International Journal of Combinatorial Optimization Problems and Informatics, 14(1), Jan-April 2023, 66-75. ISSN: 2007-1558.

www.editada.org

Detection of Risk Factors for Diabetes Mellitus with Machine Learning

Mireya Tovar Vidal[1], Juan Manuel Cancino Gordillo[1]*

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, México.
mireya.tovar@correo.buap.mx, juan.cancinogordillo@viep.com.mx

Abstract. One of the most important diseases worldwide in public health is Diabetes Mellitus (DM) since this is one of the most severe and frequent non-communicable diseases with various chronic complications. In this paper, we propose a procedure to detect the most common risk factors in patients suffering from the disease known as diabetes mellitus, through principal components analysis (PCA) and non-negative matrix factorization (NMF). We then check the results using these factors like features, through the machine learning algorithms, improving the classification results. According to the experimental results, accuracy of more than 80% was obtained.

Keywords: machine learning, PCA, NMF, classification.

Article Info
Received Oct 07, 2022
Accepted Jan 20, 2023

1 Introduction

In the medical field, diagnosis is the most critical part when treating a person, since the physician uses his knowledge to detect certain patterns in the behavior or medical studies of a patient and reach a conclusion that translates into treatment or medication. Diabetes Diabetes Mellitus (DM) is a serious, chronic illness that occurs when glucose levels in the blood exceed certain limits [1]. There are two common variants of DM, type I, and type II. In particular, type II Diabetes Mellitus (T2DM) is associated with insulin resistance (insulin secretory abnormalities), i.e., where cells respond poorly to insulin, affecting their glucose intake [2]. These abnormalities are caused by inflammation and metabolic stress, among other genetic factors [2].

Globally, DM is one of the ten leading causes of death. According to [3] since 2000, the population diagnosed with diabetes worldwide increased from 4.6% to 9.3% in 2020. If this health problem is not addressed, it is predicted that in the year 2030, 10.2% of the world's population will have diabetes and this will increase to 10.9% (700 million people) in the year 2045.

In Mexico, the death rate per DM for 2020 was of 11.95% per 10,000 inhabitants and it has been increasing since 2018. The disease occurs in all age groups, but it increases with age, affecting people over the age of 65 [4].

To detect a person with this disease, surveys known as risk factors are conducted, which are a series of questions about their daily activities and certain anthropometric measurements, based on this survey an expert can diagnose a person, given the results of this and if you are at risk of developing the disease of diabetes also known as prediabetic.

The aim of the work was to perform an analysis of the parameters used by physicians to detect T2DM, this is achieved performing data cleaning, dimension reduction, and build a mathematical model in which will let us know the relevant attributes of a dataset. In the end these translate into a study to determine whether a person has this disease. Subsequently, create an efficient classification model based on attributes obtained from the analysis to find the most representative risk factors for T2DM.

The rest of the paper is organized as follows. Section 2 presents related work on the topic of classifiers using different methods for file cleaning and the results obtained by implementing the classification algorithms.

Section 3 includes a brief explanation of the methodology, algorithms, and metrics used. Section 4 presents the experimental results and finally, includes the conclusions of the work performed.

2 Related Work

Some work related to the use of machine learning algorithms is described below.

In 2017, Yamilé et al. [5] conducted a cross-sectional study with a randomized sample design to detect the prevalence of chronic non-communicable diseases and their risk factors. The authors used a total of 2085 records of people from 14 municipalities of different ages (32-56 years) using variables such as sex, age, abdominal perimeter, glucose, insulin, triglycerides, cholesterol, among others. Using the mean and standard deviation to generalize the attributes, they presented the tables to several experts in the field to diagnose each record. Concluding that at an older age (approximately ≥ 50) hormonal and metabolic changes occur that affect several systems. Consequently, glucose intolerance, T2DM, and abdominal obesity develop.

