Early stage diabetes risk clustering

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.

In this study, different types of clustering will be carried out in order to group similar patients based on the symptoms they present (without knowing whether they are diabetic or not). In this way, new undiagnosed patients can be assigned to a cluster and this can help professionals to determine whether a patient may have diabetes or not, or to provide preventive measures or more specific controls depending on the group to which they belong. For a better visualization and understanding of the resulting groups, the analyzes corresponding to the hierarchical and partitional clustering have been carried out with R [1] and the probabilistic clustering with Weka [2].

The analyzed dataset [3] 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 [4], 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 [5], 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 [6]. This is where unsupervised classification algorithms come into play, which, if applied properly, can help professionals to detect patterns among patients and determine whether they may have diabetes or not.

3. Methodology

The dataset was preprocessed first. Subsequently, various analyzes have been performed applying different types of unsupervised classification algorithms on it: hierarchical clustering, partitional clustering and probabilistic clustering. Finally, the different clusters obtained have been interpreted and labelled. FIGURE 1. STUDY WORKFLOW shows the steps that have been followed during the study.

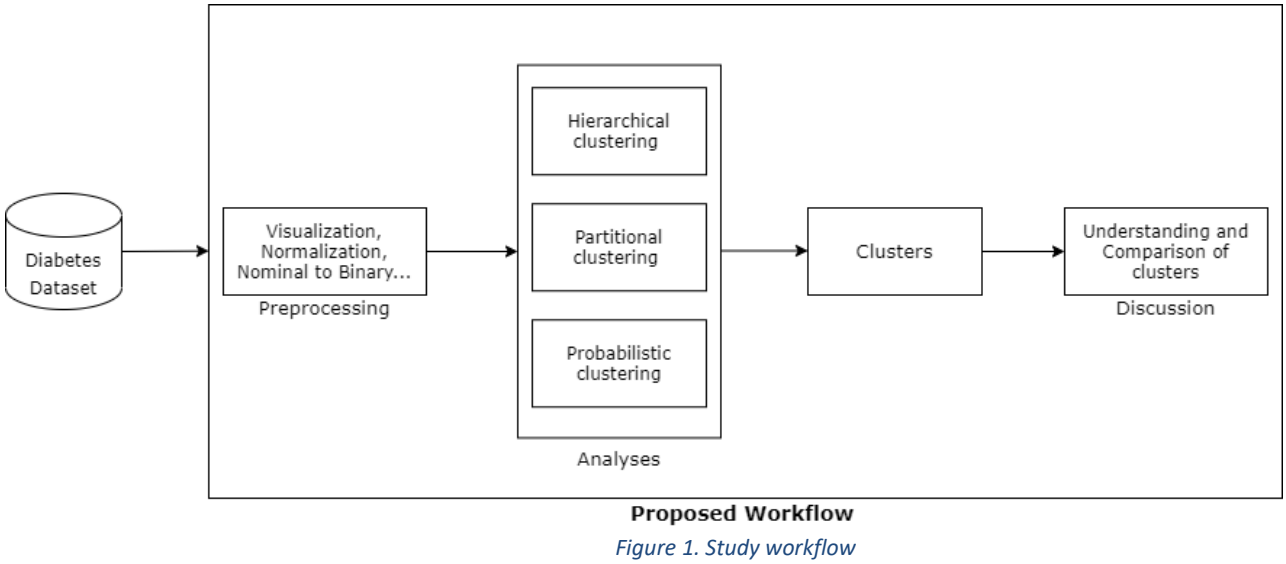

Figure 1. Study workflow

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. Subsequently, the variables have been normalized so that they all have a

common scale, without distorting differences in the ranges of values. Being an unsupervised classification study, the class variable has been removed from the dataset.

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

The different clustering algorithms performed can be seen in TABLE 2. CLUSTERING ANALYSES PERFORMED.

