COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Introduction to Statistical Learning

Author:
Hermine Tranie (CID: 01400919)

Date: February 25, 2021

1 Introduction of dataset

In this coursework, the goal is to apply distance based methods from lectures to a data set of our choice.

1.1 Rationale

Around 10% of adults in the world have diabetes, and this number is increasing everyday. The implications of the development of the disease on the body can lead to premature deaths. However the earlier the diagnosis, the better chances of longevity in life expectancy, as symptoms can be controlled. This is why I have decided to pick a dataset from the US National Institute of Diabetes and Digestive and Kidney Diseases which focuses on Pima Indians diabetes. [1] This dataset is looking at women who are at least 21 years old. The goal of this analysis is to look for particular symptoms that help with the diagnosis of this disease, in women, with a particular feature related to gestation. The space visualisation that CMS shows, allows you to see, when you have a new patient, how he/she compares to the others, and see where he/she stands in the space.

1.2 Dataset structure

This dataset is comprised of 768 patients with 8 features each (see Figure 1): 1) number of pregnancies the patient had, 2) plasma glucose concentration after 2 hours in an oral glucose tolerance test in mg/dl, 3) diastolic blood pressure in mm Hg, 4) triceps skin fold thickness in mm, 5) 2-hour serum insulin (mu U/ml), 6) BMI: body mass index (kn/m2), 7) diabetes pedigree function, 8) age in years. Clearly, all these features have different units and scales so we are going to standardize them, in order to be able to compare them.

2 Multidimensional Scaling (CMS) subject to different distance metrics

Now, we are using multidimensional scaling in order to visualize the distances between our feature data points. This method takes into account the similarity between pairs of points, such as the distance, in order to estimate their location in space between each other.

2.1 Eigenvalue plots with multiple distances

CMS can be done by analyzing which eigenvectors drive the data (the ones with highest eigenvalues) and then choose the dimension of the scaling based on this number. This is why we are first of all looking at eigenvalue plots. This is extremely useful for us because the visualization will tend to be closer to reality with scaling. I tested a few metrics including: Euclidean, Manhattan, Minkovwki and Canberra distances for the eigenvalue plots. By looking at eigenvalues, from the Euclidean plot first, there was a significant amount of negative ones which lead to believe of the presence of non euclidean error and hence try other metrics such as Canberra.

The distance metric which gave the most 2 standout eigenvalues was Canberra (see Figure 2). It is ideal to have this number of eigenvalues standing out because it means we can plot easily the data in 2D without compromising its overall meaning and it's easy to visualize data in 2 dimensions. I will be comparing to standard Euclidean distance (see Figure 3) throughout this analysis. Finally, I plotted the log abs eigenvalues (see Figure 4) on the Euclidean distance to confirm my 2D hypothesis. Indeed this plots gives us a magnitude criterion which is to reject any eigenvalue that's not greater than the magnitude of the last one. We can clearly see that a small number of eigenvalues pass this criterion so choosing dimension 2 is fine, as it's a bit more refined.

2.2 CMS with Euclidean and Canberra distances

Now that we have the dimensionality of 2, we can apply the classical multidimensional scaling to look at the similarity of these data points in space between each other. When looking at the classical multidimensional scaling plot on Canberra distance (see Figure 6), we can already see some pattern as to how the distances between our data points are made for the corresponding leading eigenvalues. We can compare this to the Euclidean CMS (see Figure 5), which is separating much less our data points, making it quite hard to interpret.

3 K-means clustering

Now, let's perform our clustering algorithm: K-means, with the goal of findings clusters of data points to predict diabetes or not with the input features. Once again, our Canberra distance is performing quite well (see Figure 8) with a clustering pattern emerging. Although we can see some triangles (from cluster number 2) in the area where there is mainly circles (from cluster number 1) it is a much better cluster than using Euclidean distance (see Figure 7) which is not separating as well our data points. One can note that it seems that it is easier to spot the non diabetes patients (in red) rather than the ones with diabetes (in blue), this is the case because in the population, as stated in the introduction, there are much more patients who don't have diabetes rather than do.

