Data Science

2020/21

Época Especial Exam

July 21st, 2021
Duration: 2 hours

Student ID: _____ Name:_____

Rules:

- **No** consultation, **but** calculator, is allowed.
- Delivery just **this** sheet, with your identification and answers inside the grid.
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.
- Each group counts at most 2 and at least 0 points. Each correct answer counts 0.4 points and each wrong one counts -0.2.

Solution

Data Profiling		
	T	**F**
1	X	
2	X	
3	X	
4		X
5		X

Data Preparation		
	T	**F**
1	X	
2		X
3		X
4	X	
5		X

Classifiers Evaluation		
	T	**F**
1		X
2	X	
3		X
4		X
5		X

Classification		
	T	**F**
1		X
2	X	
3	X	
4		X
5	X	

Pattern Mining		
	T	**F**
1		X
2	X	
3		X
4	X	
5		X

Clustering		
	T	**F**
1	X	
2		X
3		X
4	X	
5		X

Time Series		
	T	**F**
1		X
2		X
3	X	
4		X
5	X	

SNA		
	T	**F**
1		X
2	X	
3		X
4	X	
5		X

Ethics		
	T	**F**
1	X	
2		X
3		X
4		X
5	X	

Data Description

Consider the problem of predicting if some patient will survive, through the use of a dataset with 165 medical records, described by 50 variables. From these the `class` variable has two possible values `survive` (102) and `die` (63).

The tree on the left was learned through the C4.5 algorithm and the information gain criteria, when applied over 100 of the 165 records available, to learn the target variable `Class`, after applying some preparation techniques.

The tree was printed through `sklearn.tree` package. Each node in the tree shows the variable tested, the number of records satisfying the branch conditions, the number of records from `survive` and `die` classes, respectively, and the label predicted by the tree.

