TÉCNICO LISBOA

DS 2021

Data Science Project

Team nr: 12	**Student 1 :** Alexandre Miguel da Silva Pires	**IST nr:** 92414
	Student 2 : Diogo Miguel Rodrigues Cruz Fouto Silva	**IST nr:** 93705
	Student 3 : João Nuno Bastos Fonseca	**IST nr:** 92497

1 Data Profiling

In this section we peruse the data to assess its content and general quality, as well as predict what transformations are necessary and which will have a considerable impact on future models, based on the profile of the data. The mentioned figures for Data Profiling are all available in the appendix.

Data Dimensionality

By analysing the datasets' dimensionality, we aim to find some information regarding the size of the data we have, what kind of values it holds, and how many values are missing for each variable.

The 'Air Quality' (AQ) dataset presents 169273 records with 31 variables each. Most variables are **numeric** (see ***Figure 1***) and by applying some domain knowledge on these, we can infer that many of them are correlated; they refer to the max, min, or average values of some air components. ***Figure 2*** gives us the number of missing values for each variable, which are not that many when compared to the number of records. With such basic information, we are unable to make predictions about how the data will affect our models; we have no data on the relevance of each variable, the spread of missing values across records, or the quantity of missing values in each record, but we know that missing values will have to be dealt with because of `scikit-learn`*'s* limitations.

The 'NYC Collisions' (NYC) dataset presents a smaller number of records (45669) and variables (20), yet these variables are mostly **symbolic**, as shown in ***Figure 3***, which ideally have to be correctly handled to preserve any relations the symbols might have. NYC presents a considerable amount of missing values for four of the variables, but like AQ, we can't make predictions about how this will impact our models yet. Some information we found when applying domain knowledge: "VEHICLE_ID" is likely irrelevant, since most vehicles do not crash repeatedly, and a lot of different possible values will exist for variables like "CONTRIBUTING_FACTOR" and "SAFETY_EQUIPMENT;" as collisions occur for a myriad of reasons, and have to be handled cautiously to prevent an explosion in the number of variables.

Data Granularity

Data granularity refers to the study of the variance of each numeric variable and how much information is lost if we consider different bin sizes for each variable.

For the AQ dataset: some of the variables are constrained to a small number of different values, like the PM10 measurements (***Figure 5***), but with very disparaging number of records per each of those values, and they therefore would lose with generalisation, i.e., a smaller number of bins. Other variables, like the NO2 measurements (***Figure 6***), present a larger number of possible values and a clearer distribution that could be approximated with a smaller number of bins (but it would lose some specificity, nonetheless).

The NYC dataset only contains three numeric variables, but two of them are irrelevant; they are IDs. ***Figure 7*** presents the granularity for the remaining relevant numeric variable, the person's age. Domain knowledge would suggest that applying bins of 5-10 years doesn't result in a big loss of generality, although such cannot be seen in ***Figure 7***.

Data Distribution

When analysing the data's distribution, we aim to find information about the standard deviation of each variable, how many outliers exist, find trends, and estimate distributions for each variable.

The relevant AQ results are in ***Figures 8-10***. ***Figure 8*** shows the distribution of values for some variables. Most variables reside in stable ranges, but a few – PM2.5 and PM10 variables, for instance – vary quite wildly. From this we can deduce that the latter variables may be of increased importance for our analysis. ***Figure 9*** presents the outliers per variable, which are few when compared to the number of records. The trend histograms suggest possible distributions for the variables. For instance, ***Figure 10*** suggests a "Log Normal" distribution (the green curve) for CO_Min. Finally, the histogram for ALARM, the target variable, hints at the data's imbalance; records with a "Safe" ALARM outnumber the ones with "Danger" by almost a factor of 10.

The relevant NYC results are in ***Figures 11-15***. ***Figure 11*** shows the boxplot chart for the only relevant numeric variable, PERSON_AGE. Some values seem suspicious: some people are 120 years old. The histograms for the symbolic variables present much more relevant information, like what safety equipment people were wearing (***Figure 12***), what injuries they had (***Figure 13***), and which sex is involved in more collisions (***Figure 14***). Finally, the target variable's values seem highly unbalanced but realistic: there are far fewer dead people than injured. (***Figure 15***).

Data Sparsity

When studying the data's sparsity, we want to find the correlation among variables and the range of values the variables cover (i.e, domain coverage). For this we present correlation heatmaps and scatter plots that show the former and the latter, respectively.

For the AQ dataset, ***Figures 16 and 17*** show us that symbolic variables are almost totally covered; there are records for almost all cities and provinces. Regarding numeric variables, ***Figure 18 and 19*** give us the scatter plots for PM10_MEAN and NO2_Std, respectively. These also suggest good domain coverage; the most relevant values for both air components are accounted for. What's more, they go further and suggest that PM10 is a good measure for predicting the target variable; the same cannot be said for NO2_Std. Finally, ***Figure 20*** gives us the correlation

between all variables. From it, we can infer that variables which track different metrics of the same air components are highly correlated, as we predicted in the **Data Dimensionality** section. We can apply domain knowledge further: in places with high CO emissions, other pollutants, NO2 and PM, for example, are also more likely to exist in abundance.

