Prediction and Detection of Diabetes using Machine Learning

Abstract
Data mining and machine learning have become a vital part of different disease detection and prevention. One of them is diabetes. The purpose of this paper is to evaluate data mining methods and their performances that can be used for analyzing the collected data about the diabetes. We identified the most appropriate data mining methods to analyze the data by comparing them theoretically and practically. Some attributes of this dataset are: Age, Body Mass Index, Insulin, Glucose, etc. Methods are applied on these data to determine their effectiveness in analyzing and preventing diabetes. Evaluations on the data showed that the method with a higher performance is "Decision Tree". This was achieved by some performance measures, such as the number of instances correctly classified, accuracy, precision, recall and F-measure, that has brought better results compared to other methods. We come to the conclusion that the data mining methods and machine learning contribute to the predictions on the possibility of occurrence of the diabetes.

Keywords 1
Machine Learning, Prediction, Diabetes Disease, Data Mining

1. Introduction

Diabetes is a disease that is increasingly affecting the world even the most developed countries. Diabetes by the nature of its development as a globally problematic disease requires maximum commitment from medical staff, patients, family and society. Diabetes is a disease with high social, health and economic costs [1]. Diabetes is a chronic disease characterized by an increase in glucose or blood sugar levels because the body cannot produce insulin or its production is insufficient, or insulin is not able to act on the cells of the organism [2, 3]. Medics still do not know exactly why such a thing is happening and they have called the cause: x syndrome. Historically diabetes treatment has been done by fighting the symptoms and not the cause. According to the World Health Organization, Diabetes affects about 5% of the world's population and the number of patients is constantly increasing [1]. In developed countries, diabetes and the largest number of diabetics are found in people over 65 years of age. Whereas in developing countries where our country is part of the largest number of diabetics is found in the age of 45-64 years, but in recent years type 2 diabetes is more commonly encountered also in the age of 30-40 years [1]. The availability of historical data naturally leads to the application of data mining techniques for pattern discovery. The goal is to find rules that help understand diabetes and make it easier to diagnose it sooner. Prevention of diabetes is of great interest in the field of medicine. The use of data mining accelerates data analysis, and analysts can examine existing data to identify patterns and trends of diabetes.

This paper is structured as follows: Section. 2 describes the relationship that exists between

data mining, machine learning and medicine. The methodology and description of the dataset are described in Section. 3. Sections. 4 and 5, represent a theoretical description of the methods and algorithms that will be applied practically to our data. Section 6 presents the results of the application of algorithms and an explanation for the algorithm with the best results. In sect. 7 the conclusions and future work are discussed.

2. Using Data Mining and Machine Learning in Medicine

Medicine is the science and practice of establishing the diagnosis, prognosis, treatment, and prevention of disease. Medicine encompasses a variety of health care practices evolved to maintain and restore health by the prevention and treatment of illness [4]. This is one of the most important areas when applying data mining techniques can produce significant results [5].

With data mining techniques, doctors will be able to predict illnesses effectively and they will be better equipped to manage potential high-risk candidates [6]. The high volume of diseases data and the complexity of the relationships between them have made medicine an appropriate field for applying data mining techniques. Data mining can be used to examine many large datasets involving a large set of variables beyond what a single analyst or doctor, or even an analytical team can. Like any other problem solving method, the task of data mining begins with a problem definition. The identification of the data mining problem enables the determination of the data mining process and the modeling technique. Machine learning is a subfield of data science that deals with algorithms able to learn from data and make accurate predictions. Data mining gives health organizations the opportunity to learn about disease trends etc. By using data mining methods and machine learning algorithms we improve diabetes analysis and we help to reduce and prevent it.

3. Data and Methodology

We compare theoretically and practically data mining methods to discover the most appropriate method for our data. The methods were compared by applying machine learning algorithms to concrete data in the WEKA "Waikato Environment for Knowledge Analysis" [7] environment. The implemented algorithms are: Simple Logistic, Multilayer Perceptron, Logistic, Naive Bayes, Bayes Net, SMO, C4.5.In Figure 1, we explain all the stages of this study from predicting diabetes using data mining methods and machine learning algorithms of these methods.

