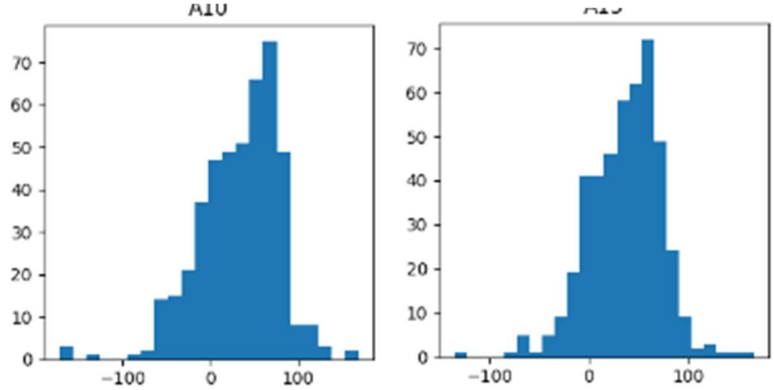
CD – Dummy Exam 2

A – Data Profiling



1 – A variable is a **false predictor** if it gives the same information as any other variable

2 – The Hughes phenomenon says that the accuracy of a model increases with the increase of the number of dimensions until it reaches a maximum value, then it starts decreasing.

3 – The two variables in the Histograms are redundant

4 – Using the Occam's Razor method means picking the simplest model

5 – We should use histograms to identify correlation between variables

B – Data Preparation

1 – We call feature selection when we use the width and height to calculate an area

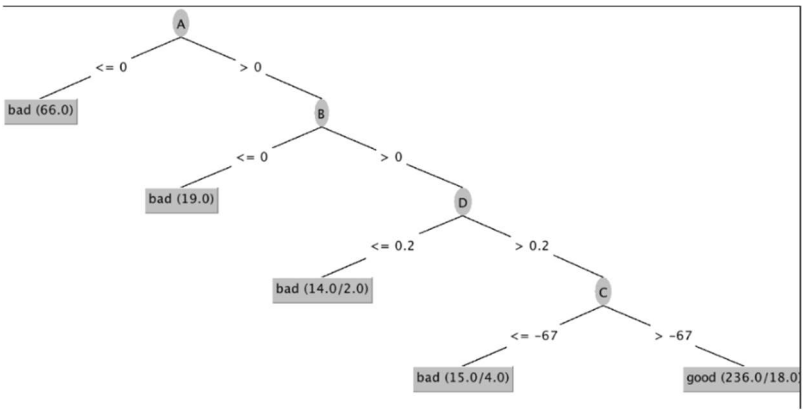
2 – If we have a dataset with lots of correlated variables we should use Sequential Forward Selection

3 – Feature selection has no impact on Naïve Bayes

4 – We use LDA for unsupervised data instead of PCA

5 – Dummification transforms symbolic data into a set of binary variables

C – Classifiers Evaluation



1 – The specificity for the tree is above 50%

2 – The accuracy for the tree is above 50%

3 – The error for this tree is above 40%

4 – We have more FN than FP in the given tree

5 – Distance matters in Naïve Bayes

CD – Dummy Exam 2

D – Classification

1 – In 3-NN if we have 4 equally-distant neighbors where 3 are positive and 1 is negative we will use 2-NN to classify the point

2 – Holdout is used for small datasets and leave-one out for big ones

3 – Boosting uses a majority vote system without assigning weights

4 – If we have a 4-NN classifier with 2 positive and 2 negative closest neighbors we will use 3-NN to classify it

5 – Bagging is less prone to overfitting and better at dealing with noise

E – Pattern Mining

1 – If an itemset is not frequent a superset that contains it will also be not frequent

2 – The lift of A->B is given by P(B|A)/P(A)

3 – If we find the patterns CGB AOPS SBDH we say that they are all maximal patterns

4 – In the patterns in 3 the only closed pattern is CGB

5 – The apriori algorithm generates all possible combinations in each step

F – Clustering

1 – The larger the Dunn index the better

2 – Separability is an intra-cluster similarity measure

3 – DB index measures how compact the clusters are, the smaller the value the better

4 – In k-means the centroid is an existing point

5 – We want big silhouette values

G – Time Series

1 – Applying a smoothing average can be compared to increasing the granularity of a TS

2 – The difference between MAE and MSE is that the first one uses squares of the error

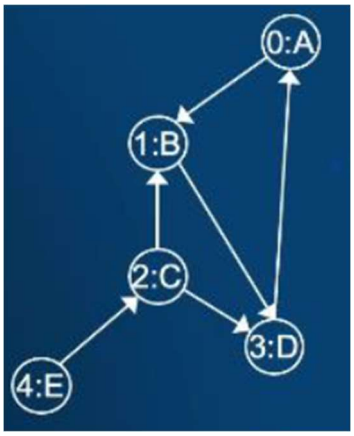
3 – Seasonal component usually uses fixed periods of time less than an year while Cyclic uses periods of time larger than an year.

4 – SAX consists in segmenting a time series to **PAA** and then convert each segment to a symbol.

5 – In SARIMA P is the order of the AR model, D the number of times to differentiate the data and Q the order of the MA model. We use p,d and q to represent the same as P D Q but for the seasonal part.

CD – Dummy Exam 2

H – Social Networks



1 – In the given Network B has a prestige of 1/2

2 – C is the node with most centrality

3 – The diameter of the Network is 3

4 – In the PageRank Algorithm all connections have the same weight

5 – In PageRank Algorithm a node has more prestige if it's pointed out by more prestige nodes

I – Ethical Concerns

1 – In data processing we should remove data from servers to prevent future release or use

2 – Data cannot be further processed under historical research purposes

3 – Large companies don't need to be accountable according to GDPR

4 – We don't need GDPR to know that Catarina will pass the exam

5 – There are 10 simple rules for responsible data research

J – Deloitte Case Study

1 – We run out of memory while using DBSCAN

2 – We removed some variables

3 – We had a false predictor in the original variables

4 – We had no missing values in the dataset

5 – Naïve Bayes was not able to discriminate between records