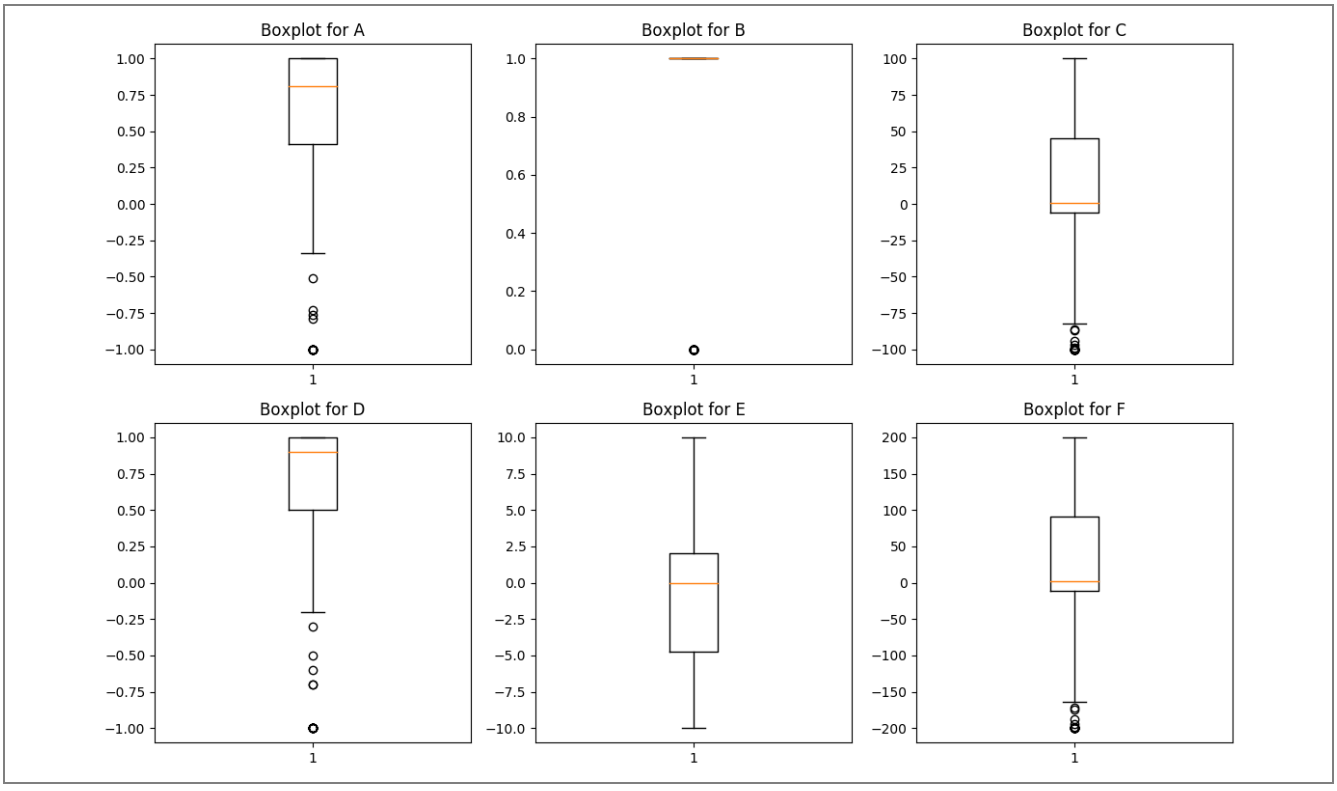
Data Science

by Cláudia Antunes

Lab Data Preparation

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.

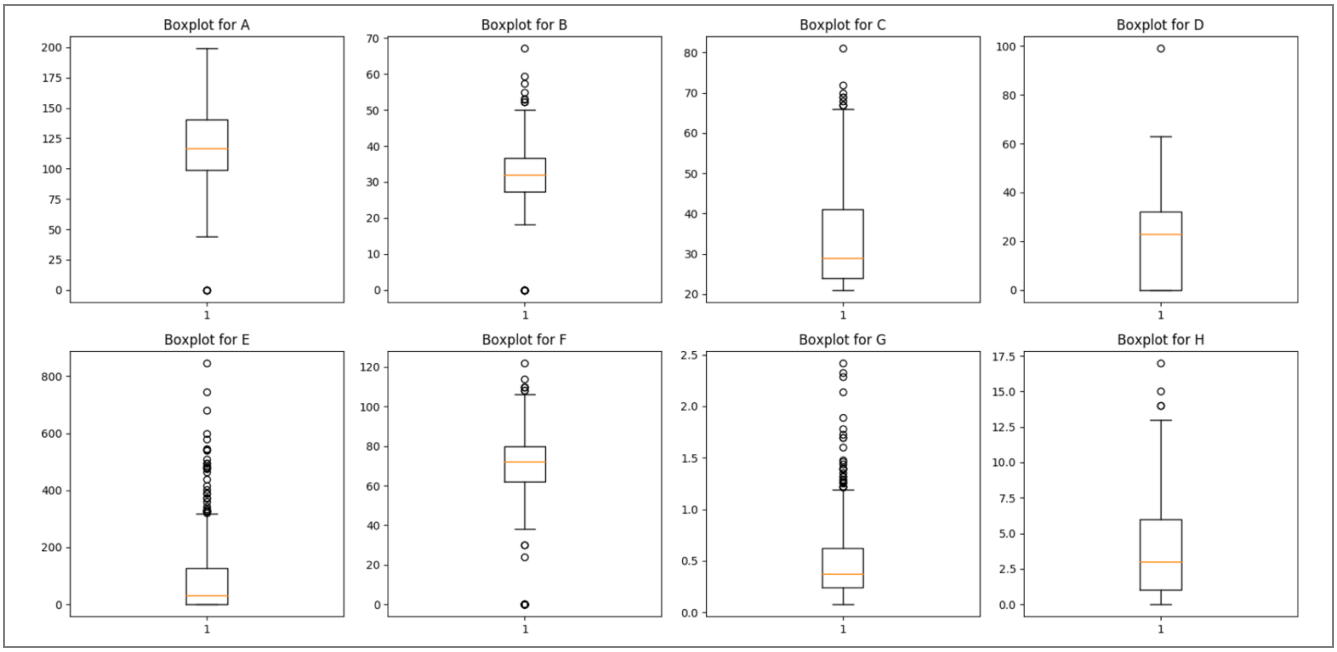


1. Normalization of this dataset could **not** have impact on a KNN classifier.
2. Discretization, independently of the technique applied, could change the performance of a decision tree trained over this dataset.
3. There is evidence in favour for sequential backward selection to select variable F previously than variable A.
4. The application of PCA generates at most six principal components where each eigenvector is necessarily given by a vector of length six.

5. Knowing that C and F are strongly correlated (correlation=1), we can say that removing one of those variables, would not have any impact on the performance of a KNN classifier.

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. Normalization of this dataset should have a high impact on naïve Bayes classifier.
2. Missing value imputation using the mean value per class improves the quality of discovered patterns.
3. Minkowski distances between time series disregard temporal dependencies.
4. Classic imputation methods over tabular data are generally adequate to impute missings in time series data.
5. Moving average can be applied to smooth a time series.

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.