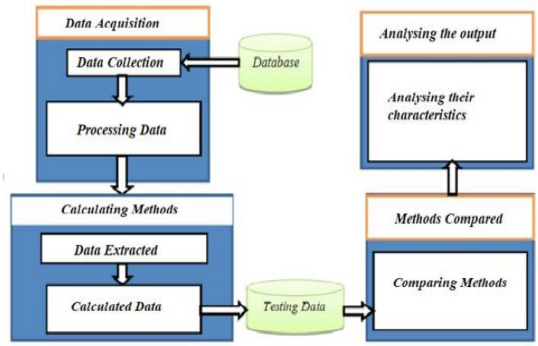

Figure 1: System design

In data gathering step we collect data from the sector of statistics of the Public Health Institute. The collected data is stored into database for further process. The dataset is made up of 270 records or instances.

Table 1
Dataset details

The name of the dataset	Number of Examples	Number of input attributes	Number of possible classes	Total number of attributes	Values that are missing
Diabetes Data	270	7	2	8	0

The variables or attributes of this dataset are: 1) Age: As you age, your risk of diabetes increases, especially when you are over 45 years old., 2) Body Mass Index: It is an indicator of weight (underweight, normal, overweight) based on length and weight. Given weight (kg) / (length m) 2. Ideal BMI values are 18.5-24.9. If we have values 25-29.9 the person is considered overweight, 30-39.9 indicates obesity and 40+ significant obesity. 3) Insulin: Serum Insulin in two hours. Values higher than 150μU/ ml mean that a person needs insulin

therapy, therefore he is pre-diabetic or diabetic., 4) Glucose: Glucose tolerance test values (glucose value mg / dl 2 hours after 75 g glucose) A person is said not to suffer from diabetes if the tolerance test value at two hours is less than or equal to 110 mg / dL (Norman 1)., 5) Skin Thickness: Triceptal Muscle Thickness (mm) - Indicative value of 23 mm overweight for women, values higher than normal indicate that the person is overweight., 6) Blood Pressure: Diastolic blood pressure (mm Hg) Normal blood pressure values are: 60-80 mm Hg, 80-89 indicates pre-hypertension and 90+ hypertension., 7) Number of pregnancies: A woman can be diagnosed with diabetes Gestational during pregnancy. Hormones produced during pregnancy can make cells more resistant to insulin. Those who are older than 25 have a higher risk. Moreover, if a woman has diabetes during one pregnancy, there is an increased risk at the next pregnancy (Diabetes-Bing Health). 8) Outcome: negative when the person is not diagnosed with diabetes and positive when the person is diagnosed with diabetes. The experiments were conducted with a female population over 19 years of age. Diabetes dataset is in CSV format.

4. Classification

Classification is a data mining technique that categorizes data in order to assist in more accurate predictions and analysis [8]. It is one of the data mining methods that aims to analyze very large datasets. It is used to derive patterns that accurately define the important data classes within the data set. Classification techniques predict the target classes for each of the present data instance. [9]. Classification algorithms attempt to detect relationships between attributes that would make it possible to predict the result. They analyze the input and produce a prediction. The classification task of data mining is generally used in healthcare industries [9].

4.1. Naïve Bayes

Bayesian classification represents a supervised learning method as well as a statistical classification method. The Naive Bayes Classifier technique is based on the Bayesian theorem and is used especially when the dimensionality of the inputs is high [10].

Bayesian classification provides practical learning algorithms and prior knowledge, here the observed data can be combined. It calculates the apparent hypothetical probability. The algorithm works as follows. Bayes' theorem offers a way to calculate the probability of a hypothesis based on our prior knowledge. It works based on conditional probability [11]. It can be represented as:

$$P(M|N = \frac{P(M|N)P(M)}{P(N)}$$

Here M and N are two events and, P(M|N) is the conditional probability of M given N. P(M) is the probability of M. P(N) is the probability of N. P (N|M) is the conditional probability of N given M.

