Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 2519–2528

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques

Francesco Mercaldo[a,*], Vittoria Nardone[b], Antonella Santone[b]

[a]*Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy*
[b]*Department of Engineering, University of Sannio, Benevento, Italy*

Abstract

Medical studies demonstrated that diabetes pathology is increasing in last decades and the trend do not tends to stop. In order to help and to accelerate the diagnosis of diabetes in this paper we propose a method able to classify patients affected by diabetes using a set of characteristic selected in according to World Health Organization criteria. Evaluating real-world data using state of the art machine learning algorithms, we obtain a precision value equal to 0.770 and a recall equal to 0.775 using the HoeffdingTree algorithm.

Peer-review under responsibility of KES International

Keywords: health; machine learning; deep learning; classification

1. Introduction

The number of people with diabetes is steadily increasing, as a study by the International Diabetes Federation demonstrated. In 2013 the total number of diabetes patients was 382 million and in 2035 that number will be expected to be 595 million.

By restricting the field to Type 2 diabetes (i.e., the one on which it can be done by preventing risk factors), in 2010 this disease already afflicted 285 million people, and could be 438 million in 2030, with a progression of 21,000 new cases per day.

Type 2 diabetes accounts for about 90% of cases of diabetes, with the remaining 10% mainly due to Type 1 diabetes mellitus and gestational diabetes. Obesity is considered the main cause of Type 2 diabetes in subjects who are genetically predisposed to the disease.

Type 2 diabetes is initially treated with increased exercise and dietary changes. If, through these measures, blood glucose levels are not adequately controlled, it may be necessary to administer drugs such as metformin or insulin. In patients requiring insulin, usually, there is an obligation to regularly check the levels of blood sugar.

* Corresponding author.
E-mail address: francesco.mercaldo@iit.cnr.it

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.193

Since 1965, Pima Indians living in the Gila River Indian Community in southern Arizona, USA have participated in a longitudinal study of diabetes and its complications. This tribe has the world's highest reported prevalence of diabetes (50% at 35 years of age)[9]. Pima Indians have diabetes which is not associated with insulin dependency, ketoacidosis, or islet-cell antibodies, and is, therefore, Type 2 diabetes[8], even when it occurs in the young[17]. Diabetic nephropathy is the predominant form of kidney disease in this population and is similar in its clinical characteristics and classical pathologic features to that described in other populations[12]. It frequently results in end-stage renal disease, which develops in nearly 15 per cent of diabetic Pima Indians by 20 years duration of diabetes[7].

Starting from these considerations in this paper we propose a method, based on machine learning techniques, able to discriminate between diabetes affected patients and not affected ones. Using a feature vector from both the distributions patients diabetes affected and not affected ones, we perform in addition a best selection step in order to involve the minor number of features to make our solution adoptable in real time using, for instance, embedded devices.

The diagnostic, binary-valued features investigate are relative whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

The paper poses the following research question:

- RQ: is it possible to discriminate between diabetes affected patients and not affected using a set of characteristic selected in according to World Health Organization criteria as a feature vector?

The rest of the paper is organized as follows: the next section provides an overview of related work; the following section illustrates the proposed features and the detection technique; the fourth section presents the results of the evaluation, and, finally, conclusion and future works are given in the last section.

2. Related Work

In this section we review the current literature related to the diabetes affected patients classification issue.

Researches in[13] proposed an integration approach between the SVM technique and K-means clustering algorithms to diagnose diabetes disease. The focus of the method was adjusted so that only the most important features received attention. They performed the T-Test order to quantify the improvements achieved by their approach before and after combination process between K-means and SVM algorithms.

Authors in[11] develop a model for the prediction of gestational diabetes mellitus (GDM) from maternal characteristics and biochemical markers at 11 to 13 weeks' gestation. They demonstrate that in the screening study, maternal age, body mass index, racial origin, previous history of GDM and macrosomic neonate were significant independent predictors of future. The detection rate was 61.6% at a false-positive rate of 20% and the detection increased to 74.1% by the addition of adiponectin and sex hormone-binding globulin.