In 2018, Orlando A. Chan et al. [6] presented an investigation conducted on a dataset of 768 patients, where all the records are based on women for the detection of gestational diabetes, mentioning that these attributes in the dataset are of high importance for the detection of T2DM. Some of the variables considered are glucose, insulin, blood pressure, and age. The authors' final objective was to create an expert system to detect diabetes, from the selected attributes of the dataset, using the classification algorithms provided by the WEKA and BigML tools. The results achieved an accuracy of 70% in the classification of patients who did not present T2DM and 63% for positive detection, achieving 73.83% accuracy, using decision trees as the classification algorithm.

Decision trees are used in [7] with a female-focused dataset frequently used for diabetes screening. This work is presented in two stages; one of them consists of improving the data with a preprocessing of the data, applying methods such as garbage in, garbage out; generally used in data mining projects, helping to eliminate instances of the dataset. By having less irrelevant information, a more accurate prediction model is generated, which managed to increase accuracy to 78.17%. Demonstrating that data preprocessing improves the classification of instances in the dataset.

In 2015, the authors of [8] presented a comparison between different data mining algorithms using a dataset called Pima Indians Diabetes Dataset. It is composed of 768 records with eight attributes (age, insulin, pressure, among others) and a field to be classified (positive and negative). The algorithms used in the work are the J48, Naïve Bayes, and RBF (Radial Basis Function) decision trees, which use three evaluation metrics (accuracy, recall, and F_1) to measure the classification results. In the study, the highest accuracy was that of the J48 algorithm, with an average of 77.1%, but with a lower instance classification than the Naïve Bayes algorithm. They list the attributes of the dataset but do not present any type of data processing. The data set was split for training and evaluation, with approximately 230 records for evaluation.

The authors of [9] proposed several algorithms used in the data mining branch, such as SMO (Sequential Minimal Optimization), random forest, treeJ48, and Naïve Bayes to compare the performance of classification algorithms and to be able to determine which algorithm has the accuracy to perform diagnosis. The authors follow the guidelines of cleaning the data not required for the study, such as interpreting the missing data in the dataset, which does not consist of a single location according to the description they provide. For the evaluation of the algorithms, they use the method known as cross-validation along with the metrics of precision, accuracy, and F_1, where they divide their dataset with a 50:50 ratio. The authors mention that the ratio they use for their dataset is not ideal, since for this type of evaluation it is better to section the dataset into three parts. In the results and conclusions of the paper, they mention the results of the accuracy of the J48 algorithm, reaching at best 73.82% accuracy when classifying.

In 2022, the authors of [10] presented a framework for a model of prediction over clinical diagnosis of DM. They used different algorithms of classification and deep neural network. They used PIMA Indian and the laboratory of the Medical City Hospital (LMCH) diabetes datasets. Achieving encouraging results with the neural network.

Diabetic foot ulcer is other severe complication of the development of diabetes mellitus. Authors [11] used machine-learning techniques to identify risk factors associated with diabetic foot ulcers. Support Vector Machines, Naïve Bayes, K-nearest neighbor, random forest among other were used for constructing prediction models.

In this paper, we will perform missing data processing on three datasets to improve the experimental results of T2DM disease classification. After preprocessing the data, non-relevant attributes are removed from the dataset using Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF). These new datasets are the input to classification algorithms, such as *TreeJ48*, Naïve Bayes, among others. Finally, a comparison, of the classification algorithms output is performed to demonstrate that data preprocessing and the use of attribute reduction with PCA and NMF is feasible without loss of classification accuracy.

3 Methodology

In this work, we have implemented a method based on four stages: data analysis, pre-processing, classification, and evaluation.

3.1 Dataset Description

In this work, different datasets were used. These have different sources.

One dataset used is known as the Pima Indians Diabetes Database. This is a resource of 768 women with the attributes of the number of pregnancies, age, weight, BMI (body mass index), blood pressure, and a field reporting the outcome to the prognosis of DM [12].