Clustering type	Algorithm or Linkage method	No. clusters
Hierarchical	Complete	2
Hierarchical	Centroid	2
Hierarchical	Ward	2
Hierarchical	Ward	3
Hierarchical	Ward	4
Partitional	K-means	2
Partitional	K-means	3
Probabilistic	EM	2

Table 2. Clustering analyses performed

To estimate the optimal number of clusters in each algorithm, different known indexes have been considered that can be observed in TABLE 3. APPLIED INDEXES. As these indexes only serve as a reference, several numbers have been tried in some algorithms where a clear majority was not observed.

To visualize the clusters, a Principal Component Analysis has been carried out. The data points have been plotted according to the first two main components that explain the majority of the variance.

As it can be seen in FIGURE 2. PCA: VARIABLES, VARIANCE AND CONTRIBUTION, the dimensions that most explain the variance in the observations are *Dim1* (*24.4%*) and *Dim2* (*13.9%*), the rest of the dimensions do not even reach *10%* of explained variances. The variables that contribute the most in *Dim1* are *Polydipsia*, *Polyuria* and *partial paresis*, and the variable that contributes the most in *Dim2* is *Alopecia*. It is observed that there are variables that do not contribute especially in any of the 2 dimensions, therefore do not seem to be important or influential variables, such as: *Irritability*, *Obesity* and *Genital thrush*. Also comment that these are the least frequent symptoms among patients.