4 Self-Organizing Map

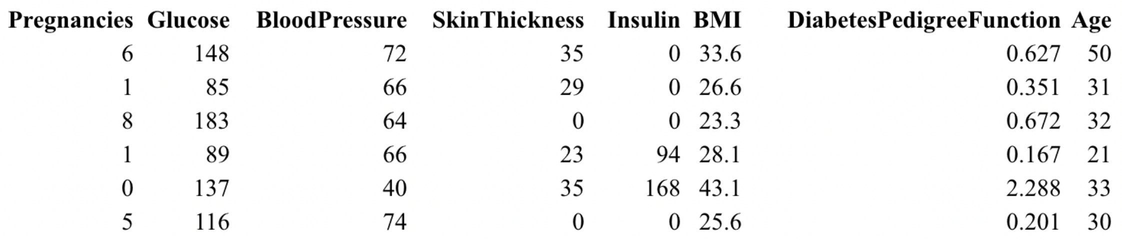
Finally, let's plot a self organizing map (see Figure 9) which allows to display some clustering in the data, which comes to confirm our hypothesis from our K-means clustering algorithm.
The bottom right corner and top left corner seem to be showing different features and clustering. The lower the values of our features it seems on top left corner go together with the non diabetes diagnosis. The higher the values of our features on the bottom right corner go with the diabetes diagnosis.
One can note that the age (in what) seems to not be playing a gigantic role in the clustering whereas pregnancies, glucose and blood pressure seem to have high importance. To emphasize on the choice on this data set with the selection of women patients, we could hence mention the gestational diabetes. Indeed as mentioned, it seems like the number of pregnancies is one of the leading factors by the SOM plot, along with other more straight forward ones as mentioned above.

5 Conclusion

In this coursework, we have explored distance based method for data visualisation (CMS) and clustering (k-Means) which have shown us some classes of data going together. The points scaled with Euclidean distance are much less clustered, but as we've show with the eigenvalue plots, there is the presence of non euclidean error, so this is why we are exploring Canberra distance too. With the latter, our clustering is quite good, whilst having some triangles in the circle area and vice versa, which is expected anyways for real life data set. We could think about looking at outliers before our analysis for example, to optimize our clustering. Finally, doing SOM analysis shows some clustering as well with the most important features (glucose, blood pressure, pregnancies) influencing the model by enforcing neighborhood relationships on the resulting cluster centroids. [2]

6 Figures

All the figures mentioned in the core analysis are displayed below:



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30

Figure 1: Dataframe for Pima Indians diabetes

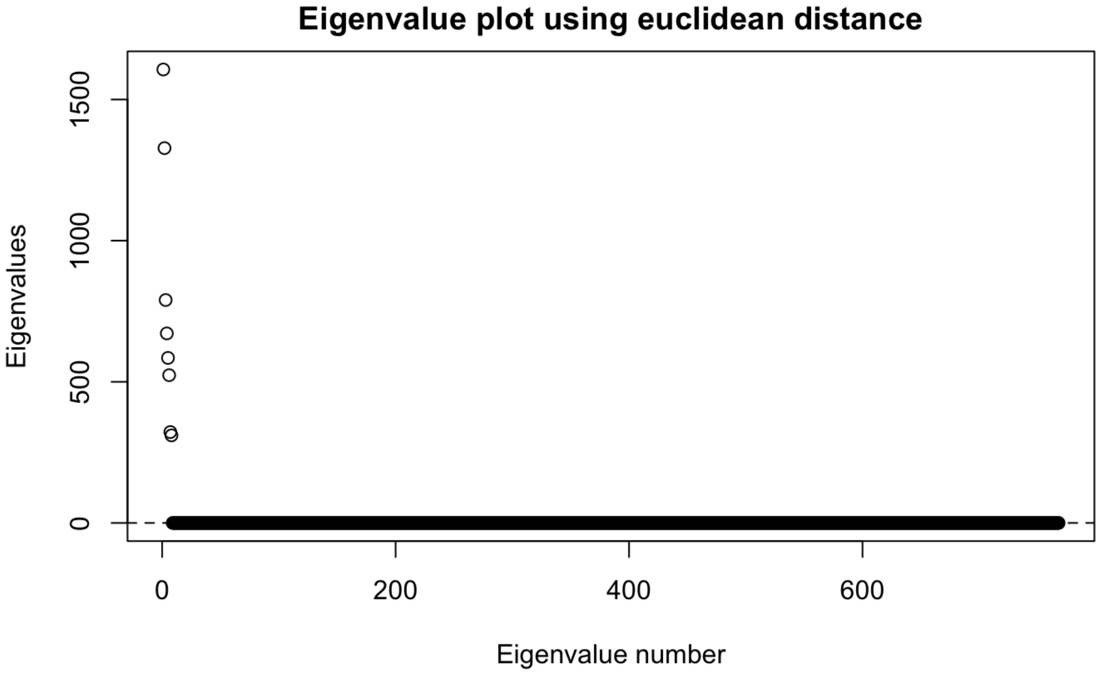


Figure 2: Eval plot for Euclidean distance

