Early stage diabetes risk prediction

José Manuel Díaz Urraco

1. Introduction

Diabetes is a disease that can cause fatal consequences such as blindness, kidney failure, myocardial infarction, stroke, amputation of the lower limbs and, in the worst cases, death. In addition, their medical treatment is long and expensive. Therefore, getting an early diagnosis is crucial, not only to prevent health problems (and possible deaths) but also to save medical expenses.

Fortunately, various studies ([1], [2], [3] and many more) show that through the use of Data Mining techniques, tools can be developed to support professionals in achieving an early diagnosis of diabetes.

In this study, various analyzes have been carried out on a dataset that contains information about hospital patients who have symptoms related to diabetes. During these analyzes, non-probabilistic supervised classification algorithms have been applied to generate models that allow predicting whether a patient has diabetes or not. The results of these models have been interpreted or explained to the extent possible. The software used to carry out this study has been Weka [4].

The analyzed dataset [5] has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet (Bangladesh) and approved by a doctor. The dataset has 520 instances, 16 predictor variables and the class variable, as shown in Table 1. Dataset variables.

No.	Variable name	Values	Type
1	Age	16-90	Numeric
2	Gender	Male, Female	Nominal
3	Polyuria	Yes, No	Nominal
4	Polydipsia	Yes, No	Nominal
5	sudden weight loss	Yes, No	Nominal
6	weakness	Yes, No	Nominal
7	Polyphagia	Yes, No	Nominal
8	Genital thrush	Yes, No	Nominal
9	visual blurring	Yes, No	Nominal
10	Itching	Yes, No	Nominal
11	Irritability	Yes, No	Nominal
12	delayed healing	Yes, No	Nominal
13	partial paresis	Yes, No	Nominal
14	muscle stiffness	Yes, No	Nominal
15	Alopecia	Yes, No	Nominal
16	Obesity	Yes, No	Nominal
17	class	Positive, Negative	Nominal

Table 1. Dataset variables

2. Problem description

Diabetes is a chronic and irreversible disease of the metabolism in which an excess of glucose or sugar is produced in the blood and in the urine. According to WHO [6], in 2019 an estimated 1.5 million deaths were directly caused by diabetes, and another 2.2 million deaths were attributable to high blood glucose in 2012. Furthermore, it also states that diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation.

The second main problem with this disease is its expensive medical treatment. According to ADA [7], the total costs of diagnosed diabetes have risen to $327 billion in 2017. This cost includes $237 billion in direct medical costs (hospital inpatient care, prescription medications to treat complications of diabetes, anti-diabetic agents and diabetes supplies, physician office visits) and $90 billion in reduced productivity (increased absenteeism, reduced productivity while at work for the employed population, reduced productivity for those not in the labour force, inability to work as a result of disease-related disability, lost productive capacity due to early mortality).

To reduce the impact of these problems, it has been proven that by identifying patients with pre-diabetes and initiating early interventions in lifestyle and/or pharmacological treatments, the progression of the disease can be delayed, or in some cases even prevented [8]. This is where supervised classification algorithms come into play, which, if applied properly, can help professionals determine whether or not a patient has diabetes.

3. Methodology

The dataset was preprocessed first. Subsequently, various analyzes have been performed applying different non-probabilistic supervised classification algorithms on it, first without performing any variable filtering and then filtering the dataset in various ways (three feature subset selection approaches: univariant, multivariant and wrapper). The results obtained from these analyzes have been honestly validated using the 10-fold cross validation method. Finally, the results obtained have been interpreted and a small comparison has been made with results obtained from other similar studies. Figure 1. Study workflow shows the steps that have been followed during the study.