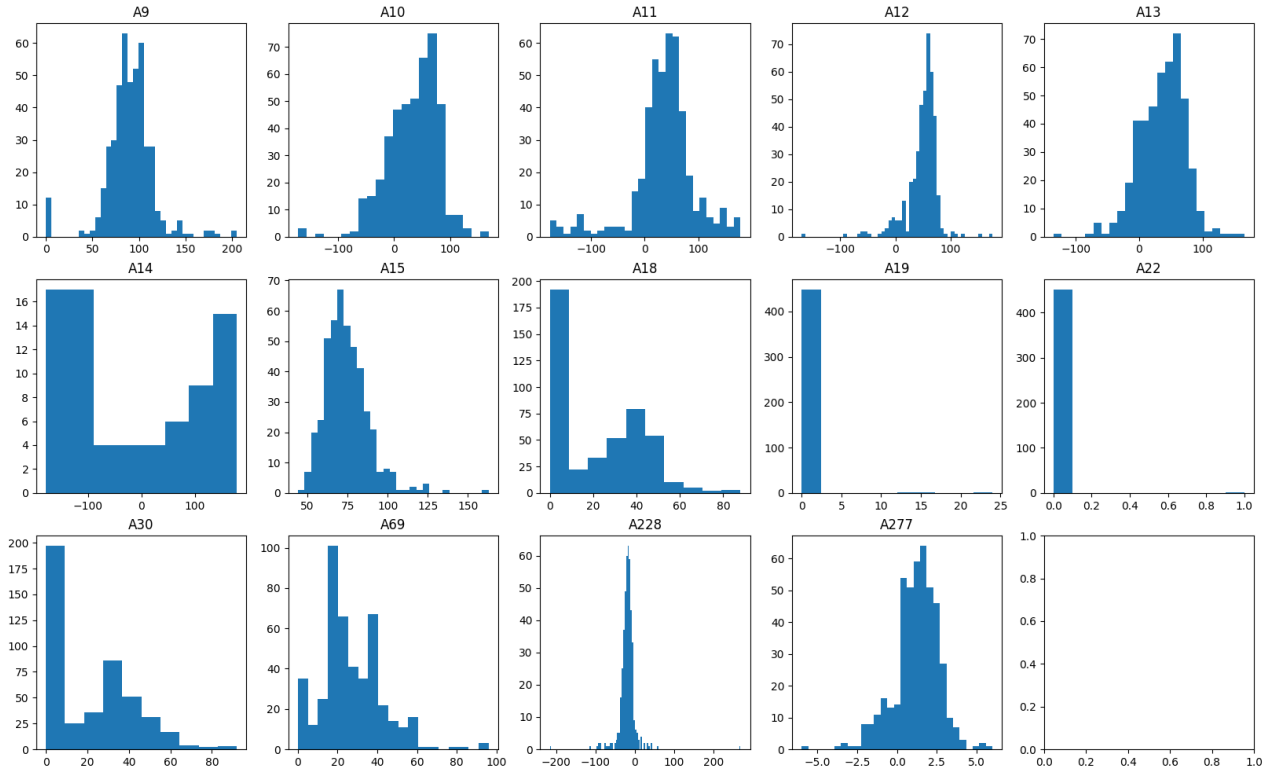


Figure 1 Histograms for some variables

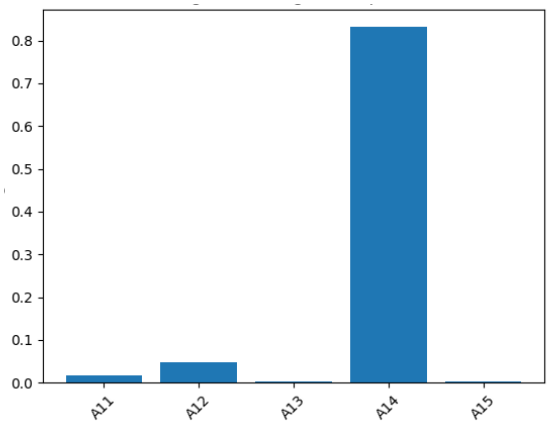


Figure 2 Variables with Missing Values – percentage of records

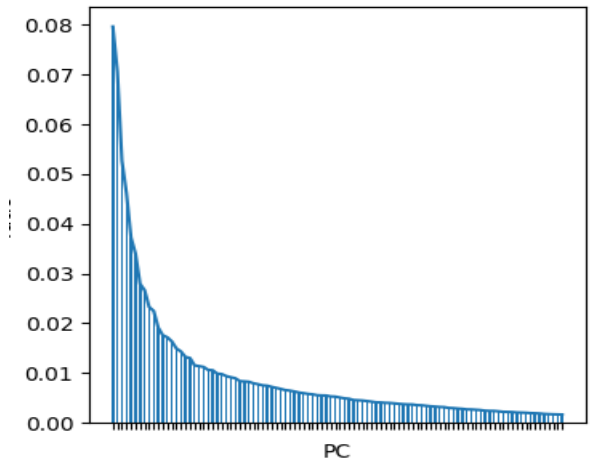


Figure 3 Explained variance for the first 100 principal components

1. It is better to drop the variable A14 than removing all records with missing values.
2. Dummifying the variables will improve the mining results.
3. Removing the Z variable from the training will improve model performance over any non-observed records.
4. The first 10 principal components are enough for explaining half the data variance.
5. Feature generation based on both variables A9 and A19 seems to be promising.

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.

