A – Data Profiling

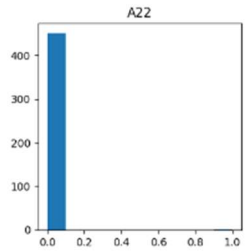
1 – If we have a dataset with 450 records and 600 variables we are in the presence of the curse of dimensionality

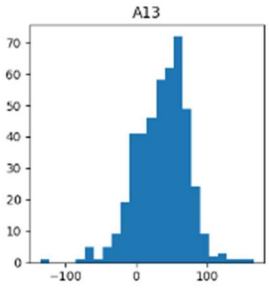
2 – Larger Bins have higher frequencies

3 – We say that two variables are redundant when it has always the same value

4 – Histograms represent the 5-number Summary

5 – If we change the granularity from year to month, we make the granularity more coarse





B – Data Preparation

1 – We should use feature generation in the two variables depicted in the histograms

2 – Using discretization in A22 would bring better results

3 – In feature engineering we reduce the complexity of data to create better variables

4 – We should use Backward selection if there are few irrelevant variables

5 – If we have 100 instances of a class and 1000 of another we should use undersampling

C – Classifiers Evaluation

1 – We call negatives to the minority class

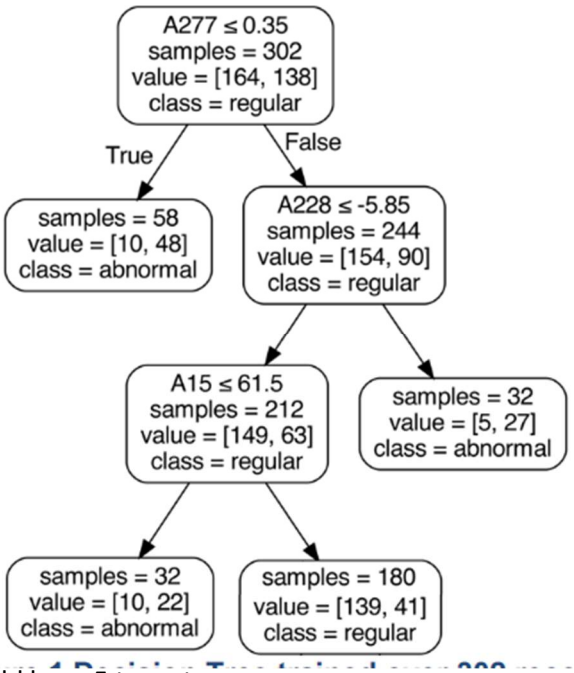
2 – The recall for the given tree is above 50%

3 – The precision for this tree is below 40%

4 – We have more TN than TP in the given tree

5 – The number of False Negatives is 41

(Assume that they occur when <=, A277 occurs if A277<=0.35 and a sup threshold of 10%)



D – Classification

1 – 30-NN would classify (~A277, A228,A15) as abnormal

2 – Naïve Bayes would classify Z = (~A277, A228) as regular

3 – P(A277|Regular) is 10/164

4 – Using the tree above, if we applied Random Forests we would have 5 tree stomps

5 – If we used the 2 best features we can say that A15 is one of them

CD – Dummy Exam 1

E – Pattern Mining

1 – ~A277 -> A228 has a support of 244/302

2 – A277 is a frequent itemset

3 – ~A277 -> ~A228 confidence is above 30%

4 – Any superset with A277 is frequent

5 – If a superset contains ~A228 it's not frequent

DTW	x_1	x_2	x_3	x_4
x_1	0	$\sqrt{9}=3$	$\sqrt{2}=1.4$	?
x_2	-	0	$\sqrt{3}=1.7$	$\sqrt{6}=2.4$
x_3	-	-	0	$\sqrt{2}=1.4$
x_4	-	-	-	0

1	5	2	1
2	4	1	2
0	0	9	10
0	0	9	10
	0	3	1

In F and G use: x1 = < 0, 0, 2, 1>, x2 = <2, 2, 0>, x3 = <1,2,1,1>, x4 = <0, 3, 1>

F – Clustering

1 – The DTW distance between x1 and x4 is given by the 2 figure

2 – The DTW distance between x1 and x4 is sqrt(1)

3 – The dendogram using a minimum link criterium would be {{x1,x4},x2,x3}

4 – Cohesion is an intra-cluster similarity measure

5 – The sillouete value for x1 given {x1,x2} {x3,x4} is -1.8/3

1 – The larger the Dunn index the better

G – Time Series

1 – The DTW path between x1 and x4 is (1,1)(2,1)(2,2)(3,3)

2 – A time series is stationary if its mean variance and autocorrelation does not change over time

3 – DK fuller test is not suitable to test the stationarity of time series

4 – PAA represents the time series as a sequence of box basis function with all boxes with fixed sizes.

5 – If we had a model that **never** fails its predictions it would predict that Catarina is going to **pass** the exam

H – Social Networks

1 – Social networks show an average diameter of seven

2 – The node owner can influence the centrality of its node in a social network

3 – The centrality of a node is given by Oi/n, being Oi the outgoing edges and n the number of nodes

4 – A node that has more ingoing edges is said to have more prestige

5 – The transition probability matrix in the PageRank Algorithm only has to be irreducible

I – Ethical Concerns

1 – The central entities on the GDPR are the data subject and the controller

2 – Data can be further processed under public interest

3 – Privacy is not a concern for GDPR

4 – GDPR is worried with adequacy

5 – According to GDPR it's allowed to collect religious beliefs along with the data

J – Deloitte Case Study

1 – We dealt with the curse of dimensionality in the Deloitte case study

2 – We computed the target variable

3 – We know if a client has churned if the sum of all variables is 0 before feature selection

4 –We created a false predictor

5 – We applied SMOTE to get the best results in decision trees