Alseema et al.[1] adopt the Finnish diabetes risk questionnaire for identification of those at risk for drug-treated Type 2 diabetes. They updated the risk questionnaire by using clinically diagnosed and screen-detected Type 2 diabetes instead of drug-treated diabetes as an endpoint and by considering additional predictors. Of the 18,301 participants, 844 developed Type 2 diabetes in a period of 5 years (4.6%), the conclusion of the study is that the predictive value of the original Finnish risk questionnaire could be improved by adding information on sex, smoking and family history of diabetes.

The goal of the study performed by Bennetts et al.[4] was to utilize k-means clustering analysis to identify typical regional peak plantar pressure distributions in a group of 819 diabetic feet. The number of clusters was varied from 2 to 10 to examine the effect on the differentiation and classification of regional peak plantar pressure distributions. The main of of their analysis is to provide an understanding of the variability of the regional peak plantar pressure distributions seen within the diabetic population and serves as a guide for the preemptive assessment and prevention of diabetic foot ulcers.

Tao Zheng et al.[20] propose a data informed framework for identifying subjects with and without Type 2 Diabetes Mellitus (T2DM) from Electronic Health Records via feature engineering and machine learning. We evaluate and contrast the identification performance of widely-used machine learning models within our framework, including k-

Nearest-Neighbors, Nave Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. Our framework was conducted on 300 patient samples (161 cases, 60 controls and 79 unconfirmed subjects), randomly selected from 23,281 diabetes related cohort retrieved from a regional distributed Electronic Health Records repository ranging from 2012 to 2014.

Rohlfing et al.[16] analyzed the Third National Health and Nutrition Examination Survey (NHANES III) for the sensitivity and specificity of HbA1c in the diagnosis of diabetes based on FPG. They concluded that Hemoglobin A1c (HbA1c) provided a specific and convenient approach to screening for diabetes and suggested a value of 6.1% or greater, 2 sd above the mean in the normal NHANES III population.

Buell et al.[5] recently completed a similar analysis based on the 1999-2004 NHANES data. The diagnosis of diabetes was considered established if FPG was 126 mg/dl or greater. Using a ROC analysis, they found that HbA1c of 5.8% or greater is the point that yielded the highest sum of sensitivity (86%) and specificity (92%).

Nakagami et al.[10] also recently assessed HbA1c vs. FPG in the diagnosis of diabetes. In a cross-sectional study of 1904 Japanese people in one town, aged 3589 yr, they found that the area of the ROC for HbA1c was almost the same as that for FPG (0.856 vs. 0.902, respectively), suggesting that each is a good diagnostic test.

Perry et al.[14], doing OGTTs on people with FPG 100125 mg/dl, found that FPG was insensitive in the detection of OGTT-defined diabetes. The addition of HbA1c greater than 6.1% to FPG greater than 100 mg/dl improved the sensitivity of screening substantially, from 45% to 61%.

At the best of authors' knowledge, this is the first paper with the aim to classify diabetes patients using the minimum feature number in order to make the proposed solution usable in real-world environment.

3. The Method

In this section we describe the approach we adopted in order to discriminate between patients affected from diabetes and not affected ones.

Diabetes was diagnosed according to World Health Organization Criteria[6,8,19,2]; that is, if the 2 hour post-load plasma glucose was at least 200 mg/dl (11.1 mmol/l) at any survey examination or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care[9]. In addition to being a familiar database to the investigators, this data set provided a well validated data resource in which to explore prediction of the date of onset of diabetes in a longitudinal manner.

Indeed, in order to discriminate between diabetes affected and not affected patients, we consider a vector composed by following features:

- Number of times pregnant (F1 feature);
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test (F2 feature);
- Diastolic blood pressure, expressed in mm Hg (F3 feature);
- Triceps skin fold thickness, expressed in mm (F4 feature);
- 2-Hour serum insulin, expressed in mu U/ml(F5 feature);
- Body mass index, expressed in weight in kg/(height in m)$\hat{2}$ (F6 feature);
- Diabetes pedigree function (F7 feature);
- Age, expressed in years (F8 feature).

The eight features were chosen in order to form the basis for forecasting the onset of diabetes within five years in Pima Indian women. Those variables were chosen because they have been found to be significant risk factors for diabetes among Pimas or other populations.