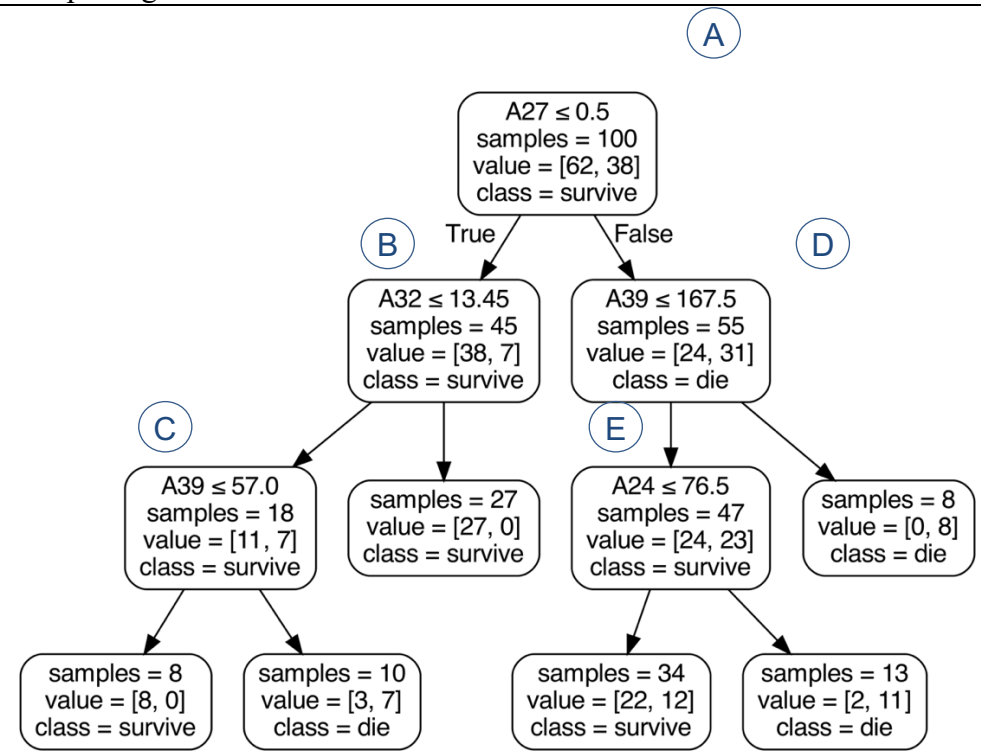


Figure 4 Decision tree trained over 100 records

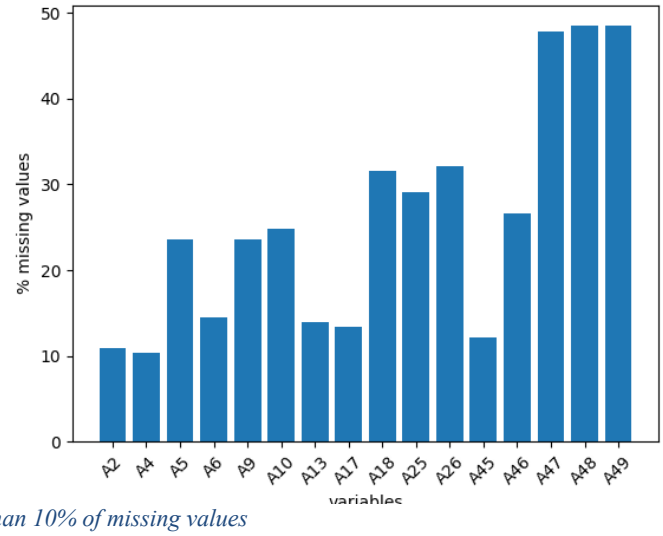


Figure 5 Variables with more than 10% of missing values

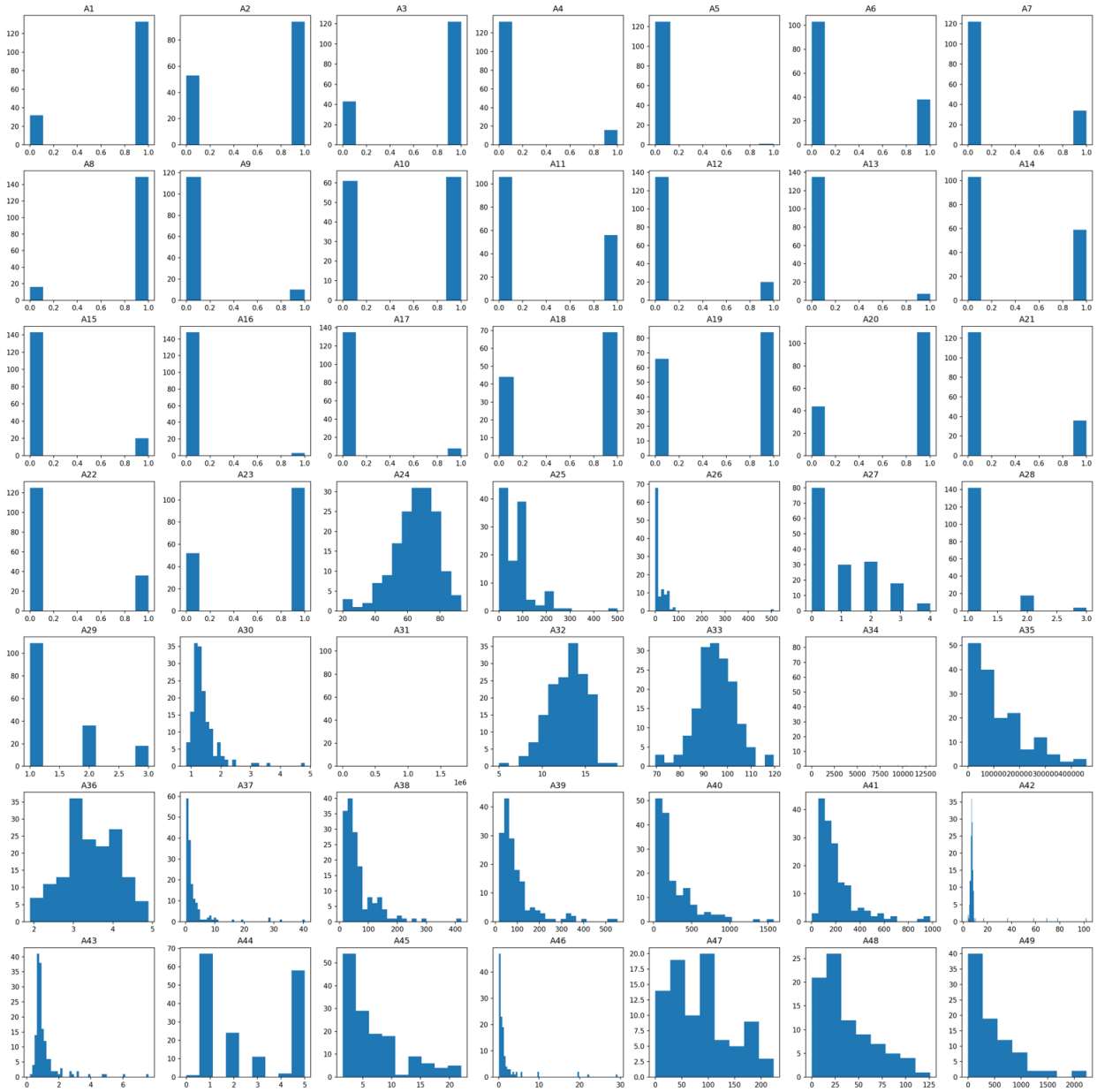


Figure 6 Histograms for all descriptive variables

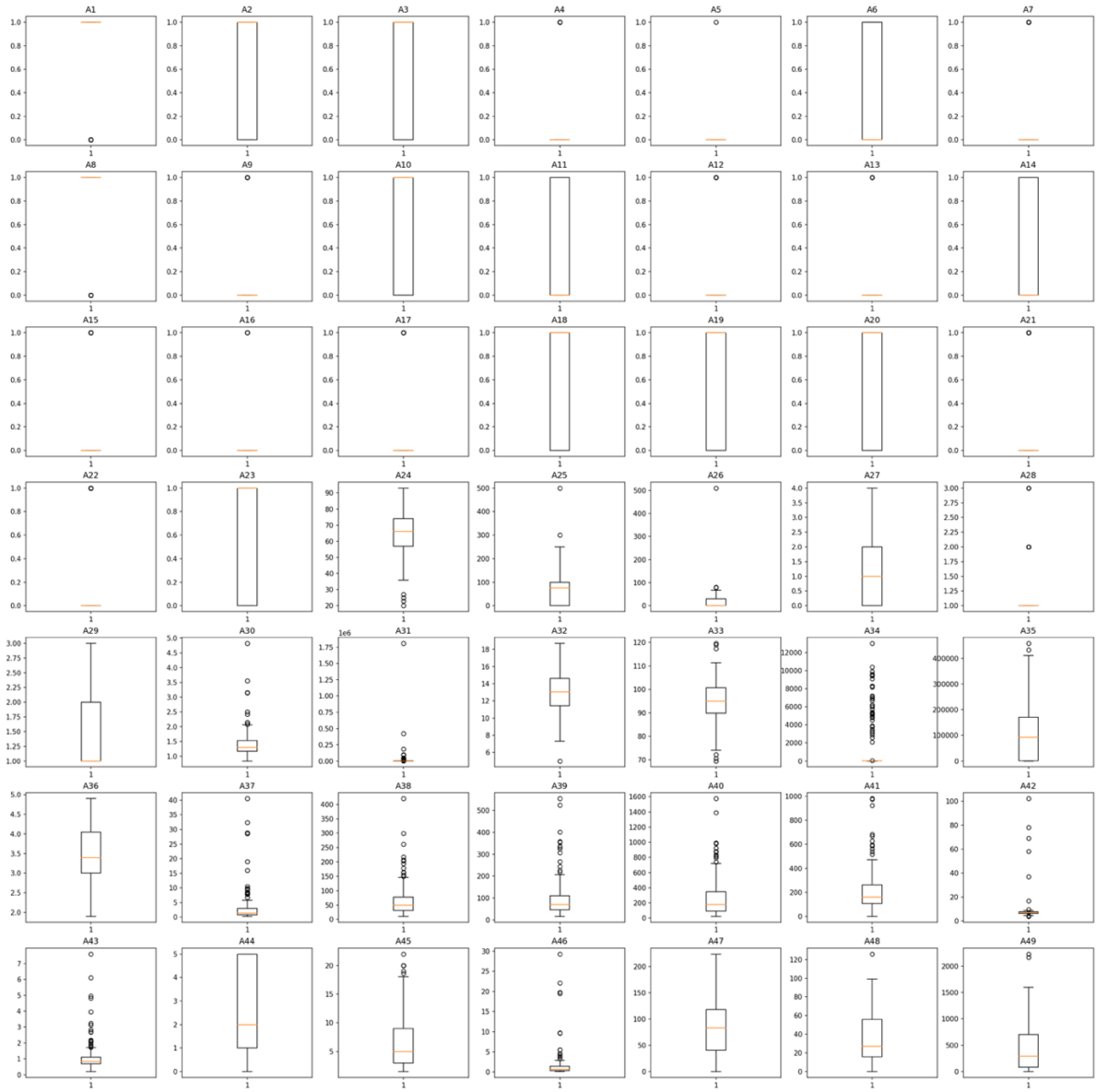


Figure 7 Boxplots for all descriptive variables

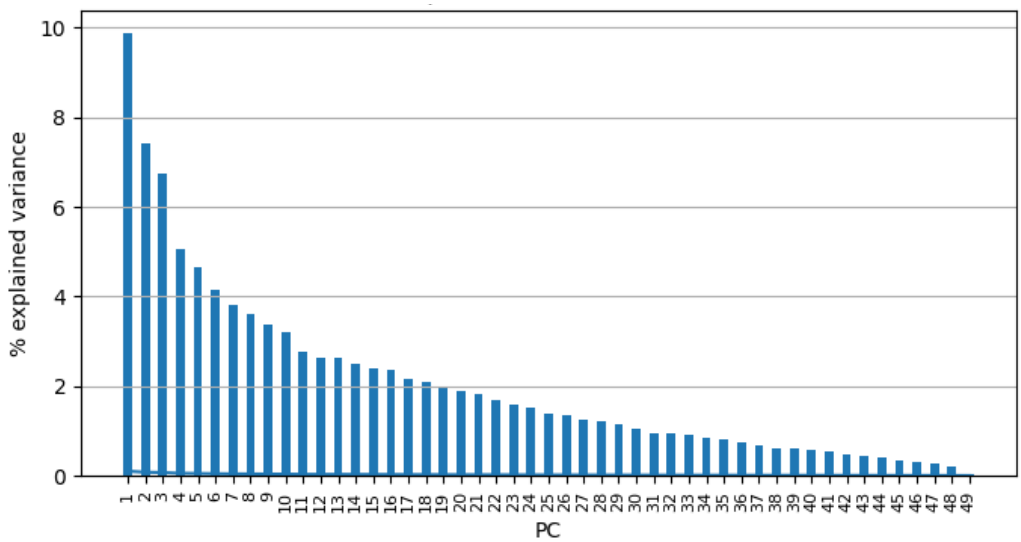


Figure 8 Explained variance for each principal components

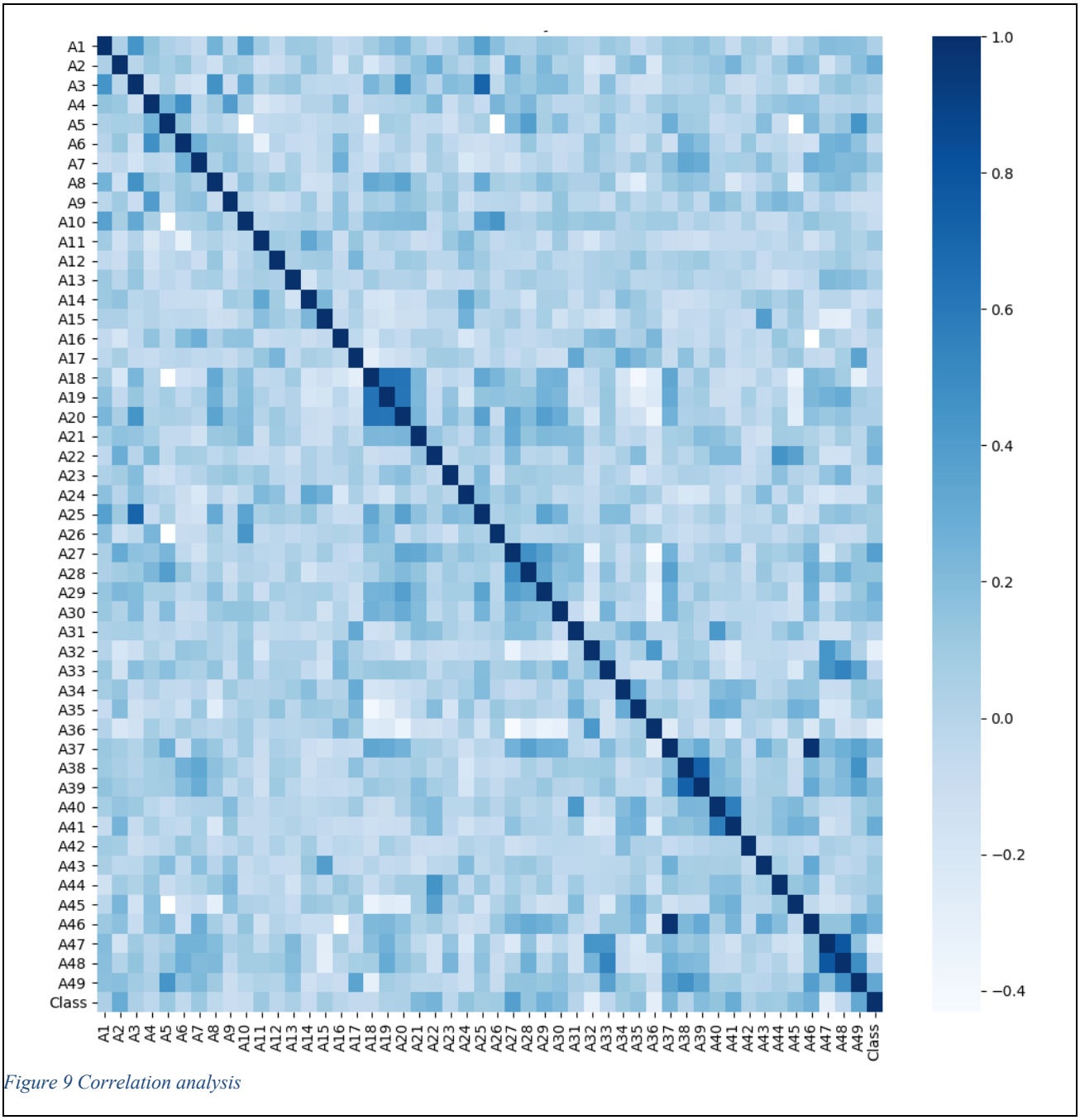


Figure 9 Correlation analysis

1. If A38 and A39 were redundant then selecting just one of them would obviously increase the accuracy of KNN.
2. Dummification is mandatory in this dataset.
3. Discarding variables A48 and A49 would be better than discarding all the records with missing values for those variables.
4. Using the first 30 principal components would imply an error between 10 and 20%.
5. Feature generation based on the use of variable A16 **wouldn't be** useful, but the use of A5 seems to **be** promising.

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.