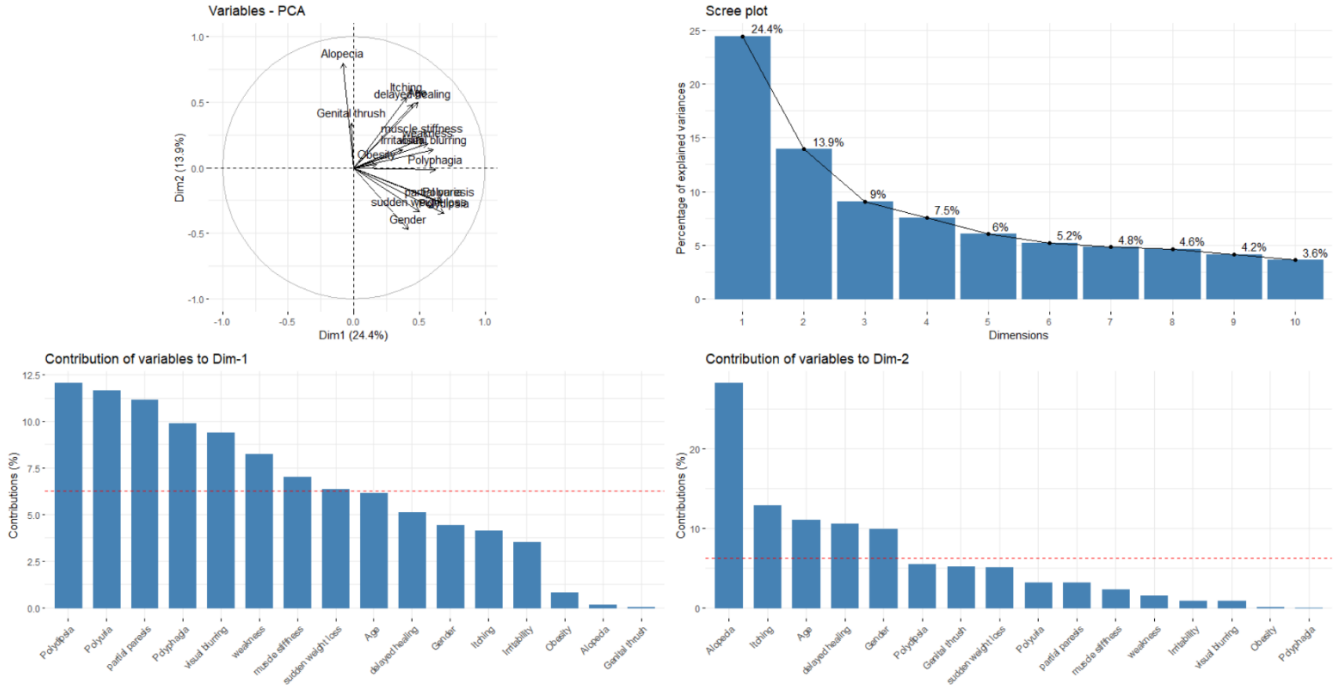

Figure 2. PCA: variables, variance and contribution

Index	Authors
kl	Krzanowski and Lai 1988
ch	Calinski and Harabasz 1974
hartigan	Hartigan 1975
ccc	Sarle 1983
scott	Scott and Symons 1971
marriot	Marriot 1971
trcovw	Milligan and Cooper 1985
tracew	Milligan and Cooper 1985
friedman	Friedman and Rubin 1967
rubin	Friedman and Rubin 1967
cindex	Hubert and Levin 1976
db	Davies and Bouldin 1979
silhouette	Rousseeuw 1987
duda	Duda and Hart 1973
pseudot2	Duda and Hart 1973
beale	Beale 1969
ratkowsky	Ratkowsky and Lance 1978
ball	Ball and Hall 1965
ptbiserial	Milligan 1980, 1981
gap	Tibshirani et al. 2001
frey	Frey and Van Groenewoud 1972
mcclain	McClain and Rao 1975
gamma	Baker and Hubert 1975
gplus	(Rohlf 1974) (Milligan 1981)
tau	(Rohlf 1974) (Milligan 1981)
dunn	(Dunn 1974)
hubert	(Hubert and Arabie 1985)
sdindex	(Halkidi et al. 2000)
dindex	(Lebart et al. 2000)
sdbw	(Halkidi and Vazirgiannis 2001)

Table 3. Applied indexes

3.2.1. Hierarchical Clustering

Complete linkage

As it can be seen in FIGURE 3. COMPLETE LINKAGE: NUMBER OF CLUSTERS, the vast majority of indices suggest *2* as the optimal number of clusters. Due to the great advantage over the rest of the numbers and the fact that the next most suggested number is *10* (many clusters, little heterogeneity between them), the clustering will only be performed in this case with $k = 2$.

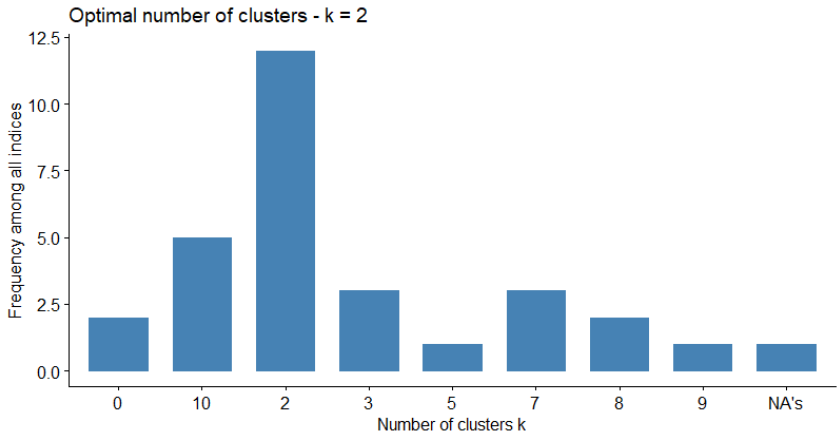

Figure 3. Complete linkage: number of clusters

Centroid linkage

As it can be seen in FIGURE 4. CENTROID LINKAGE: NUMBER OF CLUSTERS, the vast majority of indices suggest *2* as the optimal number of clusters. Due to the great advantage over the rest of the numbers and the fact that the next most suggested number is *10* (many clusters, little heterogeneity between them), the clustering will only be performed in this case with $k = 2$.