We extracted the feature vector from a freely available dataset from UCI machine repository standard dataset[1] for research purpose that include diabetes affected and not affected patients[18].

[1] https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

The considered dataset contains 768 different instances and all patients in the dataset are females at least 21 years old. The binary target variable takes (0 or 1) values, while 0 implies a negative test for diabetes, and 1 indicates a positive test. There are 500 cases in class 0 and 268 cases in class 1.

The population for this study was the Pima Indian population near Phoenix, Arizona. That population has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because of its high incidence rate of diabetes[9,3,15]. Each community resident over 5 years of age was asked to undergo a standardized examination every two years, including an oral glucose tolerance test.

The dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases [2], that is part of the United States National Institutes of Health[3], which in turn is part of the Department of Health and Human Services[4]. The mission of the institute is represented by to support research, training, and communication with the public in the topic areas of "diabetes and other endocrine and metabolic diseases; digestive diseases, nutritional disorders, and obesity; and kidney, urologic, and hematologic diseases" in order to improve people's health and quality of life.

We designed an experiment in order to evaluate the effectiveness of the feature vector we propose, expressed through the research question RQ stated in the introduction.

More specifically, our aim is to verify if the eight features are able to discriminate between patients affected from diabetes and not affected ones.

We learn several state of the art classifiers with the eight features.

The evaluation consists of three different stages: (i) we provide a comparison of descriptive statistics of the diabetes affected and not affected populations; (ii) hypotheses testing, to verify whether the features vector exhibit different distributions for diabetes affected and not affected populations; and (iii) classification analysis in order to assess if the eight features are able to discriminate between diabetes affected and not affected patients.

With regards to the descriptive statistics, we report the box plot of the distribution of diabetes affected and not affected patients in order to demonstrate that the distribution are different and the features we consider are good candidate for the discrimination between diabetes affected and not affected patients.

Relating to the hypotheses testing, the null hypothesis to be tested is:

H_0 : 'diabetes affected and not affected patients have similar values of the considered features'.

The null hypothesis was tested with Mann-Whitney (with the p-level fixed to 0.05) and with Kolmogorov-Smirnov Test (with the p-level fixed to 0.05). We run two tests in order to enforce the conclusion validity.

The goal of the tests is to determine the level of significance, i.e., the probability that erroneous conclusions be drawn: in this case, we consider the significance level equal to .05 i.e., we accept to make mistakes 5 times out of 100.

The classification analysis goal is to verify if the considered features are able to correctly classify between diabetes affected and not affected patients. Six machine learning classification algorithms were used: J48, MultilayerPerceptron (a deep learning algorithm), HoeffdingTree, JRip, BayesNet and RandomForest. These algorithms were applied to the eight features (i.e., to the feature vector).

The classification analysis is performed using the Weka[5] tool, a suite of machine learning software, employed in data mining for scientific research.

4. The Evaluation

The results of the evaluation will be discussed reflecting the data analysis' division in three phases explained in previous section: descriptive statistics, hypotheses testing and classification.

[2] https://www.niddk.nih.gov/
[3] https://www.nih.gov/
[4] https://www.hhs.gov/
[5] http://www.cs.waikato.ac.nz/ml/weka/

4.1. Descriptive statistics

Figures 1, 2, 3, 4, 5, 6, 7 and 8 show the box plots related to Number of times pregnant (F1 feature), Plasma glucose concentration a 2 hours in an oral glucose tolerance test (F2 feature), Diastolic blood pressure (F3 feature), Triceps skin fold thickness (F4 feature), 2-Hour serum insulin (F5 feature), Body mass index (F6 feature), Diabetes pedigree function (F7 feature) and Age (F8 feature) features with the two different distributions i.e., patients diabetes affected and not affected ones.

Relating to Figure 1, the negative and positive distributions seem to be different, even if a series of points in the two distributions assume similar values.

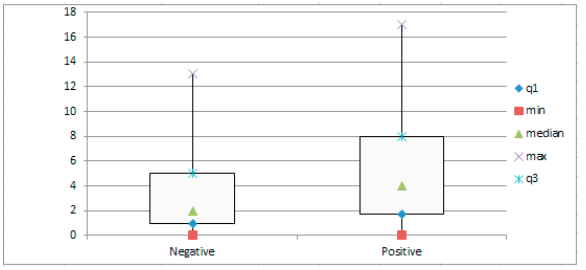

Fig. 1. Box plots related to the Number of times pregnant (F1 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

The distributions related to Figure 2 appear more different from each other when compared with the box plots represented in Figure 2.

