Report #0003

DIABETES

Anomalies search

May 9th, 2023

Executive Summary

Four clusters were obtained that achieve good performance, with elements similar to each other and different from those of other clusters.

The interpretation of each cluster, based on the results obtained, is as follows:

Cluster 0 has a demographic of highly educated, single males who are likely to be working professionals.

Cluster 1 has a demographic of low-educated, married individuals who are mostly unemployed.

Cluster 2 has a demographic of highly educated, single males who are likely to be self-employed entrepreneurs.

Cluster 3 has a demographic of low-educated, married individuals who are likely to be employed in medium-sized cities.

The recommendations for the Noche supermarket marketing team are as follows:

1-Consider the types and interests particular to each of the clusters

2-Consider clusters 0 and 2 as possible premium audiences, and depending on spending, offer special offers intrinsic to the particular taste of each of the groups.

3-Make more affordable offers for audiences 1 and 3, as well as offers on products related to couples, or more commonly consumed by couples.

5-Consider future analyses, with more experimentation at the time of segmentation, and possible corrections in the data taken as mentioned in the quality problems section.

Table of Contents

Content

Introduction

, the problem and the procedures carried out to achieve the stated objective.

Discussion

Data understanding

The source of the data for this project is a file named "diabetes.csv", which is housed within a Google Drive folder that was provided by the professor.

Attached to this file there is also a table 1 which briefly explains each attribute.

Variable	Type	Values	Description
Pregnancies	Numeric	Integer	Number of times the person was pregnant.
Glucose	Numeric	Real positive	Plasma glucose concentration at 2 hours in oral glucose tolerance test.
BloodPressure	Numeric	Real positive	Diastolic pressure (mm Hg).
SkinThickness	Numeric	Real positive	Tricep skinfold thickness (mm).
Insulin	Nominal	Real positive	Serum insulin at 2 hours (muU/ml).
BMI	Numeric	Real positive	Body mass index (weight in Kg / height in meters squared).
DiabetesPedigree Function	Numeric	Real positive	Diabetes probability function based on family history.
Age	Numeric	Integer	Age in years.
Outcome	Nominal	Boolean	People with positive or negative diabetes. 0 = negative / 1 = positive.

Table 1- Description of variables.

As we can see, the dataset provides information about National Institute of Diabetes and Digestive and Liver Diseases Patients, including their outcome, age, plasma glucose concentration, diastolic pressure, triceps skinfold thickness, insulin, body mass index, probability of diabetes based on family history and number of times the person was pregnant. This information can be

used to determine rules that make it possible to diagnose this population sector in an efficient way.

The dataset consists of 770 observations.

The attribute outcome was loaded in rapidminer as binomial type.

Units are described in table 1.

There is one duplicate patient.

For the following histograms, where applicable the number of bins was determined using the "Sturge's Rule", due to the number of observations. In this case 11 is the number of bins.

Quality problems

Although the dataset does not show missing values at first glance, when analyzing the attributes "Glucose", "Insulin", "SkinThickness", "BloodPressure" and "BMI", we were able to observe the large number of patients who have a 0 value recorded in these attributes and who actually indicate a missing value.

In the attributes "Age" and "Pregnancies" there are two values, one in each attribute, which could be considered as noise, since there is no possibility of being 390 years old, nor of having had 40 children at the age of 30.

Values of the attribute "SkinThickness" greater than 60 could be noise.

The attribute "DiabetesPedigreeFunction" has two extremely isolated values.

A row or patient has the value 78 in all attributes except for the attribute "Outcome".

Data preparation

Firstly, the duplicate row has been removed , outcome was transformed from binomial to numerical and patients whose value in any of the attributes "Glucose", "Insulin", "SkinThickness", "BloodPressure" or "BMI" was 0 were transformed to missing in RapidMiner.

Of these, 5 for glucose, 35 for blood pressure, 226 for skin thickness, 374 for insulin, and 11 for BMI were found to be missing.

Once these missing values were discarded, and considering that most of them were in skin thickness and insulin, the 393 complete observations were used to generate a linear regression for these two attributes, since these were suitable for the regression to be carried out.

In this process, performance tests were performed, along with cross validation to make sure that the regression model was performant enough.