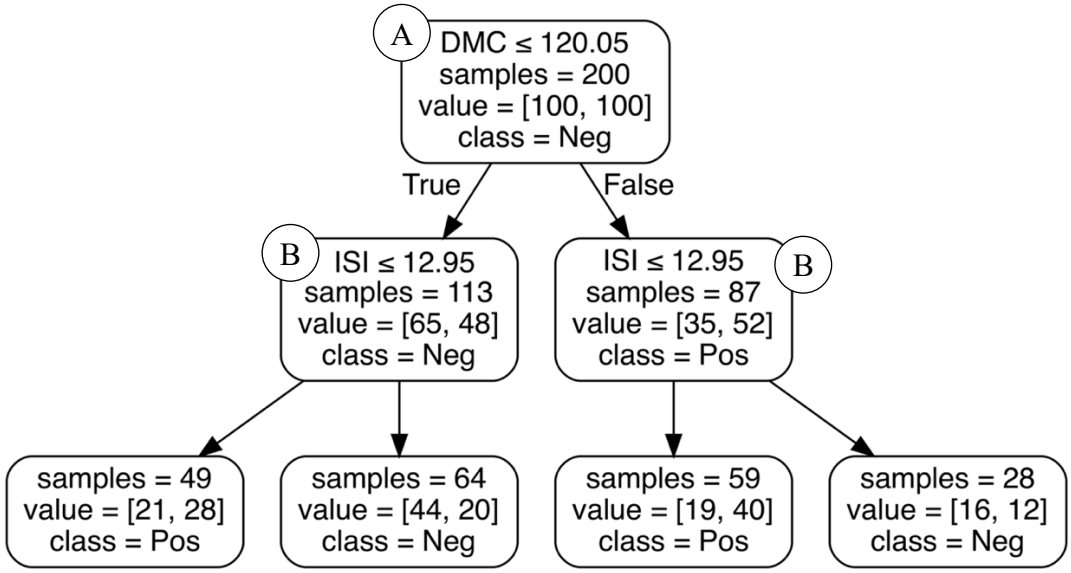


Figure 10 Decision tree trained over 200 records

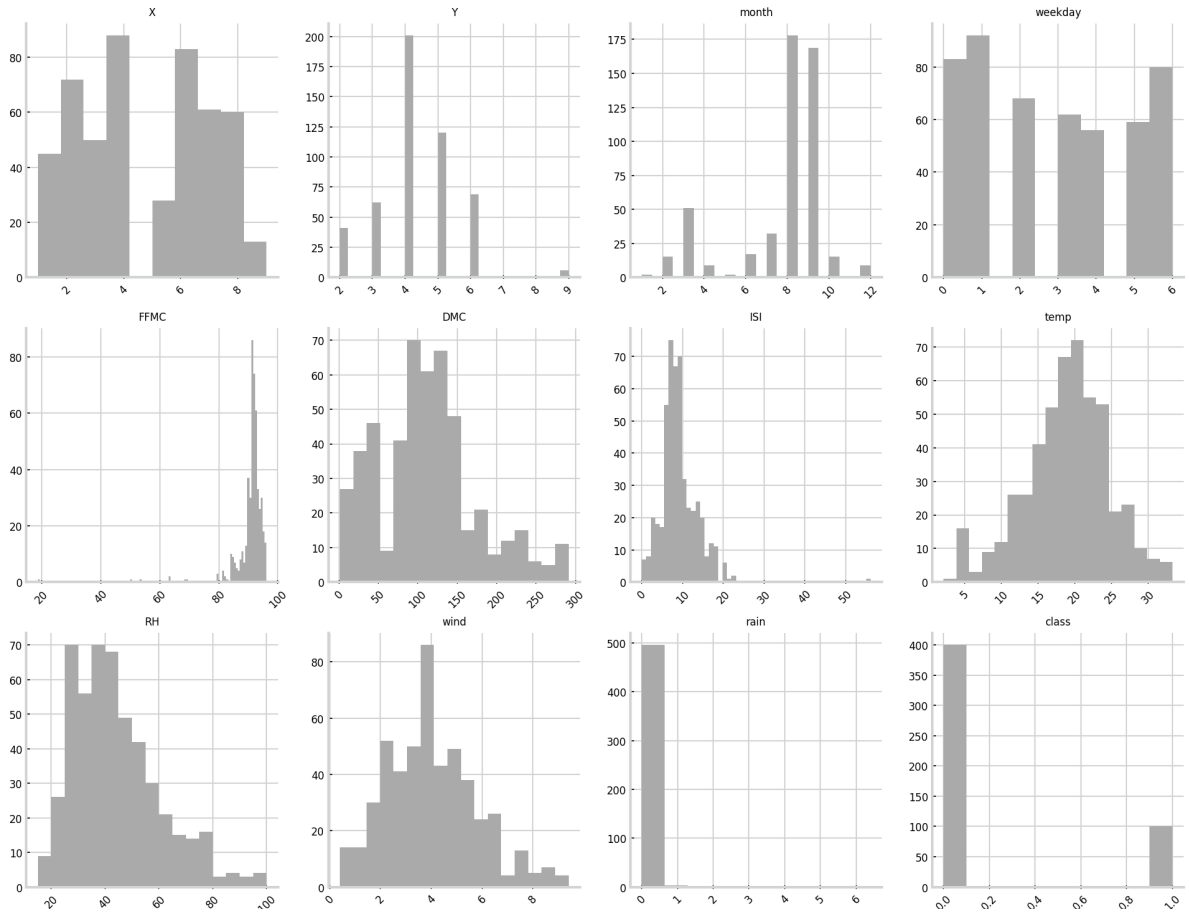


Figure 11 Histograms for all variables

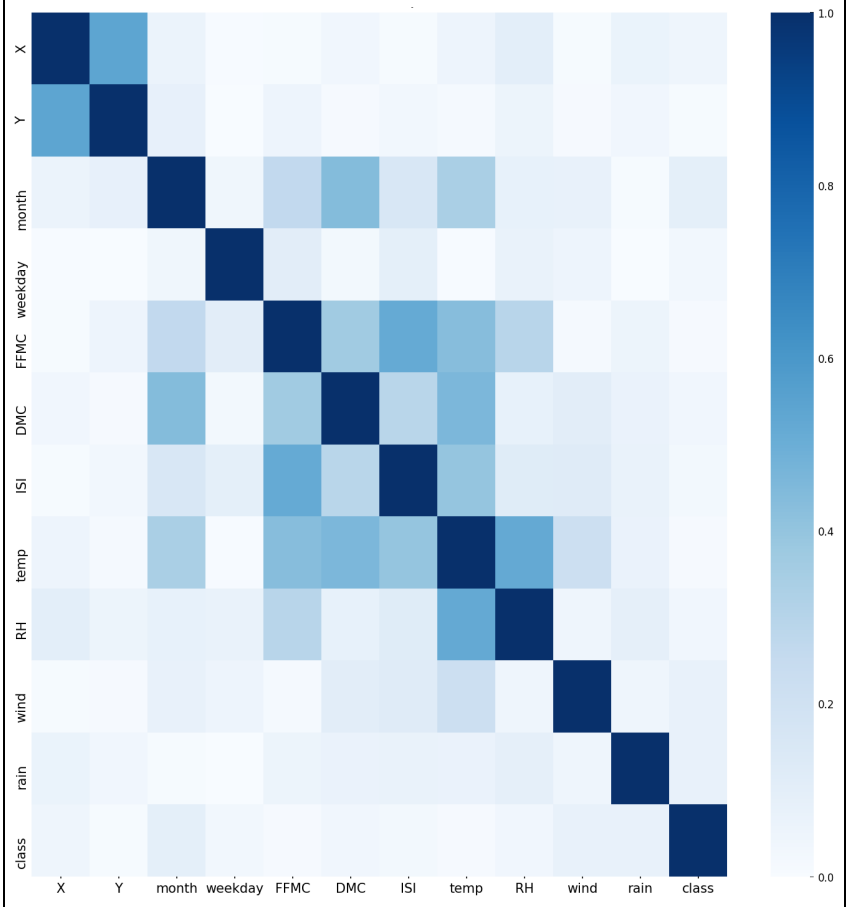


Figure 12 Correlation analysis

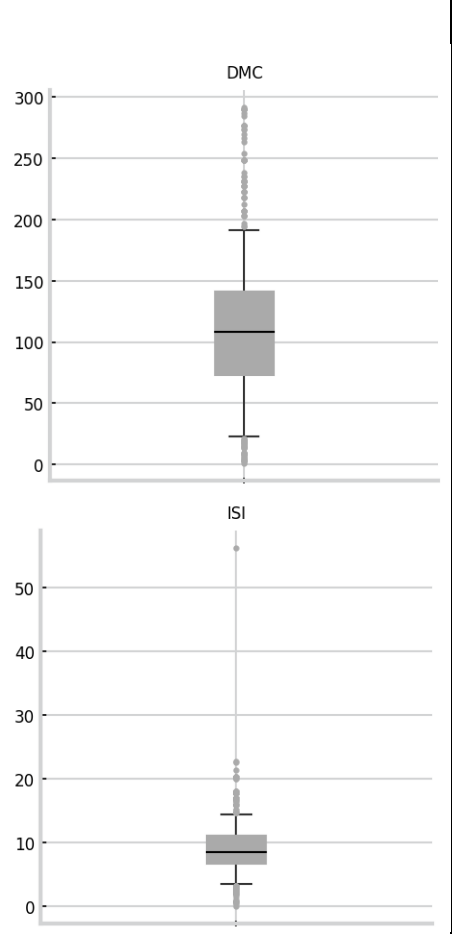


Figure 13 Boxplots for some variables

1. A scaling transformation is mandatory, in order to improve the **Naïve Bayes** performance in this dataset.
2. Removing variable ***rain*** might improve the training of **decision trees** .
3. Given the usual semantics of ***weekday*** variable, d*ummification* would have been a better codification.
4. Applying a non-supervised feature selection based on the redundancy, would **not increase** the performance of the generality of the training algorithms in this dataset.
5. Balancing this dataset by SMOTE would be **riskier** than oversampling by replication.

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.