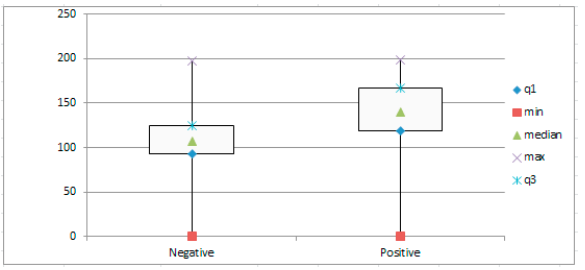

Fig. 2. Box plots related to the Plasma glucose concentration a 2 hours in an oral glucose tolerance test (F2 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

The box plots in Figure 3 exhibit that the distributions related to negative and positive patients to diabetes are very closed, this is symptomatic that the feature is probably not able to discriminate the instance of the two populations.

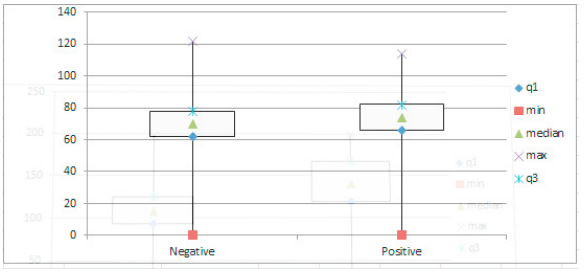

Fig. 3. Box plots related to the Diastolic blood pressure (F3 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

Relating to Figure 4, we highlight that the positive patients distribution show higher or equal instances values if compared with the negative one.

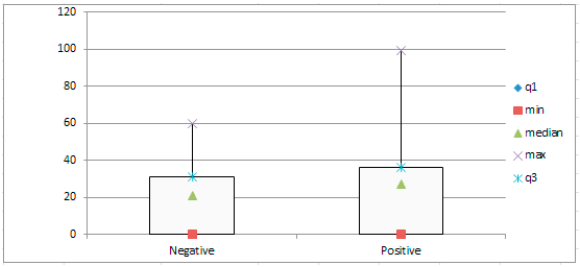

Fig. 4. Box plots related to the Triceps skin fold thickness (F4 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

Figure 5 shows that the positive distribution instances are comparable or less higher that the negative ones.

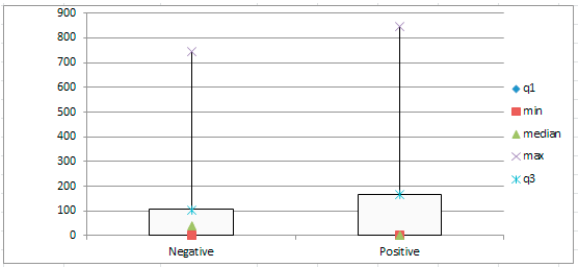

Fig. 5. Box plots related to the 2-Hour serum insulin (F5 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

Figure 6 shows that the positive distribution instances are comparable or higher that the negative ones. This is symptomatic that the feature can be a good candidate to discriminate between diabetes positive patients and negative ones.

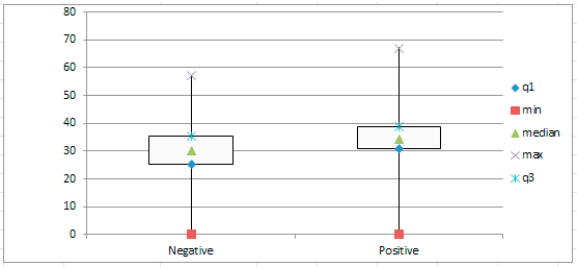

Fig. 6. Box plots related to the Body mass index (F6 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

Figure 7 shows that the positive distribution instances exhibit very closed value. From the boxplot analysis it seems that the feature is not a good candidate to identify positive and negative patients.

