TÉCNICO LISBOA

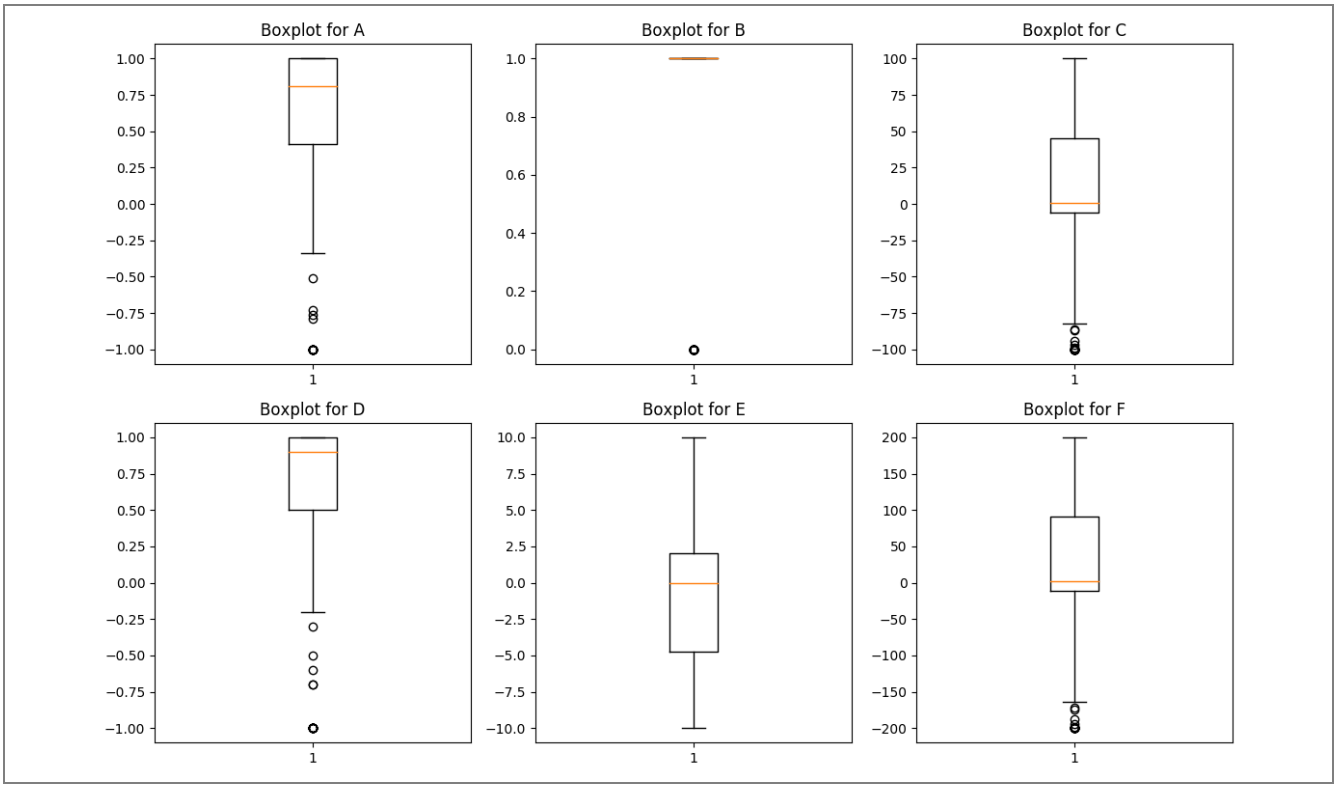
Data Science

by Cláudia Antunes

Lab Data Profiling

A. Exam 2020-01-13

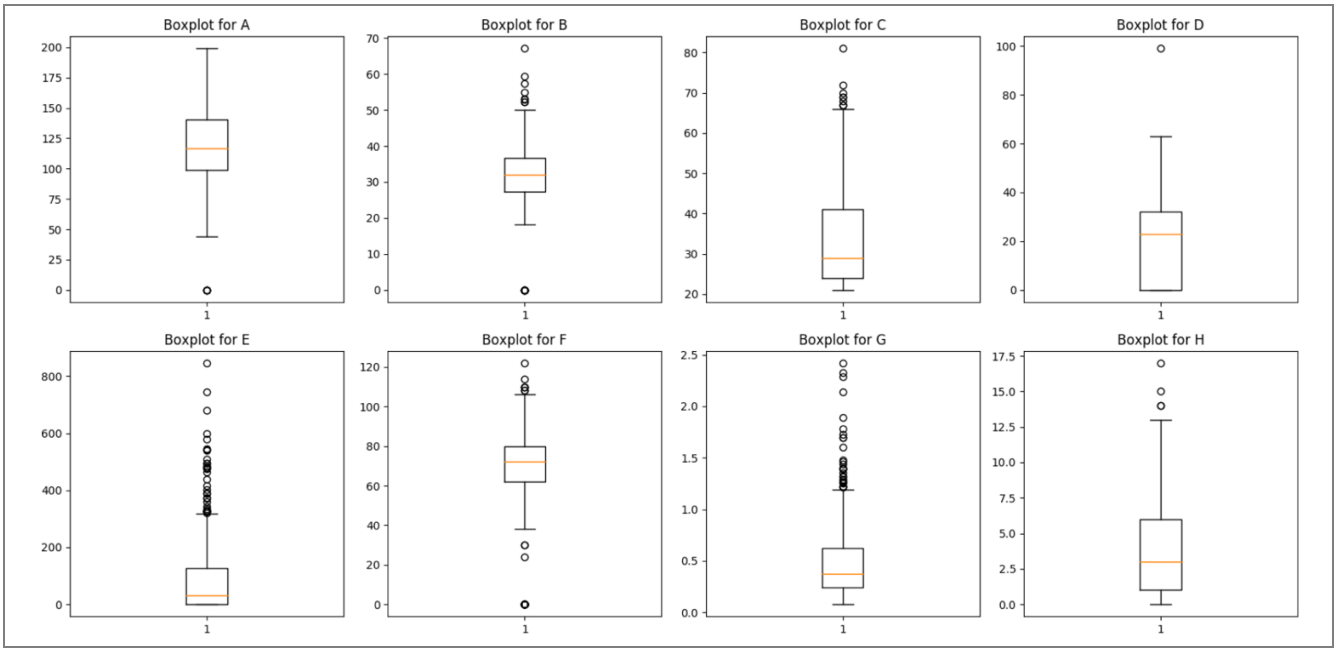
Consider a dataset composed by 350 records, described by 6 variables (B is Boolean), described by the boxplots for each one of the variables.



1. Boxplots can be used to understand variables distribution.
2. Those boxplots can prove the correlation between C and F.
3. Variable B is balanced.
4. Variable E shows a high number of outlier values.
5. Scatterplots are more adequate to identify correlation than boxplots.

B. Exam 2020-01-27

Consider a dataset composed by 768 records, described by 8 numeric variables, described by the boxplots for each one of the variables below.



1. From the boxplots above we can conclude that variables D and H have similar statistical distributions.
2. Those boxplots show that the data is not normalized.
3. Variable C shows some outlier values.
4. Histograms are more adequate than boxplots to identify outlier values.
5. Boxplots are more adequate than histograms to assess if variables are balanced.

C. Exam 2021-01-19

Consider the problem of diagnosing arrhythmia in patients, through the use of a dataset with 452 medical records, described by 250 variables. One of these variables, call it Z, contains the type of arrhythmia detected in each positive patient, and 0 if the problem was not diagnosed. From it, the variable class was derived assuming the value regular whenever Z=0 (245) and abnormal (207) otherwise.