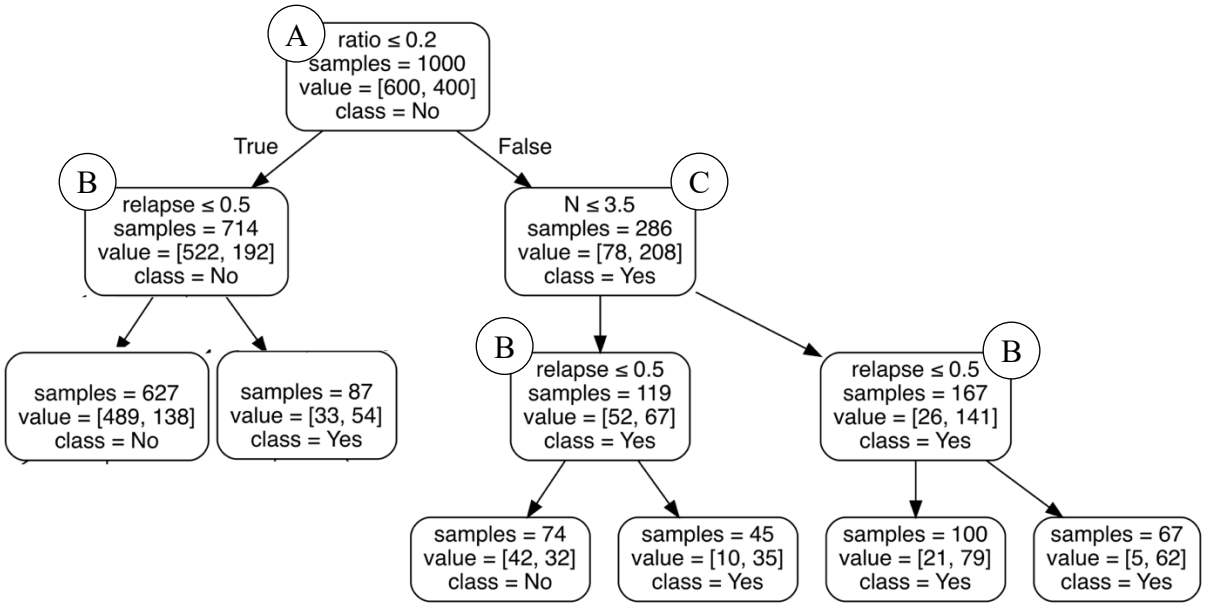


Figure 14 Decision tree trained over all records

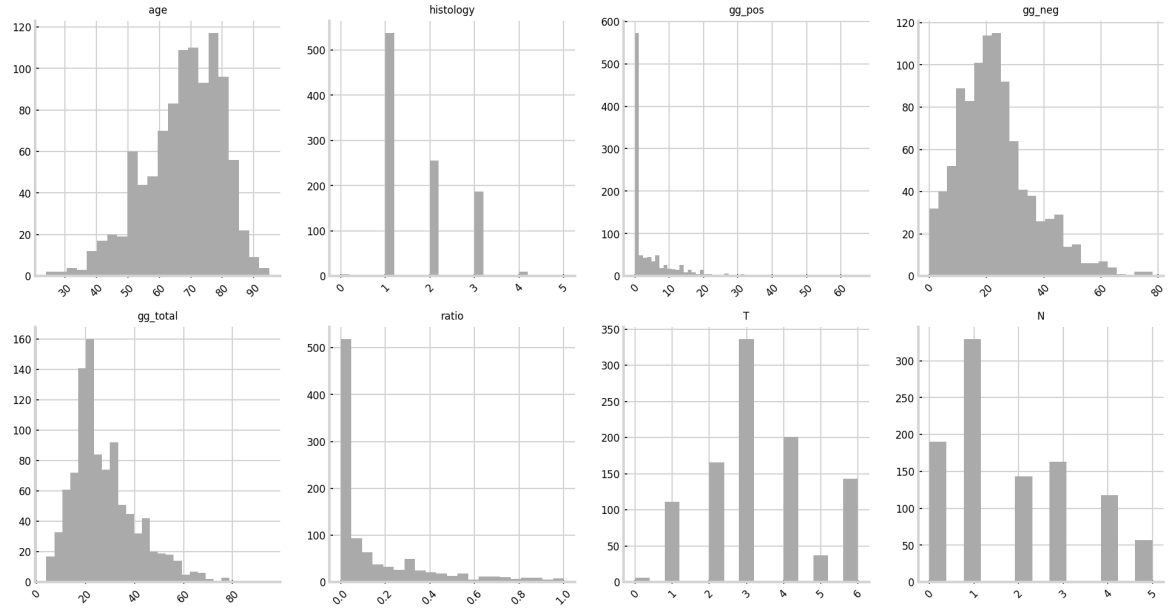


Figure 15 Histograms for non-binary variables

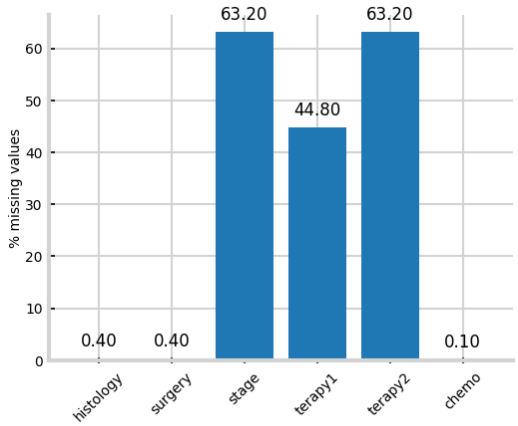


Figure 16 Number of missing values per variable (right)

1. Multiplying ***ratio*** and Boolean variables by **100**, and variables with a range between 0 and 10 by **10**, would have an impact similar to other scaling transformations.
2. Dropping **all rows** with missing values can lead to a dataset with less than **25%** of the original data.
3. Variable ***T*** and ***N*** could have result from an equal-frequency discretization operation.
4. Not knowing the semantics of ***histology*** variable, *dummification* could have been a more adequate codification.
5. Balancing this dataset by SMOTE would most probably be preferable over undersampling.

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.