Naive Bayes is a strong and powerful predictor. This technique can be useful for very large number of data sets [12]. The Naive Bayesian classifier is fast and incremental and it can deal with discrete and continuous attributes. It has excellent performance and it can explain the decisions.

4.2. Support Vector Machine

SVM classifier is a supervised learning algorithm based on statistical learning theory introduced by Vepnik (Vapnik, 1995) [13]. The main idea behind this method is to determine a hyperplane that optimally separates two classes using training dataset. SVM is a set of related supervised learning method used in medical diagnosis for classification and regression [14]. Support Vector Machine (SVM) model is the representation of examples defined as points in space that are mapped so that the examples of the different categories can be divided by a clear gap that is as large as possible [15]. SVM also supports regression and classification techniques and can handle multiple continuous and categorical variables. The efficiency of SVM-based classification is not directly dependent on the dimension of the classified entities. This algorithm achieves high discriminative power by using special nonlinear functions called kernels to transform the input space into a multidimensional space [16]. It can be seen that the choice of kernel function and best value of parameters for particular kernel is critical for a given amount of data [16]. It also normalizes all attributes by default.

4.3. The decision tree

Decision tree model has a tree structure, which can describe the process of classification instances based on features [17]. It splits the data in the database into subsets based on the values of one or more fields. This process will be repeated for each subgroup recursively until all instances are a node in a single class. The result of the decision tree is a tree-shaped structure that describes a series of decisions given at each step [17]. Decision trees are easy to interpret and understand. They provide white box structure for each provided dataset and can be combined with any other data mining techniques [18]. The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the C4.5 algorithm. The C4.5 is a fraction between information gain and its splitting information. It selects the attribute value of the data that most effectively separate the tested data into subset data which enriched the class. The tree is generated by the normalized information gain [19]. The C4.5 inductive algorithm generates rules from a single tree. It can transform multiple decision trees and create a set of classification rules. Such features of this algorithm can be used to scale general rules, instruction time, size, and number of rules. This algorithm fits to medical records because it copes with missing values. Furthermore the algorithm handles continuous data which is common in medical symptoms. Random Forest is a method of classification which combines hundreds or thousands of decision trees and it trains each one of them on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features [20]. The final predictions of the Random Forest method are made by averaging the predictions of each individual tree. It is fast and easy to implement, and it produces highly accurate predictions and also it can handle a very large number of input variables without over-fitting [21].

4.4. Artificial Neural Network

Neural networks are an area of Artificial Intelligence (AI), where based on the inspiration we have from the human brain [22]. Applying neural network techniques, a program can learn from the examples and create an internal set of rules for classifying different inputs. All processes of a neural network are performed by this group of neurons or units [22]. Each neuron is a separate communication device, making its operation relatively simple. The function of one unit is simply to receive data from other units, as a function of the inputs it receives to calculate an output value, which it sends to other units. In artificial neural networks, neurons are organized in layers which process information using dynamic state responses to external inputs [17]. Artificial neural network is an example of supervised learning [23]. Artificial neural networks (ANNs) are capable of predicting new observations from existing observations. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. One of the most used Neural Networks is the Multilayer Perceptron (MLP), in which its neurons apply a nonlinear activation function to calculate their outputs [24]. The activation function includes a sigmoid function ($f(x) = 1 / (1 + \exp(-x))$) in the hidden layer and a linear function ($fj(x) = \Sigma p\ i=1 w_{ij} x_i$, where x_i's are predictor variables and w_{ij}'s are input weights) in the output layer. The functional form of the MLP can be written as:

$$y = f\left(\sum_{i=1}^{N} W_{\mathrm{ji}}\, Xi + b_{\mathrm{j}}\right)$$

where x_i is the i-th nodal value in the previous layer, y_j is the j-th nodal value in the present layer, b_j is the bias of the j-th node in the present layer, w_{ji} is a weight connecting xi and y_j, N is the number of nodes in the previous layer, and f is the activation function in the present layer [24].