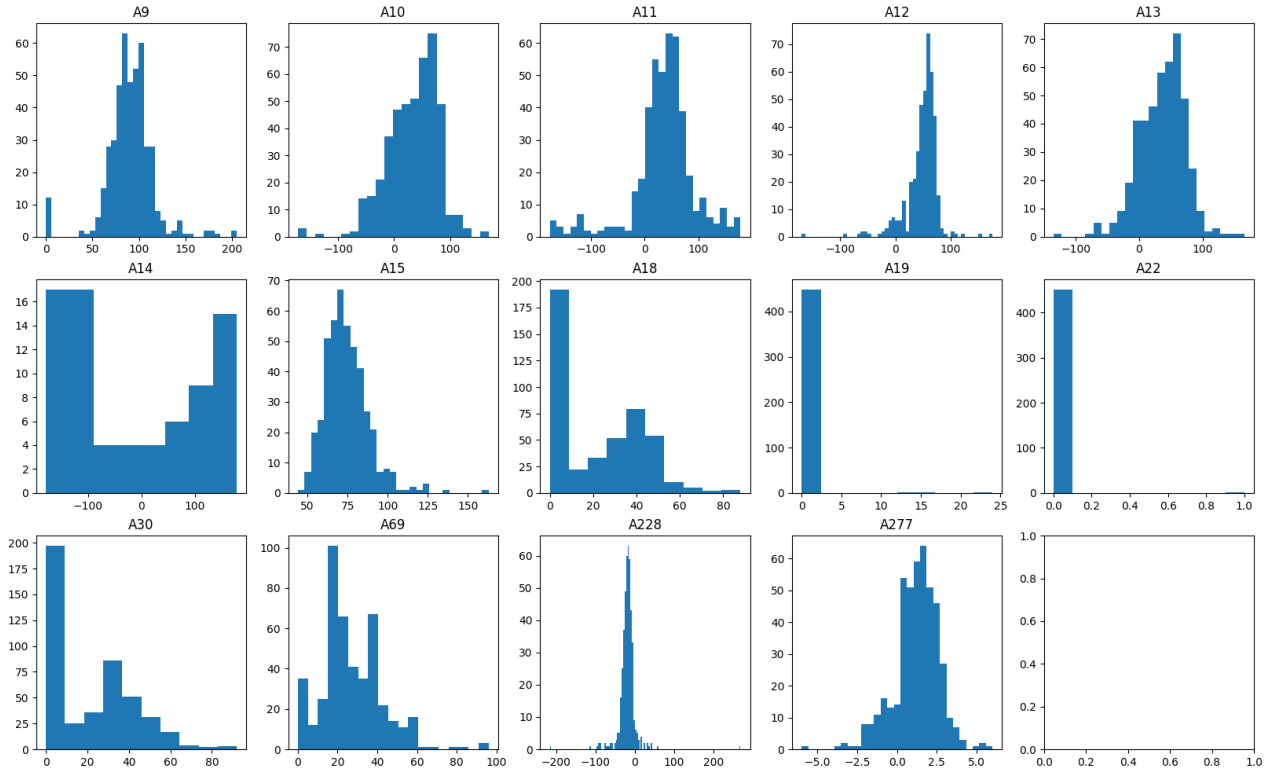


Figure 1 Histograms for some variables

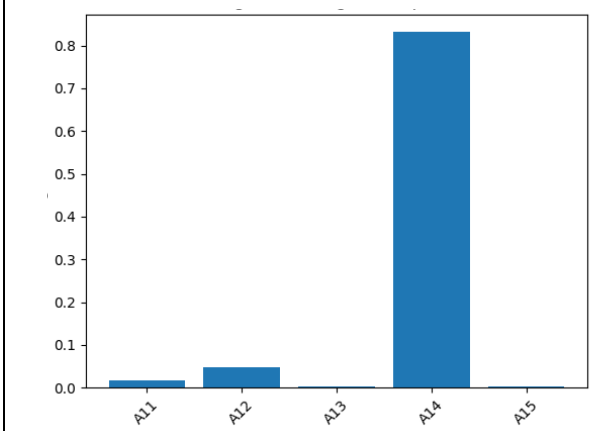


Figure 2 Variables with Missing Values – percentage of records

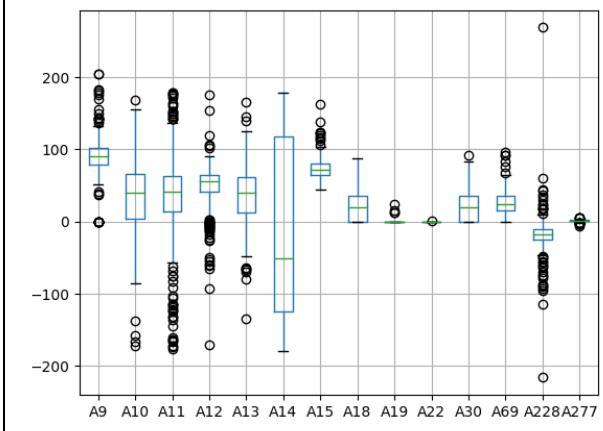


Figure 3 Boxplots for some variables

1. We face the curse of dimensionality when training a classifier with this dataset.
2. Variable Z is a false predictor.
3. Variables A19 and A22 are redundant, but we can't say the same for the pair A18 and A30.
4. The figure doesn't show any missing values for A9, but these may be hidden as some non-pre-identified value.
5. Variables A14 and A228 seem to be useful for classification tasks.

D. Exam 2021-02-05

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.