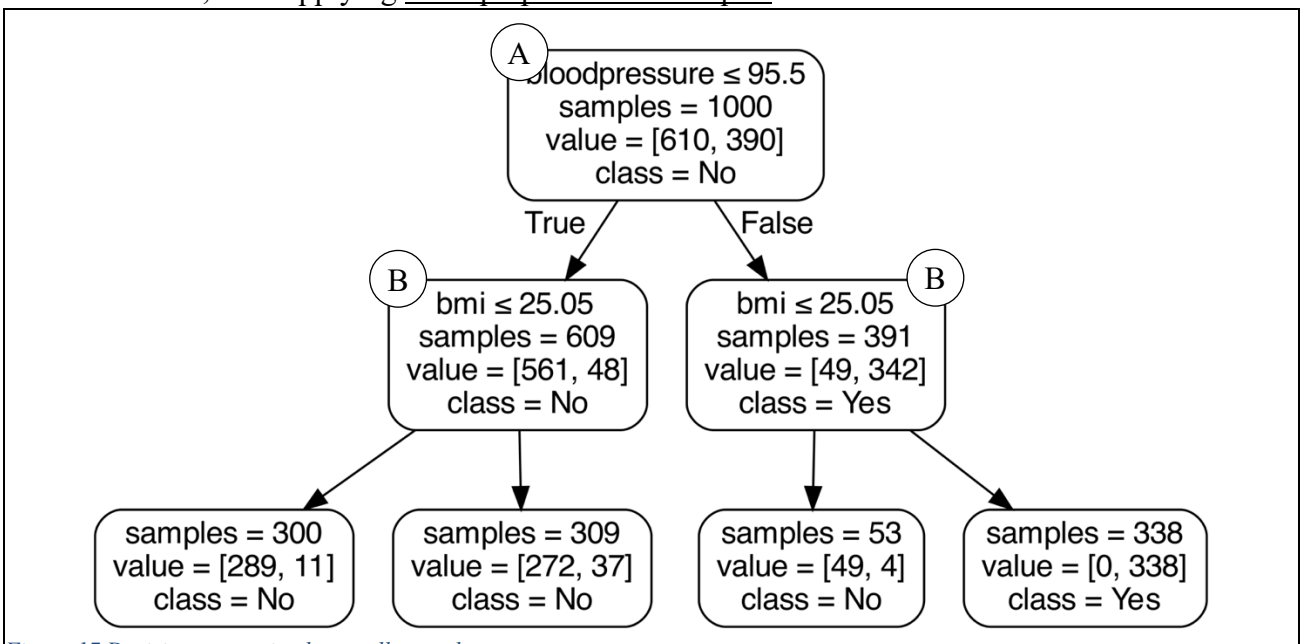


Figure 17 Decision tree trained over all records

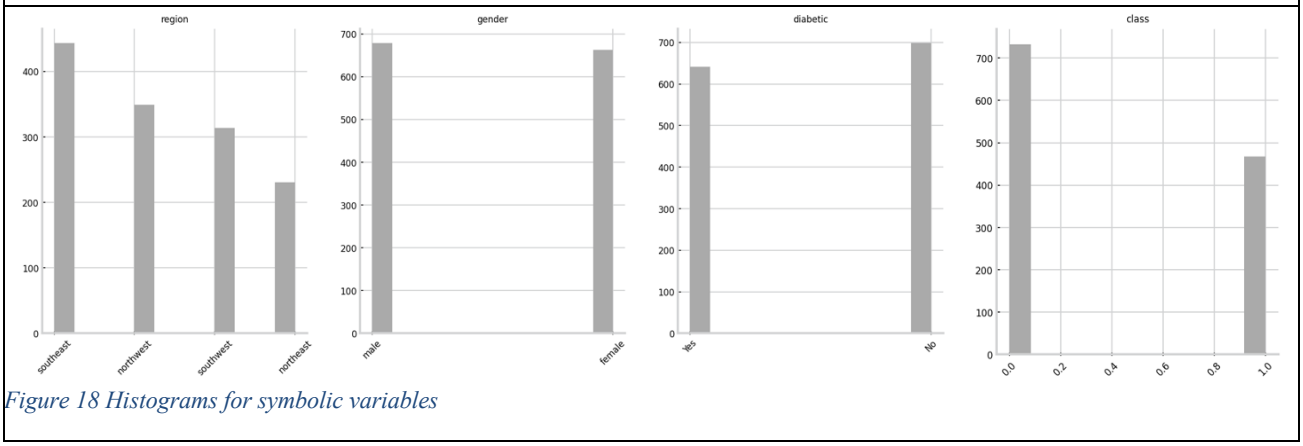


Figure 18 Histograms for symbolic variables

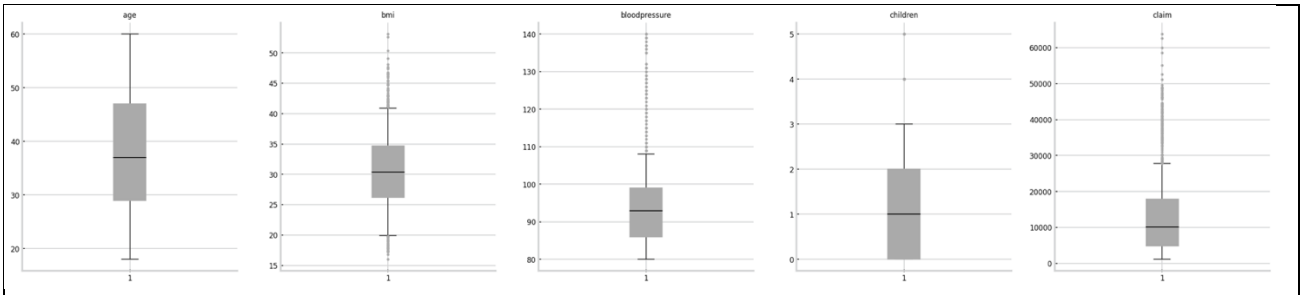


Figure 19 Boxplots for numeric variables

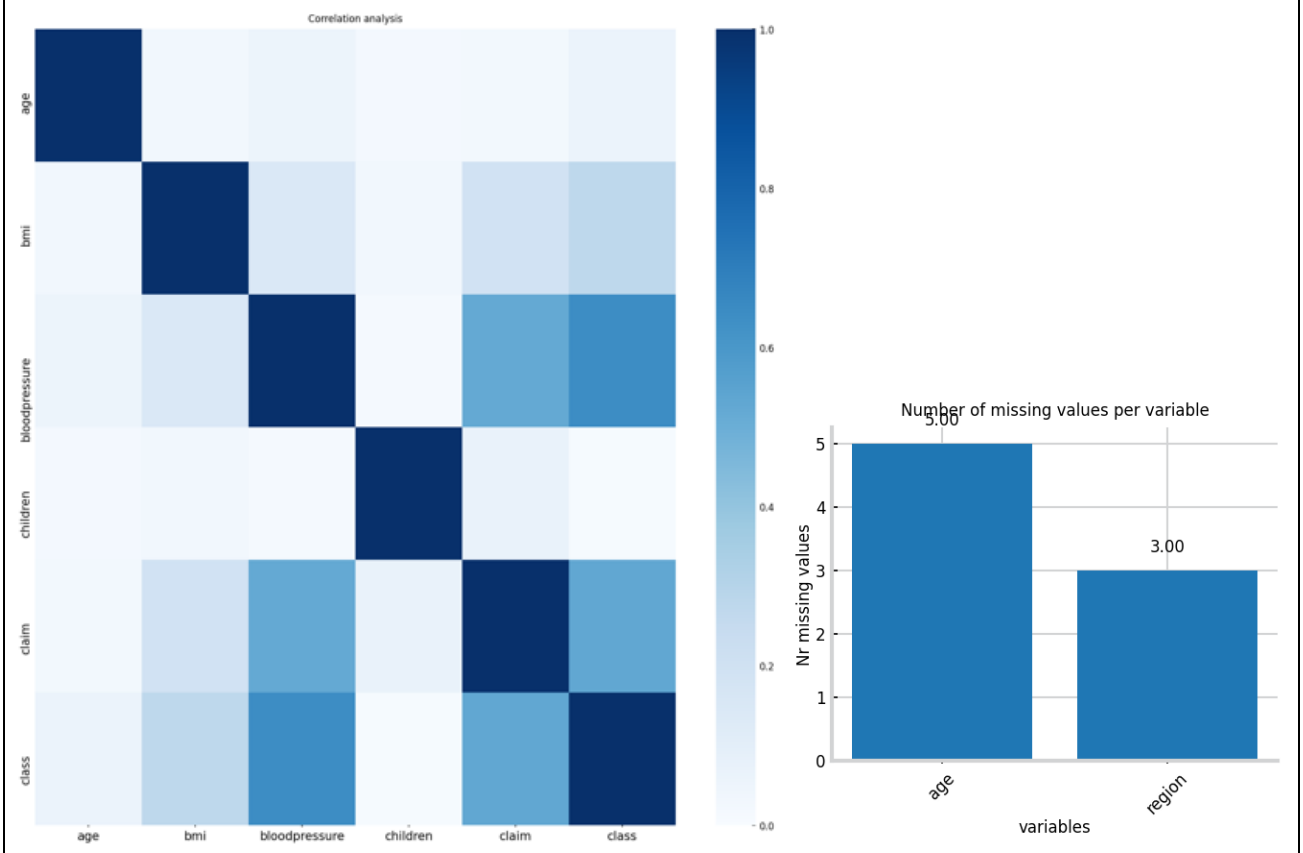


Figure 20 Correlation (left) and variables with missing values (right)

1. Given ***claim*** has a larger range than the rest of variables, applying a scaling transformation **could** reduce the performance of KNN algorithm.
2. **Dropping all records** with missing values would be better than to **drop the variables** with missing values.
3. Knowing ***bmi*** establishes a threshold for obesity determination, it can be used as the basis for feature generation.
4. Considering the common semantics for ***region*** variable, *dummification* would be the most adequate encoding.
5. Balancing this dataset would be mandatory to improve the results.

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`.