5. Association Rules and Regression

Association Rule is one of the most important canonical tasks in data mining and probably one of the most studied techniques for pattern discovery. Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository [25]. Association Rules identify the arguments found together with a given, event or record: "the presence of one set of arguments brings the presence of another set". This is how rules of

type are identified: "if argument A is part of an event, then for a certain probability argument B is also part of the event" [26]. Association also has great impact in the health care industry to discover the relationships between diseases, state of human health and the symptoms of disease [27]. It can be used to detect and study the etiological pathways in the populations as they suggest interconnections of various risk factors responsible for a disease and are easily interpretable [26]. The objective of the association rule was to discover interesting association or correlation relationships among a large set of data items. Support and confidence are the most known measures for the evaluation of association rule.

While classification provides categorical, discrete labels, regression has continuous function values. So regression is used mainly to predict missing numeric data values rather than discrete class labels. Regression analysis is a statistical technique for examination of connection between the dependent variable and independent variable, which aims to predict the dependent variable from the independent variable or variables [28]. Regression also involves identifying the distribution of trends based on available data. For this purpose regression trees can be used as well as decision trees whose nodes have numerical values instead of categorical values. Logistic regression used to estimate the probability of occurrence of a specific event and the dependent variable is odds ratio which is another way of expressing possibility. This model can be taken into account as the generalized linear model as a link function and its mistake following of the polynomial distribution [28].

This model as:

$$E = \log it(p) = ln\frac{p}{1-p} = \alpha + \beta 1X1,\mathrm{j} + \cdots + \beta \mathrm{k}X\mathrm{k}, \quad i = 1 \ldots n$$

$$p = P\mathrm{r}(\mathrm{Y}\mathrm{i}=1)$$

$$p = P\mathrm{r}(\mathrm{Y}\mathrm{i}=1|\mathrm{X}) = \frac{e^{\alpha+\beta 1X_{1,\mathrm{i}}+\cdots+\beta \mathrm{k}X_{\mathrm{k},\mathrm{i}}}}{1+e^{\alpha+\beta 1X_{1,\mathrm{i}}+\cdots+\beta \mathrm{k}X_{\mathrm{k},\mathrm{i}}}}$$

Is that

Where: P = is the probability that an example belongs to a particular category,

e = base of natural algorithm (~ 2.72),
α = constant of the equation,
β = coefficient of the predictor variables.

6. Experimental Results

To conduct this study we used WEKA [7] software based on the approach and familiarity with its use. WEKA is an open source tool for data mining, which allows users to apply pre-processing algorithms but it does not provide assistance in terms of which one to apply. However, since different data mining algorithms have different requirements regarding the dataset, some preprocessing is applied by default inside some of the algorithms. Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction, selection, etc. Data preprocessing affects the way in which outcomes of the final data processing can be interpreted. WEKA software package has different programs for different techniques and algorithms.

Experiments are done by using Cross-validation on default option folds= 10. Cross validation helps to improve the model results. The 10-fold cross validation technique has been used for better predictions. We have divided our dataset in to 10 samples. Each sample had to go from the process of retained as a validation data, where the rest 9 samples acted as a training data. This was a 10 times vice versa process. That's why it is call 10-fold cross validation. The advantage gained by this process step is that it cuts down the bias association with random sampling methods. Different classification algorithms were applied on our dataset, and the results for all methods were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of correctly classified instances, accuracy, precision, recall and f-measure. Performance indicators are given on the following Table 2 and Table 3

Table 2
Comparison of the results of the algorithms applied in WEKA

Method	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Recall	Precision	F-Measure
Artificial Neural Network	Multilayer Perceptron	190 (70.34%)	80(29.63 %)	0.702	0.704	0.703
Bayes Classifier	Naive Bayes	190 (70.37%)	80(29.63%)	0.704	0.701	0.701
SVM	SMO	188(69.63 %)	82(30.37 %)	0.696	0.697	0.675
Decision Tree	C4.5	214 (79.26%)	56(20.74 %)	0.793	0.796	0.794
Regression	Logistic	192(71.11%)	78(28.89%)	0.711	0.706	0.704
	Simple Logistic	194(71.85%)	76(28.15%)	0.719	0.714	0.712
Bayes Classifier	Bayes Net	182(67.41%)	88(32.59%)	0.674	0.673	0.674
Decision Tree	Random Forest	179(66.30%)	91(33.70%)	0.663	0.657	0.659

Table 3
The Accuracy of models

Algorithm	Multilayer Perceptron	Naive Bayes	SMO	C4.5	Logistic	Simple Logistic	Bayes Net	Random Forest
Accuracy	68.70%	68.40%	64.70%	79.00%	68.05%	68.95%	65.75%	63.75%

The algorithm with the best results according to Table 2 and Table 3 is C4.5algorithm.