Figure 5 shows that the positive distribution instances are comparable or higher that the negative ones. As a matter of fact the positive box plot exhibit higher value than the negative one: for this reason this feature can be a good candidate to discriminate between positive patients and negative ones.

4.2. Hypothesis testing

The hypothesis testing aims at evaluating if the features present different distributions for the populations of diabetes affected patients and not affected ones with statistical evidence.

We assume valid the results when the null hypothesis is rejected by both the tests performed.

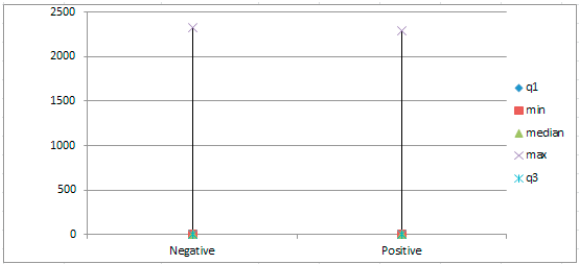

Fig. 7. Box plots related to the Diabetes pedigree function (F7 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

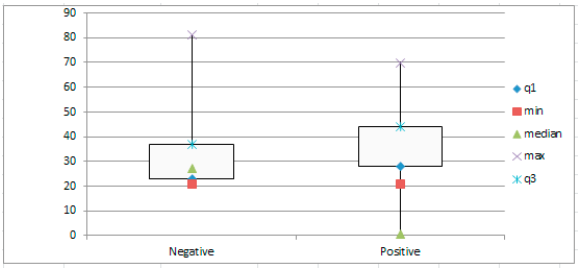

Fig. 8. Box plots related to the Age (F8 feature) for diabetes affected patients (positive distribution) and not affected ones (negative distribution).

Table 1 shows the results of hypothesis testing: the null hypothesis H_0 can be rejected for all the eight features. This means that there is statistical evidence that the feature vector is a potential candidate for correctly classifying between diabetes affected patients and not affected ones.

Variable	Mann-Whitney	Kolmogorov-Smirnov
F1	0,00000	$p < .001$
F2	0,00000	$p < .001$
F3	0,00000	$p < .001$
F4	0,00000	$p < .001$
F5	0,00000	$p < .001$
F6	0,00000	$p < .001$
F7	0,00000	$p < .001$
F8	0,00000	$p < .001$

Table 1. Results of the null hypothesis H_0 test.

This result will provide an evaluation of the risk to generalize the fact that the selected features produce values which belong to two different distributions (i.e., the one related of the diabetes affected patients and not affected one): the null hypothesis H_0 test confirms that the features can distinguish those observations. With the classification analysis we will be able to establish the accuracy of the features in associating any feature vector to the diabetes affected patients or to the not affected patients distribution.

4.3. Classification analysis

We used five metrics in order to evaluate the results of the classification: Precision, Recall, F-Measure and ROC Area.

The precision has been computed as the proportion of the examples that truly belong to class X among all those which were assigned to the class. It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved:

Algorithm	Precision	Recall	F-Measure	Roc Area
A1	0.735	0.738	0.736	0.751
A2	0.750	0.754	0.751	0.793
A3	0.757	0.762	0.759	0.816
A4	0.755	0.760	0.755	0.739
A5	0.741	0.743	0.742	0.806
A6	0.754	0.758	0.755	0.820

Table 2. Classification results: Precision, Recall, F-Measure and RocArea for classifying the evaluated feature vector, computed with six different classification algorithms i.e. J48 (A1), MultilayerPerceptron (A2), HoeffdingTree (A3), JRip (A4), BayesNet (A5) and RandomForest (A6).

$$Precision = \frac{tp}{tp+fp}$$

where *tp* indicates the number of true positives and *fp* indicates the number of false positives.

The recall has been computed as the proportion of examples that were assigned to class X, among all the examples that truly belong to the class, i.e., how much part of the class was captured. It is the ratio of the number of relevant records retrieved to the total number of relevant records:

$$Recall = \frac{tp}{tp+fn}$$

where *tp* indicates the number of true positives and *fn* indicates the number of false negatives.

The F-Measure is a measure of a test's accuracy. This score can be interpreted as a weighted average of the precision and recall:

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

The Roc Area is defined as the probability that a positive instance randomly chosen is classified above a negative randomly chosen.