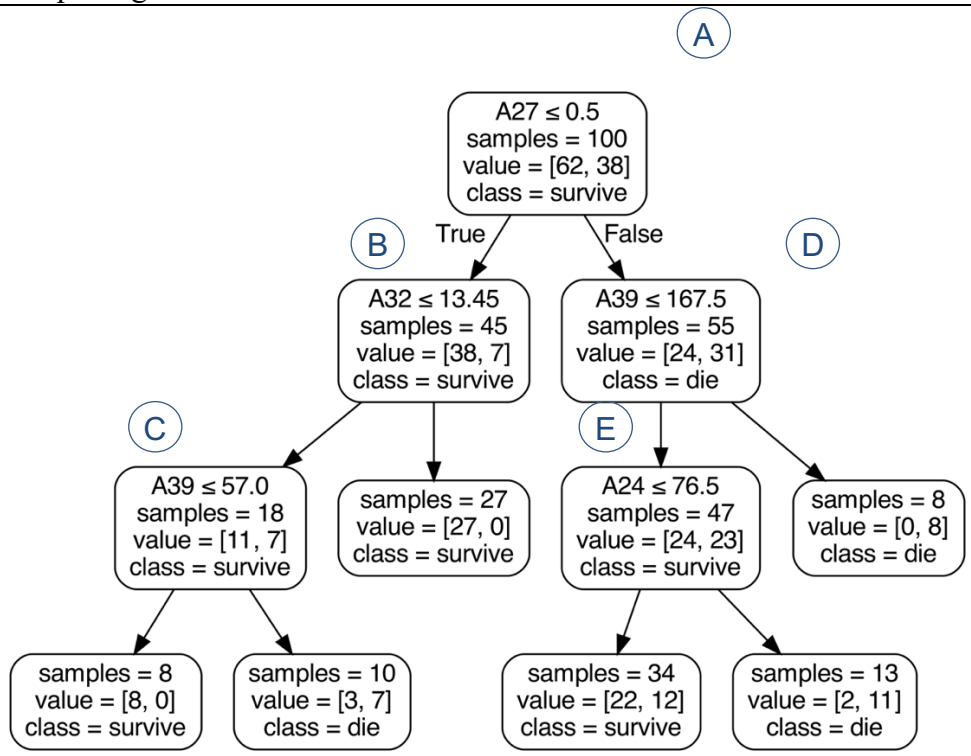


Figure 4 Decision tree trained over 100 records

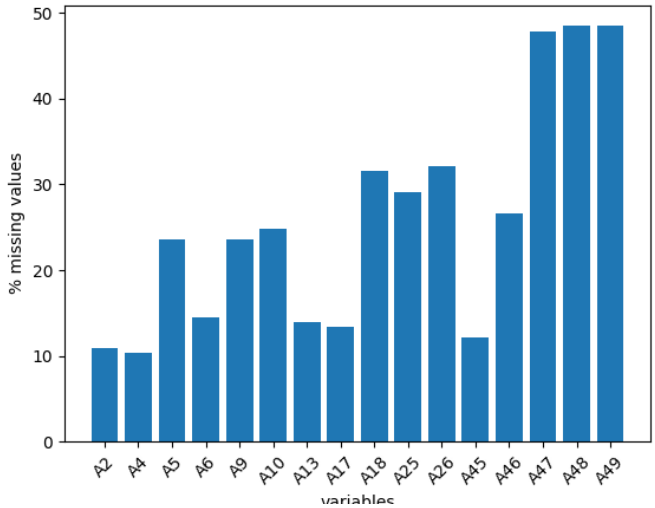


Figure 5 Variables with more than 10% of missing values

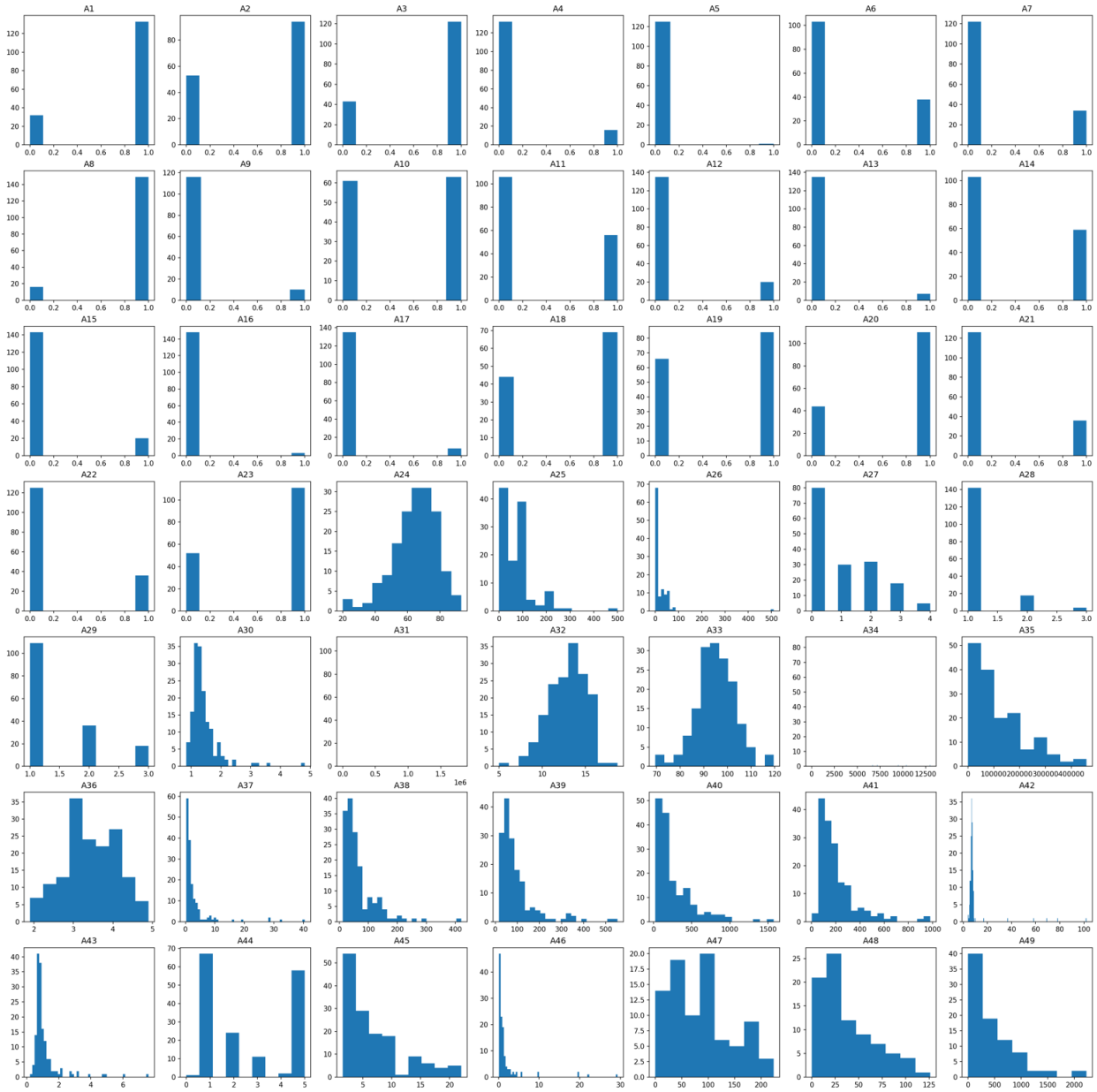


Figure 6 Histograms for all descriptive variables

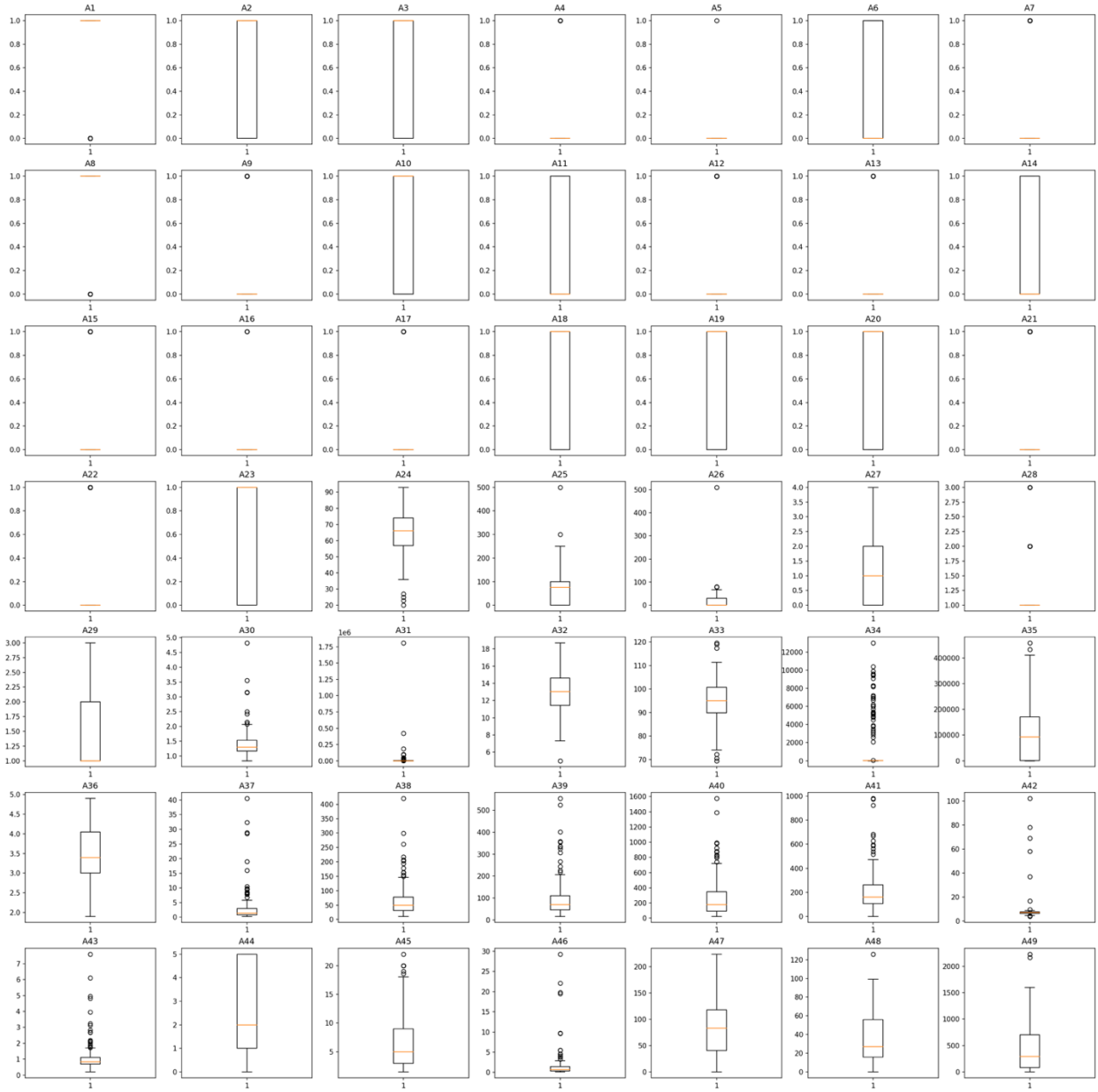


Figure 7 Boxplots for all descriptive variables

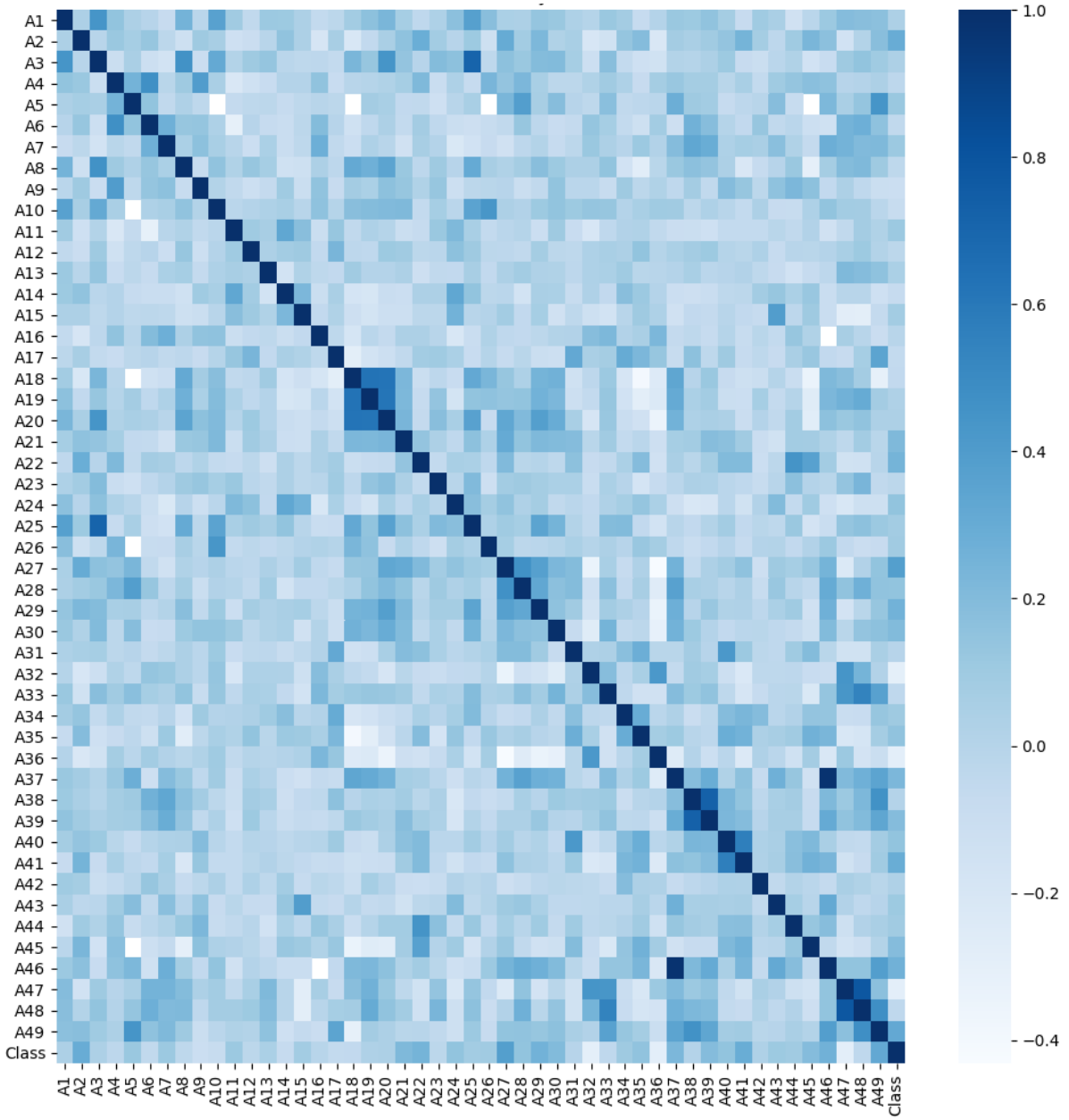


Figure 8 Correlation analysis

1. All variables, but the class, should be dealt with as numeric.
2. Outliers seem to be a problem in the dataset.
3. Variable A27 is one of the most relevant variables.
4. The intrinsic dimensionality of this dataset is 30.
5. The number of existing missing values in this dataset highly impairs the learning process.

E. Exam 2022-02-10

Consider a classification task approached through the exploration of a dataset with 500 records, described by 12 variables. From these the `class` variable has two possible values `Pos` (100) and `Neg` (400). The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **200** of the 500 records available, to learn the target variable `class`, after applying some preparation techniques.