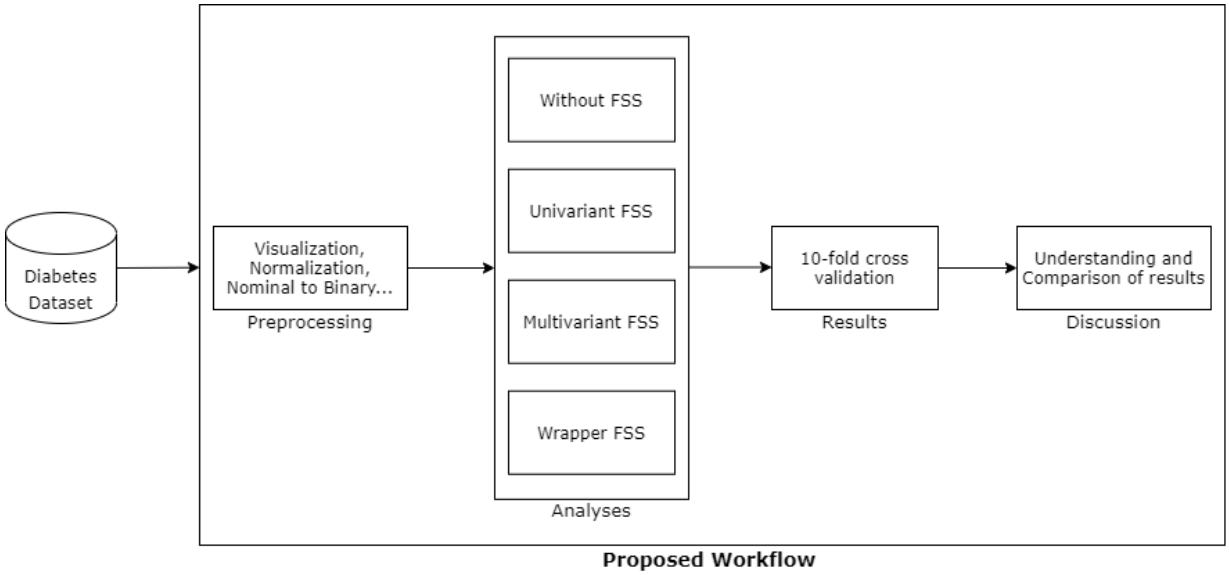

Figure 1. Study workflow

3.1. Preprocessing

First, the dataset has been visualized, the variables it presents can be observed in Table 1. Dataset variables. No missing values or outliers have been observed. Nominal predictor variables have been transformed to numeric using the Weka filter *nominalToBinary*. Subsequently, the variables have been normalized so that they all have a common scale, without distorting differences in the ranges of values.

Also, during this phase, a series of observations were also made about the patients registered in the dataset:

1. Regarding **age**, there are 143 people between 18 and 39 years old, 281 people between 40 and 59 years old, 95 people between 60 and over, and only 1 person under 18 years old.
2. Regarding **gender**, 328 patients are male and 192 are female.
3. With regard to **symptoms**, the most frequent are: weakness (305 patients), polyuria (258 patients), itching (253 patients), delayed healing (239 patients), polyphagia (237 patients), polydipsia (233 patients) and visual blurring (233 patients as well). In turn, the least frequent are: obesity (88 patients), genital thrush (116 patients), irritability (126 patients) and alopecia (179 patients). Also note that there are 53 people (of the 520 in total) who do not have any symptoms and, of those 53 people, there are 6 who do have diabetes (class -> Positive).
4. Finally, regarding the **class** variable, there are 320 patients with diabetes and 200 without diabetes (class -> Negative).

3.2. Analyses

Different non-probabilistic supervised classification algorithms have been applied on the original dataset after preprocessing (without feature subset selection) and then applying: univariant feature subset selection, multivariant feature subset selection and wrapper feature subset selection.

In each analysis, the 5 algorithms observed in Table 2. Used classifiers and modified parameters have been applied. The performance of the generated models has been improved by modifying the values of the parameters of all the classifiers and by testing various combinations. Table 2. Used classifiers and modified parameters also shows the parameters that have been modified for each classifier, the rest of the parameters keep their default values.

Classifier	Weka function	Modified parameters
k-NN	IBk	KNN = 20, crossValidate = True
RIPPER	JRip	folds = 2, seed = 100, usePruning = False
MLP	MultilayerPerceptron	hiddenLayers = t, seed = 20, trainingTime = 1000
SVM	SMO	c = 2.0, kernel = Puk, randomSeed = 20
C4.5	J48	confidenceFactor = 0.8

Table 2. Used classifiers and modified parameters

3.2.1. Without FSS

The different classifiers have simply been applied to the dataset (520 instances, 16 predictor variables and the class variable) after preprocessing.