The classification analysis consisted of building classifiers in order to evaluate the feature vector accuracy to distinguish between diabetes affected patients and not affected ones.

For training the classifier, we defined T as a set of labeled messages *(M, l)*, where each M is associated to a label $l \in \{IM, NM\}$. For each M we built a feature vector $\mathrm{F} \in R_y$, where y is the number of the features used in training phase ($y = 8$).

For the learning phase, we use a k-fold cross-validation: the dataset is randomly partitioned into k subsets. A single subset is retained as the validation dataset for testing the model, while the remaining $k-1$ subsets of the original dataset are used as training data. We repeated the process for $k = 10$ times; each one of the k subsets has been used once as the validation dataset. To obtain a single estimate, we computed the average of the k results from the folds.

We evaluated the effectiveness of the classification method with the following procedure:

1. build a training set $T \subset D$;
2. build a testing set $T' = D \div \mathrm{T}$;
3. run the training phase on T;
4. apply the learned classifier to each element of T'.

Each classification was performed using 20% of the dataset as training dataset and 80% as testing dataset employing the full feature set.

The results that we obtained with this procedure are shown in table 2.

We obtain the best precision (0.757) and the best recall (0.762) using the HoeffdingTree algorithm, the remaining algorithms obtain a precision ranging between 0.735 and 0.755 and a recall ranging between 0.738 and 0.760.

In addition we perform a principal components analysis (PCA) in order to identify, from the 8 features involved, the best features discriminating the positive and negative patients. We employ two different algorithms: BestFirst and GreedyStepwise. The PCA results are shown in Table 3.

Table 3. Feature Selection Results.

F2	*Plasma glucose concentration*
F6	*Body mass index*
F7	*Diabetes pedigree function*
F8	*Age*

Algorithm	**Precision**	**Recall**	**F-Measure**	**Roc Area**
A1	0.742	0.749	0.743	0.7914
A2	0.752	0.755	0.753	0.809
A3	0.770	0.775	0.769	0.824
A4	0.755	0.749	0.755	0.709
A5	0.749	0.755	0.742	0.802
A6	0.743	0.747	0.744	0.813

Table 4. Classification results with best the features: Precision, Recall, F-Measure and RocArea for classifying the evaluated feature vector, computed with six different classification algorithms i.e. J48 (A1), MultilayerPerceptron (A2), HoeffdingTree (A3), JRip (A4), BayesNet (A5) and RandomForest (A6).

Both the feature selection algorithms employed, the BestFirst and the GreedyStepwise confirm that 4 features on the 8 considered in the full vector dataset are the most discriminatory in positive and negative patients identification.

As expected all the 4 features resulting from the feature selection step successfully passed the null hypothesis H_0 test, as shown in Table 1.

Table 4 shows the classification analysis considering the features retrieved from the feature selection step. The obtained results in terms of the analyzed metrics are closed to the previous ones, confirming that the excluded feature were not useful in the classification task.

The best features classification obtains following results: we reach a best precision value equal to 0.770 and a recall equal to 0.775 using the HoeffdingTree algorithm, increasing the precision of 0.757 and the recall of 0.762 obtained in the full feature vector classification.

5. Conclusions and Future Work

In this paper we propose a method able to discriminate between diabetes affected patients and not affected ones using machine learning algorithm. We evaluate our method on a real-world data extracted by the Pima Indian population near Phoenix in Arizona. Training the model using six different classification algorithms, we obtain a precision equal to 0.757 and the recall of 0.762 after the best features selection step.

We plan to improve our work in several ways: the first one, is represented by the application of model checking techniques in order to formulate diabetes patients specific properties to investigate whether the generated properties are able to discriminate between negative and positive patients with better precision and recall values. The second future work consists in the application of our method in order to identify others common diseases.