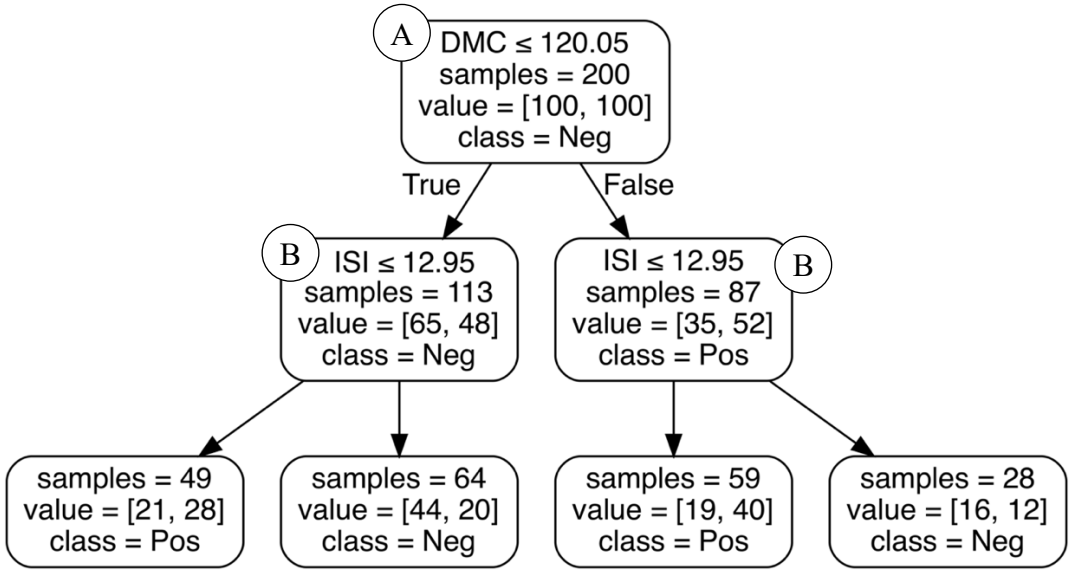


Figure 9 Decision tree trained over 200 records

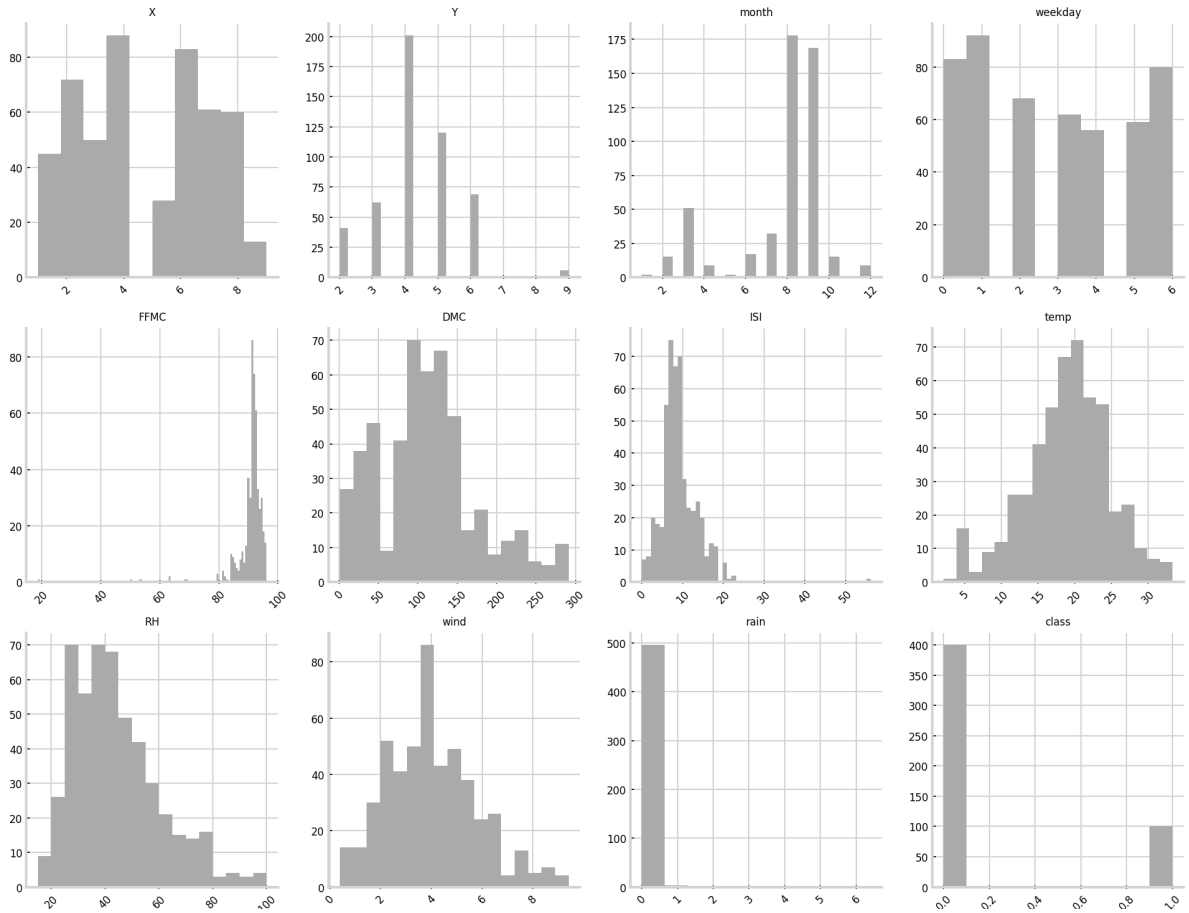


Figure 10 Histograms for all variables

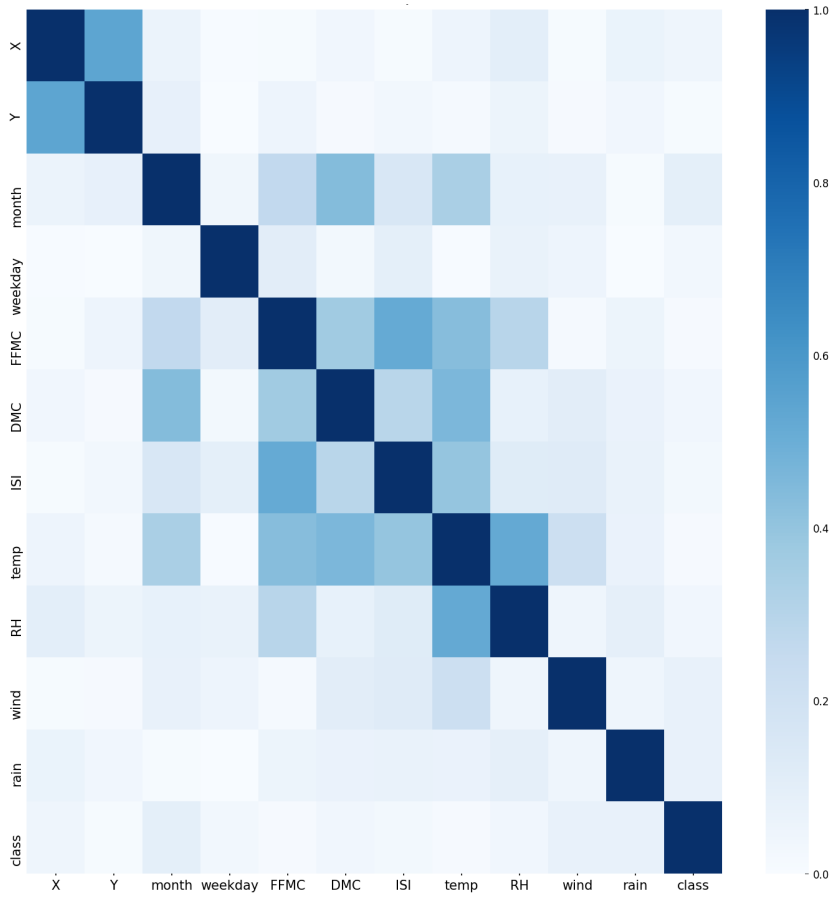


Figure 11 Correlation analysis

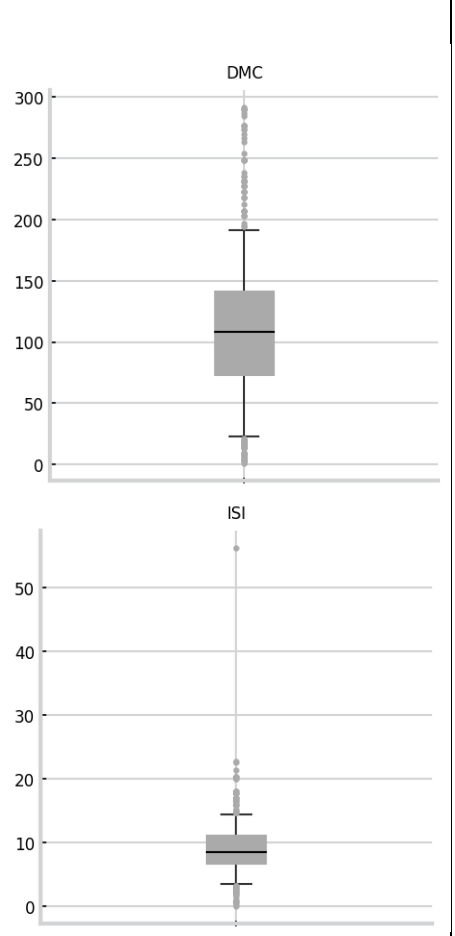


Figure 12 Boxplots for some variables

1. The variable month can be seen as ordinal.
2. Variable rain seems to be relevant for the majority of mining tasks.
3. It is clear that variable ISI shows some outliers, but we can't be sure of the same for variable DMC.
4. From the correlation analysis alone, it is clear that there are relevant variables.
5. Variable month presents some outliers.

F. Exam 2022-02-26

Consider a classification task, whose goal is to determine a survival model. The task was approached through the exploration of a dataset with **1000 records**, described by **16 variables**. From these the `class` variable represents survival, and it has two possible values `Yes` (400) and `No` (600). The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **all the 1000** records available, to learn the target variable `class`, after applying some preparation techniques.