3.2.2. Univariant FSS

The *Mutual Information* parametric method (*InfoGainAttributeEval* in Weka) has been applied. As it can be seen in Figure 2. Univariant FSS. Ranked variables, the classifiers have been applied with 2 different subsets of variables. They were first applied to 8 variables (threshold set at 0.05) and then applied to 10 variables (threshold set at 0.04).

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.36225    3 Polyuria=Yes
 0.35906    4 Polydipsia=No
 0.16342    2 Gender=Female
 0.14877    5 sudden weight loss=Yes
 0.14465   13 partial paresis=Yes
 0.08784    7 Polyphagia=Yes
 0.07287   11 Irritability=Yes
 0.05116   15 Alopecia=No                    first threshold = 0.05
 0.04661    9 visual blurring=Yes
 0.04267    6 weakness=No                    second threshold = 0.04
 0.02239    1 Age
 0.01097   14 muscle stiffness=No
 0.00905    8 Genital thrush=Yes
 0         10 Itching=No
 0         12 delayed healing=No
 0         16 Obesity=No

Selected attributes: 3,4,2,5,13,7,11,15,9,6,1,14,8,10,12,16 : 16
```

Figure 2. Univariant FSS. Ranked variables

Better results were obtained with the subset of 10 variables, which is not unreasonable to think since *weakness* and *visual blurring* are 2 of the most frequent symptoms among the patients in the dataset, so their information may be important for the class variable.

3.2.3. Multivariant FSS

The *Correlation-based feature selection* method (*CfsSubsetEval* in Weka) has been applied. As can be seen in Figure 3. Multivariant FSS. Selected variables, the resulting dataset has 6 predictor variables.

```
Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.465

Attribute Subset Evaluator (supervised, Class (nominal): 17 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4,5,11,13 : 6
                     Gender=Female
                     Polyuria=Yes
                     Polydipsia=No
                     sudden weight loss=Yes
                     Irritability=Yes
                     partial paresis=Yes
```

Figure 3. Multivariant FSS. Selected variables

3.2.4. Wrapper FSS

The *WrapperSubsetEval* method has been applied in Weka. As seen in Table 3. Wrapper FSS. Classifiers, filtering has been tested with various classifiers. The performance of the models generated with each classifier has been compared by changing the values of the parameter *No. xval folds*. It has been tested with the values: 2, 3, 4, 5, 6 and 7. In all classifiers the model with the best performance is obtained with *No. xval folds = 3*.

Wrapper FSS		
Wrapper Classifier	**No. xval folds**	**No. selected variables**
IBk	3	12
JRip	3	10
SNO	3	11
J48	3	12

Table 3. Wrapper FSS. Classifiers

The best result was obtained for the *IBk* with *k = 1*. Figure 4. IBk wrapper resulting variables shows the resulting variables after applying this filtering.

