral Hypoglycemic drugs
No	1997.00	Type 2	Insulin
No	2007.00	Type 2	Diet, Insulin
No	2002.00	Type 2	Diet, Oral Hypoglycemic drugs
No	1992.00	Type 2	Diet, Oral Hypoglycemic drugs, Insulin
No	2009.00	Type 2	Diet
No	2006.00	Type 2	Diet, Oral Hypoglycemic drugs
No	2012.00	Type 2	Oral Hypoglycemic drugs
No	2007.00	Type 2	Oral Hypoglycemic drugs, Insulin
No	2002.00	Type 2	Oral Hypoglycemic drugs
No	2002.00	Type 2	Oral Hypoglycemic drugs
No	2007.00	Type 2	Oral Hypoglycemic drugs
No	2000.00	Type 2	Oral Hypoglycemic drugs
No	1992.00	Type 2	Diet, Oral Hypoglycemic drugs

Handling missing data remains a statistical issue and requires special approaches [17]. However, ignoring them can lead to significant bias in addition to a loss of precision in the analysis model.

From our observation, the distribution of missing data in our database is arbitrary, accordingly, K-Nearest-Neighbor (K-NN) algorithm is used as an imputation tool since it is considered as one of the appropriate algorithms for the nature of our data [18, 19].

In order to improve the quality of representation for variables with multi-value modalities a transformation into indicator variables is achieved. Finally, categorical variables encoding is needed for the preparation of our data for a Principal Component Analysis (PCA); each modality of a given variable takes a unique and very specific numerical value.

4. Data analysis and data mining

4.1. Principal Component Analysis

Principal Component Analysis (PCA) is a multidimensional method of data analysis that allows the comparison of relationships between observed variables on a large number of individuals. PCA summarizes the information by replacing the original variables by their linear combinations. For a better understanding of our large dataset PCA is carried out on cleaned data providing an understandable illustration of relationships between variables, between individuals and an individual-variable analysis. A reduced centered PCA is used since our variables are on different scales. The correlation matrix is given in table (see Table 4).

Table 4. Correlation matrix.

	DM	Y_diagnosis	Duration	Type	Admission_G	Fasting_G	HbA1C	Cholest	LDL	HDL	Systolic	Diastolic	HT	BMI
DM	1													
Y_diagnosis	0.933	1												
Duration	0.618	0.622	1											
Type	1.000	0.933	0.618	1										
Admission_G	0.518	0.468	0.335	0.518	1									
Fasting_G	0.516	0.488	0.370	0.516	0.544	1								
HbA1C	0.650	0.594	0.404	0.650	0.613	0.555	1							
Cholest	−0.128	−0.067	−0.044	−0.128	0.062	0.028	0.011	1						
LDL	−0.189	−0.124	−0.087	−0.189	−0.004	−0.035	−0.055	0.852	1					
HDL	−0.036	−0.041	−0.051	−0.036	0.018	−0.010	−0.037	0.182	0.043	1				
Systolic	0.041	0.029	0.012	0.041	0.026	0.005	0.012	0.079	0.048	0.052	1			
Diastolic	−0.031	−0.003	−0.028	−0.031	0.007	−0.033	−0.012	0.157	0.137	0.011	0.711	1		
HT	0.327	0.281	0.184	0.327	0.118	0.150	0.169	−0.130	−0.174	0.047	0.200	0.075	1	
BMI	0.115	0.071	0.048	0.115	.086	0.085	0.107	0.017	−0.031	0.028	0.061	0.056	0.122	1

Actually, when the correlations are high (> 0.8), one of the variables is removed from the analysis.

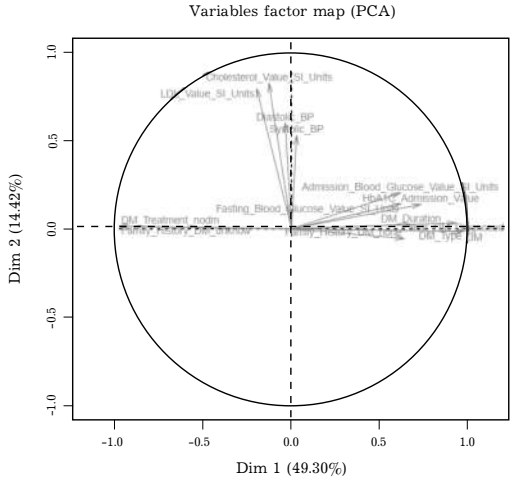

Fig. 3. Projection of 13 variables on first plane.

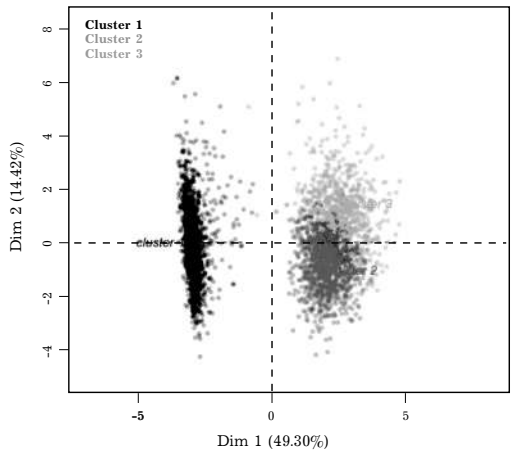

Fig. 4. Projection of individuals on first plan.