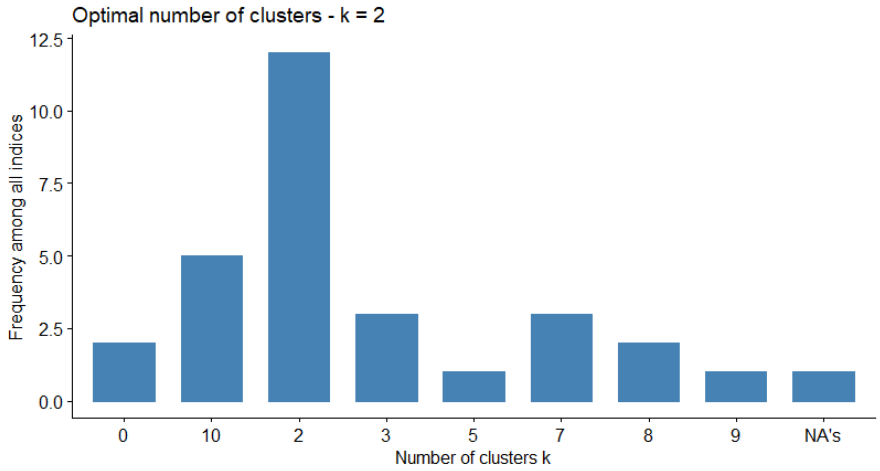

Figure 4. Centroid linkage: number of clusters

Ward linkage

As it can be seen in FIGURE 5. WARD LINKAGE: NUMBER OF CLUSTERS, in this case the largest number of clusters suggested is *3*. However, *2* and *4* are also quite suggested so clustering with these numbers will also be performed.

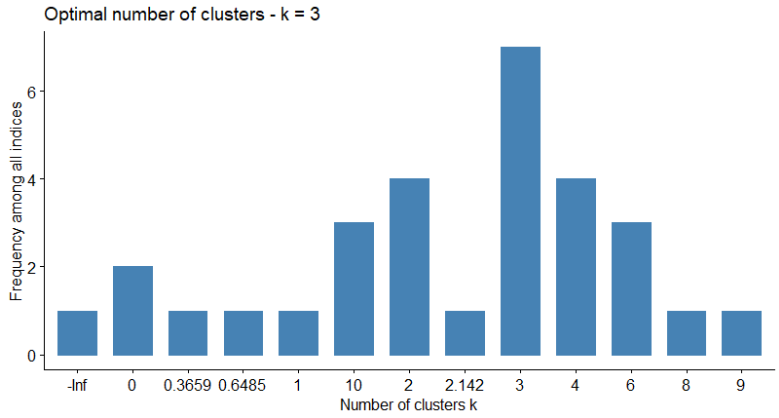

Figure 5. Ward linkage: number of clusters

3.2.2. Partitional Clustering

In this case, as it can be observed in FIGURE 6. K-MEANS ALGORITHM: NUMBER OF CLUSTERS the highest suggested number is *3*. But clustering will also be done with *k = 2* since it is the next most suggested number.

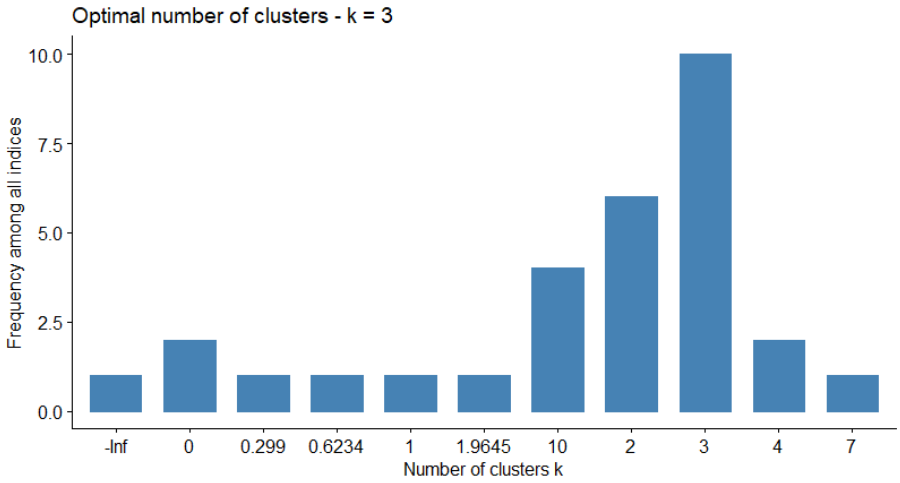

Figure 6. K-means algorithm: number of clusters

3.2.3. Probabilistic Clustering

Regarding probabilistic clustering, it can be seen in FIGURE 7. EM ALGORITHM: NUMBER OF CLUSTERS that the optimal number of clusters has been found automatically using cross validation (*numClusters = -1* in Weka) and is *2*.

```
EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 4
```