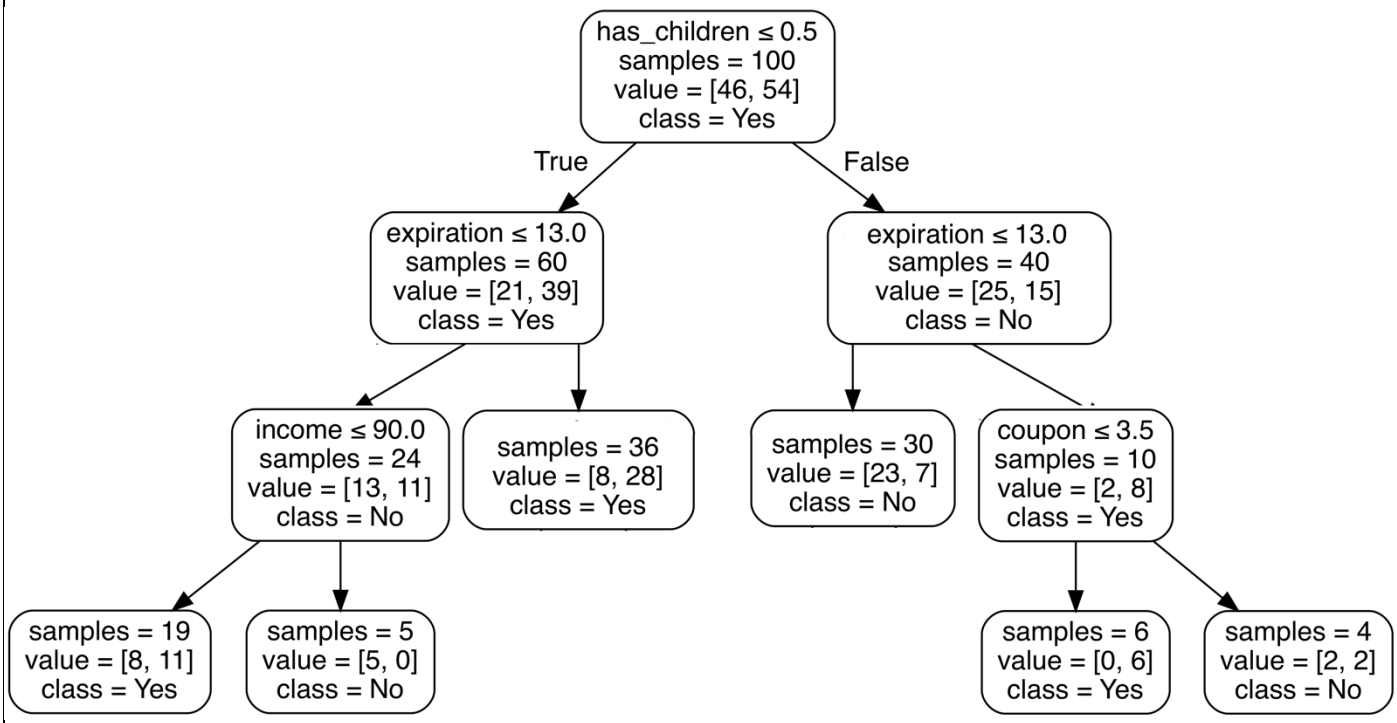

Figure 1 Decision tree trained over 100 records

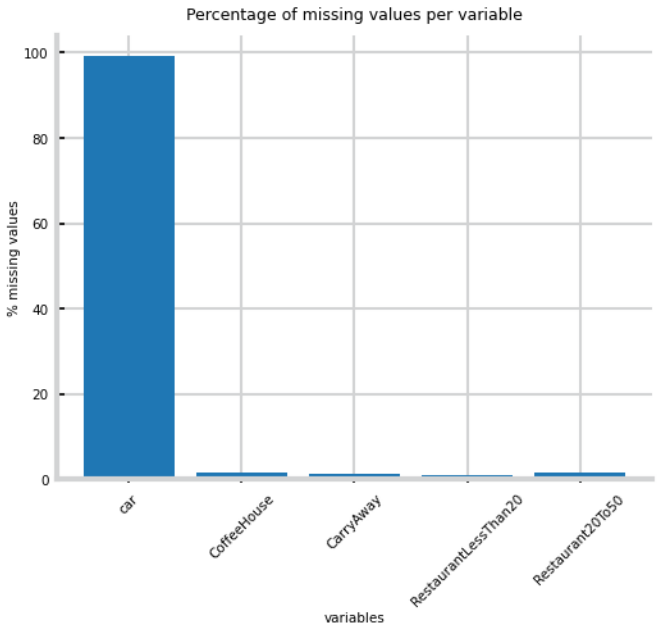

Figure 2 Variables with missing values