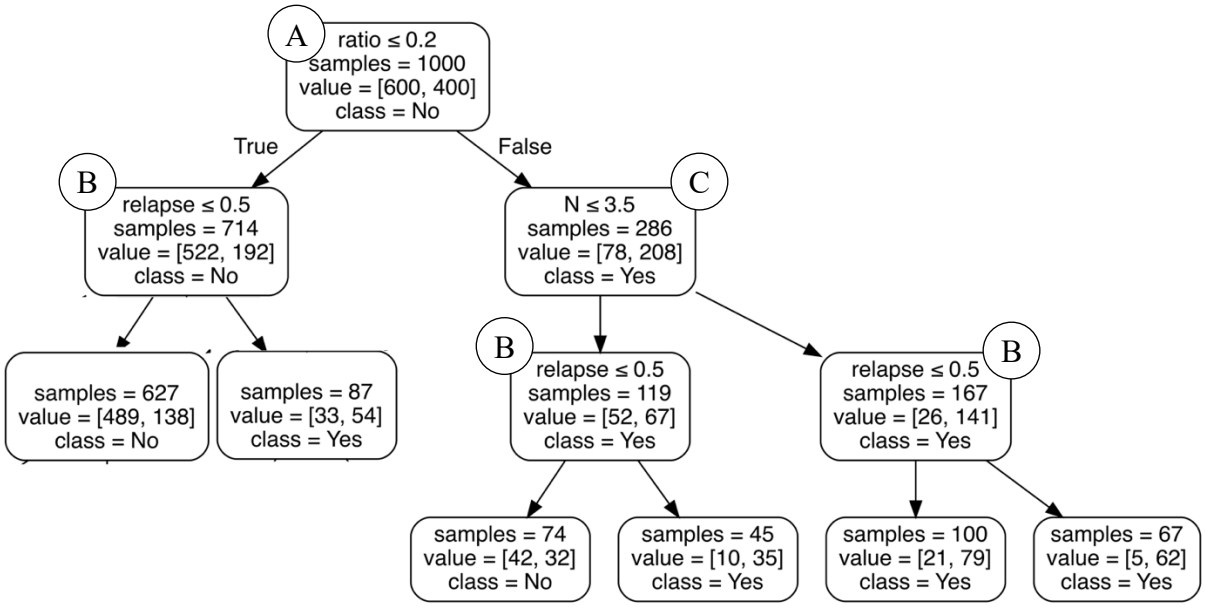


Figure 13 Decision tree trained over all records

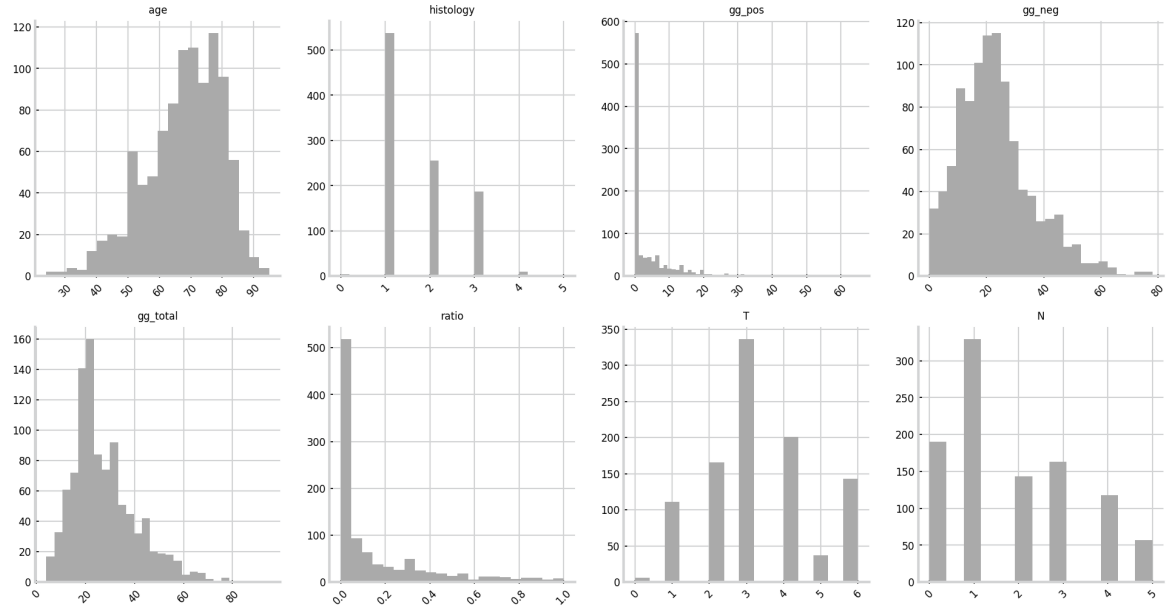


Figure 14 Histograms for non-binary variables

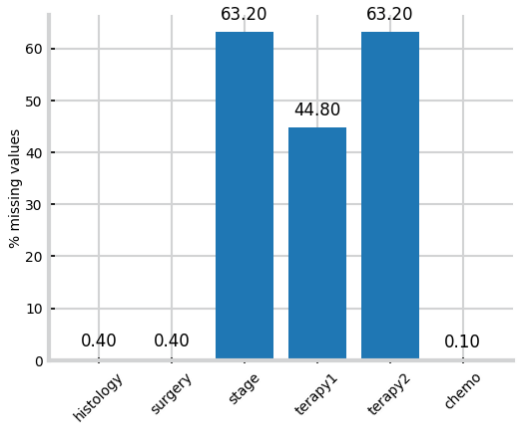


Figure 15 Number of missing values per variable (right)

1. The variable histology can be seen as ordinal.
2. Knowing that variable gg_total is defined as the sum of variables gg_pos and gg_neg, we can say that variable gg_neg and gg_total are redundant.
3. It is clear that variable relapse is one of the three most relevant features.
4. Variable gg_pos don't have any outliers.
5. Given the number of records and that some variables are numeric, we might be facing the curse of dimensionality.

G. Exam 2023-01-23

Consider a classification task, whose goal is to determine if some patient will make an insurance claim above a threshold (`class` variable). The task was approached through the exploration of a dataset with **1200 records**, described by **9 variables**. There are 490 records `Yes` (41%) and 710 records `No` (59%) regarding the `class` variable. The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **only 1000** records from the available, to learn the target variable `class`, after applying some preparation techniques.

