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, its 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, 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 probabilistic supervised classification algorithms (including metaclassifiers) 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. Figure 1. Study workflow shows the steps that have been followed during the study.

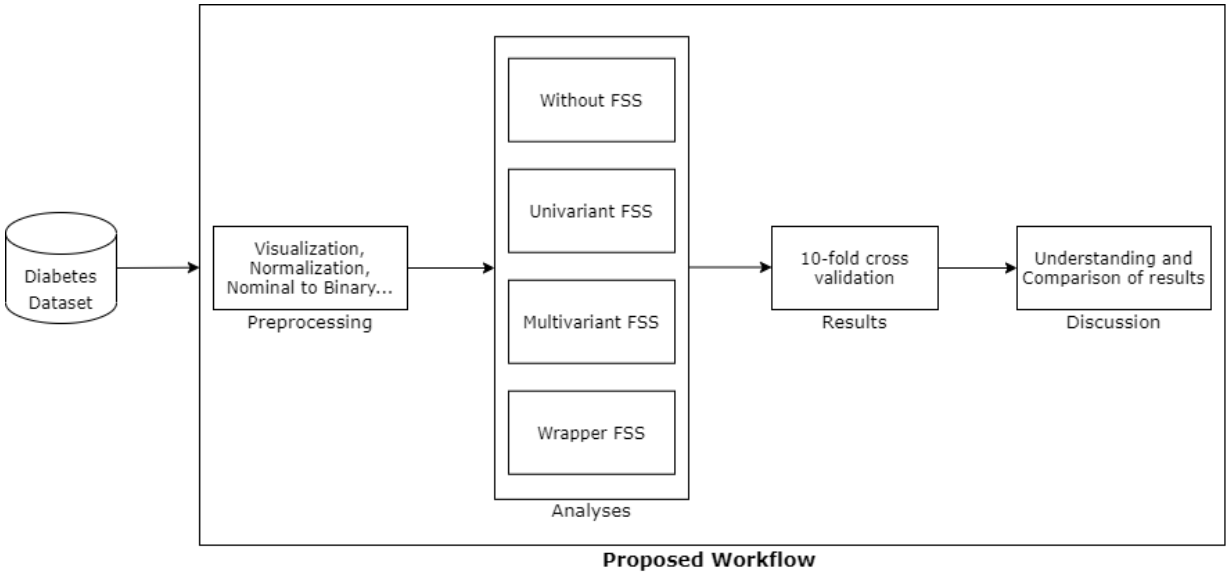

Figure 1. Study workflow

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. On this occasion, to facilitate the understanding of the correlation matrix (generated with Python) shown in section 3.2 and the outputs of the models, the transformation has been carried out directly on the *arff* file instead of using the Weka *nominalToBinary* filter. 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 probabilistic supervised classification algorithms, which can be observed in Table 2. Used classifiers and modified parameters, 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.

Classifier	Weka function	Modified parameters
Logistic Regression	Logistic	–
Linear Discriminant Analysis	LDA	–
Naïve Bayes	NaiveBayes	–
Tree Augmented Naïve Bayes	BayesNet	searchAlgorithm = TAN
Stacking	Stacking	metaClassifier = J48, classifiers = [IBk, JRip, MLP, Logistic, LDA, NaiveBayes, TAN]
Boosting	AdaBoostM1	classifier = TAN weightThreshold = 50
Random Forest	RandomTree	–
Fusion (mean)	Vote	classifiers = [IBk, MLP, JRip, Logistic, LDA, NaiveBayes, TAN]
Naïve Bayes Tree	NBTree	–
Logistic Model Trees	LMT	–

Table 2. Used classifiers and modified parameters

Most of probabilistic classifiers require the predictor variables to be independent of each other (not redundant) for a better performance. Figure 2. Correlation matrix displays Pearson's correlation values, which measure the degree of linear relationship between each pair of variables. Correlation values can be between *-1* and *+1*. As it can be observed, the strongest relationship is found between

the variables *Polyuria* and *Polydipsia* with a value of *0.6*.

Before performing the different feature subset selections, it has been verified whether the generated models present better performance by removing one of these 2 variables or if, on the contrary, the models are more accurate considering both variables. By eliminating *Polyuria* variable, accuracy decreases in all models except in *LDA* where it increases by almost *0.5%* and the number of *False Negatives* goes from *50* to *42*. By eliminating *Polydipsia* variable, accuracy decreases in all models except *LDA* where it increases by almost *0.6%* and the number of *False Negatives* goes from *50* to *40*.