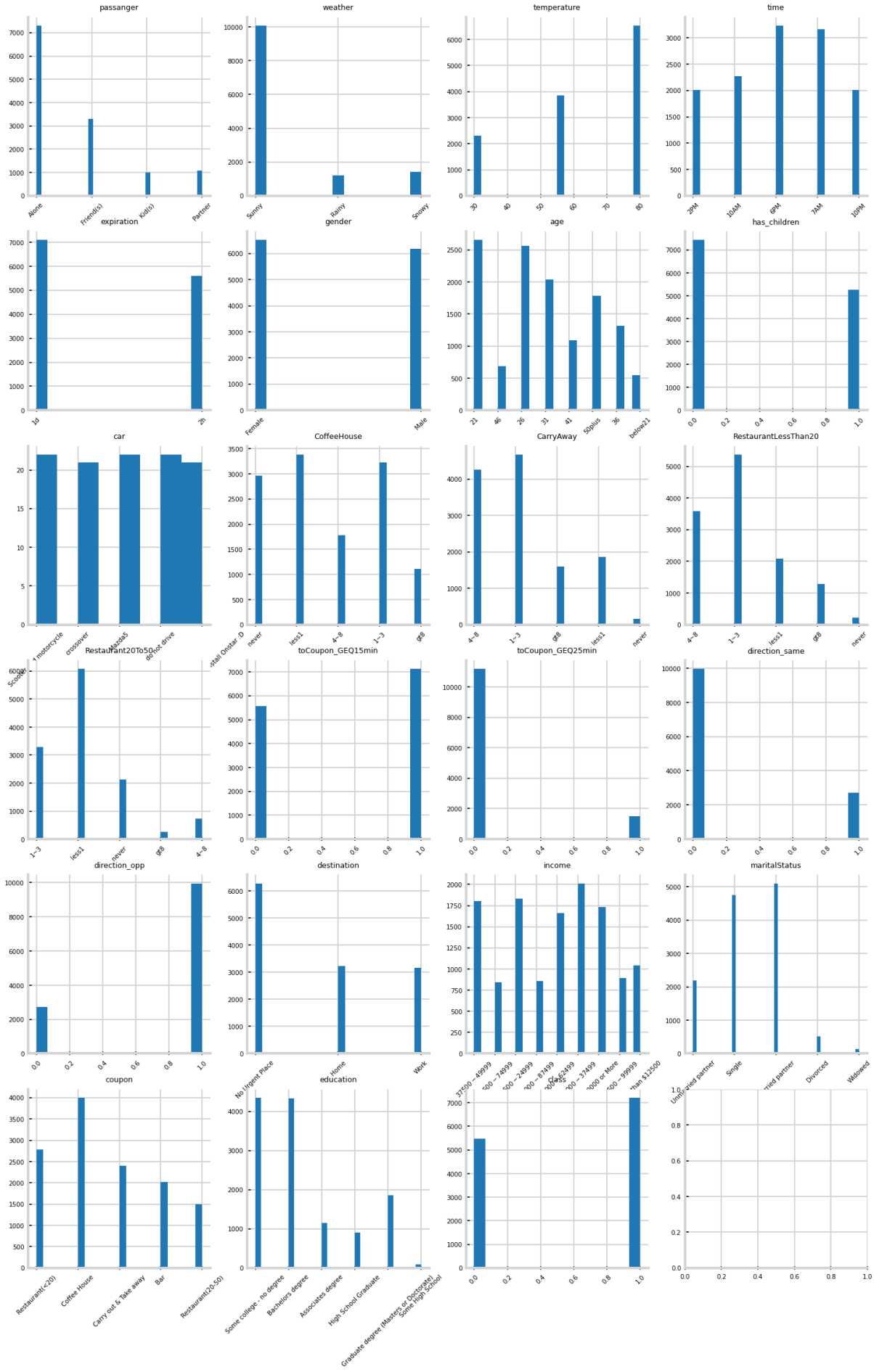

Figure 3 Histograms for all descriptive variables **before** transformation

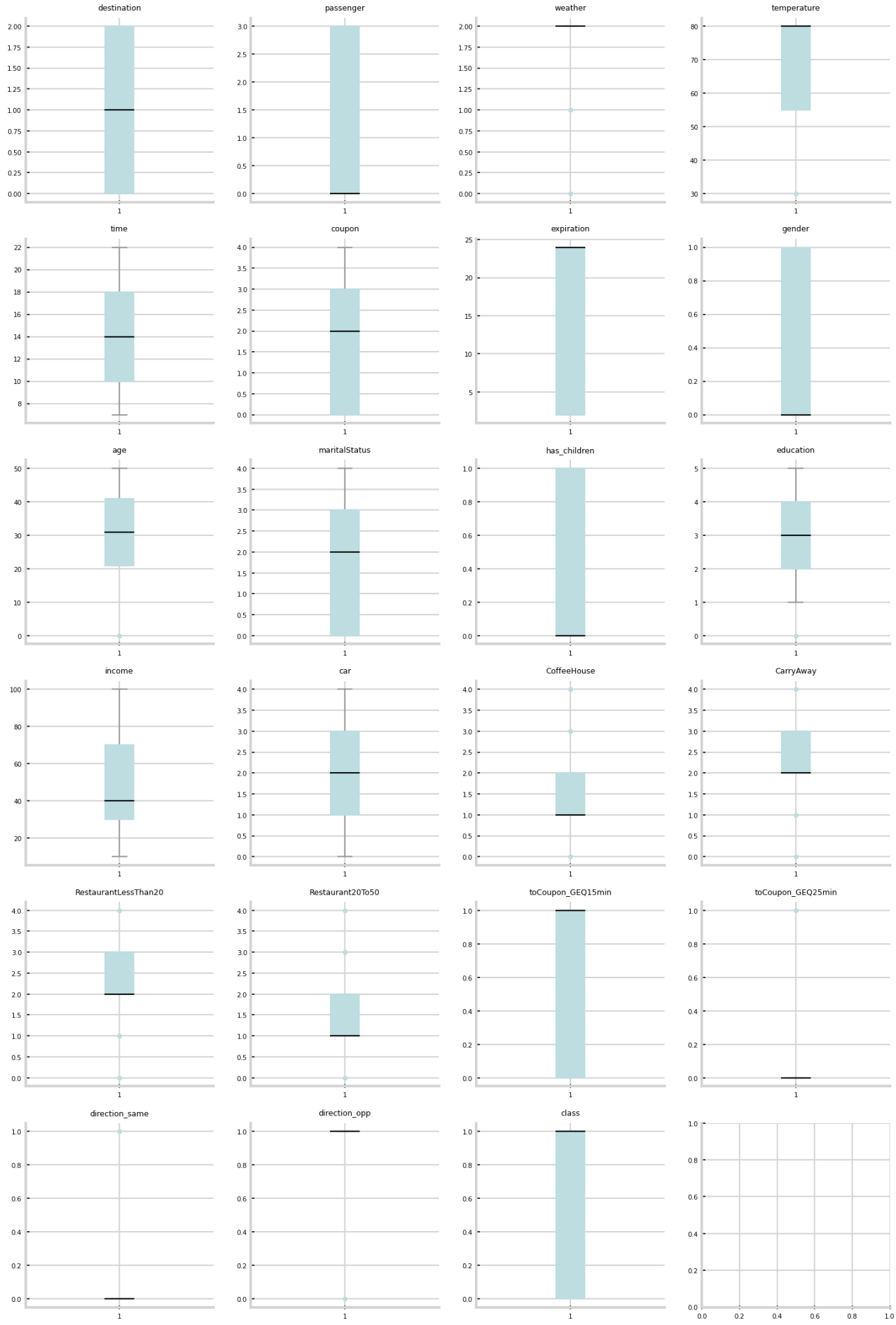

Figure 4 Boxplots for all descriptive