Figure 7. EM algorithm: number of clusters

4. Results

4.1. Hierarchical Clustering

Complete linkage

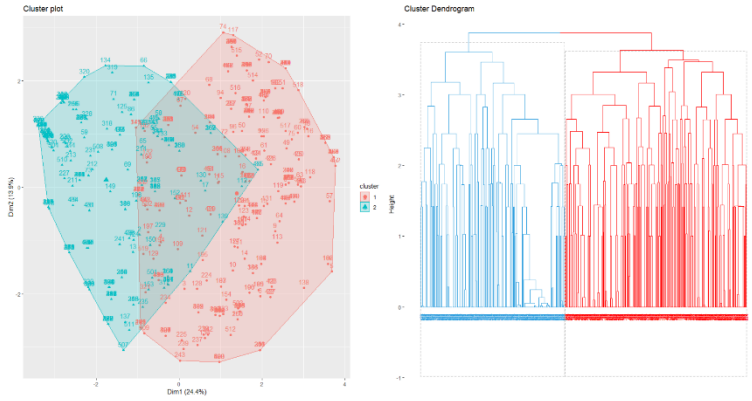

Figure 8. Complete linkage: formed clusters

In FIGURE 8. COMPLETE LINKAGE: FORMED CLUSTERS the two resulting clusters after using the complete linkage can be seen, together with the corresponding dendrogram. It is observed that clustering in this case is not very appropriate since there are many patients that belong to both clusters, that is, the clusters are not heterogeneous at all with each other.

Centroid linkage

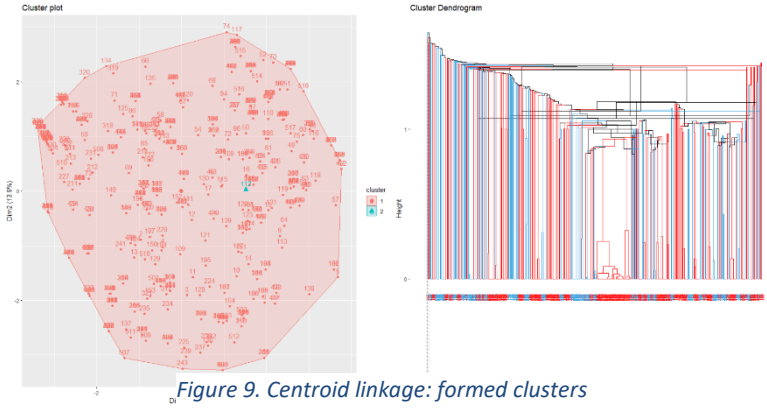

Figure 9. Centroid linkage: formed clusters

In this case, when using the centroid linkage, the result is much worse than in the previous case. In FIGURE 9. CENTROID LINKAGE: FORMED CLUSTERS it can be seen that *cluster2* is within *cluster1* and it seems to consist of only one patient. It can also be observed in the dendrogram how the separation between clusters has not been carried out correctly.