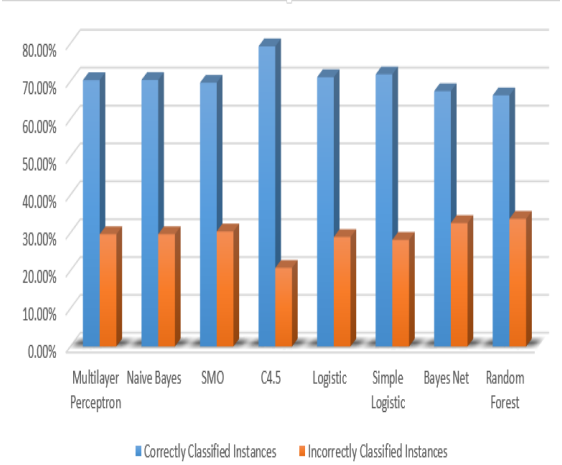

Figure 2: Correctly and Incorrectly Classified Instances in %

Figure 2 shows the correctly and incorrectly classified instances in percentage. These values show that the algorithm with the highest percentage of 79.3% is C4.5. This is also shown in Figure 6, where a clearer picture of the results of this algorithm is shown.

This algorithm is clear and easy when we use it to interpret the results. It selects the attribute value of the data that most effectively separates the tested data into subset data which enriches the class. The model construction is done by modifying the parameter values and this algorithm classifies diabetes disease data with a higher accuracy than other algorithms of data mining methods. This is shown in Table 3, it is the comparison of Accuracy of models after the implementation of algorithms.

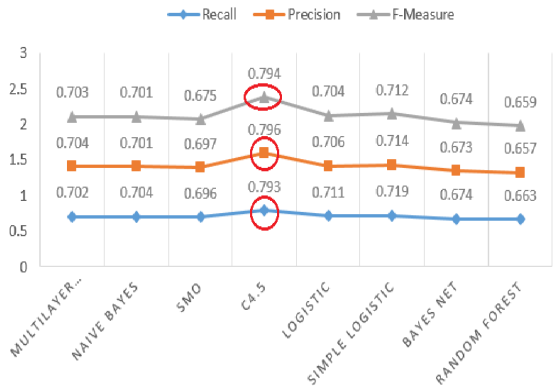

Figure 3: Performance of algorithms

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data. F-measure is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score: precision is the number of correct positive results divided by the number of all positive results returned by the classifier, and recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

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

$$Recall = \frac{TP}{TP + FN}$$

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

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

We converted our data to CSV format. The C4. 5 algorithm for building decision trees is implemented in WEKA as a classifier called J48. J48 has the full name weka.classifiers.trees.J48. What came out of this algorithm: the visualization and the decision tree are presented in Figure 4 and Figure 5.

```
=== Summary ===

Correctly Classified Instances         214               79.2593 %
Incorrectly Classified Instances        56               20.7407 %
Kappa statistic                          0.572
Mean absolute error                      0.2668
Root mean squared error                  0.3652
Relative absolute error                 55.7339 %
Root relative squared error             74.6674 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.776     0.196     0.722     0.776     0.748      0.573     0.876     0.802     pozitive
                 0.804     0.224     0.845     0.804     0.824      0.573     0.876     0.889     negative
Weighted Avg.    0.793     0.213     0.796     0.793     0.794      0.573     0.876     0.855

=== Confusion Matrix ===

   a   b   <-- classified as
  83  24 |   a = pozitive
  32 131 |   b = negative
```