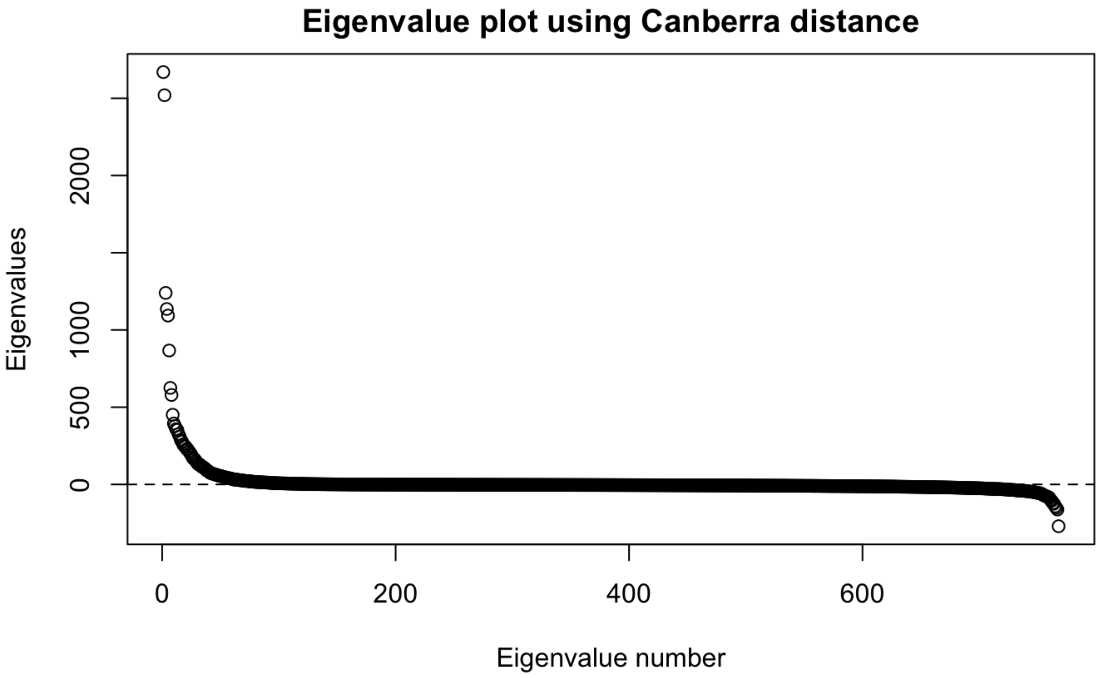


Figure 3: Eval plot for Canberra distance

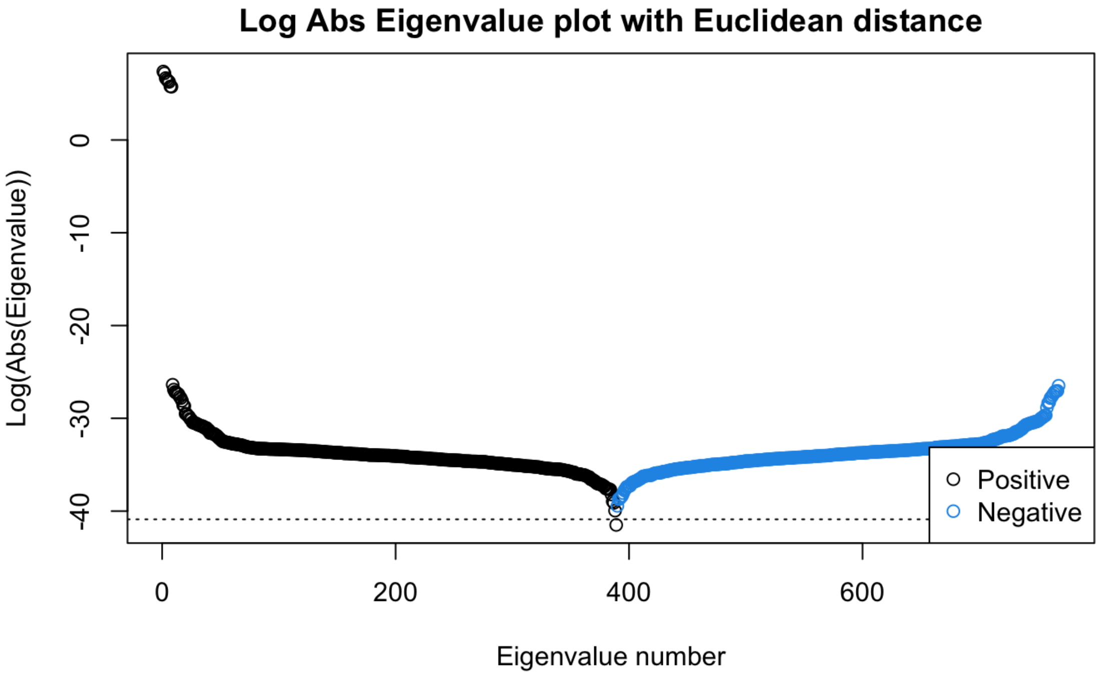


Figure 4: Log Abs Eval plot for Euclidean distance

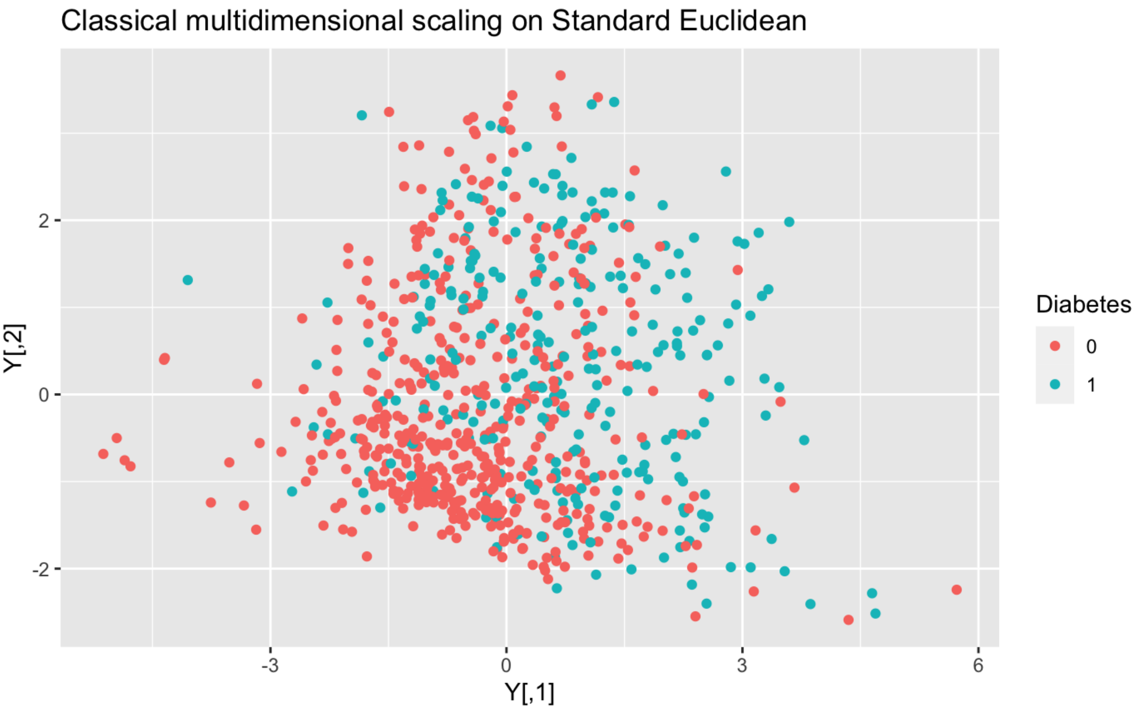