For the NYC dataset, the results are present in ***Figures 20-23***. Again, the numerical variables give little insight, but from the scatter plots of the symbolic variables, we get good domain coverage (***Figures 20 and 21***); the relevant values for the variables are present. In the correlation matrix (***Figure 22***), we see some irrelevant correlations, like PERSON_ID x VEHICLE_ID; as the average person only drives one or two cars. But there are also relevant correlations, like between PED_ACTION and PED_LOCATION; the combo action-location is likely very much to blame for accidents caused by pedestrians. Finally, we can conclude that this dataset presents fewer highly correlated variables than the AQ one.

2 Data Preparation

In this section we clean and prepare the data for future processing, by applying different techniques to remove unnecessary values, impute missing values, and transform the datasets to ease the learning of our models.

Missing Value Imputation

The AQ dataset had few records with missing values, but we decided to fill those in because we had the tools to do it well: in Data Profiling we got good domain coverage, a simple distribution for most variables, and a lot of numeric variables. For the imputation, experiments were made using the average value for all missing numeric variables because some, like CO, behaved more or less like Log Normal distributions, so the average value was a close approximation of the most common value, and using the most frequent value for both symbolic and binary ones. This, of course, impacted the data's distribution, because the most frequent value isn't always the best choice – in cities or provinces, for instance.

We took the same approach with NYC; simply reducing the number of records would not work because there were very few records in which people were killed, which would further unbalance the data.

Dummification and other transformations

Due to the numeric nature of the AQ dataset, very few transformations were made at this phase: the city and province names were dropped, as they were paired with a numeric identifier (GbCity and GbProv respectively), and the date was transformed to a numeric value where the first date was considered 0, and each date would be numbered based on the number of days since the first date. We do note that this date transformation was not a good solution, as the measurements on the dataset were made on specific days each, which left the variable being unique for each record, similarly to an ID, which is irrelevant for modelling.

The NYC dataset provided more of a challenge, not only due to the number of symbolic variables, but the vast amount of values each variable could take, which would mean a big increase in the data size if they were all dummified. Due to this fact, we created handcrafted taxonomies that ideally would have some reasoning behind them, as to give an interpretable value to measures, like averages. The changes applied were as follows: bodily injuries were turned into numbers, where a lower number means an injury in a lower part of the body; a person's sex was classified into a simple number; lastly, safety equipments got a general number, since they could not be taxonomized. Similarly to AQ's

date, NYC's crash date and time were also transformed to the number of days and hours, respectively. Irrelevant variables were also dropped, which corresponded to ID's: 'PERSON_ID', 'ID', 'UNIQUE_ID', 'VEHICLE_ID'. The remaining variables were dummified, which were: PED_ACTION, PERSON_TYPE, PED_LOCATION, CONTRIBUTING_FACTOR_1 and CONTRIBUTING_FACTOR_2, which led to an undesirable increase in the number of variables, due to the possible values all of them could take and not reflecting possible relationships between them; they should've, instead, been taxonomized.

In both datasets, the target variables were also transformed into a binary digit, since only two values were possible.

Outliers Imputation

For both datasets, no outlier imputation was made besides the suspicious ages found in NYC. When analysing the histogram for that variable, we found that the strange scale of the x-axis was due to noise in the data. As such, records with ages above 120 and below 0 were removed. Now that the noise was removed, considering the boxplot for the same variable (***Figure 11***) we can see that the IQR considers a lot of legitimate values as outliers, and as such we didn't remove them.

As for the AQ dataset, no outlier imputation was made, since an extremely high number of pollutant gases should always be considered dangerous, therefore there was no reason for us to limit these numbers.

Scaling

Two scaling techniques, min-max and z-score, were applied to both datasets, which together with the unscaled dataset were then compared with both Naïve-Bayes and KNN classifiers, in order to assess which produced better results. The best results were picked by taking into consideration the results from both models, which proved sometimes uncertain, as two models can disagree on what is the best scaling. In such cases, we considered the parameters being tested (accuracy, recall, specificity and precision), taking recall into special consideration, as both our datasets were considerably unbalanced, so accuracy could prove unreliable.

In the case of the AQ dataset, the best results for the various scalings can be seen in ***Figure 23*** (we only present results for the best KNN classifier for simplicity's sake). While all scalings produced well-rounded results, with a few exceptions, the z-score scaling was the best.

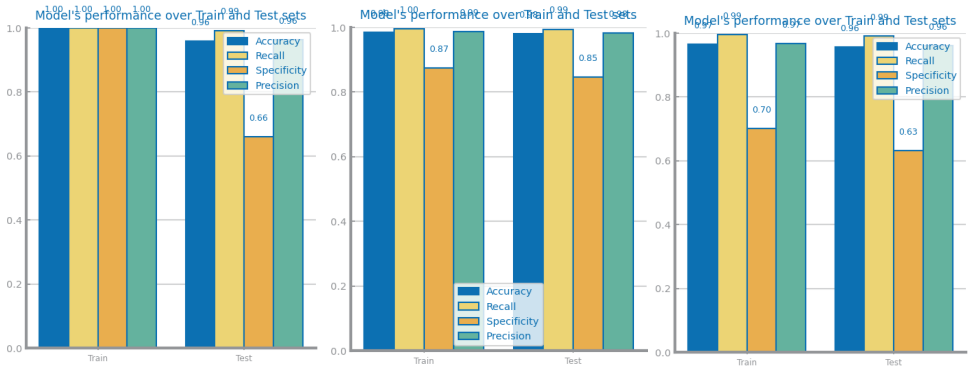

Figure 23 – Scaling Results AQ KNN - From left to right: No-Scaling, Z-Score, Min-Max

For the NYC dataset, the best results for the various scalings are in ***Figure 24*** (we only present results for the best NB classifier for simplicity's sake). These also proved well-rounded, yet, in this case, the min-max scaling that was the best.