The results of the Skin thickness model:

Performance:

- root_mean_squared_error: 8.623 +/- 1.639 (micro average: 8.772 +/- 0.000)

Linear Regression:

- 0.505 * Outcome = 1 - 0.505 * Outcome = 0+ 0.314 * Pregnancies+ 0.009 * Glucose+ 0.002 * BloodPressure+ 0.969 * BMI+ 0.083 * DiabetesPedigreeFunction- 0.004 * Age- 4.396

The results of the Insulin model:

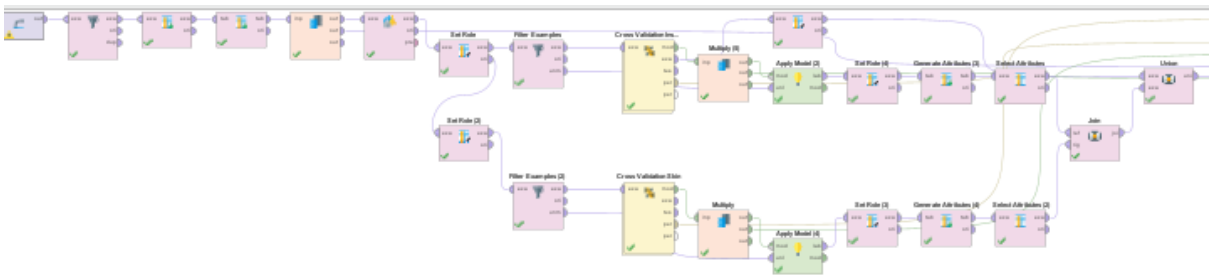
Performance:

- root_mean_squared_error: 110.177 +/- 44.710 (micro average: 117.971 +/- 0.000)

Linear Regression:

- 0.904 * Outcome = 1 - 0.905 * Outcome = 0- 0.789 * Pregnancies + 2.108 * Glucose - 0.460 * BloodPressure + 2.220 * BMI - 0.942 * DiabetesPedigreeFunction + 0.023 * Age- 139.262

Finally, once the model has been applied to the missing values for the attributes that were regressed, the dataset unified, and the remaining missing values discarded, the total number of observations is 723.

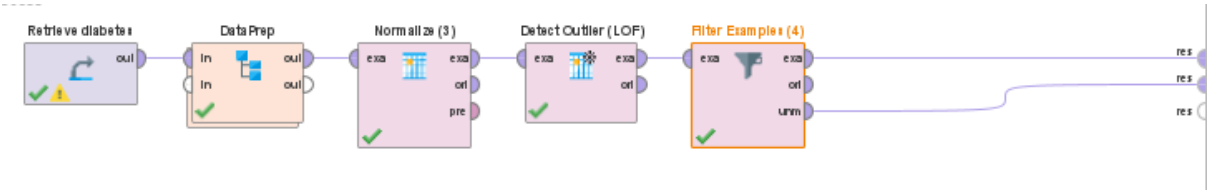


Outliers

First of all, a normalization was carried out due to the different units that handle the attributes.

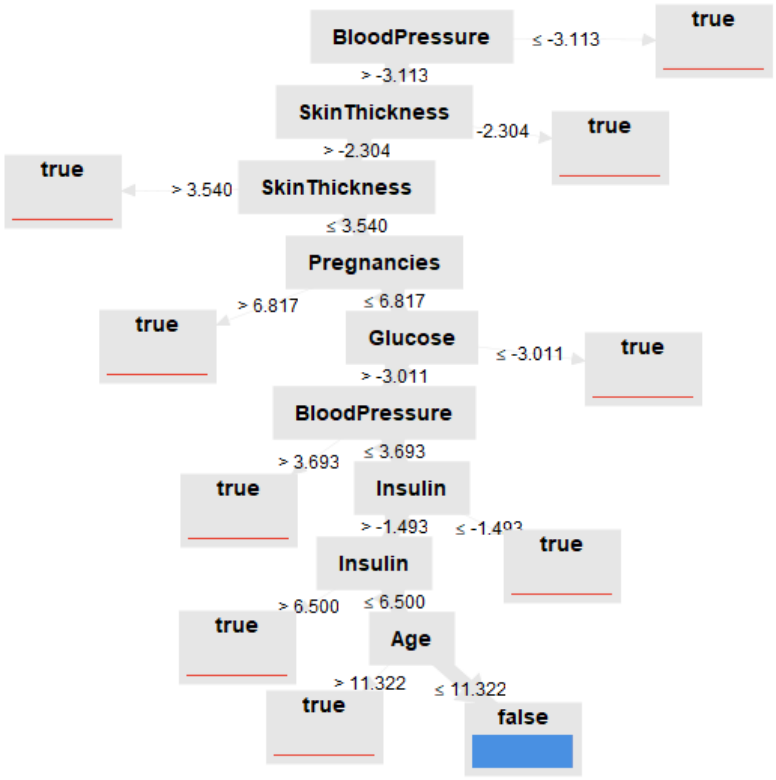
The Local Outlier Factor (LOF) algorithm was used to find outliers in the dataset. LOF is an unsupervised anomaly detection method that calculates the local density deviation of a given data point relative to its neighbors. He considers as outliers the samples that have a substantially lower density than their neighbors.