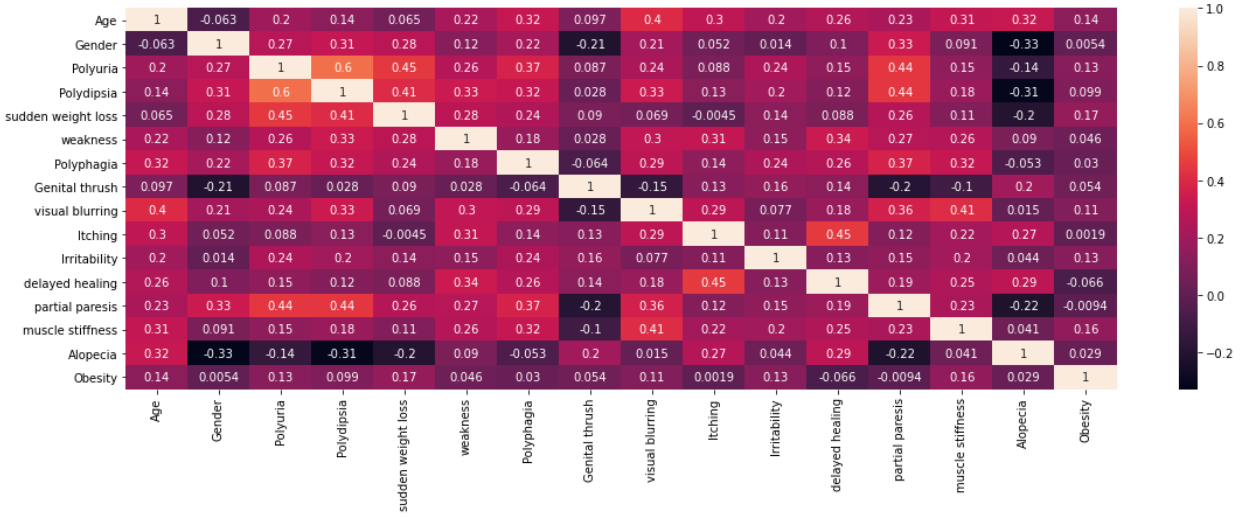

Figure 2. Correlation matrix

Therefore, because performance is not improved in the vast majority of classifiers and improvement is minimal in the *LDA* classifier, neither of these two variables has been eliminated at this moment. In fact, the decrease in performance in the models when eliminating one of these 2 variables seems to indicate that they are important together in predicting whether a patient has diabetes or not.

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 and analysing the effect of correlated variables.

3.2.2. Univariant FSS

The *Gain Ratio* parametric method (*GainRatioAttributeEval* in Weka) has been applied. As it can be seen in Figure 3. Univariant FSS. Ranked variables, the classifiers have been applied with 3 different subsets of variables. They were first applied to 2 predictor variables (threshold set at *0.3*), then applied to 5 predictor variables (threshold set at *0.1*) and then applied to 8 predictor variables (threshold set at *0.05*).

As it can be seen, the variables *Polydipsia* and *Polyuria* are the ones that provide most of the information when it comes to predicting the class variable.

After building all the probabilistic models for each subset of variables, it has been observed that each classifier performs better with a different subset, as shown in Table 3. Different Univariant FSS.

```
Ranked attributes:
 0.3623    1 Polyuria
 0.3619    2 Polydipsia
 0.172     3 Gender
 0.1518    4 sudden weight loss
 0.1467    5 partial paresis
 0.0912    7 Irritability
 0.0883    6 Polyphagia
 0.0551    8 Alopecia
 0.047     9 visual blurring
 0.0436   10 weakness
 0.042    11 Age
 0.0118   13 Genital thrush
 0.0115   12 muscle stiffness
 0        14 Itching
 0        15 delayed healing
 0        16 Obesity
```

first threshold = 0.3

second threshold = 0.1

third threshold = 0.05

Figure 3. Univariant FSS. Ranked variables

Classifier	Subset size where it performs best
Logistic Regression	8
Linear Discriminant Analysis	2
Naïve Bayes	5
Tree Augmented Naïve Bayes	5
Stacking	5
Boosting	5
Random Forest	8
Fusion (mean)	5
Naïve Bayes Tree	8
Logistic Model Trees	8

Table 3. Different Univariant FSS

3.2.3. Multivariant FSS

The *Correlation-based feature selection* method (*CfsSubsetEval* in Weka) has been applied. As can be seen in Figure 4. Multivariant FSS. Selected variables, the resulting dataset has 6 predictor variables.

```
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
                     Polyuria
                     Polydipsia
                     sudden weight loss
                     Irritability
                     partial paresis
```

Figure 4. Multivariant FSS. Selected variables

3.2.4. Wrapper FSS

The *WrapperSubsetEval* method has been applied in Weka. As seen in Table 4. 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*.