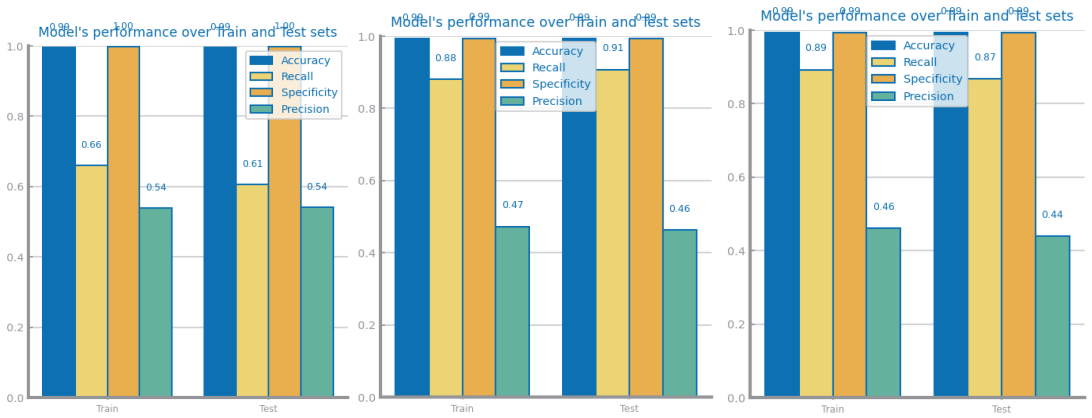

Figure 24 – Scaling Results NYC NB - From left to right: No-Scaling, Min-Max, Z-Score

Balancing

Similarly to the scaling, the resulting datasets from the previous section were tested with the following balancing techniques: SMOTE, undersampling, and oversampling. All these techniques were once again tested using Naïve-Bayes and KNN classifiers, with a selection criteria similar to the one explained in the previous section.

In both datasets the SMOTE technique yielded the best results, even if just by a slim margin. The AQ and NYC results after applying SMOTE are displayed in ***Figure 25***, for both the Naïve-Bayes and KNN models.

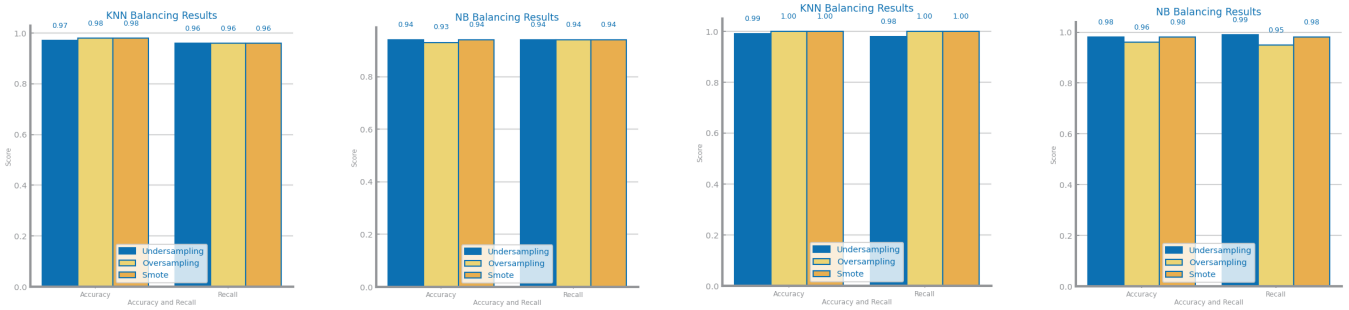

Figure 25 – Balancing Results - From left to right: AQ KNN, AQ NB, NYC KNN, NYC NB

It is important to note the risk of applying SMOTE to a dataset, as it will result in the creation of new records that are tagged as the least occurring value of the target class, that is, “Danger” for AQ’s “ALARM”, and “Killed” for NYC’s “PERSON_INJURY”. Since it is possible that these new records could never *truly* appear in real data, we are compromising the correctness of the data in order to hopefully increase the accuracy of future models. The result is that the two datasets now contain an equal number of records for every value the target class can take, and each of these records is unique, leaving future models with more data to process, even if some of this data was generated.

Finally, it is important to address that, due the linear nature of the project’s development, it was unfeasible for us to test if these were indeed the best choices for future models. It is likely that a better combination of data preparation techniques could be found.

3 Classification

In this section we compare the classification performance of various machine learning models. We split the original datasets with a 70/30 ratio to make the train and test sets, respectively. We feed the algorithms with the train data. We

prepare it as described in 'Data Preparation,' or choose and remove redundant and irrelevant features with feature selection and feature extraction on our scaled dataset. In feature selection we studied the variance, correlation, and F-score, separated and combined. For AQ, a simple analysis of the variance with a threshold of 2.0 removed 20 irrelevant features and delivered the best results. For NYC, discarding highly correlated (threshold = 0.8) features produced the best results. For feature extraction, we applied PCA on both AQ and NYC. The explained variance ratios are presented in ***Figures 27 and 28***. We decided to keep 80% of the variance for both datasets – this is the best value for the AQ dataset, according to the Elbow Method, but not for NYC, where we decided to keep the same explained variance ratio nonetheless, because we were scared to reduce the variance of the dataset as much as the Elbow Method suggested. We test the algorithms with the test data, which we prepare the same ways as described above, but with no balancing. ***Figure 26*** concisely describes the data preparation flow.