The number of neighbors considered (parameter n_neighbors) is normally set to (1) greater than the minimum number of samples that a cluster must contain, so that other samples can be local outliers with respect to this cluster, and (2) less than the maximum number of nearby neighbors for samples that can potentially be local outliers. In practice, this information is often not available, and taking n_neighbors=20 seems to work well in general.



Those rows with outlier scores greater than 2 were discarded, these were 14, so the resulting dataset consists of 709 observations.

After this, we proceeded to make a decision tree to detect these outliers.



accuracy: 98.48% +/- 1.38% (micro average: 98.48%)

	true false	true true	class precision
pred. false	705	7	99.02%
pred. true	4	7	63.64%
class recall	99.44%	50.00%	

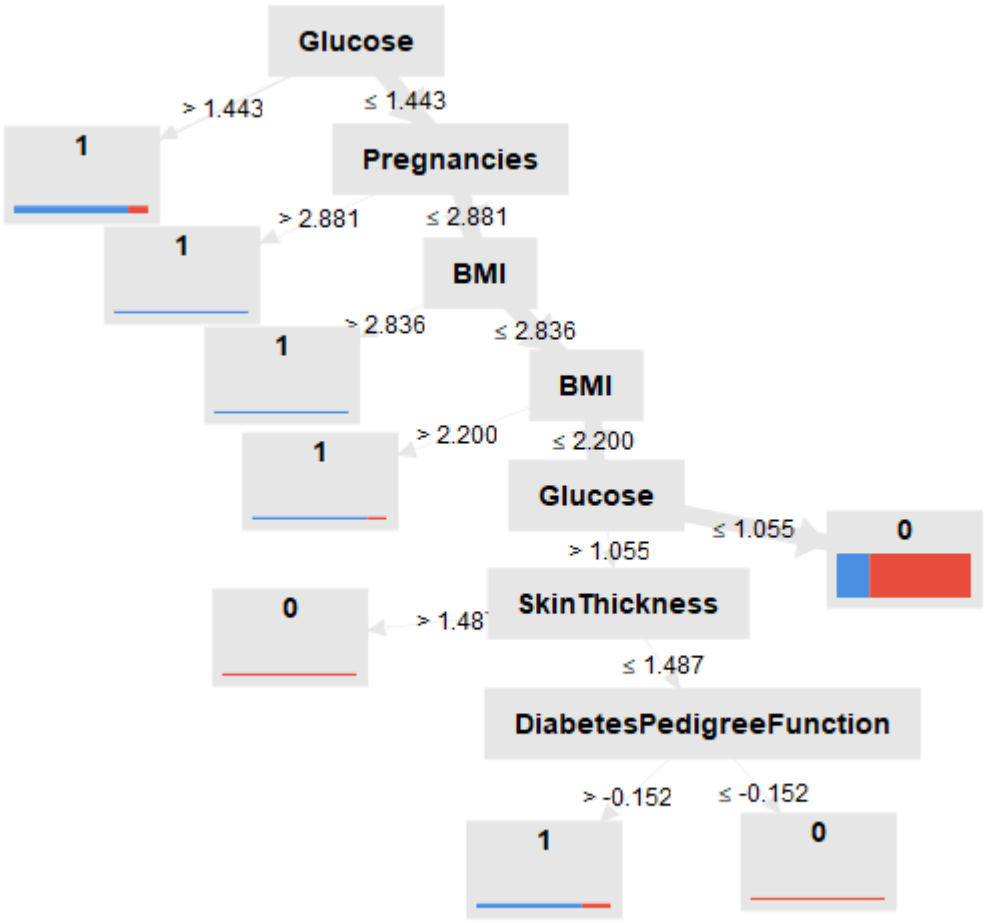
Tuples to be analyzed

During the preparation process, an id was assigned at the beginning, this can be seen if you open the process attached to this document. This is the list by id, of those tuples that must be analyzed and corrected by an operator.

ID
14
19
118
126
255
282
329
446
521
599
676
11
107
539
8
10
16
50
61
63
73
76
79
82
146
173
183
194
223
262
267
270
301
333
337
343

348
350
358
372
427
431
436
454
469
485
495
504
524
535
537
591
603
606
621
645
687
700
706
709

Diacnostic of diabetes



accuracy: 73.56% +/- 1.97% (micro average: 73.56%)

	true 1	true 0	class precision
pred. 1	77	21	78.57%
pred. 0	167	446	72.76%
class recall	31.56%	95.50%	

Modeling

Finally, to help in the diagnostic process, a decision tree was developed, we seek to identifying two types of patients, those who have diabetes and those who do not, with the data considered good.

Validation and feedback

The performance of the model, was evaluated using the Cluster distance performance process as shown in Fig 16, evaluating it for the criteria:

-Avg. within centroid distance

-Davies Bouldin

Because the first measures how close the data points within each cluster are to the centroid of that cluster, and the second measures the similarity between clusters. Better results for these two metrics generally indicate better clusters.