The second dataset comes from research conducted by Chen et al., [13]. This one presents us with a wide range of structured anthropometric data of citizens residing in China from the years 2010 to 2016, with the purpose of highlighting a relationship between body mass index (BMI) and T2DM with the age of the patients. This dataset presents several fields that provide insight into whether the record (patient) has a relative who suffered from diabetes, the diagnosis of T2DM disease, anthropometric values, and certain medical studies. Some fields are missing or null, such as HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein), and glucose measurements, among others.

Finally, the third dataset is comprised of records provided by the health sector located in Carmen Xhan, near the Mexico-Guatemala border, which includes diabetic data, with attributes such as age, BMI, blood pressure, HDL cholesterol, LDL cholesterol, and glucose indices [14]. These are complemented by the medical records of a surgeon who treats several people with and without T2DM disease in the city of Chiapas, Mexico. This is given that the original dataset for the health sector was not sufficient to perform an acceptable binary classification.

3.2 Pre-processing

Data pre-processing, in general, consists of data substitution, elimination, or grouping before applying any classification algorithm. Proper preprocessing is of great importance because the information within the sets is flexibly controlled, leading to capture errors, abnormal results, or missing data, which is a problem to be solved. To solve this problem, a more common proposal is to eliminate records that are not complete, but this causes the omission of a lot of data and will not provide a good classification. Another proposal is the replacement by the average, which works by analyzing the existing complete data and using the average of each attribute to complete the records and thus having a more complete knowledge base, which results in better pre-processing.

Subsequently, attribute reduction PCA and NMF algorithms are applied to the data resulting from the pre-processing. The purpose of term reduction is to identify the most relevant attributes of the dataset, which are referred to as risk factors in the dataset. The eliminated data were not relevant and were not related to any other attribute.

To carry out this process the language called *Python* is used, which offers several tools to manipulate large amounts of data efficiently (like pandas or numpy) to conclude in a comma-separated file (*CSV*) where the new data will be saved.

3.3 Classification algorithms

The objective of the implementation of the classification algorithms is to discover which attributes are responsible for performing the classification of the data, compare these attributes with the results of PCA and NMF, and determine a list of main attributes of the risk factors.

Some of the most used classification algorithms are SVM (Support Vector Machine) [15,24], Random Forest [16], J48 Decision Trees [17,24], and Naïve Bayes [18]. Once the most relevant risk factors are obtained from the dataset, the dataset will be re-classified to see if there is an improvement in its classification with fewer attributes.

The algorithm known as J48 Decision Trees aims to create a model that predicts the value of a variable using inductive learning from observations and logical constructions by learning if-then-if-not decision rules inferred from data features, it is generally composed of two stages.

In the first stage, the tree is built from a training data set where each internal node is composed of a test attribute and the portion of the training set present in the node is divided according to the values of the attribute. When there are more objects of one class in a node, an internal node is generated; when it contains objects of a class, a leave with the class assignment is generated. In the second stage, each new object is classified going from the root node to a leaf, generating a path based on the decisions of the internal nodes

The Naïve Bayes classifier is a simple but powerful learning algorithm, since it is based on conditional probability and Bayes' theorem. The Conditional probability can be defined has the possibility of an event, which we call A, occurring as a consequence of another event, which we call B, having taken place.

The implementation of the Random Forest classifier solves one of the problems caused by classification trees known as overfitting, which occurs when a flexible model begins to memorize the training data, losing the variance of the data set. This method combines hundreds or thousands of decision trees, training each of them with a slightly different data set by dividing the nodes of each tree, limiting the number of features available, the number of records available.

Support machine vectors is a classification and regression algorithm developed in the 90s, within the field of computational science. It was originally used for binary classification, but its application has been extended to multiple classification. This algorithm is based on the maximum classification margin that is based on the concept of hyper-plane at the same time.

3.4 Classification algorithms

The object of the evaluation is to measure the efficiency of the model created on new records. This efficiency is measured in percentages that may vary depending on the dataset used.