References

1. M Alssema, D Vistisen, MW Heymans, G Nijpels, Charlotte Glümer, PZ Zimmet, JE Shaw, Mats Eliasson, CDA Stehouwer, AG Tabák, and others. 2011. The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes. *Diabetologia* 54, 5 (2011), 1004–1012.
2. Leanne Bellamy, Juan-Pablo Casas, Aroon D Hingorani, and David Williams. 2009. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet* 373, 9677 (2009), 1773–1779.
3. PeterH Bennett, ThomasA Burch, and Max Miller. 1971. Diabetes mellitus in American (Pima) indians. *The Lancet* 298, 7716 (1971), 125–128.
4. Craig J Bennetts, Tammy M Owings, Ahmet Erdemir, Georgeanne Botek, and Peter R Cavanagh. 2013. Clustering and classification of regional peak plantar pressures of diabetic feet. *Journal of biomechanics* 46, 1 (2013), 19–25.

5. Catherine Buell, Duclie Kermah, and Mayer B Davidson. 2007. Utility of A1C for diabetes screening in the 1999–2004 NHANES population. *Diabetes care* 30, 9 (2007), 2233–2235.
6. Femmie de Vegt, Jacqueline M Dekker, Agnes Jager, Ellen Hienkens, Pieter J Kostense, Coen DA Stehouwer, Giel Nijpels, Lex M Bouter, and Robert J Heine. 2001. Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: The Hoorn Study. *Jama* 285, 16 (2001), 2109–2113.
7. Stephen A Kamenetzky, Peter H Bennett, Stephen E Dippe, Max Miller, and Philip M LeCompte. 1974. A clinical and histologic study of diabetic nephropathy in the Pima Indians. *Diabetes* 23, 1 (1974), 61–68.
8. Hilary King and Marian Rewers. 1991. Diabetes in adults is now a Third World problem. The WHO Ad Hoc Diabetes Reporting Group. *Bulletin of the World health Organization* 69, 6 (1991), 643.
9. William C Knowler, Peter H Bennett, Richard F Hamman, and Max Miller. 1978. Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *American Journal of Epidemiology* 108, 6 (1978), 497–505.
10. Tomoko Nakagami, Makoto Tominaga, Rimei Nishimura, Nobuo Yoshiike, Makoto Daimon, Toshihide Oizumi, and Naoko Tajima. 2007. Is the measurement of glycated hemoglobin A1c alone an efficient screening test for undiagnosed diabetes?: Japan National Diabetes Survey. *Diabetes research and clinical practice* 76, 2 (2007), 251–256.
11. Surabhi Nanda, Mina Savvidou, Argyro Syngelaki, Ranjit Akolekar, and Kypros H Nicolaides. 2011. Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. *Prenatal diagnosis* 31, 2 (2011), 135–141.
12. RG Nelson, JM Newman, WC Knowler, ML Sievers, CL Kunzelman, DJ Pettitt, CD Moffett, SM Teutsch, and PH Bennett. 1988. Incidence of end-stage renal disease in type 2 (non-insulin-dependent) diabetes mellitus in Pima Indians. *Diabetologia* 31, 10 (1988), 730–736.
13. Ahmed Hamza Osman and Hani Moetque Aljahdali. 2017. Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM. *International Journal of Advanced Computer Science & Applications* 1, 8 (2017), 236–244.
14. R Clark Perry, R Ravi Shankar, Naomi Fineberg, Janet McGill, and Alain D Baron. 2001. HbA1c measurement improves the detection of type 2 diabetes in high-risk individuals with nondiagnostic levels of fasting plasma glucose. *Diabetes Care* 24, 3 (2001), 465–471.
15. DJ Pettitt, MF Saad, PH Bennett, RG Nelson, and WC Knowler. 1990. Familial predisposition to renal disease in two generations of Pima Indians with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia* 33, 7 (1990), 438–443.
16. Curt L Rohlfing, Hsiao-Mei Wiedmeyer, Randie R Little, Jack D England, Alethea Tennill, and David E Goldstein. 2002. Defining the relationship between plasma glucose and HbA1c. *Diabetes care* 25, 2 (2002), 275–278.
17. S.L. Salas and Einar Hille. 1978. *Calculus: One and Several Variable*. John Wiley and Sons, New York.
18. Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 261.
19. Joint Who and FAO Expert Consultation. 2003. Diet, nutrition and the prevention of chronic diseases. *World Health Organ Tech Rep Ser* 916, i-viii (2003).
20. Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. 2017. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics* 97 (2017), 120–127.