As shown in figure (see Fig. 3), the PCA gives a clear illustration of our variables grouped according to axis into three groups.

The first group is composed by the following variables:

DM	:	diabetes mellitus
$Admission\_blood\_Glucose$	:	rate of blood glucose on admission
$Hba1c$	:	rate of hba1c test
$DM\_duration$	:	duration of diabetes
$Family\_history$	:	family history of diabetes
$Fasting\_blood\_glucose$	:	rate of fasting blood glucose
$Year\_DM\_diagnosed$	:	year when diabetes was diagnosed

The second group includes the variables:

$DM\_treatment$	:	diabetes being treated
$Family\_history\_unknown$	:	absence of family history

The third group involves the following variables :

$Cholesterol - value$	:	rate of cholesterol in the blood
LDL_value	:	rate of LDL
$Diastolic$	:	diastolic blood pressure
$Systolic$	:	systolic blood pressure

Figure (see Fig. 4), shows that the studied population is divided into three groups (clusters), the first cluster marked by black dots represent non diabetic patients, the second cluster noted by red dots corresponds to first stage diabetics or pre-diabetic patients while the third cluster marked in green represents diabetic patients.

4.2. Data mining: Decision Tree

A comparative study by Abdar et al., of KNN, SVM, C5.0, Logistic Regression and Neural Network algorithm applied on a biomedical dataset, shows that Decision Tree gives the best results in term of specificity, sensitivity, precision and accuracy [20].

Our cleaned dataset contain both discrete and ordinal attributes. In consequence, the prediction is applied using Cart algorithm with output variable "DM" labeled with two classes: "No" refereeing to non diabetics and "YES" refereeing to diabetics [21].

A meta-analysis published in the Journal of the American College of Cardiology (JACC) in 2015 considered data of more than 4 million adults to find evidence of link between hypertension and diabetes. It concluded that people with high blood pressure have a higher risk of developing type 2 diabetes [22]. Figure (see Fig. 5) confirms that: The major risk factors are hypertension and dyslipidemia followed by less important ones as: high rate of triglyceride, advanced age and low rate of cholesterol-LDL. In fact, a patient with hypertension and dyslipidemia has a high risk of developing diabetes. Also, an hypertensive middle aged person may develop diabetes if its rate of cholesterol-LDL is

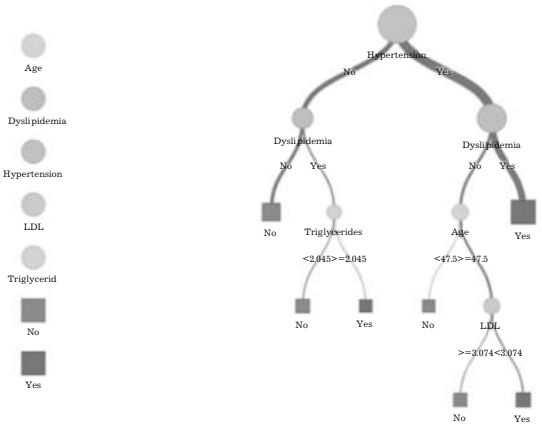

Fig. 5. Decision tree: Cart algorithm.

less than 3.07. However, a patient has reduced chances of developing diabetes if he doesn't have any associated problems with hypertension or dyslipidemia.

5. Conclusion

The objective of our proposed work is to determine the main risk factors of Type 2 Diabetes Mellitus (T2DM) in Gulf countries, from "Gulf COAST" dataset, by using Data mining (Decision Tree) and Data analysis (Principal Component Analysis) tools. In order to have a representative and cleaned sample, time and effort were spent in data preparation by: transforming the data and cleaning up types and values of our variables, treating anomalies and replacing missing values.

A Principal Component Analysis was carried out in order to analyze our variables and class our individuals (patients). It allowed to divide our population into three distinctive groups namely: non diabetic persons, pre-diabetics and type 2 diabetic persons.

Our results show pragmatically how to avoid T2DM. Decision Tree algorithm predicted the major risk factors of T2DM. Therefore, acting on permanent controls, dietary habits and physical activity can help avoiding the progression towards T2DM.

Acknowledgments

We gratefully thank the principal investigator of Gulf COAST **Professor Mohammad Zubaid** for providing the **Gulf COAST dataset**.

In addition, we would like to express our sincere gratitude to "The Mohammed Bin Rashed University" and "Gulf COAST registry" for giving the permission to use **Gulf COAST dataset**.

[1] Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. World Health Organization (1999).

[2] Meo S. A., Usmani A. M., Qalbani E. Prevalence of type 2 diabetes in the Arab world: impact of GDP and energy consumption. Eur. Rev. Med. Pharmacol. Sci. **21** (6), 1303–1312 (2017).