In the evaluation, it is necessary to consider the confusion matrix generated by the classification algorithms. It compares the prediction of the classes with the labeled results. Four metrics are used [19]:

- Precision: The number of relevant cases retrieved divided by the number of cases retrieved (see Equation 1).
- Recall: Expresses the proportion of relevant cases recovered, compared to the total number of existing relevant cases, regardless of whether they are recovered or not. (See Equation 2).
- Accuracy: Measures the percentage of cases that the model got right or classified correctly (see Equation 3).
- F_1: Used to combine the precision and recall measures into a single value (see Equation 4).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{\mathrm{TP} + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Accuracy}{Precision + Accuracy} \tag{4}$$

where:

- True positives (TP): Positive results correctly classified.
- True negatives (TN): Negative results correctly classified.
- False positives (FP): Negative results classified as positive.
- False negatives (FN): Positive results classified as negative.

To determine the attributes known as risk factors, it is necessary to eliminate the attributes that are not relevant for classification, it is recommended to use term reduction methods as they provide attribute scoring using probability and statistics [20]. This help to improve the results of the evaluation metrics in a dataset with fewer attributes, the attributes present in these are interpreted as the risk factors.

4 Experimental Results

This section will briefly present the dataset used to show the treatment that was carried out and how the results presented were reached.

All the data used in this work is in a GitHub (*https://github.com/ViepJuan/Article-Data*), in this repository are the datasets and more importantly the programs and logic used to fill the datasets with the media of each column.

4.1 Datasets and Result Pre-processing

In the first dataset, PIDD, all patients were female, at least 21 years of age, and of Indian descent, with a total of 500 records for non-diabetic patients and 268 from diabetic patients [12]. Table 1 describes the dataset which includes eight attributes and one outcome indicating whether the patient has diabetes (class 1) or not (class 0).

Table 1. Description of the PIDD dataset [12]

Attribute	Description	*Min-Max*
Pregnancies	Number of pregnancies	0-17
Glucose	Plasma glucose concentration for two hours in an oral glucose tolerance test.	0-199
Blood pressure	Blood pressure diastolic (mm Hg)	0-122
Skin thickness	Triceps skinfold thickness (mm)	0-99
Insulin	2-hour serum insulin (mu U/ml)	0-846
BMI	Body Mass Index (Kg/m^2)	0-67.1
PedigreeFunction	Diabetes Family Tree Function	0.08-2.42
Age	Age in years	21-81
Result	Patient outcome to diabetes disease	0-1

After pre-processing the dataset, 763 fields with invalid values were detected distributed among the different attributes. The attribute with the least missing data was glucose, while with insulin almost half of the data was missing. Subsequently, the PCA and NMF attribute reduction algorithms were applied independently to the pre-processed dataset. Table 2 presents the list of attributes retained after attribute removal by these algorithms.

Table 2. Reduction of attributes of the PIDD dataset

Attribute	Pre-processing	*Pre-processing + PCA*	*Pre-processing + NMF*
Pregnancies	√	√	
Glucose	√	√	√
Blood pressure	√		√
Skin thickness	√		√
Insulin	√	√	√
BMI	√	√	√
PedigreeFunction	√	√	
Age	√	√	

The second dataset attempts to identify a relationship between age, body mass index (BMI), and diabetes in general in the Chinese population [13]. The dataset comes from different cities in China (Shanghai, Beijing, Nanjing, Suzhou, among others) with patients older than 20 years, with a total of 4,174 diabetics out of 211,835 registered patients. Table 3 describes the attributes of this second dataset.

Table 3. Description of the China dataset [13]

Attribute	Description	*Min-Max*
Age	Age in years	20-99
BMI	Body Mass Index (Kg/m^2)	0-846
WC	Waist circumference in centimeters	13-116.6
Systolic pressure	Systolic pressure (mm)	59-222
Diastolic pressure	Diastolic pressure (Hg)	38-164
HbA1c	Glycosylated hemoglobin test	0.59-6.99
Total cholesterol	Lipoprotein test	0.02-17.84
LDL cholesterol	Low-density lipoproteins	0-10.07
HDL cholesterol	High-density lipoproteins	0-10.4
Triglyceride	Triglyceride test	0-32.64
Result	Patient outcome to diabetes disease	0-1