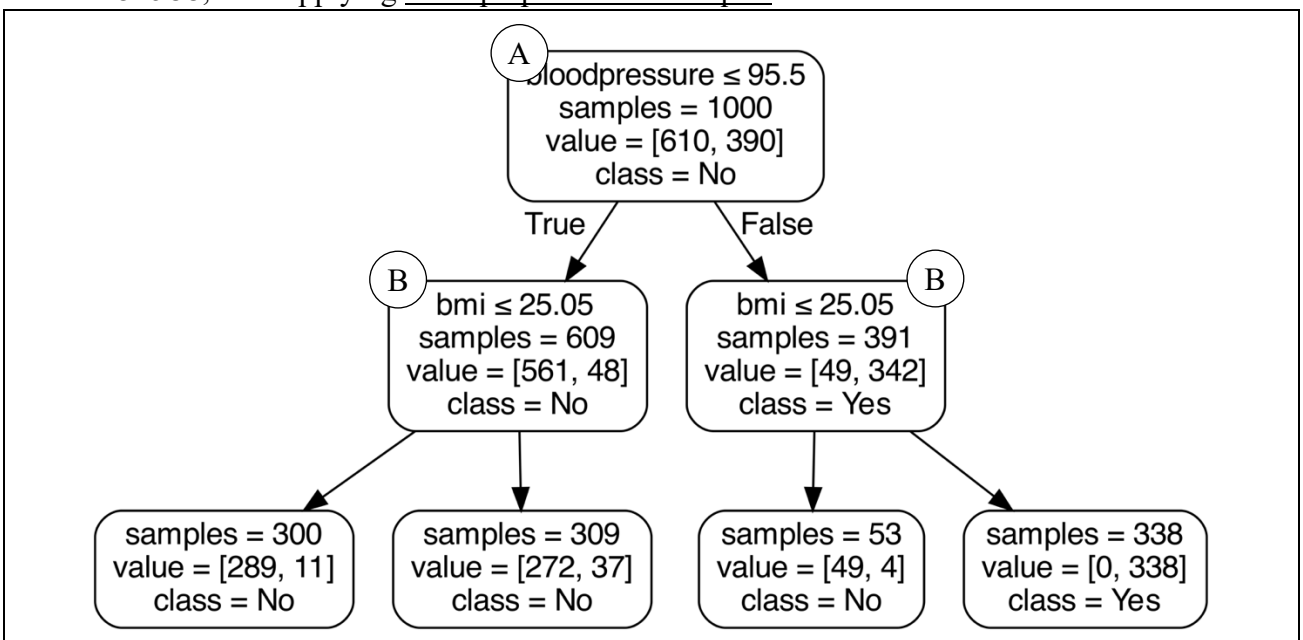


Figure 16 Decision tree trained over all records

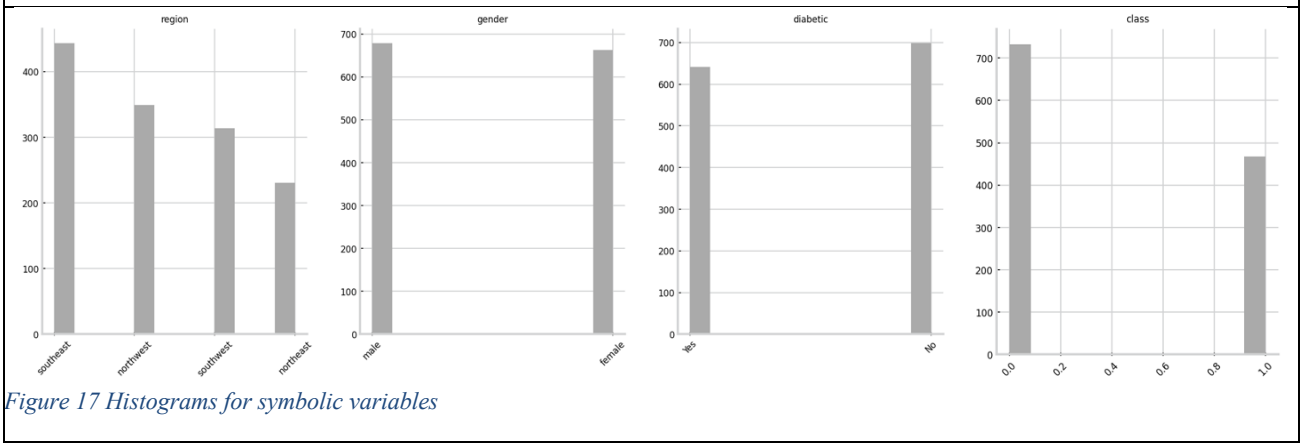


Figure 17 Histograms for symbolic variables

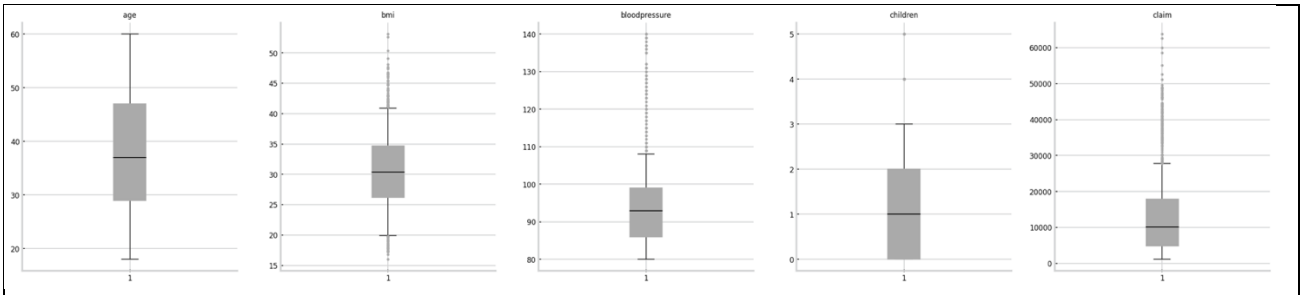


Figure 18 Boxplots for numeric variables

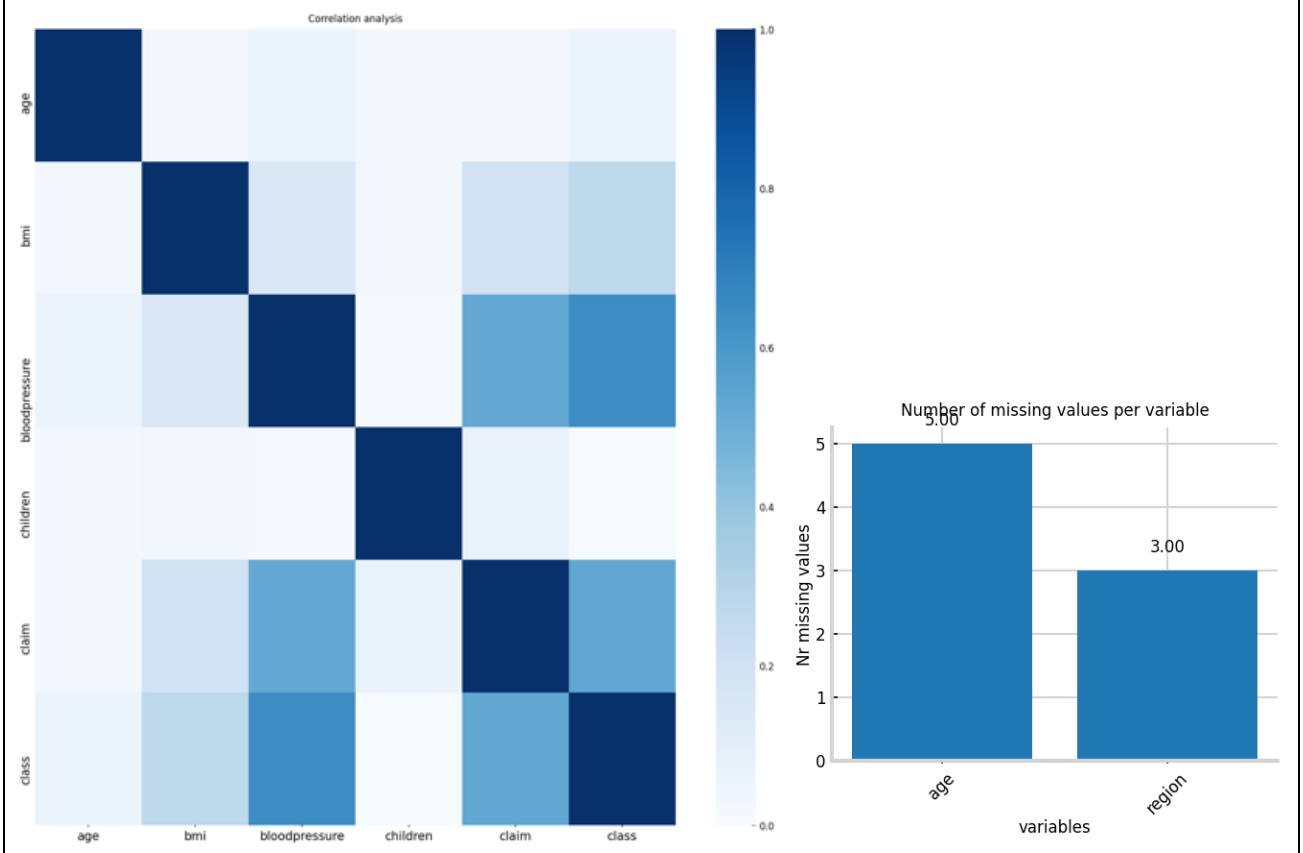


Figure 19 Correlation (left) and variables with missing values (right)

1. The variable region can be seen as ordinal without losing information.
2. Variables bmi and bloodpressure are redundant.
3. The variable claim seems to be one of the three most relevant features.
4. The boxplots presented show a large number of outliers for most of the numeric variables.
5. Knowing the claim variable expresses the amount claimed by each patient, it represents a data leakage.

H. Exam 2023-02-10

Consider a classification task, whose goal is to determine if the driver in a tesla car accident will die in the accident (class=`driver_death`). The task was approached through the exploration of a dataset curated from **235 records**, described by **11 variables** plus the class, where there were 96 records `Yes`

(41%) and 139 records `No` (59%) regarding the `driver_death` variable. One of the eleven variables available contained a description of the accident, "*car collides with tesla, both drivers die*" is one of the 200 descriptions provided. The tree below was **learned through** the C4.5 algorithm and the information gain criteria, when applied over **the curated dataset**, to learn the target variable `driver_death`.