[3] Boutayeb W., Lamlili M., Boutayeb A., Derouich M. Mathematical modelling and simulation of β-cell mass, insulin and glucose dynamics: Effect of genetic predisposition to diabetes. Journal of Biomedical Science and Engineering. **7** (6), 330–342 (2014).

[4] Elhayany A., Lustman A., Abel R., Attal-Singer J., Vinker S. A low carbohydrate Mediterranean diet improves cardiovascular risk factors and diabetes control among overweight patients with type 2 diabetes mellitus: a 1-year prospective randomized intervention study. Diabetes, Obesity and Metabolism. **12** (3), 204–209 (2010).

[5] Pulgaron E. R., Delamater A. M. Obesity and type 2 diabetes in children: epidemiology and treatment. Current Diabetes Reports. **14** (8), Article number: 508 (2014).

[6] Donaghue K. C., Chiarelli F., Trotta D., Allgrove J., Dahl-Jorgensen Knut. Microvascular and macrovascular complications. Pediatric Diabetes. **8**, 163–170 (2007).

[7] De Luis D., Fernandez N., Arranz M., Aller R., Izaola O., Romero E. Total homocysteine levels relation with chronic complications of diabetes, body composition, and other cardiovascular risk factors in a population of patients with diabetes mellitus type 2. Journal of Diabetes and its Complications. **19** (1), 42–46 (2005).

[8] Lei-Da C., Toru S., Frolick M. N. Data mining methods, applications, and tools. Information systems management. **17** (1), 65–70 (2000).

[9] Koh H. C., Tan G., and others. Data mining applications in healthcare. Journal of healthcare information management. **19** (2), 64–72 (2011).

[10] Parvez A., Saqib Q., Syed R., Afser Q. Techniques of data mining in healthcare: a review. International Journal of Computer Applications. **120** (15), 38–50 (2015).

[11] Jothi N., Rashid Nur'Aini Abdul, Husain W. Data mining in healthcare–a review. Procedia computer science. **72**, 306–313 (2015).

[12] Yoo I., Alafaireet P., Marinov M., Pena-Hernandez K., Gopidi R., Chang J.-F., Hua L. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems. **36** (4), 2431–2448 (2012).

[13] Tapak L., Mahjub H., Hamidi O., Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. Healthcare informatics research. **19** (3), 177–185 (2013).

[14] Mukesh K., Rajan V., Anshul A. Prediction of Diabetes Using Bayesian Network. International Journal of Computer Science and Information Technologies. **5** (4), 5174–5178 (2014).

[15] Azrar A., Ali Y., Awaisl M., Zaheer K. Data mining models comparison for diabetes prediction. Int. J. Adv. Comput. Sci. Appl. **9** (8), 320–323 (2018).

[16] Tramunt B., Smati S., Grandgeorge N., Lenfant F., Arnal J.-F., Montagner A., Gourdy P. Sex differences in metabolic regulation and diabetes susceptibility. Diabetologia. **63** (3), 453–461 (2020).

[17] Graham J., Cumsille P. E., Shevock A. E. Methods for handling missing data. Handbook of Psychology, Second Edition & Computer Engineering. Vol. 2 (2012).

[18] Aljuaid T., Sasi S. Proper imputation techniques for missing values in datasets. IEEE: 2016 International Conference on Data Science and Engineering (ICDSE). 1–5 (2016).

[19] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. B. Missing value estimation methods for DNA microarrays. Bioinformatics. **17** (6), 520–525 (2001).

[20] Abdar M., Kalhori N., Sutikno T., Subroto I. M. I., Arji G. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical & Computer Engineering. **5** (6), 1569–1576 (2015).

[21] Lavanya D., Rani K. U. Performance evaluation of decision tree classifiers on medical datasets. International Journal of Computer Applications. **26** (4), 1–4 (2011).

[22] Emdin C. A., Anderson S. G., Woodward M., Rahimi K. Usual Blood Pressure and Risk of New-Onset Diabetes: Evidence From 4.1 Million Adults and a Meta-Analysis of Prospective Studies. Journal of the American College of Cardiology. **66** (14), 1552–1562 (2015).

Використання методу добування даних для прогнозування факторів ризику цукрового діабету другого типу в країнах Перської затоки

Бутаєб В.[1], Бадауї М.[2], Аль Алі Х.[3], Бутаєб А.[1], Ламлілі М.[1]

[1] *Університет Мохаммеда Першого, Факультет наук, Уджда, Марокко*
[2] *Університет Мохаммеда Першого, Вища школа технологій, Уджда, Марокко*
[3] *Еміратський авіаційний університет, Дубай, Об'єднані Арабські Емірати*

Поширеність діабету в країнах Перської затоки значно зростає через різні фактори ризику, такі як: ожиріння, нездорове харчування, фізична бездіяльність та куріння. Метою цього дослідження є використання засобів добування даних та інтелектуального аналізу даних для визначення різних факторів ризику розвитку цукрового діабету другого типу (ЦД2) у країнах Перської затоки на основі бази даних Gulf COAST.

Ключові слова: *база даних Gulf COAST, добування даних, дерево прийняття рішень, аналіз основних компонентів, діабет 2-го типу.*