```
Selected attributes: 1,2,3,4,8,9,11,12,13,14,15,16 : 12
                     Age
                     Gender=Female
                     Polyuria=Yes
                     Polydipsia=No
                     Genital thrush=Yes
                     visual blurring=Yes
                     Irritability=Yes
                     delayed healing=No
                     partial paresis=Yes
                     muscle stiffness=No
                     Alopecia=No
                     Obesity=No
```

Figure 4. IBk wrapper resulting variables

4. Results

All analyzes have been carried out using a **k-fold cross validation** with *k = 10*. Table 4. Most relevant results in each analysis shows the most relevant results obtained in each analysis. It has been decided to include the number of false negatives since, in the case of diabetes, although the accuracy of the model is very good, it is important that this number (false negatives) is as low as possible, since otherwise, patients with diabetes would be being diagnosed as patients without diabetes, and this could cause serious health risks or even death in those patients.

It should be noted that, in general, worse results have been obtained after applying the different variable filters, with the exception of the wrapper filtering, where better results have been obtained for each classifier in question.

As it can be seen in Table 4. Most relevant results in each analysis, the *IBk* algorithm after applying wrapper filtering shows exceptionally good accuracy. It is also observed that the *SMO* algorithm, by not applying any variable filtering, presents *0* false negatives and a fairly good accuracy as well, the bad thing about this model is that its interpretability is complex compared to *IBk* and others.

Weka classifier	FSS	Correctly Classified Instances	No. False Negatives	F-Measure	ROC Area
IBk	Wrapper	99.04%	4	0.990	0.991
JRip	–	95.96%	15	0.960	0.962
MultilayerPerceptron	Multivariant	90.19%	33	0.903	0.960
SMO	–	97.69%	0	0.977	0.970
J48	Univariant	94.04%	20	0.941	0.948

Table 4. Most relevant results in each analysis

5. Discussion

5.1. IBk (Wrapper FSS)

IBk algorithm is a method that simply searches among the closest observations to the one that is being tried to predict and classifies the point of interest based on most of the surrounding data.

In Table 5. Different k values together with their root mean squared error different values of *k* are observed along with their root mean squared error for the wrapper filtering. The best *k* is the one that minimizes this error (prediction error) that corresponds to the square root of the average difference between the observed known outcome values and the predicted values. The lower the root mean squared error, the better the model. So, in this case, the results are shown for *k = 1*.

k	Root mean squared error
1	0.0944
2	0.1577
3	0.1835
4	0.2042
5	0.2117
6	0.2154
7	0.2321

Table 5. Different k values together with their root mean squared error

5.2. JRip (Without FSS)

As it is seen in Figure 5. JRip rules, the *JRip* algorithm has generated 10 rules. For example, the first rule is read as: if the patient does not have *Polyuria*, the patient is *Male*, the patient does not have *Polydipsia*, the patient does not have *Irritability*, does not have *weakness* and does not have *partial paresis*, then the patient is classified as not having diabetes (*Negative* class). It is observed that the number of instances supported by this rule is *88* and *0* instances are incorrectly classified. Also, if we look at the last rule, we see that 320 patients are classified as *Positive* in diabetes if none of the 9

previous rules are met. The rest of the rules are interpreted in the same way as the previous ones.

```
JRIP rules:
===========

(Polyuria=Yes <= 0) and (Gender=Female <= 0) and (Polydipsia=No >= 1) and (Irritability=Yes <= 0) and (weakness=No >= 1) and (partial paresis=Yes <= 0) => class=Negative (88.0/0.0)
(Polyuria=Yes <= 0) and (Alopecia=No <= 0) and (Polydipsia=No >= 1) and (Itching=No <= 0) and (weakness=No <= 0) => class=Negative (59.0/0.0)
(Polydipsia=No >= 1) and (Gender=Female <= 0) and (Polyuria=Yes <= 0) and (Alopecia=No >= 1) and (Polyphagia=Yes <= 0) and (delayed healing=No >= 1) => class=Negative (16.0/0.0)
(Polydipsia=No >= 1) and (delayed healing=No <= 0) and (Alopecia=No <= 0) and (Itching=No <= 0) and (Age >= 0.378378) and (Gender=Female <= 0) => class=Negative (15.0/0.0)
(Polyuria=Yes <= 0) and (Polyphagia=Yes <= 0) and (Itching=No <= 0) and (sudden weight loss=Yes <= 0) and (Age >= 0.445946) and (Irritability=Yes <= 0) => class=Negative (7.0/0.0)
(Polyuria=Yes <= 0) and (Age <= 0.324324) and (sudden weight loss=Yes >= 1) and (Gender=Female <= 0) => class=Negative (3.0/0.0)
(Polyuria=Yes <= 0) and (visual blurring=Yes <= 0) and (Age <= 0.243243) and (Gender=Female >= 1) => class=Negative (5.0/0.0)
(Polyuria=Yes <= 0) and (Age >= 0.513514) and (Age <= 0.540541) and (Irritability=Yes <= 0) => class=Negative (5.0/0.0)
(Polyuria=Yes <= 0) and (Irritability=Yes >= 1) and (Age <= 0.391892) and (Polyphagia=Yes <= 0) and (Age >= 0.27027) => class=Negative (2.0/0.0)
 => class=Positive (320.0/0.0)

Number of Rules : 10
```

Figure 5. JRip rules

5.3. MultilayerPerceptron (Multivariant FSS)

MLP networks are made up of an input layer, one or more hidden layers and an output layer. Each layer is made up of a series of nodes. As it can be seen in Figure 6. MLP. Nodes (until Sigmoid Node 4) and weights, the *Sigmoid Node 0* and *Sigmoid Node 1* nodes are output nodes (*Positive* and *Negative*, respectively) and show the value of the weights associated with each node of the previous layer (hidden layer) after performing the backpropagation algorithm. While the rest of the nodes (*Sigmoid Node 2* to *Sigmoid Node 9*) are hidden nodes. These nodes show the value of the weights associated with each node of the input layer (the attributes). The threshold (bias) is also shown together with the value of the weights.