Table 4. Reduction of attributes of the China dataset

Attribute	Pre-processing	*Pre-processing + PCA*	*Pre-processing + NMF*
Age	√	√	√
BMI	√	√	√
WC	√		√
Systolic pressure	√		√
Diastolic pressure	√	√	√
HbA1c	√	√	√
Total cholesterol	√	√	√
LDL cholesterol	√	√	
HDL cholesterol	√	√	
Triglyceride	√	√	

The second dataset was also pre-processed, identifying 212,546 invalid fields spread across the different attributes. After the application of the attribute reduction algorithms, it is identified that the attributes to be eliminated by PCA are WC and systolic pressure. On the other hand, the attributes to be eliminated by NMF are LDL cholesterol, HDL cholesterol, and triglycerides. Table 4 shows the list of attributes that are preserved after attribute removal by these algorithms.

The third dataset was provided by the health sector located in Carmen Xhan, on the border between Mexico and Guatemala. The dataset was supplemented with information from a surgeon in the state of Chiapas, increasing the number of records to a total of 623 Mexican patients over 21 years of age, where only 87 people have T2DM and 536 do not. Table 5 describes the attributes of this third dataset.

Table 5. Description of the Mexico dataset

Attribute	Description	*Min-Max*
Age	Age in years	21-81
Weight	Patient's weight in kilograms	33.5-120
BMI	Body Mass Index (Kg/m^2)	15.24-41.5
Systolic pressure	Systolic pressure (mm)	70-180
Diastolic pressure	Diastolic pressure (Hg)	40-113
Total cholesterol	Lipoprotein test	0-299
LDL cholesterol	Low-density lipoproteins	0-133
HDL cholesterol	High-density lipoproteins	0-51
Triglyceride	Triglyceride test	0-1982
Result	Patient outcome to diabetes disease	0-1

The third dataset, after pre-processing, had the least amount of missing data, with a maximum of 678 missing data, with the highest concentration of these data in HDL cholesterol and the least amount in the total cholesterol test. Table 6 shows the list of attributes retained after attribute removal by these algorithms.

Table 6. Reduction of attributes of the Mexico dataset

Attribute	Pre-processing	*Pre-processing + PCA*	*Pre-processing + NMF*
Age	√	√	√
Weight	√	√	√
BMI	√	√	√
Systolic pressure	√	√	
Diastolic pressure	√	√	√
Total cholesterol	√	√	√
LDL cholesterol	√		√
HDL cholesterol	√		
Triglyceride	√		

4.2 Classification Results

To perform the classification of the datasets, WEKA tool is used, as it provides us with a wide variety of algorithms used in the data mining environment, which use classification algorithms [22]. Training set (80%), test set (20%) and tenfold cross-validation. The datasets were split into: training set (80%), test set (20%) and we used a tenfold cross-validation for the evaluation.

Using the first dataset, PIDD, several tests were performed with the TreeJ48, *Naïve Bayes*, and *SMO* classifiers. However, the best results were obtained with the *TreeJ48* classifier; they are reported in Table 7.

Table 7. Experimental results of the TreeJ48 algorithm with the PIDD dataset. The best results are highlighted in bold. Not Reported (N/R)

Data treatment	Author	*Precision* %	*Recall* %	*Accuracy* %	F_1 %
All attributes	[6]	N/R	N/R	66.50	N/R
Cleaning of missing data	[7]	N/R	N/R	78.17	N/R
N/R	[9]	72.15	58.96	77.14	64.89
N/R	[23]	N/R	N/R	78.50	N/R
N/R	[8]	78.60	76.50	N/R	77.10
Pre-processing	This work	82.40	78.70	86.70	80.5
Pre-processing + PCA	**This work**	**86.5**	**85.00**	**87.20**	**80.9**
Pre-processing + NMF	This work	86.1	86.2	86.00	86.1

In the case of the second dataset, China, the *TreeJ48*, *Naïve Bayes*, *SMO*, and *Random Forest* the classifiers were used. The best results were obtained with the Random Forest classifier. Table 8 shows the experimental results, as can be seen, the best results are obtained with the pre-processed data and with the PCA algorithm.