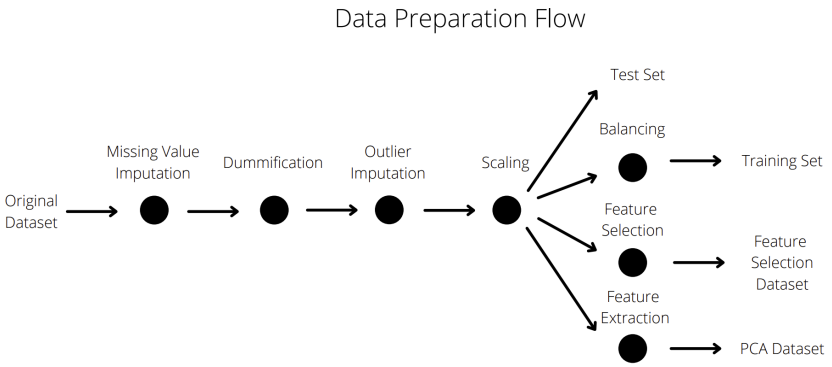

Figure 26: Data preparation flow

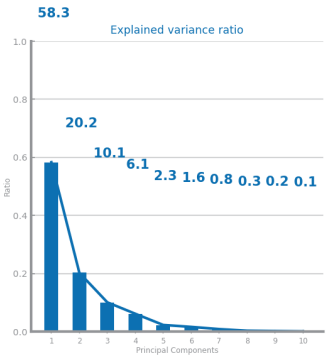

Figure 27 – AQ - Explained Variance Ratio

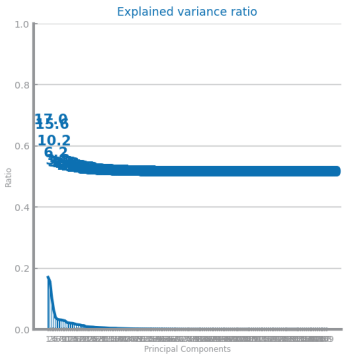

Figure 28 – NYC - Explained Variance Ratio

3.1 Naïve Bayes

The Naïve Bayes model was studied by varying different types of estimators: Gaussian, Multinomial, and Bernoulli.

As ***Figure 29*** shows the results for the different implementations of Naïve Bayes are very similar to each other on the AQ dataset, with GaussianNB having the best accuracy. This is likely because the assumption that the features fit, approximately, a Gaussian distribution is a good assumption compared to the assumptions the other implementations make.

As ***Figure 30*** shows that, in NYC, GaussianNB has a poor performance when compared to the other two. This is likely because most variables do not fit a Gaussian distribution. The multinomial and Bernoulli implementations of the NB algorithm should have the same performances, since the classification, in this case, is binary. The small difference in performance may be due to implementation details.

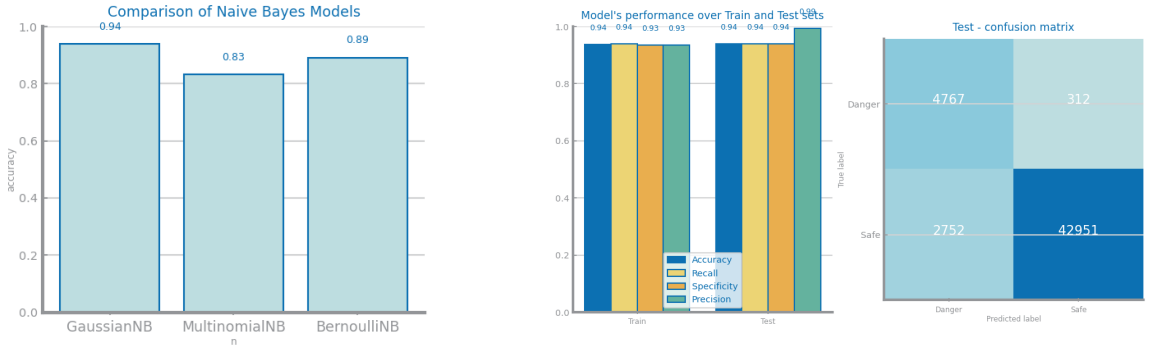

Figure 29 – AQ – accuracy for the various Naïve Bayes models and best model results

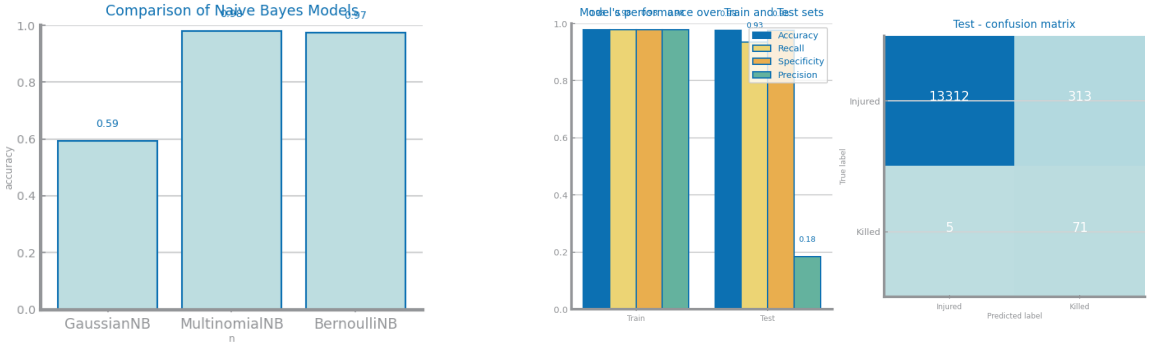

Figure 30 – NYC – accuracy for the various Naïve Bayes models and best model results

The impacts of feature engineering here are prominent, with substantial accuracy increases for both Feature Selection and PCA. However, we saw a drastic drop in recall after all these transformations, which could be attributed to the dimensionality reduction impacting the assumption of conditional independence. This may also be due to the imbalanced datasets.