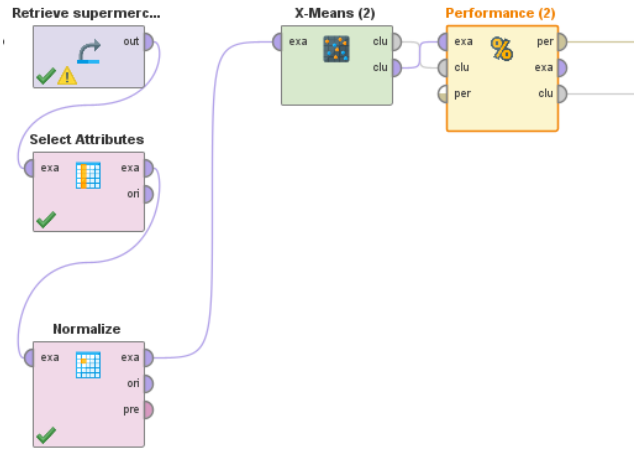


Fig 1-Persormance process

Avg. within centroid distance: -5.255

Avg. within centroid distance_cluster_0: -6.042

Avg. within centroid distance_cluster_1: -4.638

Davies Bouldin: -1.640

Although there is a clear separation between the clusters, the experimental results obtained with X-means still show some mixtures between the variables. Additionally, while the achieved segmentation is good, the marketing team could benefit from a larger number of clusters that would allow for even more specific advertising. Therefore, to address this concern, a second approach will be adopted that involves selecting a larger number of

clusters using techniques that strike a balance between cluster quantity and quality.

Conclusions

In general, first approach is quite computationally convenient, and has a good segmentation that could be used in a broader marketing campaign.

However, for the use of market intelligence, the second approach seems more interesting. This is because one of the most common ways of performing market segmentation is how much they spend with the company, in this case, how much they have in annual revenue (Gustavo Gomes, “3 ways to classify customers to guide strategies”). This is because it is possible to make a correlation with the Pareto Principle, and thus create differentiated marketing campaigns for each level of audience, or even special campaigns for 'premium' audiences.
So in order each cluster seems to represent:

-Cluster 0 has a demographic of highly educated, single males who are likely to be working professionals.

-Cluster 1 has a demographic of low-educated, married individuals who are mostly unemployed.

-Cluster 2 has a demographic of highly educated, single males who are likely to be self-employed entrepreneurs.

-Cluster 3 has a demographic of low-educated, married individuals who are likely to be employed in medium-sized cities.

Recommendations

1-Consider the types and interests particular to each of the clusters

2-Consider clusters 0 and 2 as possible premium audiences, and depending on spending, offer special offers intrinsic to the particular taste of each of the groups.

3-Make more affordable offers for audiences 1 and 3, as well as offers on products related to couples, or more commonly consumed by couples.

4-Consider future data of average expenditure made in the market and the number of children. Thus, it is possible to send even more targeted offers to audiences with a higher propensity for acceptance.

5-Consider the more in-depth specifications for each data explained in the 'Quality Issues' section

References

Çankaya, M. F. (2022, March 22). *engineering.teknasyon.com*. Retrieved from How To Normalize Your Unsupervised Data For Clustering Methods: https://engineering.teknasyon.com/how-to-normalize-your-unsupervised-data-for-clustering-methods-9389298d20d5

Ergen, B. (2022, Nov 25). *Linkedin*. Retrieved from How to detect the strongest outliers with Local Outlier Factor ?: https://www.linkedin.com/pulse/how-determine-strongest-outliers-local-outlier-factor-b%C3%BCnyamin-ergen?trk=public_post

Gareth James, D. W. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Gomez, G. (n.d.). *Blog do Agendor*. Retrieved from 3 formas de classificar clientes para direcionar estratégias: https://www.agendor.com.br/blog/formas-de-classificar-clientes/

Gupta, A. (2021, May 27). *medium.com*. Retrieved from X-Means — A Complement to the K-Means Clustering Algorithm: https://medium.com/geekculture/x-means-algorithm-a-complement-to-the-k-means-algorithm-b087ae88cf88

Jeffrey W. Lockhart, P. (2022). *Gender, Sex, and the Constraints of Machine Learning Methods*. Department of Sociology, University of Chicago.

Jiawei Jan, M. K. (2012). *Data mining concepts and techniques*. 225Wyman Street,Waltham, MA 02451, USA: Elsevier Inc.

Larson, B. N. (n.d.). *Proceedings of the First Workshop on Ethics in Natural Language Processing*. 686 Cherry St. MC 0165, Atlanta, GA 30363 USA.

Markus M. Breunig, H.-P. K. (2000). LOF: Identifying Density-Based Local Outliers.

Data

The data for this project “diabetes.csv” was given by the teacher, if needed you should contact him.