```
Sigmoid Node 0
    Inputs    Weights
    Threshold    -1.3010780675941211
    Node 2    2.9478938314046217
    Node 3    3.614189417979384
    Node 4    0.5828583010467799
    Node 5    4.216276336615465
    Node 6    0.5649803858367154
    Node 7    1.2596825108955356
    Node 8    2.326570562716632
    Node 9    -3.577154426597468
Sigmoid Node 1
    Inputs    Weights
    Threshold    1.300995347886262
    Node 2    -2.947864810798639
    Node 3    -3.614897409712262
    Node 4    -0.5828159395922242
    Node 5    -4.216536168742597
    Node 6    -0.5646836377593564
    Node 7    -1.2596376388907635
    Node 8    -2.324558225883477
    Node 9    3.5774686767933956
Sigmoid Node 2
    Inputs    Weights
    Threshold    -1.0273779357788793
    Attrib Gender=Female    5.345232779983547
    Attrib Polyuria=Yes    2.0937231979497346
    Attrib Polydipsia=No    1.9246396087632698
    Attrib sudden weight loss=Yes    3.9684058158505477
    Attrib Irritability=Yes    1.8008608567875601
    Attrib partial paresis=Yes    4.618889526407677
Sigmoid Node 3
    Inputs    Weights
    Threshold    -2.043548596232467
    Attrib Gender=Female    4.026533572222278
    Attrib Polyuria=Yes    0.7207028483992454
    Attrib Polydipsia=No    0.9812486192386908
    Attrib sudden weight loss=Yes    6.528871539426574
    Attrib Irritability=Yes    4.103715718538774
    Attrib partial paresis=Yes    2.5814646371481484
Sigmoid Node 4
    Inputs    Weights
    Threshold    -0.6477097750372207
    Attrib Gender=Female    3.378288879144346
    Attrib Polyuria=Yes    1.4271081356118662
    Attrib Polydipsia=No    -3.0480449090768293
    Attrib sudden weight loss=Yes    -0.430480049442153
    Attrib Irritability=Yes    -0.202616456611122
    Attrib partial paresis=Yes    -2.616304515517422
```

Figure 6. MLP. Nodes (until Sigmoid Node 4) and weights

5.4. SMO (Without FSS)

A trained *Support Vector Machine* has a scoring function which computes a score for a new input. A Support Vector Machine is a binary (two class) classifier; if the output of the scoring function is negative then the input is classified as belonging to class *y = -1*. If the score is positive, the input is classified as belonging to class *y = 1*.

In Figure 7. SMO. Scoring function, part of the scoring function can be seen. The numbers in angle brackets are the support vectors. The coefficient beside each support vector is the computed *alpha* value for that data point and the sign of the coefficient comes from the class label. For example, the penultimate vector *<0.621622 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0>* belongs to class *y = -1*, and the last vector *<0.5 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1>* belongs to class *y = 1*. The final value in the expression, *0.589*, is *b*.