The best model, as discussed above, was the GaussianNB for the AQ dataset, and the MultinomialNB for the NYC dataset, both without feature engineering. For NYC, the MultinomialNB had very poor precision (18%) on the test set, in spite of good results according to the other metrics (each of them above 93%). For AQ, the GaussianNB had a very good performance overall (every metric was above 94%). Both model's results are visible in ***Figures 29-30.***

3.2 KNN

For the study of KNN, we tried using the Manhattan distance, the Euclidean distance, and the Chebyshev distance, as well as their weighted counterparts, for a varying number of K.

Figure 31 shows us the results for both AQ and NYC. In AQ one can see a general decline of recall with increasing values of K, and with every chosen distance. In this case, the lowest number of neighbours ($K = 10$) is ideal, alongside the Weighted Manhattan distance. In the case of NYC, this behaviour is similar to AQ, but with the best K being either $K = 1$ or $K = 5$. The decrease in recall may be due to, again, the imbalance of the datasets, which leads to misclassifications, as explained above. The best scalings and models were the same as before.

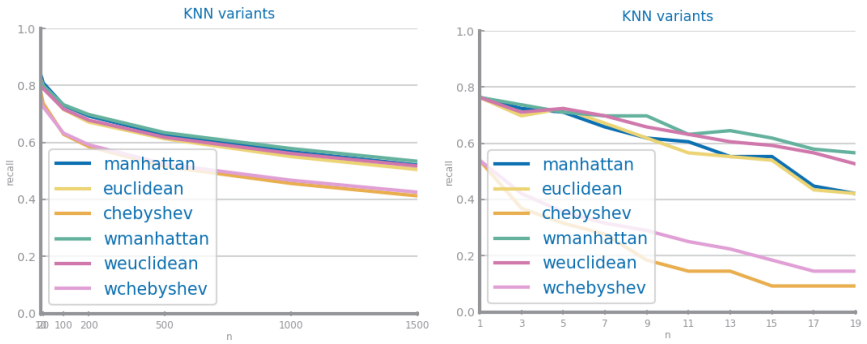

Figure 31: KNN Study: AQ Z-Score and NYC Min-Max, respectively.

We also studied the effects of PCA and Feature Selection on the classifier, but the results proved similar. This is expected given the nature of the transformations, as PCA should keep similar records close, and irrelevant features should provide only a slight change of distance between otherwise similar records.

We didn't find that the best models overfitted the training set with varying values of K, since their accuracy on the training set never increased while the one from the test set decreased.

NYC's best model had good results in terms of accuracy and specificity, but had poor recall. This may be mainly due to the unbalanced dataset, since it's much more likely for a datapoint to be near an "injured" datapoint than to a "killed" datapoint, as nothing points that "killed" data points should be clustered, and hence the majority of points will be classified as "injured". The precision may be lower than the accuracy and specificity due to the same reason.

AQ's best model had very good performance overall, with the worst measure being the recall, once again. The reason is probably the same as NYC's: There are many more "safe" datapoints than "danger" datapoints, hence it's much more likely that a node is classified as "safe".

3.3 Decision Trees

The decision trees were trained with various combinations of the parameters used for the criterion to measure the quality of the splits (entropy and Gini coefficient), the height of the tree, and minimum impurity decrease.

We had very stable (and odd) results for both datasets, in terms of accuracy. Each of them had a recall (and accuracy) score that neared perfect, varying very slightly mostly due to differences in height, but not much due to the other parameters. Our best decision tree models achieved perfect (or very near perfect) scores on all of the studied measures when used on the dev and test sets, on both datasets.

We didn't find that the model overfitted the training set with varying values of the tree height, since the model had near-perfect performance on both sets, with the performance on the test set never becoming worse than on the training set. We should note, however, that the only criterion of split quality and minimum impurity decrease used were the ones that yielded the best results on the training set.

The NYC decision tree **(Figure 48)** makes decisions regarding its target mainly based on the emotional status of the person after the collision: if the person is not conscious and is apparently dead, then it's likely that the person will die; if the person is conscious after the collision, there's a big chance that they won't die, in this case, we can consider Emotional Status to be a False Predictor, since we discovered later that it's highly correlated with the class. After that, it bases its answers on the complaints, ejection status, safety equipment, age and the role in the crash, which all seem like valid choices. The decision tree is not without its flaws, as one adicional branch also considered the crash date.

The AQ decision tree **(see Figure 49)** makes decisions regarding the safety of the air mainly through different intervals of two main measures: the mean of the PM10 measure and the mean of the PM2.5 measure. If the mean of the PM10 is below 0.317 and the mean of the PM2.5 is below 0.909, then it's very likely that the air is safe to breathe; if the mean of the PM10 is between 0.92 and 1.099, then the air is unsafe to breathe; if the mean of the PM10 is between 0.317 and 0.92, then if the mean of PM2.5 is lower than 0.972, the air is likely safe, and is otherwise likely unsafe.