variables **after** transformation

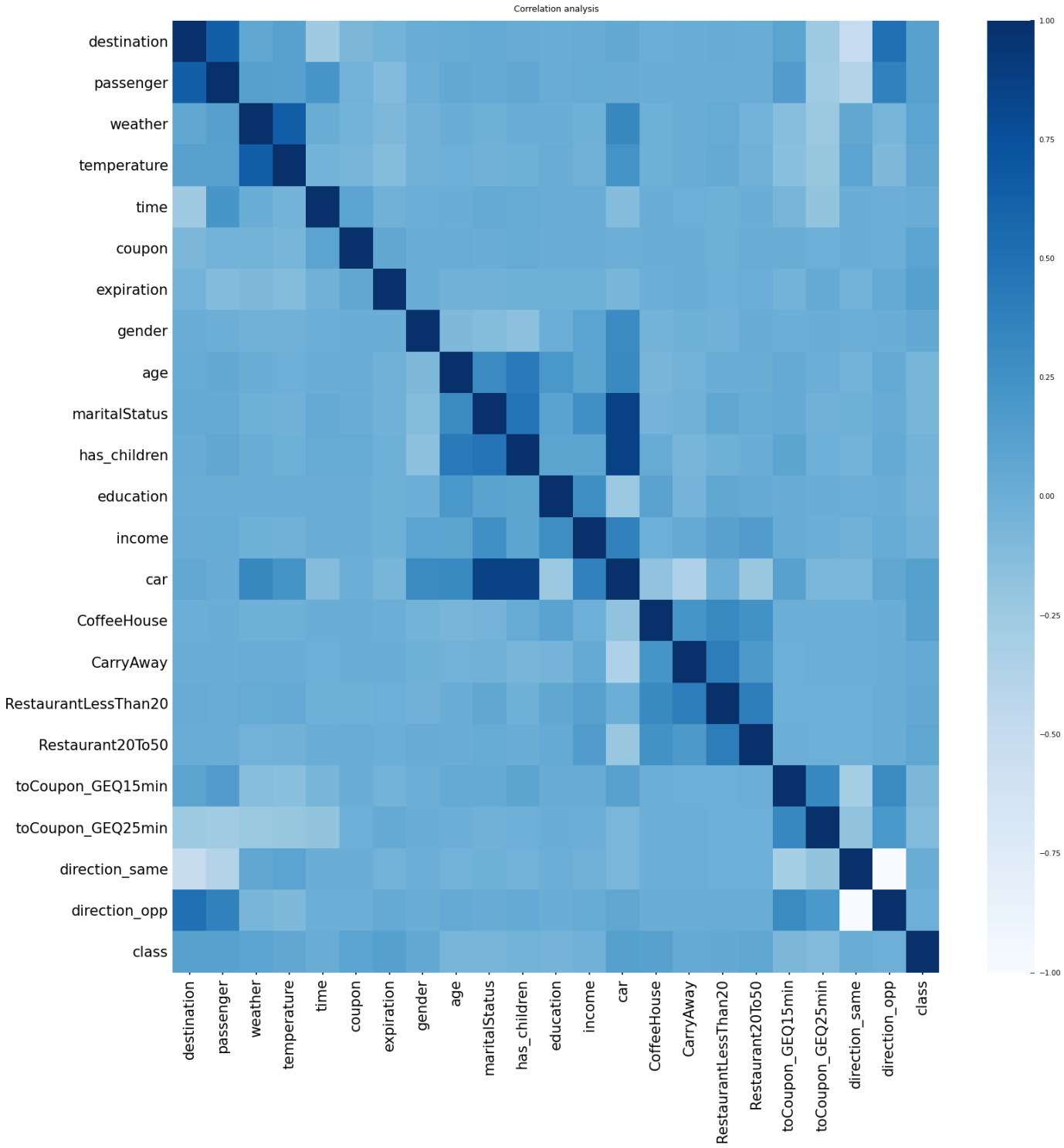

Figure 5 Correlation analysis

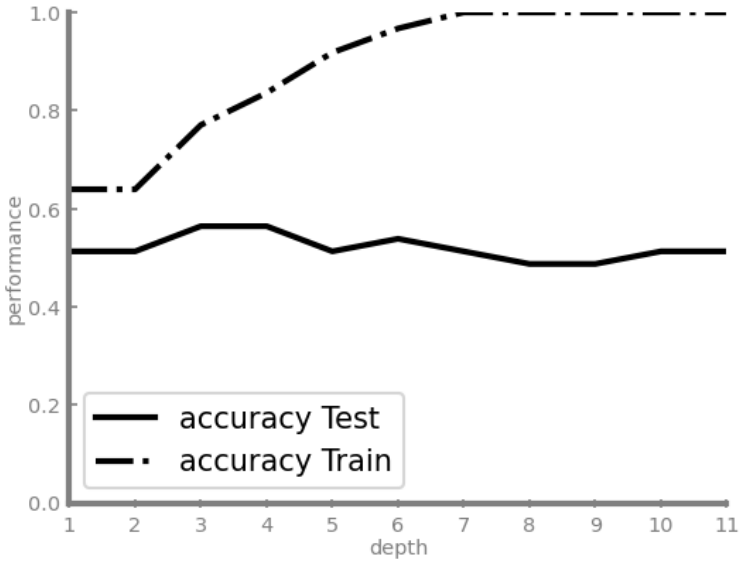

Figure 6 Accuracy for different