```
 +       0.0311 * <0.513514 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 > * X]
 -       0.0915 * <0.594595 1 1 0 1 0 1 0 1 0 0 0 1 0 1 0 > * X]
 -       0.1301 * <0.554054 0 1 0 0 0 1 1 0 1 0 0 1 1 1 1 > * X]
 -       0.1094 * <0.189189 1 1 0 1 0 0 0 0 0 0 0 1 0 1 1 > * X]
 -       0.3526 * <0.121622 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 > * X]
 -       0.2504 * <0.716216 1 1 1 1 0 1 1 1 0 0 0 0 1 0 1 > * X]
 -       0.3206 * <0.324324 1 1 1 1 1 1 0 0 1 0 1 0 1 1 1 > * X]
 +       1.4827 * <0.445946 0 0 0 0 1 0 0 0 0 0 1 1 1 1 0 > * X]
 +       0.04   * <0.418919 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 > * X]
 -       0.9298 * <0.337838 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 > * X]
 +       0.0048 * <0.540541 0 0 1 1 0 1 0 1 0 0 0 0 0 0 1 > * X]
 -       0.2235 * <0.594595 0 1 0 0 0 1 0 1 0 0 0 1 1 1 1 > * X]
 -       0.0004 * <1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 1 > * X]
 -       0.2266 * <0.256757 1 0 1 0 1 0 0 0 1 0 1 0 1 1 1 > * X]
 -       0.7762 * <0.256757 1 0 1 0 1 0 0 0 1 0 1 0 1 1 1 > * X]
 -       0.2742 * <0.527027 1 1 1 1 1 0 1 1 0 0 0 1 1 1 1 > * X]
 -       0.0299 * <0.202703 1 1 0 1 0 1 0 0 0 1 1 1 0 1 1 > * X]
 -       0.0173 * <0.364865 1 1 0 1 0 1 0 1 1 0 1 1 0 1 0 > * X]
 -       0.3536 * <0.432432 0 1 0 1 0 0 0 0 0 0 1 1 1 0 1 > * X]
 -       0.1838 * <0.635135 0 1 0 1 0 1 0 1 1 0 1 0 0 0 0 > * X]
 -       0.0976 * <0.189189 1 1 0 1 1 1 0 0 1 1 1 1 0 1 1 > * X]
 -       0.0036 * <0.364865 1 1 0 1 0 1 0 1 1 0 1 1 0 1 0 > * X]
 -       0.0157 * <0.554054 0 1 0 1 0 1 0 1 1 0 1 1 1 1 1 > * X]
 -       0.0106 * <0.310811 1 1 0 0 1 1 1 1 0 0 0 1 1 1 1 > * X]
 -       0.0309 * <0.527027 0 1 0 1 0 1 0 1 1 0 0 1 1 1 1 > * X]
 -       0.1758 * <0.189189 0 1 0 1 0 0 1 0 1 0 0 0 1 1 1 > * X]
 +       0.7368 * <0.662162 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 > * X]
 -       0.0167 * <0.621622 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 > * X]
 +       0.2005 * <0.5 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 > * X]
 -       0.589