Table 8. Experimental results of the Random Forest algorithm with the China dataset. The best results are highlighted in bold

Data treatment	*Precision* %	*Recall* %	*Accuracy* %	F_1 %
All attributes	97.20	98.00	98.00	97.20
Pre-processing	98.60	98.70	98.70	98.50
Pre-processing + PCA	**98.60**	**98.70**	**98.70**	**98.50**
Pre-processing + NMF	97.20	98.00	98.00	97.30

The *TreeJ48*, *Naïve Bayes*, *SMO*, and *Random Forest* classifiers were also applied to the third dataset. The best results were also obtained with the Random Forest classifier. Table 9 shows the experimental results, as can be seen, the best results are obtained only with the pre-processed data. However, it can also be seen that in second place are the results of the pre-processed data and with the NMF algorithm.

Table 9. Experimental results of the Random Forest algorithm with the Mexico dataset. The best results are highlighted in bold

Data treatment	*Precision* %	*Recall* %	*Accuracy* %	F_1 %
All attributes	90.70	91.30	91.10	90.50
Pre-processing	**94.50**	**94.40**	**94.30**	**93.90**
Pre-processing + PCA	93.90	93.90	93.90	93.30
Pre-processing + NMF	94.20	94.20	94.20	93.70

Finally, Table 10 shows a compilation of the best results for each dataset using the Accuracy and F_1 metrics, including the attributes used to achieve the results presented and the classification algorithm. The attributes presented in Table 10 are considered relevant risk factors because, with these attributes, robust classification is achieved with results above 80% classification accuracy.

Table 10. The best result for each dataset

Dataset	*Accuracy %*	*F_1 %*	*Used attributes*	*Classification algorithm*
First dataset (PIDD)	87.20	80.90	Pregnancies, glucose, insulin, BMI, PedigreeFunction, age	TreeJ48
Second dataset (China)	98.70	98.50	Age, BMI, diastolic pressure, HbA1c, total cholesterol, LDL cholesterol, HDL cholesterol, triglyceride	Random Forest
Third dataset (Mexico)	94.20	93.70	Age, weight, BMI, diastolic pressure, total cholesterol, LDL cholesterol	Random Forest

5 Conclusions

Early diagnosis of chronic diseases is one of the problems in the healthcare field that has been solved with the help of machine learning, providing the physician with an accurate and verifiable second opinion. In this work, we collected information from different sources to compare various algorithms that have been used for the task of T2DM classification based on different datasets from different countries, for the correct identification of patients suffering from this disease. With the intention of finding the risk factors by reducing the number of non-relevant attributes using term reduction algorithms.

Based on the experimental results obtained with each dataset, the attributes in common among the three sets are BMI and age, which are attributes that represent a person's sedentary lifestyle. Dimension reduction is one of the most powerful tools in data mining, as it allows data to be extracted without losing information, as the results of this work show. Due to this data reduction, together with correlation analysis between the attributes, we can discard the non-relevant information from the datasets and thus focuses on the relevant information, which in this case are the risk factors that determine that a patient suffers from T2DM.

Improvements are not only found in data processing but also apply to classification algorithms, where more advanced methods, such as neural networks, can increase the accuracy of correctly classified records, as some rely on established rules of the dataset getting conclusions from the results.