With regard to probabilistic classifiers, the best result was obtained with *Logistic* classifier. Figure 5. Logistic wrapper resulting variables shows the resulting variables after applying this filtering. With regard to metaclassifiers, the most interpretable model was obtained with *RandomTree* classifier. Figure 6. Random Tree wrapper resulting variables shows the resulting variables after applying this filtering.

Wrapper FSS		
Wrapper Classifier	**No. xval folds**	**No. predictor variables**
Logistic	3	7
LDA	2	4
NaiveBayes	3	6
BayesNet	5	8
Stacking	2	10
AdaBoostM1	3	15
RandomTree	4	12
Vote	4	5
NBTree	5	14
LMT	3	11

Table 4. Wrapper FSS. Classifiers

```
Selected attributes: 2,3,4,6,11,12,15 : 7
                     Gender
                     Polyuria
                     Polydipsia
                     weakness
                     Irritability
                     delayed healing
                     Alopecia
```

Figure 5. Logistic wrapper resulting variables

```
Selected attributes: 1,2,3,4,8,9,10,11,12,13,14,15 : 12
                     Age
                     Gender
                     Polyuria
                     Polydipsia
                     Genital thrush
                     visual blurring
                     Itching
                     Irritability
                     delayed healing
                     partial paresis
                     muscle stiffness
                     Alopecia
```

Figure 6. Random Tree wrapper resulting variables

4. Results

All analyzes have been carried out using a **k-fold cross validation** with *k = 10*. Table 5. 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.

Weka classifier	FSS	Correctly Classified Instances	No. False Negatives	F-Measure	ROC Area
Logistic	Wrapper	92.12%	19	0.921	0.960
LDA	–	88.27%	50	0.884	0.967
NaiveBayes	Univariant	89.81%	30	0.898	0.949
TAN (BayesNet)	Wrapper	90.58%	22	0.906	0.948
Stacking	–	97.69%	6	0.977	0.975
RandomTree	Wrapper	97.12%	10	0.971	0.972
AdaBoostM1	–	94.81%	14	0.948	0.976
NBTree	–	97.31%	9	0.973	0.989
LMT	Wrapper	98.08%	7	0.981	0.991

Table 5. Most relevant results in each analysis

As it can be seen, *Logistic* classifier under wrapper filtering presents the best performance among all the probabilistic models generated with *92.12%* accuracy and *19* False Negatives. Regarding metaclassifiers, it can be seen that they have better performance than probabilistic classifiers. However, they are more difficult to interpret. In this case, the metaclassifier with the best performance is *LMT* under wrapper filtering, but *RandomTree* also shows good results and is easier to interpret.

5. Discussion

5.1. Logistic (Wrapper FSS)

In Figure 7. Odds Ratios, the odd ratios can be observed. Odd ratios are used in logistic regression models to compare the influence of the independent variables (predictors) on the dependent variable (class). In such a way that they can be compared to each other to find out which variable is more explanatory of the dependent variable or is associated in a stronger way.

In this case, it seems that *Polyuria* is the variable that has the most influence in predicting whether a patient has diabetes, along with *Polydipsia* and *Gender* variables. For example, if the odds ratio for *Polyuria is 53.7332*, then those patients with polyuria have *53.7332* the odds of having diabetes as those without polyuria. Similarly, the *Alopecia* and *delayed healing* variables seem to have almost no influence on the class variable.

```
Odds Ratios...
                          Class
Variable                Positive
==============================
Gender                   36.3267
Polyuria                 53.7332
Polydipsia               36.7113
weakness                  1.6265
Irritability               8.368
delayed healing           0.3252
Alopecia                  0.4897
```

Figure 7. Odds Ratios

5.2. Naïve Bayes (Univariant FSS)

Attribute	Class Positive (0.61)	Negative (0.39)
Polyuria		
mean	0.7594	0.075
std. dev.	0.4275	0.2634
weight sum	320	200
precision	1	1
Polydipsia		
mean	0.7031	0.04
std. dev.	0.4569	0.196
weight sum	320	200
precision	1	1
Gender		
mean	0.5406	0.095
std. dev.	0.4983	0.2932
weight sum	320	200
precision	1	1
sudden weight loss		
mean	0.5875	0.145
std. dev.	0.4923	0.3521
weight sum	320	200
precision	1	1
partial paresis		
mean	0.6	0.16
std. dev.	0.4899	0.3666
weight sum	320	200
precisi		

Figure 8. Naive Bayes output

Figure 8. Naive Bayes output shows the means (and other measures) of each attribute of the subset for each of the two values of the class.