Number of support vectors: 378

Number of kernel evaluations: 131997 (92.5% cached)
```

Figure 7. SMO. Scoring function

5.5. J48 (Without FSS)

In Figure 8. J48 tree, part of the generated tree of the *J48* model is observed. For example, the path *A* is read as follows: if the patient does not have *Polyuria*, does not have *Polydipsia*, is *Male*, does not have *Irritability*, does not have *partial paresis*, has *delayed healing*, and is over *40* years old (normalized value is *0.324324*), then, the class is *Negative* (the patient is classified as not having diabetes). It is seen that *43* instances reach this path and only *1* instance is misclassified. Rest of the branches are interpreted in the same way.

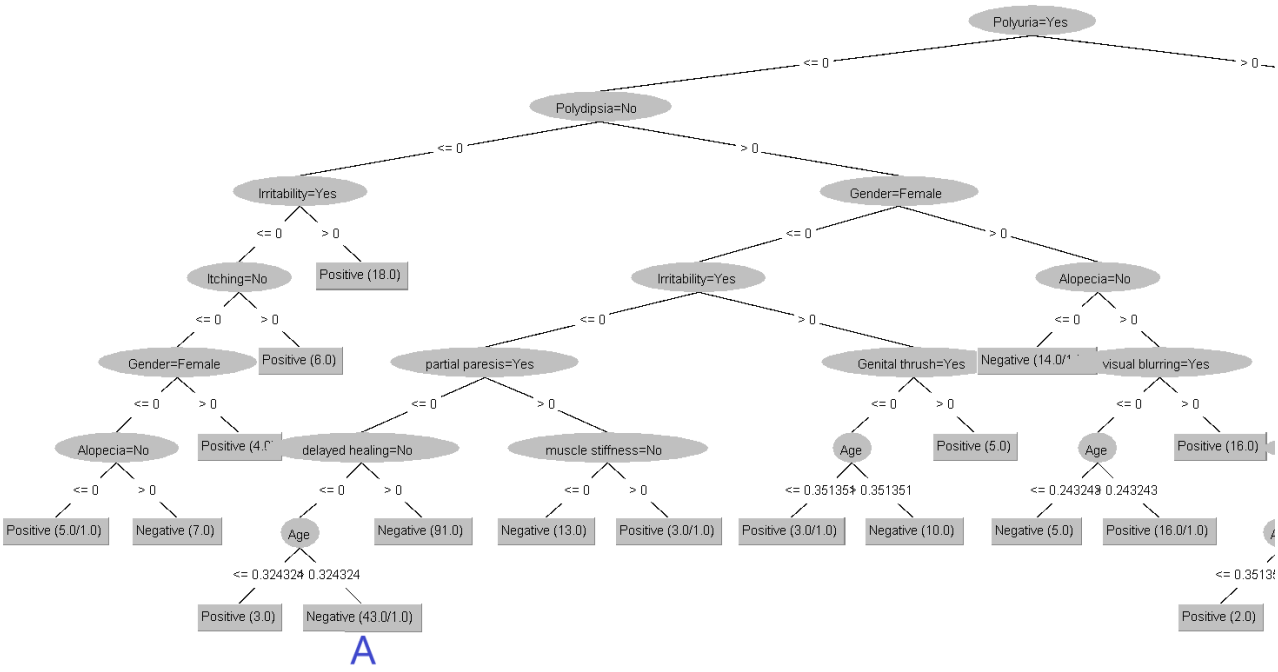

Figure 8. J48 tree

5.6. Result comparison

L. Chaves and G. Marques conducted a study similar to this one [9] where they compared their results after having applied different supervised classification algorithms with the results of other similar studies as well.

Just as a curious fact (which perhaps is worth studying and deepening) comment that, during this study, a model has been generated with an accuracy higher than those of all the models that appear in the comparation of the study mentioned above, without having been the goal of this study at any time. Specifically, it is the *IBk* model after *wrapper filtering*, with an accuracy of *99.04%*, as shown in Table 6. Comparison of results between different studies .

Study	Classifier with more accuracy	Accuracy
M. M. Faniqul Islam et al.	Random Forest	97.4%
K. Alpan and G. S.	k-NN	98.07%
H. Naz and S. Ahuja	ANN	90.34%
N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia	ANN	88%
M. Peker, O. Özkaraca, and A. Sasar	ANN	93.85%
S. Malik et al.	Random Forest	98.8%
L. Chaves and G. Marques	ANN	98.08%
This study	k-NN	99.04%

Table 6. Comparison of results between different studies

6. Conclusion

In this study, several analyzes have been carried out where different non-probabilistic supervised classification algorithms have been applied. The model that presents the best performance of all those generated is a priori the *IBk* after the wrapper filtering. In addition, it is an intuitive and easy-to-

understand model. However, the *SMO* model is also interesting (although it is difficult to interpret) since it has *0 false positives*. So, when choosing a model that helps doctors determine whether a patient has diabetes or not, maybe it would be interesting to take both models into account, for example, combining their predictions.

Finally, going forward, it would be fascinating to add more instances to the dataset for a more adequate and accurate classification. For example, more patients under 18 could be added since right now there is only 1 person. Also, more female patients could be added and other factors that may enhance the onset of diabetes such as hereditary factors or smoking. Once the dataset is completer and more balanced, another study could be carried out and the results obtained could be compared with those of the first study.

7. References

[1] Kumari, S., & Singh, A. (2013, January). A data mining approach for the diagnosis of diabetes mellitus. In *2013 7th International Conference on Intelligent Systems and Control (ISCO)* (pp. 373-375). IEEE.
[2] Shivakumar, B. L., & Alby, S. (2014, March). A survey on data-mining technologies for prediction and diagnosis of diabetes. In *2014 International Conference on Intelligent Computing Applications* (pp. 167-173). IEEE.
[3] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.
[4] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
[5] UCI Machine Learning Repository *Early stage diabetes risk prediction dataset.* https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
[6] Diabetes, World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/diabetes
[7] The Cost of Diabetes, American Diabetes Association (ADA). https://www.diabetes.org/resources/statistics/cost-diabetes
[8] Importance of Early Diabetes Diagnosis and Screening. https://www.apollodiagnostics.in/blog/importance-of-early-diabetes-diagnosis-and-screening
[9] *Chaves, L., & Marques, G. (2021). Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study.* Applied Sciences, 11*(5), 2218.*