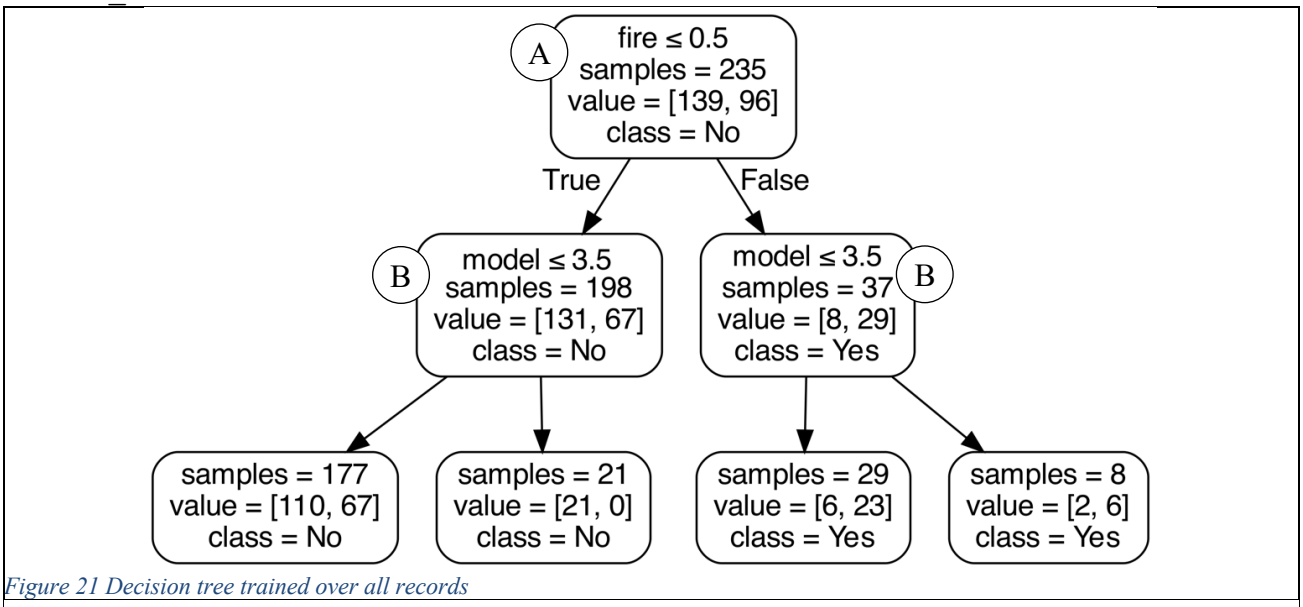


Figure 21 Decision tree trained over all records

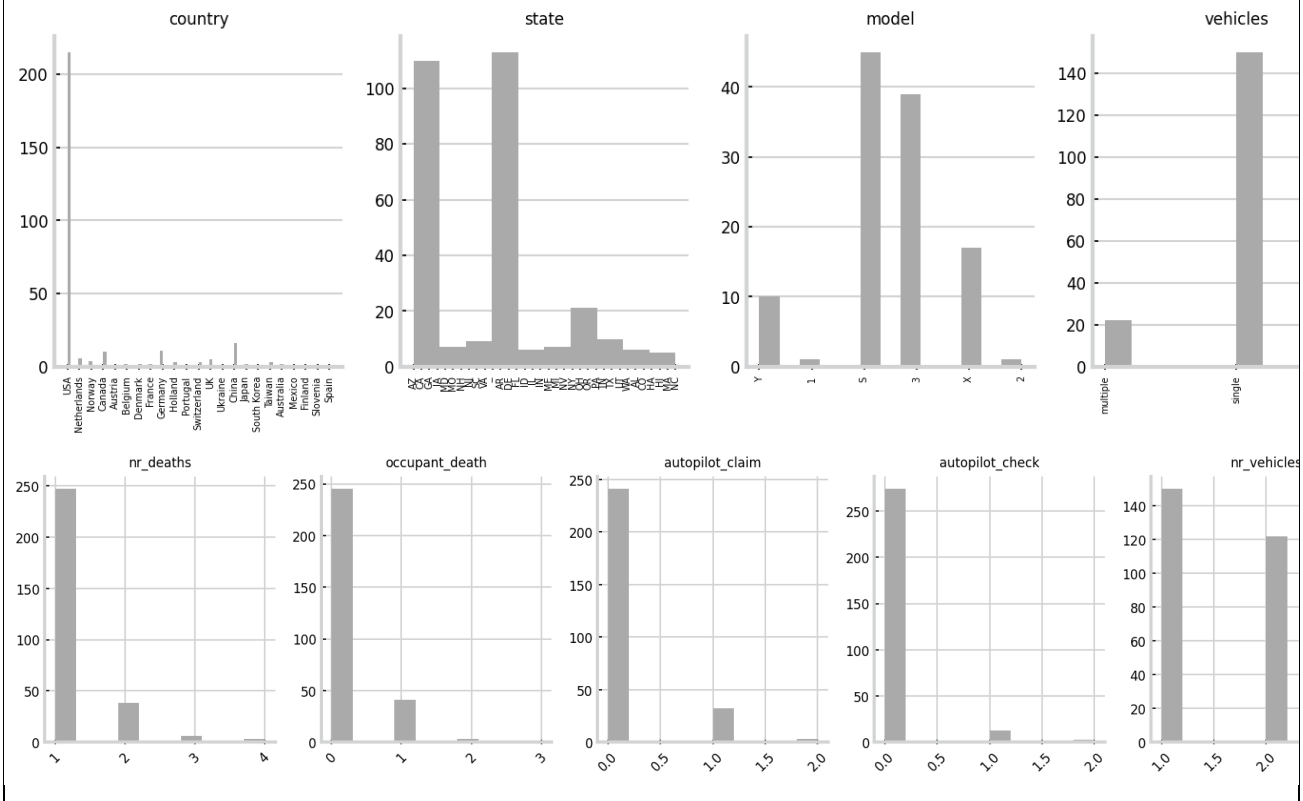


Figure 22 Histograms for original variables (excluding description and driver_death)

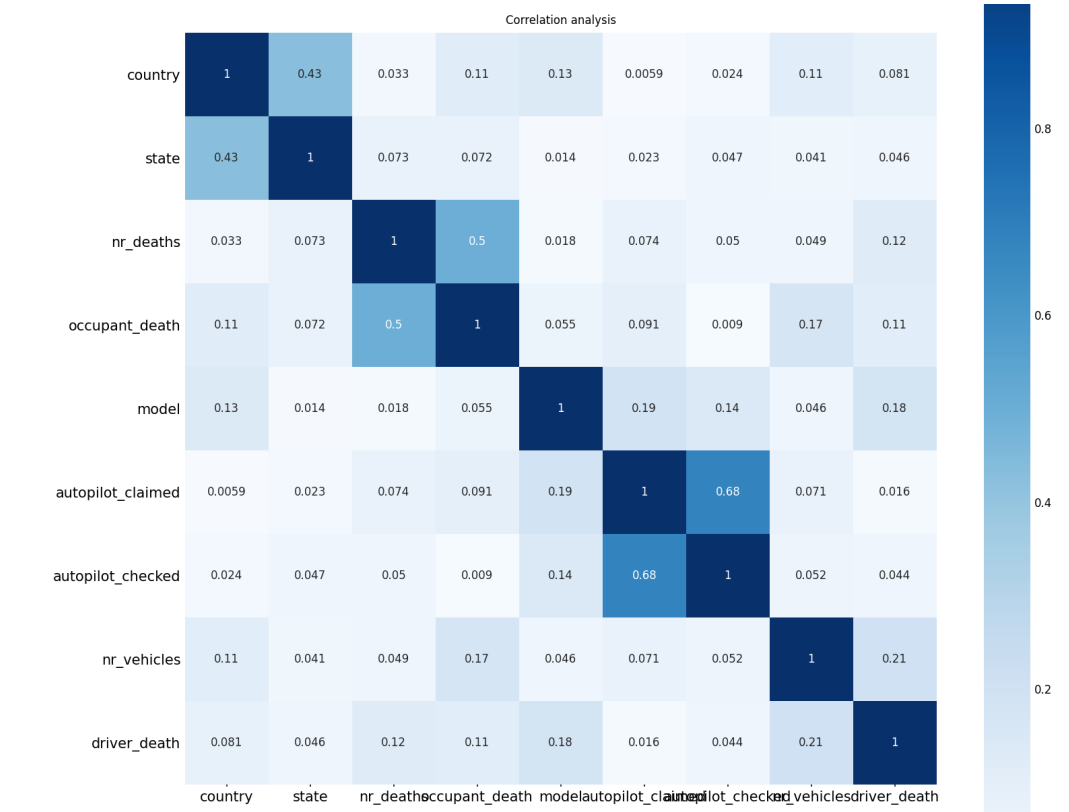


Figure 23 Correlation in the original data

1. Scaling this dataset would be **mandatory** to improve the results with distance-based methods.
2. Considering the common semantics for *country* and *state* variables, dummification if applied **would increase** the risk of facing the curse of dimensionality**.**
3. Both *country* and *state* variables **could be** used to derive a new variable using a concept hierarchy.
4. Knowing that the *description* variable was used to extract some new variables (*fire* for example), we can say that feature extraction played an important role in solving the classification problem.
5. Knowing that *fire*, *collision* and *control_loss* were terms extracted from the original *description* variable, they **can be** seen as dummy variables.