decision trees specializations

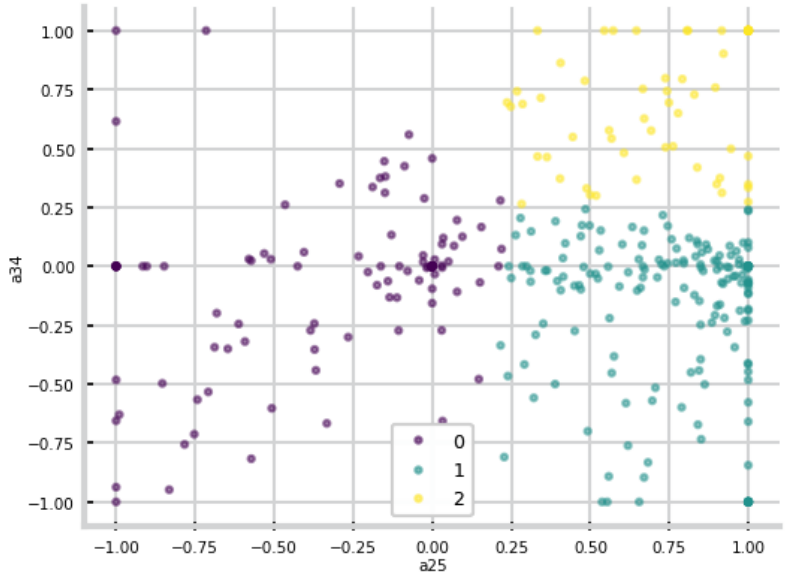

Figure 7 Clustering results with a specific algorithm and distance measure, over a two-dimensional dataset