Ward linkage

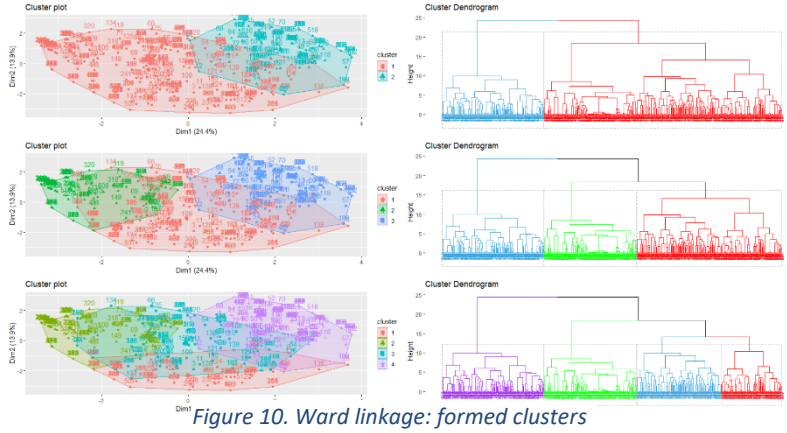

Figure 10. Ward linkage: formed clusters

In FIGURE 10. WARD LINKAGE: FORMED CLUSTERS, the corresponding clusters and dendrograms can be seen after applying the ward linkage with *k* equal to *2*, *3* and *4* respectively. In the three cases, it is observed that the clusters share a considerable number of patients, again, the clusters are not heterogeneous between them. As the

number of clusters increases, it can be observed that the patients in *cluster1* (red) are the ones that are considered to form the rest of the clusters. It is also observed that in all cases there is a cluster that always remains constant (its size does not change).

After having carried out several hierarchical clusters, it seems that the most suitable groupings are carried out using the **ward linkage** with **two clusters**. Despite the fact that a perfect separation of patients is not carried out, it seems that it is the one that achieves more homogeneity within each cluster and more heterogeneity between the two clusters. Therefore, that model is the one to be discussed later.

4.2. Partitional Clustering

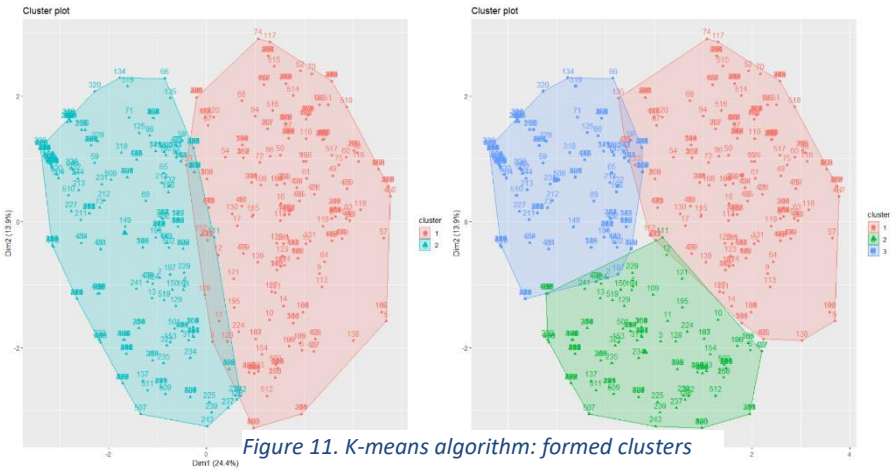

Figure 11. K-means algorithm: formed clusters

With respect to partitional clustering, it is observed in FIGURE 11. K-MEANS ALGORITHM: FORMED CLUSTERS that the groupings are done more correctly than in hierarchical clustering for both $k = 2$ and $k = 3$. The interpreted model will be the one with **three clusters** since it is observed that the separation between the clusters is a bit better (shared area between clusters seems slightly smaller).

4.3. Probabilistic Clustering

```
Clustered Instances

0       301  ( 58%)
1       219  ( 42%)


Log likelihood: -5.42635
```

Figure 12. EM algorithm: formed clusters

FIGURE 12. EM ALGORITHM: FORMED CLUSTERS shows the **two clusters** formed by the *EM* algorithm. It is observed that the clusters are balanced (they have a similar number of patients). The differences between these two clusters will be discussed later.

5. Discussion

5.1. Hierarchical (Ward, 2 clusters)

The mean of each symptom by cluster can be observed in FIGURE 13. WARD LINKAGE: MEAN OF FEATURES PER CLUSTER. In the case of *cluster2*, it is observed that the mean is higher in practically all symptoms except *Alopecia* and *Genital thrush*. Therefore, it can be affirmed that the patients belonging to that cluster present more symptoms related to diabetes than the patients in *cluster1*. Moreover, it is observed that variables *Polyuria*, *Polydipsia* and *partial paresis* present the highest means for this cluster. In previous studies it was found that when these three symptoms were had together, patients were very likely to have diabetes. Due to this, it seems that patients in *cluster2* (blue) are more likely to have diabetes than those in *cluster1*.