References

1. AD Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2020. Diabetes Care (2019). https:// doi. org/10.2337/dc20-S002
2. International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation (2019)
3. ID Federation. IDF DIABETES ATLAS 9th Edition 2019. Accessed: Oct. 6, 2022. [Online]. Available: https://diabetesatlas.org/en/
4. INEGI. Estadística a propósito del día mundial de la Diabetes, 2021. [Online]. Available: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/EAP_Diabetes2021.pdf
5. Yamilé *et al.* "Prevalencia de enfermedades crónicas no transmisibles y factores de riesgo en adultos mayores de Holguín", *Rev. Finlay*, vol. 7, num. 3, pp. 155--167 (2017).
6. Chan O., Peña J., Vianne J., and Zapata M.: Construcción de un modelo de predicción para apoyo al diagnóstico de diabetes (Construction of a Prediction Model to Support the Diabetes Diagnosis). *Pist. Educ.*, vol. 40, num. 130, pp. 2105--2122 (2018).
7. AlJarullah A. A. Decision tree discovery for the diagnosis of type II diabetes. *2011 Int. Conf. Innov. Inf. Technol. IIT 2011*, pp. 303–307 (2011). doi: 10.1109/INNOVATIONS.2011.5893838.
8. Sa'di S., Maleki A., Hashemi R., Panbechi Z., and Chalabi K. Comparison of Data Mining Algorithms in the Diagnosis of Type Ii Diabetes. *Int. J. Comput. Sci. Appl.*, vol. 5, num. 5, pp. 1--12 (2015). doi: 10.5121/ijcsa.2015.5501.

9. Hemant P. and Pushpavathi T. A novel approach to predict diabetes by Cascading Clustering and Classification", *2012 3rd Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2012*, (2012). doi: 10.1109/ICCCNT.2012.6396069.
10. Chollette C. O., Lyndon S., Melvyin S.: Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine. Vol. 220 (2022). doi: 10.1016/j.cmpb.2022.106773
11. Nanda R., Nath A., Patel S., Mohapatra E. Machine learning algorithm to evaluate risk factors of diabetic foot ulcer and its severity. Medical & Biological Engineering & Computing. Vol. 60, pp. 2349-2357 (2022). doi: 10.1007/s11517-022-02617-w
12. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care. Pp. 261--265. IEEE Computer Society Press (1988). https://www.kaggle.com/uciml/pima-indians-diabetes-database.
13. Chen Y., *et al.* Association of body mass index and age with incident diabetes in Chinese adults: A population based cohort study. BMJ Open. Vol. 8, num. 9. Pp.1--3 (2018). doi: 10.1136/bmjopen-2018-021768.
14. Secretaría de Salud, Carmen Xhan, Chiapas, México. Expedientes clínicos, (2020). https://saludchiapas.gob.mx/unidades-medicas/CSSSA007301
15. Amat R. J. (2021) "Máquinas de Vector Soporte (SVM) con Python", *cienciadedatos.net*. https://www.cienciadedatos.net/documentos/py24-svm-python.html
16. Amat R. J. (2020). "Random Forest con Python", *cienciadedatos.net*. https://www.cienciadedatos.net/documentos/py08_random_forest_python.html. Accessed 18 February 2021
17. Amat R. J. (2020) "Árboles de decisión con Python: regresión y clasificación", *cienciadedatos.net*. https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html. Accessed 18 February 2021
18. Pedregosa F. *et al.* Scikit-learn: Machine Learning in Python", *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830 (2011).
19. Heras J. M. "Precision, Recall, F1, Accuracy en clasificación" (2020). *IArtificial.net*. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/ Accessed 1 October 2021
20. Garrett G. and Wickham H. "R for Data Science". O'Reilly Media, Inc. https://www.oreilly.com/library/view/r-fordata/9781491910382/ (2016).
21. U. Machine Learning, "Pima Indians Diabetes Database", *Predict the onset of diabetes based on diagnostic measures*, 2016. https://www.kaggle.com/uciml/pima-indians-diabetes-database.
22. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and I. H. (2009) "The WEKA Data Mining Software: An Update", *SIGKDD Explor.*, 11(1):10–18.
23. Vijayan V. V. and Anjali C. (2015). Prediction and diagnosis of diabetes mellitus - A machine learning approach. *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. pp. 122-127, doi: 10.1109/RAICS.2015.7488400.
24. Fregoso-Aparicio, L., Noguez, J., Montesinos, L. et al. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr 13, 148 (2021). https://doi.org/10.1186/s13098-021-00767-9