Figure 5: CMS lot for Euclidean distance

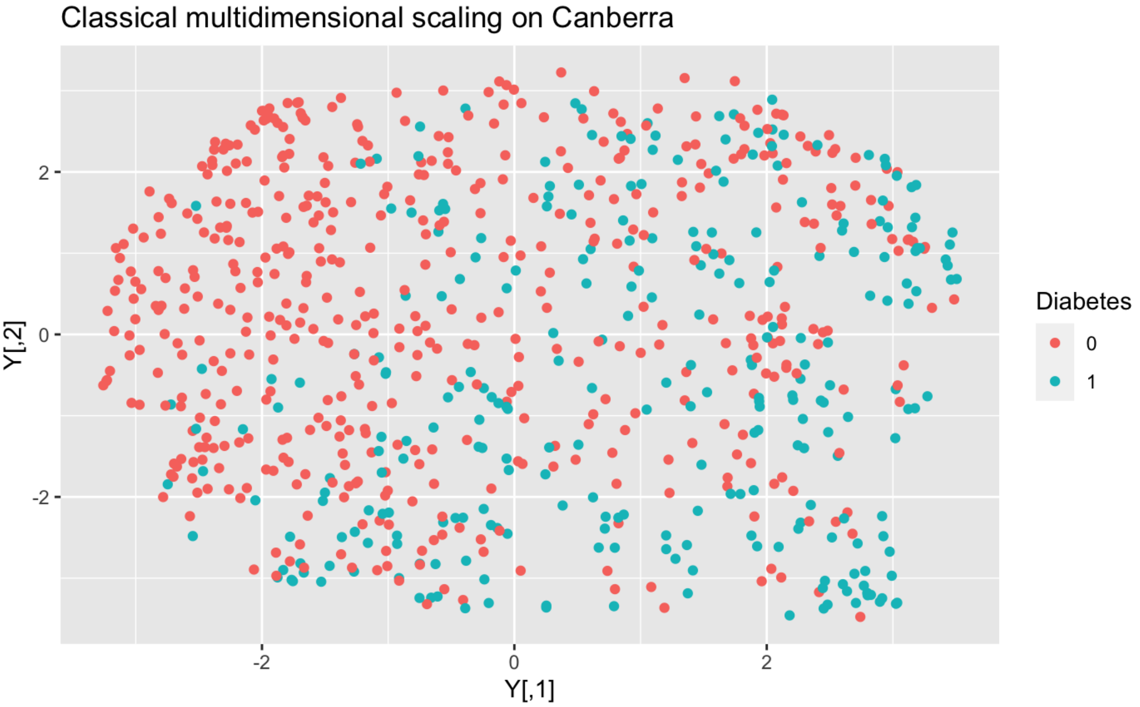


Figure 6: CMS plot for Canberra distance

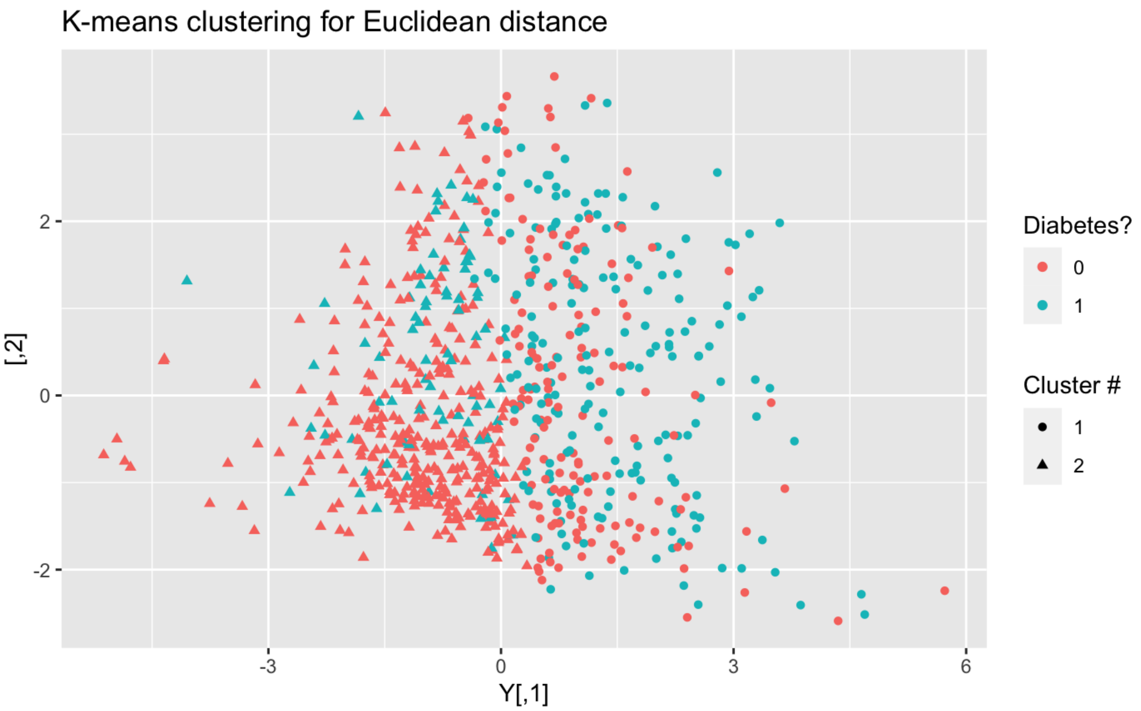


Figure 7: kMeans clustering for Euclidean distance