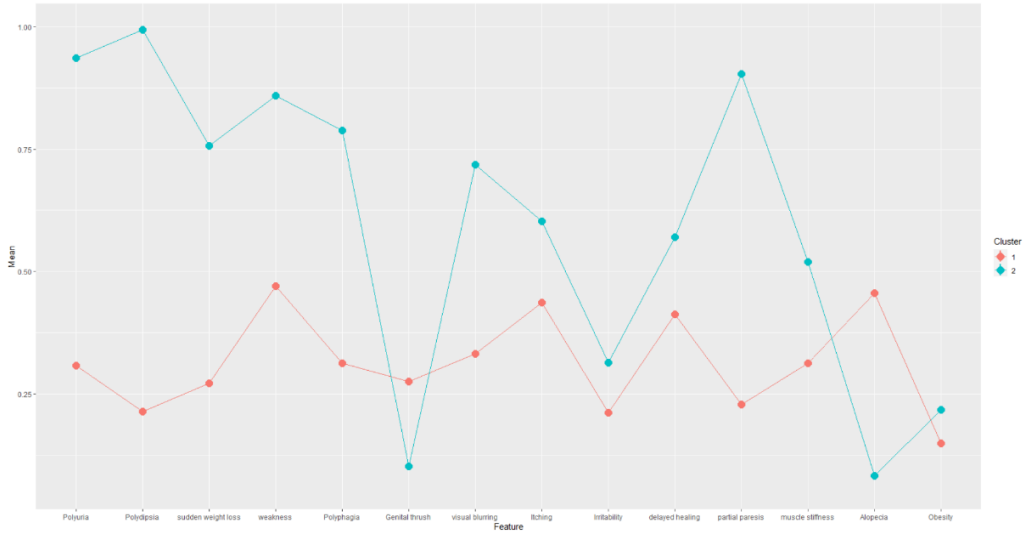

Figure 13. Ward linkage: mean of features per cluster

5.2. Partitional (K-means, 3 clusters)

As in the previous case, in FIGURE 14. K-MEANS ALGORITHM: MEAN OF FEATURES PER CLUSTER a cluster with the highest means in most symptoms (red) and a cluster with the lowest means in most symptoms (blue) can be seen. In addition, it is observed that there is a cluster (green) whose means are found between the two previous clusters in the majority of symptoms (they are not lower than any mean of the blue cluster and they are higher than some means of the red cluster in some symptoms).