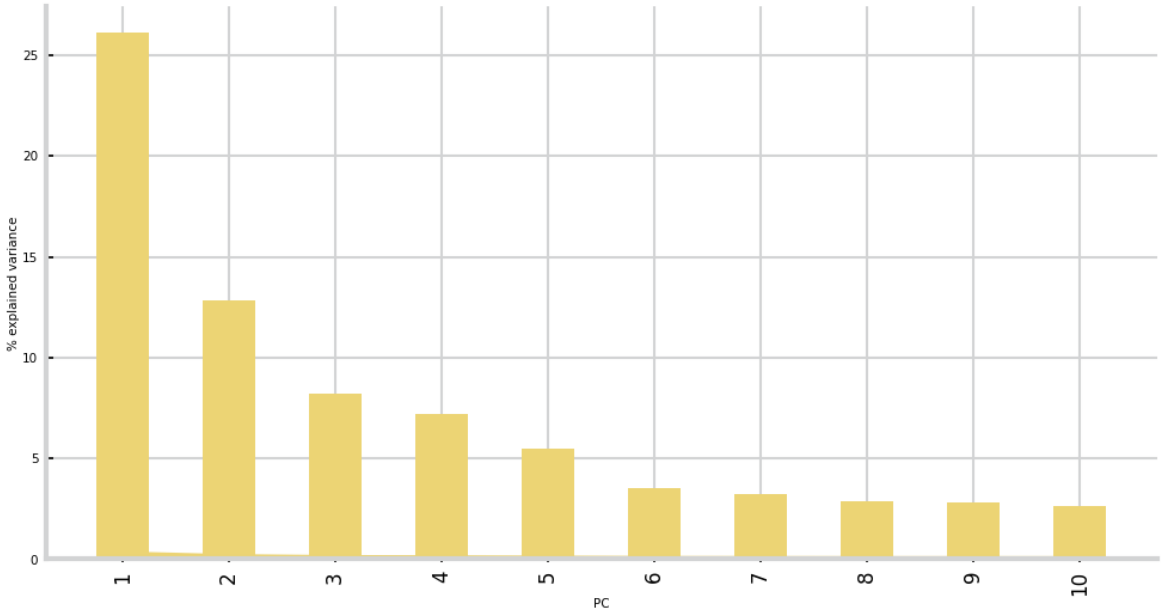

Figure 8 Explained variance for each principal component

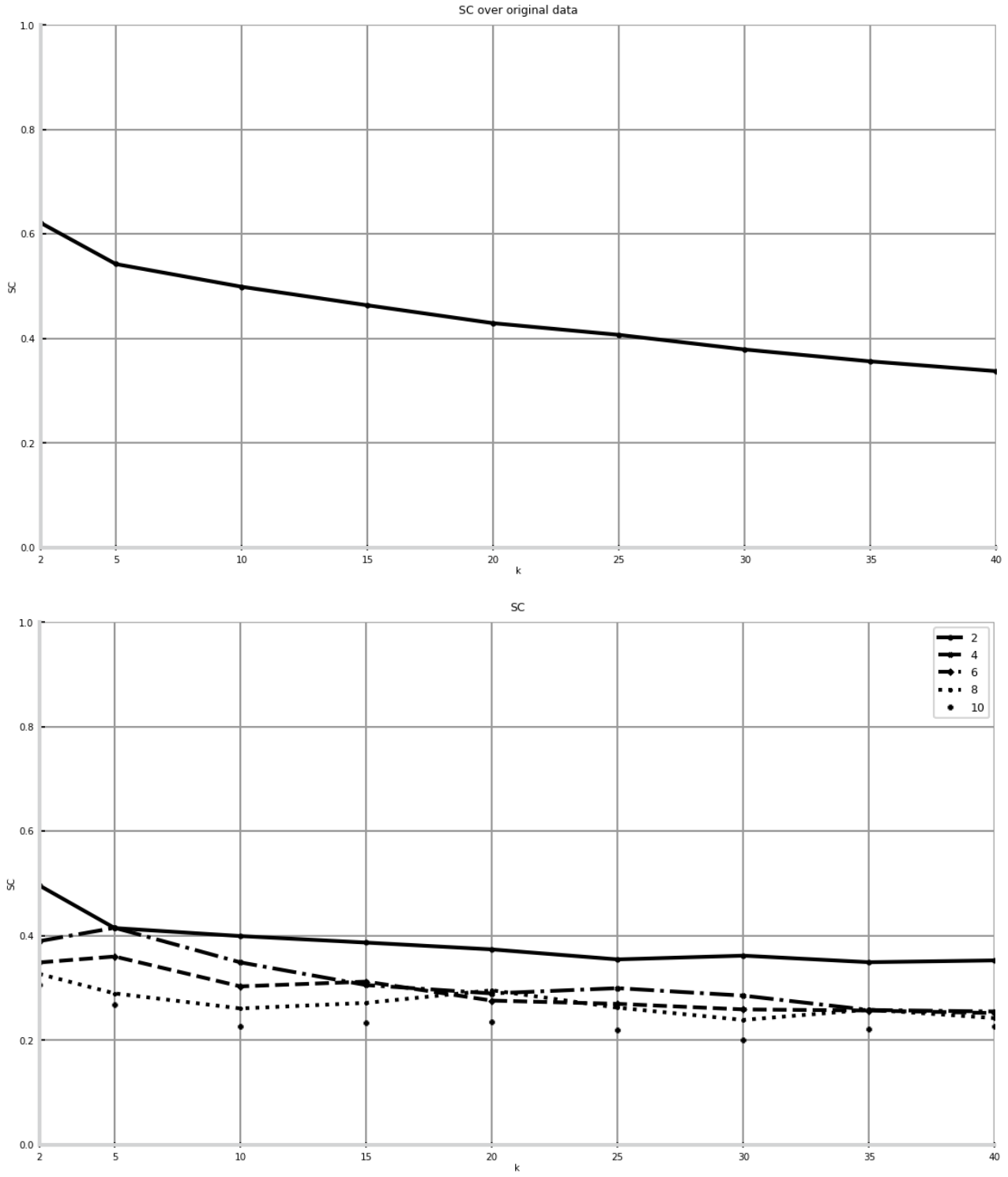

Figure 9 Silhouette coefficient (SC) along different number of clusters: over original data (top) and after applying PCA (bottom)