3.4 Random Forests

To evaluate the Random Forests' classification, we varied the height, the number of estimators, and the percentage of features used. The minimum impurity decrease was kept at 0.0005 and the split quality criterion was always the entropy criterion.

As we can see from the results of ***Figure 32*** and ***33,*** the results for both datasets stabilised after the maximum height reached 5, with the results of varying the number of estimators and percentage of features being very stable, and with, once again, a nearly perfect recall performance for certain percentages of features.

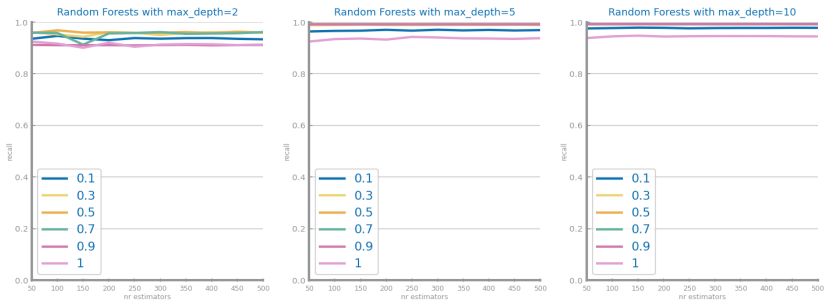

Figure 32: Random Forest Study - AQ - Recall for varying depth and number of estimators.

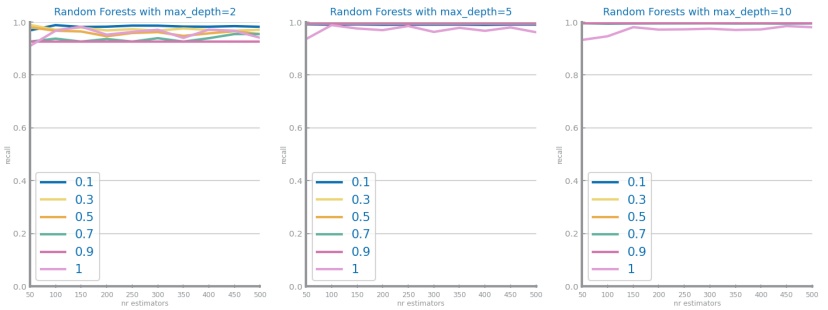

Figure 33: Random Forest Study - NYC - Recall for varying depth and number of estimators.

The results over PCA and Feature Selection, however, were less positive, as both saw a large drop across all metrics, in particular recall, which could originate from not having enough variables to conduct a varied ensemble, which is expected considering the number of variables available in these processed datasets.

We didn't find any overfitting using this classifier, with its accuracy and recall performance being nearly perfect on the training and test sets of both datasets. To study this, we only varied the number of estimators, and maintained all of the other parameters fixed at the best value we found in the previous stage.

The random forests achieved perfect (or very near perfect) scores on all of the studied measures when used on the train and test sets, on both AQ and NYC.

For the NYC dataset, the most important feature was a Conscious Emotional Status (45% of importance), next to Apparent Death (23%), Internal damage Complaints (10%), and Unconscious Emotional Status (10%). The remaining features had less than 3% of importance.

For the AQ dataset, the most important feature was the mean of the PM10 measure (58% of importance), next to the mean of the PM2.5 measure (29%), and the maximum value for the latter (8%). The remaining features had less than 4% of importance.

3.5 Gradient Boosting

To study the best parameters for classification in the case of gradient boosting, we varied their decision tree height, the number of estimators, and the learning rate. The other parameters were kept as default.

The accuracy was very stable and nearly perfect in both of the datasets, as can be seen in ***Figure 34*** and ***35.*** The only case where there was instability was when the tree height reached 5 and we varied the other parameters. This is very similar to the results of the random forests.

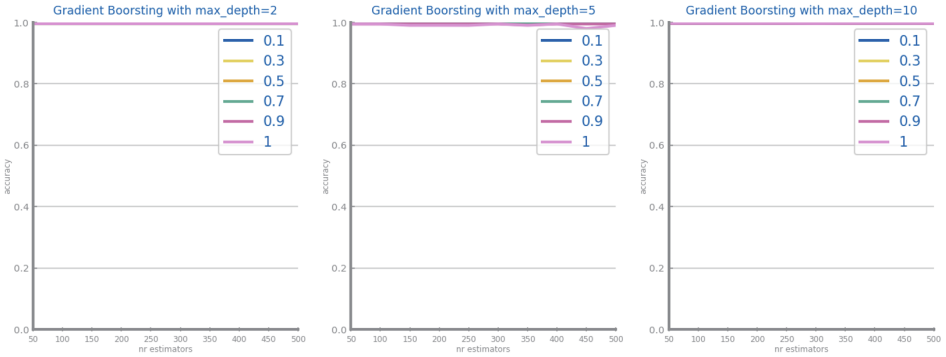

Figure 34: Gradient Boosting Study - AQ - Accuracy with varying depth and number of estimators.