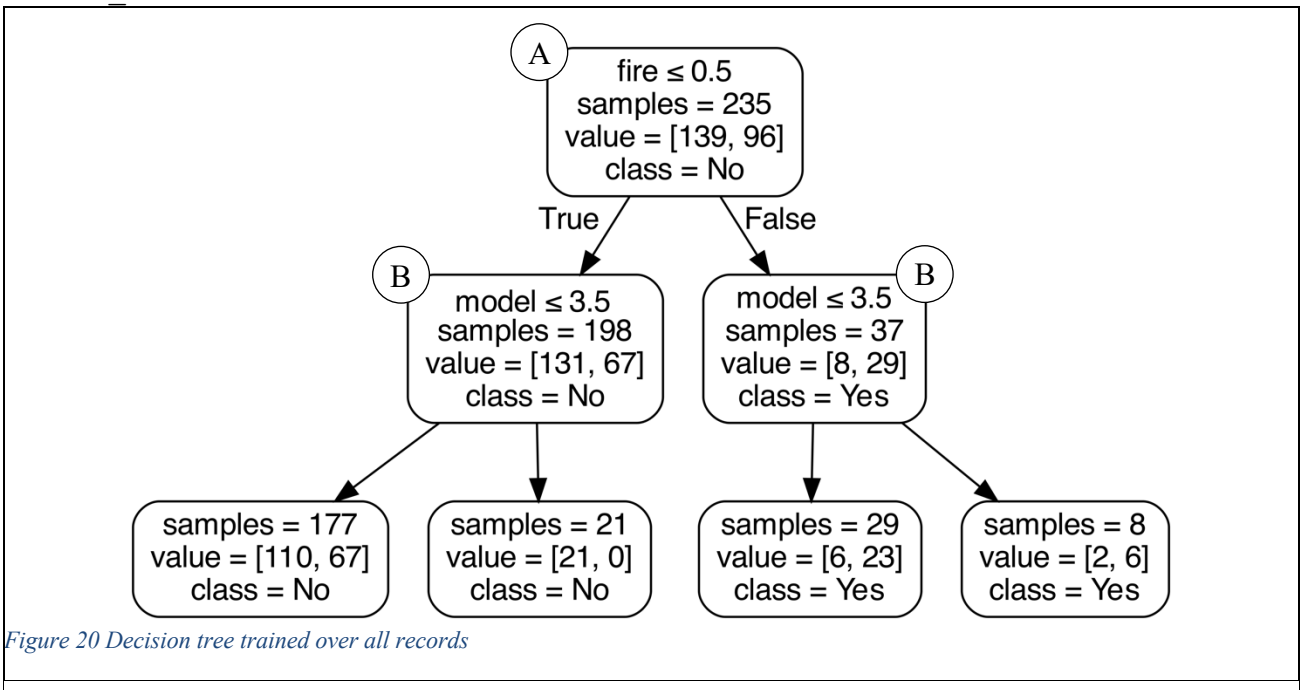


Figure 20 Decision tree trained over all records

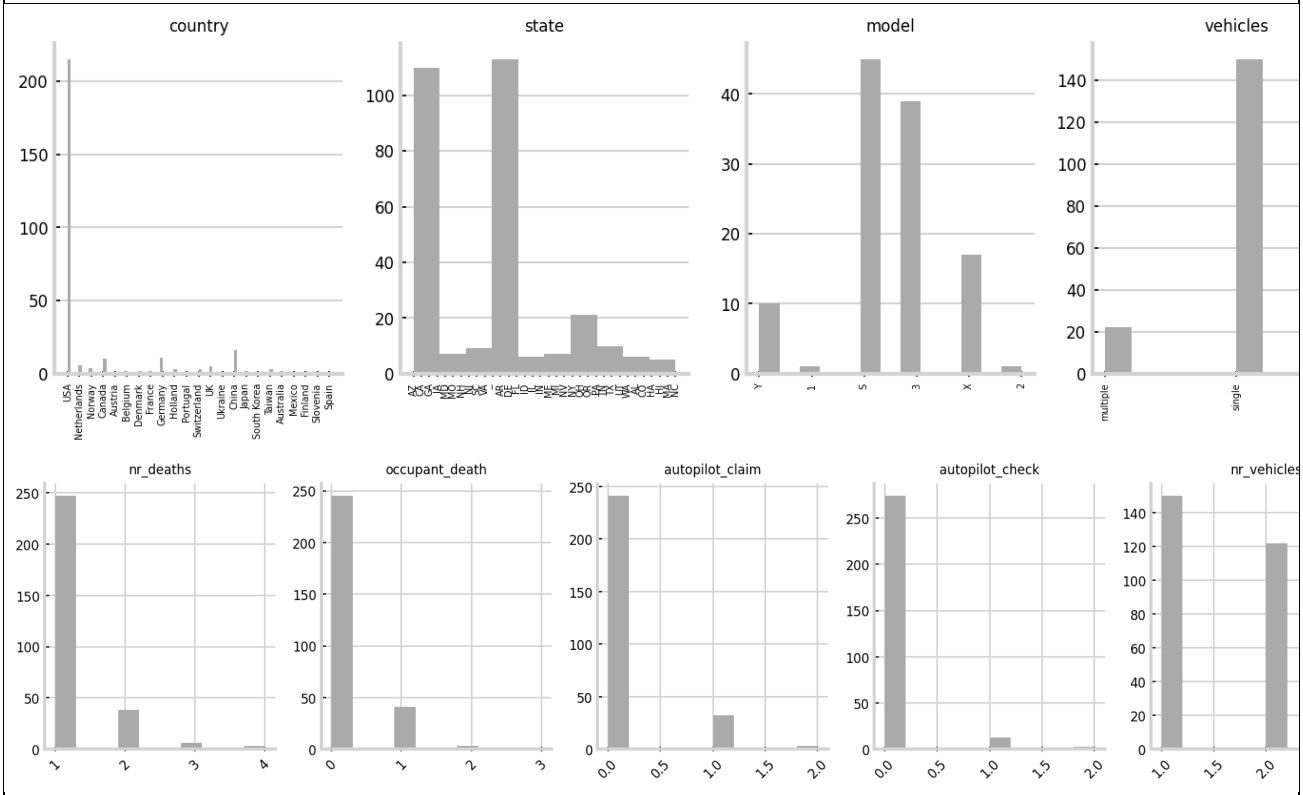


Figure 21 Histograms for original variables (excluding description and driver_death)

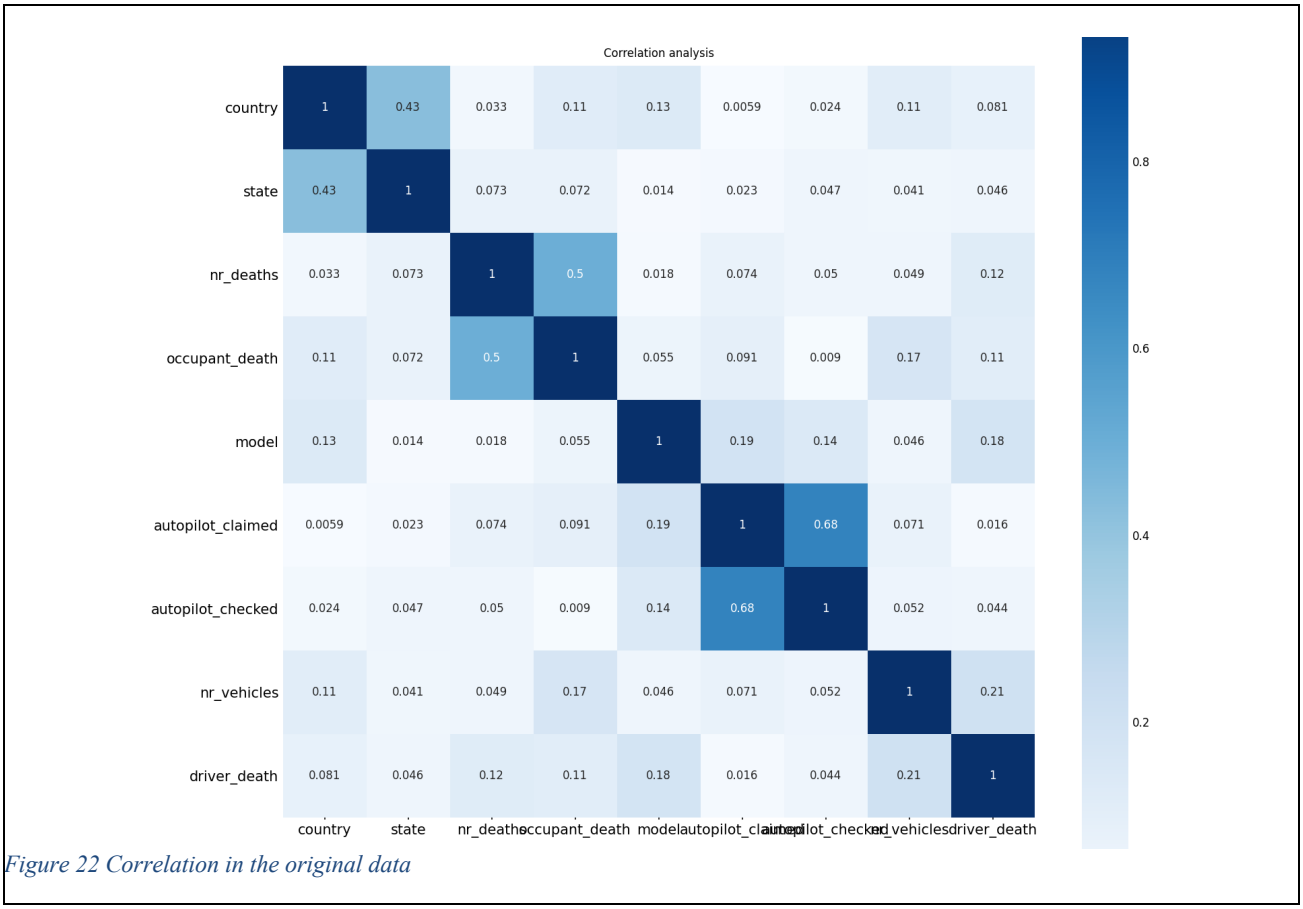


Figure 22 Correlation in the original data

1. The variable vehicles can be seen as ordinal without losing information.
2. One of the variables autopilot_claim or autopilot_checked can be discarded without losing information.
3. The variable model discriminates between the target values, as shown in the decision tree.
4. The existence of outliers is one of the problems to tackle in this dataset.
5. Variable description could be used as is for training a classifier.