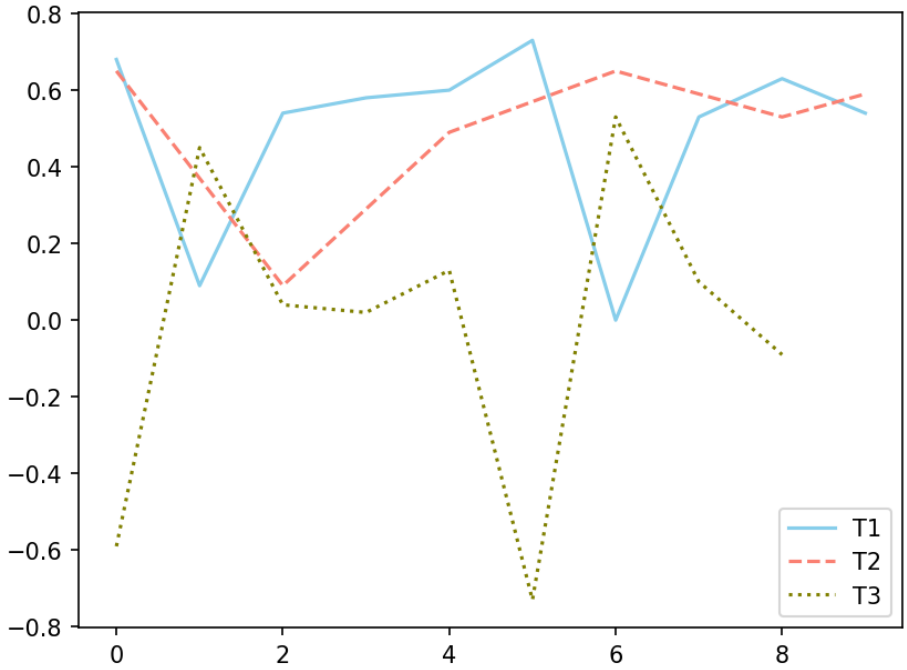

Figure 10 Time series

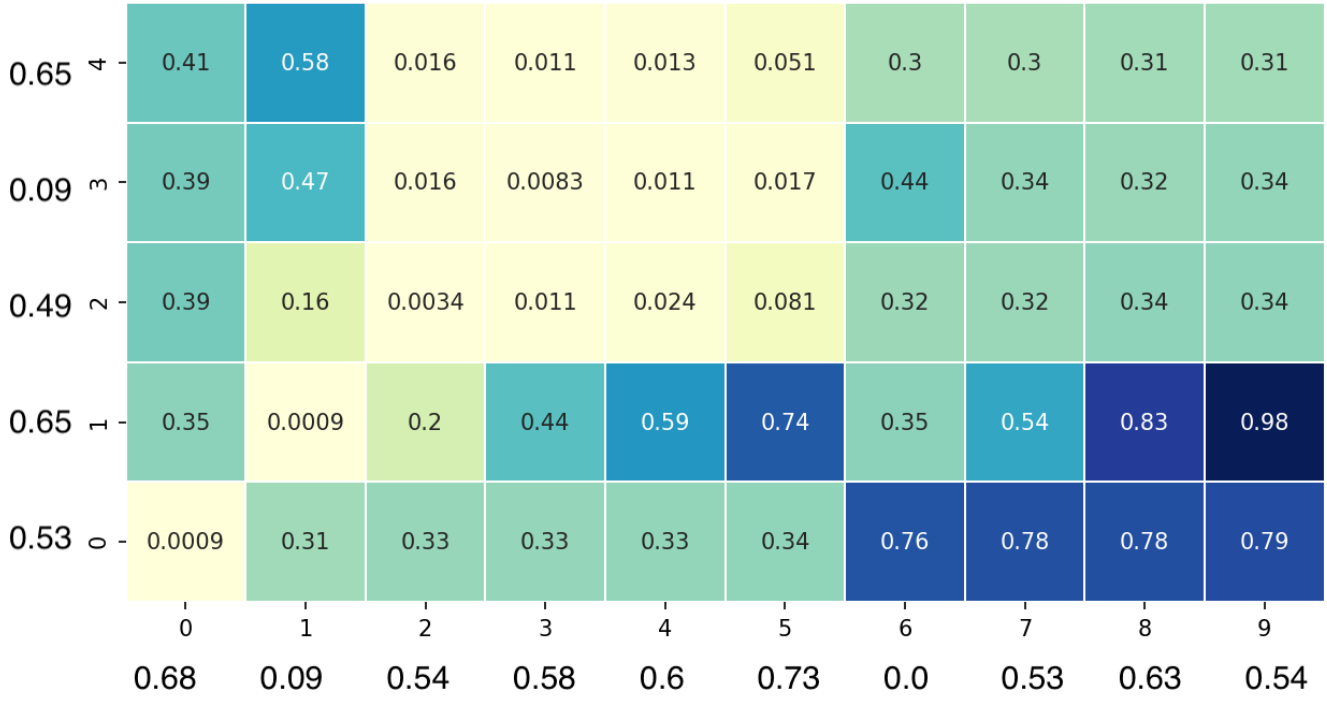

Figure 11 Accumulated cost matrix between two series

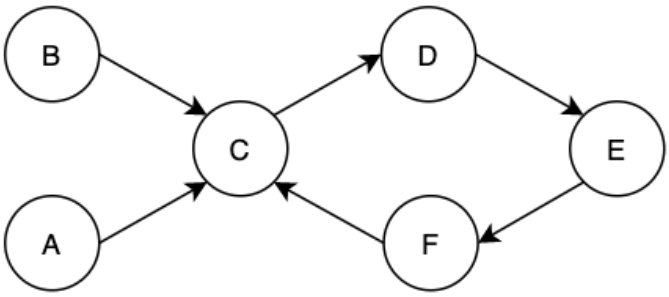

Figure 12 Social network

Statements

Pick the truth value for each statement, and fill it in the grid on the first page.