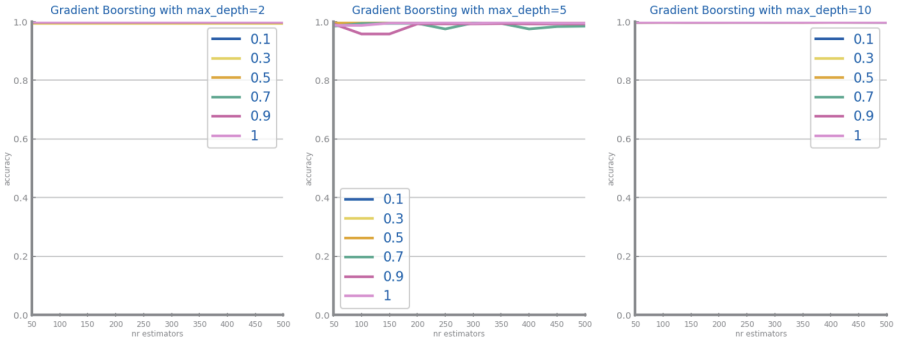

Figure 35: Gradient Boosting Study - NYC - Accuracy with varying depth and number of estimators.

We didn't find any kind of overfitting in either of the datasets. To study this, we used the parameters for the best model, changing only the number of estimators.

Regarding the NYC dataset, we can see that our best model, despite perfect accuracy, had poorer recall and precision both in the dev and test set. As we can see from the confusion matrix, this was due to killed people classified as injured, in the case of recall, and injured people classified as killed, in the case of precision. On the AQ dataset, all of the measures were nearly perfect both in the train and the test set. With both FS and PCA, we saw a general decrease of recall, which might be due to the same factors that conditioned our Random Forest models.

For the NYC dataset, the most important variable was Apparent Death (47% of importance), next to No Visible Complaints (17%) and an Unconscious Emotional Status (14%). The remaining features had less than 5% of importance.

For the AQ dataset, the most important variable was the mean of the PM2.5 measure (80% of importance), next to the mean of the PM10 measure (16%). The remaining features had less than 0.6% of importance.

3.6 Multi-Layer Perceptrons

Our MLP experiments were conducted by trying various learning rates, varying the maximum number of possible iterations, and by using different learning rate types (between constant, adaptive and invscaling).

The results for the AQ dataset show really solid results across all types of learning rates and learning rate types, with only a small drop in specificity for the best model, whose results are shown in ***Figure 36***. These results are compatible with our predictions, as the dataset features only numerical values, making the network fairly simple to learn, as irrelevant variables should be given lower weights during training.

The NYC results tell a different story, since even though the accuracy remains high for all parameter combinations tested, the best model results, present on ***Figure 37***, show only modest values for recall and precision, which are substantially more valuable metrics considering the unbalanced dataset. The results may come as a result of the dummification done, as many variables have to be learned, and most of them present little to no information.

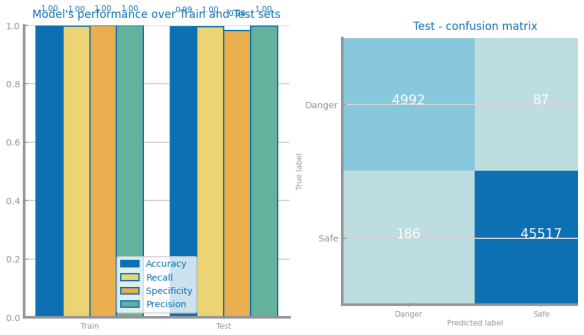

Figure 36: MLP Study - AQ - Best model.

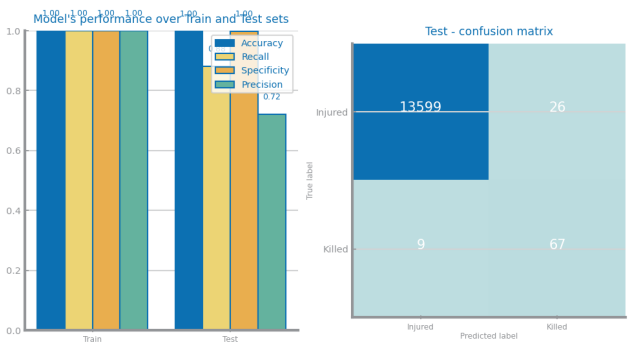

Figure 37: MLP Study - NYC - Best model.

In both models, overfitting was studied by picking the adaptive learning rate type, fixing the learning rate at 0.9, and changing only the number of epochs the model trained for. In this scenario, overfitting appears not to be a concern, as the performance remained almost perfect for both the training and the test set, in regards to accuracy and recall, and when it didn't, the training and set set performances proved similar. The resulting loss curves reinforce this, with all our configurations having a non-increasing loss in respect to the iteration.

Using PCA and FS to train our model presents considerably worse results than when using the scaled and balanced datasets: while FS didn't affect the AQ's results in a particularly negative manner, there's a drop in both accuracy and recall. This drop is particularly more noticeable in NYC, having achieved 0 specificity, which could be the result of the number of variables being insufficient to confidently decide the output. PCA provides a similarly bad result, only this time in both datasets. This may be due to the considerably lower number of dimensions for the input, which would only be beneficial if the resulting data had well-defined borders. This, however, appears not to be the case, as yet again our specificity drops to zero, suggesting that PCA's components are not enough to clearly define the data for use in

MLP. We can, however, discard the possibility of overfitting, as the low specificity remained true during the model's training.

Our best model was achieved in the datasets without feature engineering, with Inverse Scaling learning rate type, 0.3 learning rate and 750 maximum iterations for AQ, and Constant learning rate type, 0.1 learning rate, and 1000 maximum iterations for NYC.