In this case, it is observed that all the variables of the dataset are influencing the class variable since the dataset was previously filtered.

Diabetic patients seem to present especially polyuria and polydipsia, and also, although to a lesser extent, sudden weight loss and partial paresis. It is also observed that gender is influential, with the majority of non-diabetic patients being male (Male = 0). This may be because there are many more male than female patients in the dataset.

5.3. LDA (Without filtering)

```
Estimates for class value Positive
Natural logarithm of class prior probability: -0.49
Class prior probability: 0.62
Mean vector:

Age: 0.45
Gender: 0.54
Polyuria: 0.76
Polydipsia: 0.7
sudden weight loss: 0.59
weakness: 0.68
Polyphagia: 0.59
Genital thrush: 0.26
visual blurring: 0.55
Itching: 0.48
Irritability: 0.34
delayed healing: 0.48
partial paresis: 0.6
muscle stiffness: 0.42
Alopecia: 0.24
Obesity: 0.19

Estimates for class value Negative
Natural logarithm of class prior probability: -0.96
Class prior probability: 0.38
Mean vector:

Age: 0.41
Gender: 0.1
Polyuria: 0.07
Polydipsia: 0.04
sudden weight loss: 0.15
weakness: 0.44
Polyphagia: 0.24
Genital thrush: 0.17
visual blurring: 0.29
Itching: 0.5
Irritability: 0.08
delayed healing: 0.43
partial paresis: 0.16
muscle stiffness: 0.3
Alopecia: 0.51
Obesity: 0.14
```

Figure 9. LDA output

Mean vectors for each class value are shown in Figure 9. LDA output. A mean vector consists of the means of each variable of the dataset.

Variables that present similar means in both classes can be seen, which means that these variables have not been differentiators in this model to perform the classification. This is the case of the variables: *Age*, *Genital thrush*, *Itching*, *delayed healing*, *muscle stiffness* and *Obesity*. On the contrary, there are variables whose means differ remarkably from one class value to another and which therefore have been the differentiators when making the classification. The widest differences are observed with the variables: *Gender*, Polyuria, *Polydipsia*, *sudden weight loss*, *partial paresis* and *Polyphagia*.

Therefore, it seems that patients classified as diabetic in this model tend to have as symptoms: polyuria, polydipsia, polyphagia, partial paresis and sudden weight loss.

5.4.TAN (Wrapper FSS)

In Figure 10. TAN output, it can be seen the Naive Bayes tree generated by Weka along with the probability distribution tables for some variables in the subset.

Looking at the tree, it seems that the variable *Polyuria* depends on the variable *Polydipsia* (or at least they are related) and also the variable *sudden weight loss* depends on the variable *Polyuria*. These relationships also seem to be confirmed in Figure 2. Correlation matrix. Furthermore, knowing the meaning of these 3 symptoms, apparently it does not seem unreasonable to think that these relationships are not correct.