A. Data Profiling

Consider the original dataset described

1. Variables *age* and *income* are numerical, without assuming real values (their domain is not the R dataset).
2. The variable *weather* is ordinal.
3. Variables *maritalStatus* and *car* are redundant.
4. None of the binary variables is relevant.
5. We are facing the *curse of dimensionality*.

B. Data preparation

Consider the original dataset described

1. Considering the semantics usually related to the *income* variable (*salary*, *revenue*), we know that it was submitted to a process of discretization (during the data collection or data preparation steps).
2. D*ummification* is the best transformation for the *weather* variable.
3. A scale transformation can benefit the performance of Random Forests in this dataset.
4. Applying feature selection to this dataset, using the information gain criterion, may **increase** the performance of naïve Bayes algorithm.
5. Balancing by SMOTE is mandatory for better evaluate the quality of the models learnt over this data.

C. Classifiers Evaluation

1. The accuracy for the presented tree is **higher** than 90%.
2. The precision for the presented tree is **higher** than its recall
3. The chart on Figure 6 reporting the accuracy for different trees shows that the model is **only** in overfitting for models with 8 or more nodes of **depth**.
4. According to the chart on Figure 6, the **best** tree is the one with 6 nodes of depth.
5. The decision stump resulting from pruning the presented tree would have an error increase **lower** than 5 percentual points.

D. Classification

(Consider X=(*has_children*=0, *expiration*=2, *income*=50, *coupon*=1), and that the values for the remaining variables are missing, and the dataset described by the tree)

1. The decision tree presented classifies X as positive.
2. If we just consider *has_children* and *expiration* variables for describing records, there is at least a value for k for which the KNN algorithm classifies X as ***No***.
3. If we just consider the *has_children* variable for describing records, Naive Bayes algorithm classifies X as ***Yes***.
4. In random forests context, the maximum number of different *decision stumps*, trained with C4.5 and information gain, where the *income* variable is used is 4.
5. The tree pruned to have just 2 nodes of depth (root + one test for each branch) defines a linear separable space.

E. Pattern Mining

Consider 10% as the minimum support threshold and dataset defined through the decision tree.

1. We may state that **(has_children≤0.5, expiration≤13, income≤90)** is frequent and a pattern.
2. Knowing **(has_children>0.5, expiration≤13, income≤90)** is frequent, then **(has_children>0.5, income≤90)** has to be frequent.
3. The confidence for the rule **has_children>0.5 ⇨ expiration≤13** is higher than 80%
4. The *lift* for the rule **has_children>0.5 ⇨ expiration≤13** is smaller than 2.5.
5. **The *lift* for has_children≤0.5 ⇨ expiration≤13** is the same as for the rule **has_children>0.5 ⇨ expiration≤13**

F. Clustering

1. The clusters presented on Figure 7 might be discovered through the **kmeans** algorithm.
2. Cohesion for cluster 0, on Figure 7 is **better** than the cohesion for cluster 2.
3. Consider the chart on Figure 8: the first four components allow to reduce the data dimensionality loosing less than 10% of information.
4. Consider the data shown in Figure 9: we can say that the partition with 2 clusters in the original data is **reasonable**.
5. Consider the clustering results shown on Figure 9, the clustering presents better results after applying PCA, independently of the number of clusters considered.

G. Time Series

1. The series T1 in Figure 10 exhibits a seasonal component.
2. The series T2 in Figure 10 is a smooth transformation of T1.
3. The series T3 in Figure 10 may be a differentiation of T1.
4. Consider the accumulated cost matrix between two series in Figure 11: the dynamic time warping path between them is (0,0) (1,1) (2,1) (3,0) (4,1) (5,2) (6,3) (7,4) (8,4) (9,4).
5. Consider T1=[0.68, 0.09, 0.54, 0.58, 0.6, 0.73, 0, 0.53, 0.63, 0.54] and T2=[0.65, 0.37, 0.09, 0.29, 0.49, 0.57, 0.65, 0.59, 0.53, 0.59]. If T2 were the prediction of T1 according some regression model, its MAE would be less than 1.5.

H. Social Network Analysis

Consider the social network presented in
Figure 12

1. Node D is more central than node F.
2. Node C is more prestigious than node E.
3. The smallest path from node A to node F is 3 edges.
4. All non-null entries on the corresponding graph adjacency matrix are equal.
5. If a **new edge** would be added from E to C, then PageRank scores would **improve for all nodes but F**.

I. Anomaly Detection

J. Ethical Concerns

1. It seems the dataset described have been submitted to an anonymization process, for all its variables.
2. Given the dataset described and that the purpose of this data processing is to study customers' behavior, according to the GDPR, it may be performed unconditionally.
3. According to the GDPR, the described data can be processed if under the controller consent.
4. If personal data were used in the described data, then the resulting models would be manipulative.
5. Given the dataset described, according to the GDPR and considering the semantics given by variable names, none of them violate the list of prohibited data to be processed.

Good work!!!