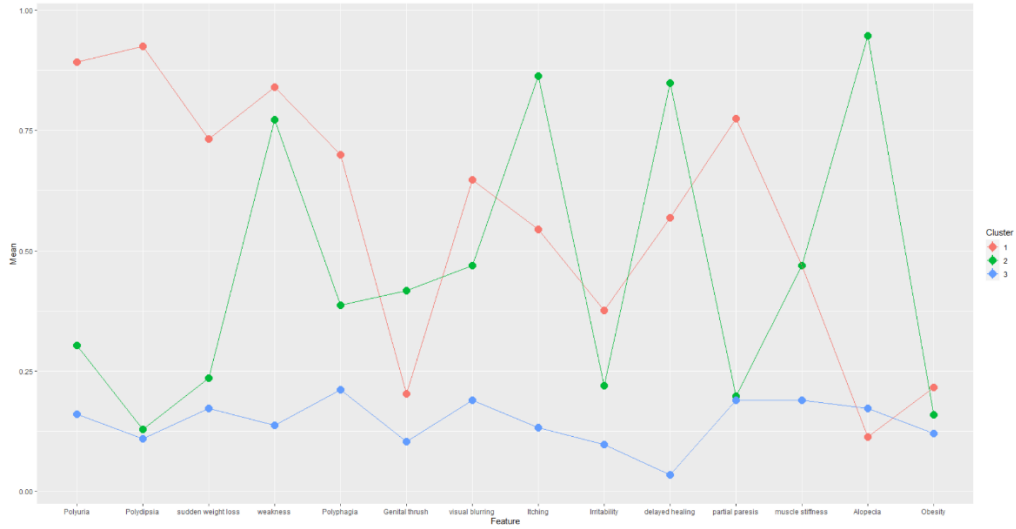

Figure 14. K-means algorithm: mean of features per cluster

Specifically, it is observed that the green cluster has higher means than the red cluster in the variables: *Genital thrush*, *Itching*, *delayed healing* and *Alopecia*. Previous studies found that these symptoms were not especially significant in most of the models when predicting diabetes. Therefore, it seems that in the **blue cluster** are the patients who present fewer symptoms and, therefore, are less likely to have diabetes, in the **green cluster** it seems that there are those patients who present more symptoms related to diabetes but perhaps the symptoms they present are not influential or decisive enough to know with any certainty whether they have diabetes or not, and in the **red cluster** it seems that there are those patients who present the most decisive or influential symptoms when it comes to predicting diabetes and, therefore, they are very likely to have diabetes.

5.3. Probabilistic (EM, 2 clusters)

Attribute	0 (0.61)	1 (0.39)
Polyuria		
mean	0.7891	0.0364
std. dev.	0.4079	0.1872
Polydipsia		
mean	0.7336	0
std. dev.	0.4421	0.0003
partial paresis		
mean	0.6127	0.1452
std. dev.	0.4871	0.3523
sudden weight loss		
mean	0.5828	0.1575
std. dev.	0.4931	0.3643
weakness		
mean	0.7322	0.3579
std. dev.	0.4428	0.4794
Polyphagia		
mean	0.6326	0.1783
std. dev.	0.4821	0.3828
Genital thrush		
mean	0.2531	0.176
std. dev.	0.4348	0.3808
visual blurring		
mean	0.6105	0.1932
std. dev.	0.4876	0.3948
Alopecia		
mean	0.3175	0.3862
std. dev.	0.4655	0.4869
Obesity		
mean	0.2145	0.0982

In this case, it is observed in FIGURE 15. EM ALGORITHM: MEANS OF FEATURES PER CLUSTER that *Cluster0* presents much higher means in the variables *Polyuria* and *Polydipsia* (very influential symptoms). In addition, it presents higher means (but not with a difference as big as with the two variables mentioned above) in the variables: *weakness*, *visual blurring*, *Polyphagia* (not influential symptoms at all). Finally, there are variables whose mean is similar in both clusters: *Genital thrush*, *Alopecia*, *Obesity* (symptoms that do not seem influential or important). As in the K-means clustering with *k = 2*, it seems that there is a cluster that encompasses patients who present all or most of the symptoms and, therefore, it is very possible that they have diabetes (*Cluster0*) and there is another cluster with patients who have little or no symptoms and are therefore less likely to have diabetes (*Cluster1*).

Figure 15. EM algorithm: means of features per cluster

6. Conclusion

In conclusion, after having applied all the clustering algorithms, if I had to choose a model to apply in real life, I would choose the K-means model with *k = 3* since, unlike the rest of the models, there is a cluster that considers those doubtful cases of patients who present symptoms related to diabetes but is not clear if they really have it, as often happens in the health domain, where the same or similar symptoms can occur for many diseases. Therefore, I consider it is important to have a cluster that encompasses these patients to pay more attention to them and avoid possible tragedies.

Finally, like in the supervised classification studies, maybe it is interesting to balance the dataset by adding more instances in order to find more differences between the patients and make the groupings more precisely. 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 help to find patterns 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 this study.

7. References

[1] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

[2] 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.

[3] UCI Machine Learning Repository *Early stage diabetes risk prediction dataset.* https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

[4] Diabetes, World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/diabetes

[5] The Cost of Diabetes, American Diabetes Association (ADA). https://www.diabetes.org/resources/statistics/cost-diabetes

[6] Importance of Early Diabetes Diagnosis and Screening. https://www.apollodiagnostics.in/blog/importance-of-early-diabetes-diagnosis-and-screening