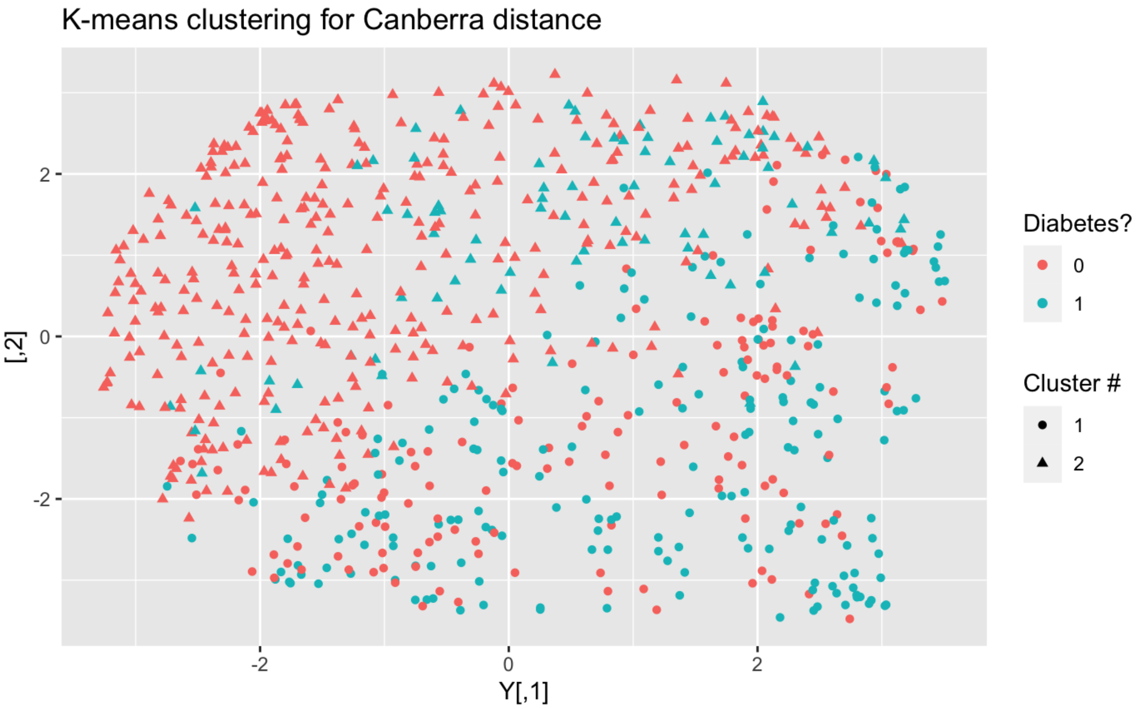


Figure 8: kMeans clustering for Canberra distance

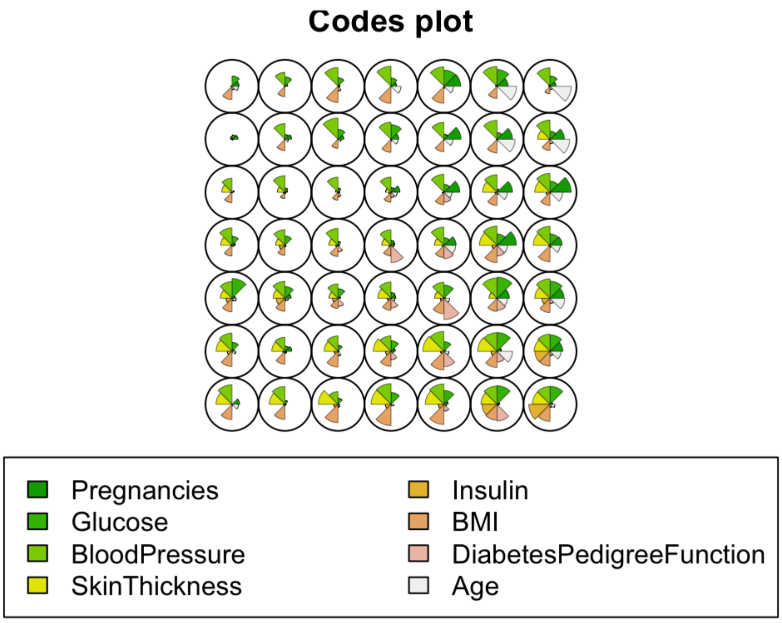


Figure 9: SOM analysis

7 Reference

[1] Pima Indians Diabetes dataset found on: `https://www.kaggle.com/uciml/pima-indians-diabetes-database`
[2] Self Clustering Organizing maps found on: `https://en.m.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Self-Organizing_Maps_(SOM)`