Figure 4: C4.5 (J48) Classifier

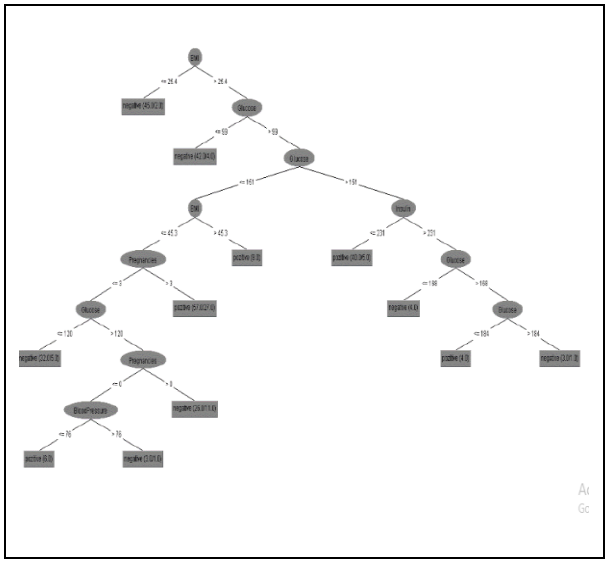

Figure 5: Decision tree

The implementation of this algorithm has classified the diabetes data based on the dataset attributes where precision, recall and f-measure have the highest values compared to other algorithms of data mining methods. This is shown in Figure 3. Figure 5 shows the visualization of the decision tree which is generated by the implementation of the C4.5 algorithm.

```
BMI <= 25.4: negative (45.0/2.0)
BMI > 25.4
|   Glucose <= 99: negative (42.0/4.0)
|   Glucose > 99
|   |   Glucose <= 151
|   |   |   BMI <= 45.3
|   |   |   |   Pregnancies <= 3
|   |   |   |   |   Glucose <= 120: negative (32.0/5.0)
|   |   |   |   |   Glucose > 120
|   |   |   |   |   |   Pregnancies <= 0
|   |   |   |   |   |   |   BloodPressure <= 76: pozitive (6.0)
|   |   |   |   |   |   |   BloodPressure > 76: negative (3.0/1.0)
|   |   |   |   |   |   Pregnancies > 0: negative (26.0/11.0)
|   |   |   |   Pregnancies > 3: pozitive (57.0/27.0)
|   |   |   BMI > 45.3: pozitive (8.0)
|   |   Glucose > 151
|   |   |   Insulin <= 231: pozitive (40.0/5.0)
|   |   |   Insulin > 231
|   |   |   |   Glucose <= 168: negative (4.0)
|   |   |   |   Glucose > 168
|   |   |   |   |   Glucose <= 184: pozitive (4.0)
|   |   |   |   |   Glucose > 184: negative (3.0/1.0)
```

Figure 6: The practical rules derived from C4.5 algorithm.

Figure 6 shows the practical rules derived from C4.5 algorithm.

Through the generated decision tree we understand the characteristics of people who are diabetic. Getting this information helps health centers, hospitals, etc. create policies or make decisions about diabetes by preventing it.

7. Conclusion

The purpose of this article was to create a decision-making structure for diagnosing diabetes. This structure was realized through the study of classification data mining methods such as Naive Bayes, Decision Tree, Support Vector Machine (SVM), Logistic Regression and their evaluation to show the highest performing method on the dataset. The results of experiments conducted in this research by implementing algorithms of data mining methods have revealed that these methods are applicable in the process of diabetes prediction. The decision tree as a data mining classification method has classified diabetes data at an

accuracy rate of 79%. This method has shown promising results for the problem of diabetes prediction as the accuracy rate is high in the experiments performed. Furthermore, the decision tree seems more viable due to the fact that in contrast to other algorithms, it expresses the rules explicitly. These rules can be expressed in human language so that anyone can understand them. Decision trees are easy to interpret and understand. The use of machine learning in analysis diabetes is important because data mining methods and machine learning can be used in the decision making process. In the future extension of this study some models will be created for predicting the diabetes that will help health centers, hospitals, etc. to create policies or make decisions about diabetes by preventing it. Algorithms' behavior changes will be looked at when more data is added. In the future we plan to do the same study but this time not only on women but on all persons regardless of gender.