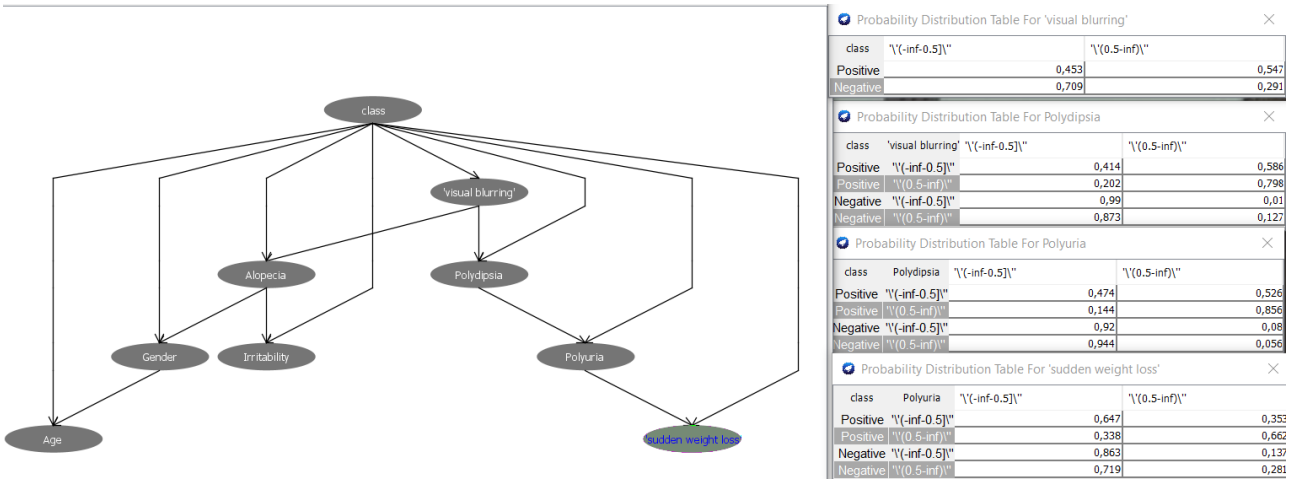

Figure 10. TAN output

Regarding the distribution tables, if we look at the table for *visual blurring*, it seems that diabetic patients have a similar probability of having or not having visual blurring (*0.547* and *0.453*), so it does not seem to be one of the most important variables to detect diabetic patients. However, non-diabetic patients appear to have a high probability of not having visual blurring (*0.709*).

If we look at the table for *Polydipsia*, it can be seen that diabetic patients with visual blurring have a high probability of having polydipsia and non-diabetic patients without visual blurring have a *0.99* probability of not having polydipsia.

Looking at the table for *Polyuria*, it can be seen that diabetic patients with polydipsia have a high probability of having polyuria. Similarly, non-diabetic patients without polydipsia have a high probability of not having polyuria. Finally, non-diabetic patients with polydipsia have a *0.944* probability of not having polyuria. What can be learned from this is that when polyuria and polydipsia are taken together, they become differentiating symptoms for the class variable.

Finally, the table for *sudden weight loss* shows that patients without diabetes have a high probability of not having sudden weight loss whether they have polyuria or not.

5.5. RandomTree (Wrapper)

This classifier selects a *K* number of attributes by using gain ratio at each node. If no value for *K* is given, *K = int (log_2 (#predictors) + 1)*. In this case there are 12 predictor variables, then *K = 4* (*K* is not *5* since Java Type Casting rounds down). Therefore, the algorithm chooses 4 attributes randomly at each node and then compute the gain ratio only for the 4 chosen to perform the split, selecting the attribute with the highest gain ratio.
In Figure 11. RandomTree sample branches it can be seen that, of the 4 attributes randomly selected

for the root node, *Polyuria* is the one that provides the most information when predicting the class.

```
Polyuria < 0.5
|   Gender < 0.5
|   |   Alopecia < 0.5
|   |   |   delayed healing < 0.5
|   |   |   |   Polydipsia < 0.5
|   |   |   |   |   Itching < 0.5
|   |   |   |   |   |   Genital thrush < 0.5
|   |   |   |   |   |   |   Irritability < 0.5 : Negative (69/0)
Polyuria >= 0.5
|   delayed healing < 0.5 : Positive (120/0)
|   delayed healing >= 0.5
|   |   Alopecia < 0.5 : Positive (98/0)
```

Figure 11. RandomTree sample branches

Looking at the left branch, it appears that patients without polyuria, male, without alopecia, without delayed healing, without polydipsia, without itching, and without genital thrush do not have diabetes (69 patients, 0 incorrectly classified). Regarding the branch on the right, patients with polyuria and without delayed healing are classified as diabetic (120 patients, 0 incorrectly classified). It is also observed that patients with polyuria, with delayed healing and without alopecia are classified as diabetic (98 instances, 0 incorrectly classified).

6. Conclusion

In conclusion, in most of the models it appears that when a patient has polyuria and polydipsia together it is usually a clear sign of having diabetes. Moreover, it is possible that many of the other common symptoms of diabetes are due to these two symptoms. Nevertheless, having certain common symptoms of diabetes separately does not guarantee that a patient has diabetes. In addition, it seems that the gender of the patient is also a discriminating factor in all or most of the models, but this should be verified after balancing the dataset by adding more female patients.

Regarding which model would be the most interesting to choose, the *RandomTree* classifier has generated a decision tree in a random way that results in classifying the patients in a fairly optimal way (fairly good accuracy and a low number of false negatives). These trees are easy to interpret, so I consider that it may be feasible to choose the model generated by this classifier under wrapper filtering although the most interesting model is still the *IBk* under wrapper filtering generated in the previous study.

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