4 Clustering

In order to apply clustering no balancing techniques were applied to the datasets, as it would generate (or erase) records. The scaled datasets were instead used, and feature selection was applied. Clustering was also applied to the datasets resulting from PCA.

Various clustering algorithms were used: KMeans, Expectation Maximisation, Hierarchical, and DBSCAN with varying EPS, minimum sample, and distance metrics. All of these algorithms were evaluated using MAE, MSE, Silhouette and Davies-Bouldin Index metrics. It is worth noting that, of all these metrics, the Silhouette score is most relevant when comparing different algorithms, as it is bounded and objective. We present the different metrics for the best clustering algorithm for each dataset:

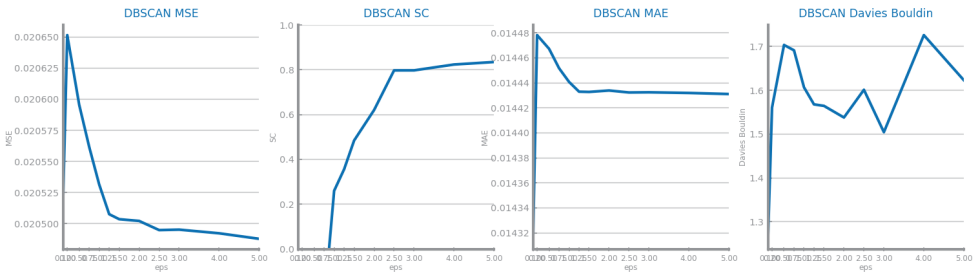

Figure 38 – AQ - Metrics using DBSCAN with euclidean distance, minimum samples = 2 and varying EPS.

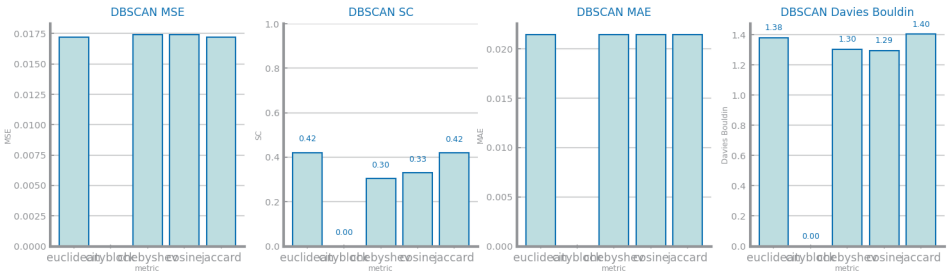

Figure 39 – NYC - Metrics using DBSCAN, with EPS = 0.6 * average distance, and varying distance metrics.

For AQ, the Silhouette is maximised at DBSCAN, using EPS=5, with a score of 0.82. This states that a considerably good clustering was found. We see that increasing EPS produces better results in both MSE and Silhouette, which implies the points in each cluster are being found further apart, which would explain the jumps in Davies-Bouldin, as an increasing EPS increases the distance of what is considered a point in a neighbourhood (increasing the distance between points in the cluster) but also provides more separated clusters, leading to opposite forces in the Davies-Bouldin score calculation. It is possible that DBSCAN produced the best results due to the data points producing non-linearly separable shapes with clear defined (by distance) borders. The resulting clustering scatter plots with DBSCAN can be seen in ***Figure 44*** in the appendix.

As for the NYC dataset, the best Silhouette was achieved using DBSCAN with both the euclidean and the Jaccard metric, where we see a maximum score of 0.42, which implies weak clustering. Jaccard might've had good results because the existing clusters are relatively small - if the clusters were very large, the denominator during the expansion phase would be enormous, which would make the score become very tiny. The Euclidean distance might've performed well because it takes into account the various dimensions of each register while also normalising them, contrary to cosine (only takes angle into account, not magnitude), Manhattan (does not normalise the distance), and Chebyshev (only takes into account *one* dimension). In this case, DBSCAN also provided the best results, for reasons one can assume are similar to the ones for the AQ dataset. The resulting clustering scatter plots can be seen in ***Figure 45*** in the appendix.

The results of the different metrics for the best clustering on the datasets with PCA applied are presented in ***Figures 40 and 41***, for AQ and NYC respectively.

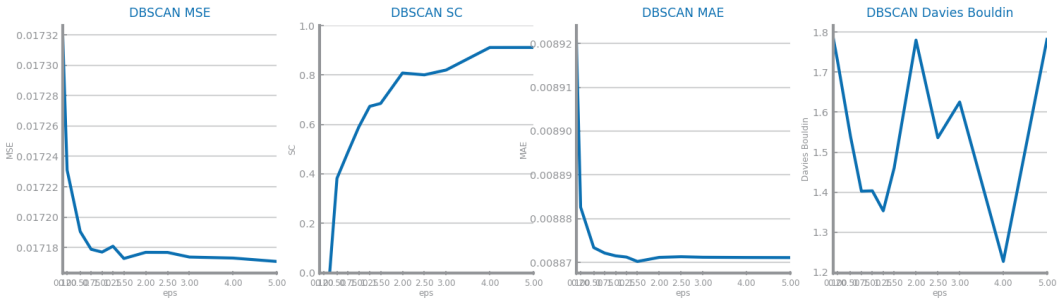

Figure 40 – AQ w/PCA - Metrics using DBSCAN with euclidean distance and varying EPS.

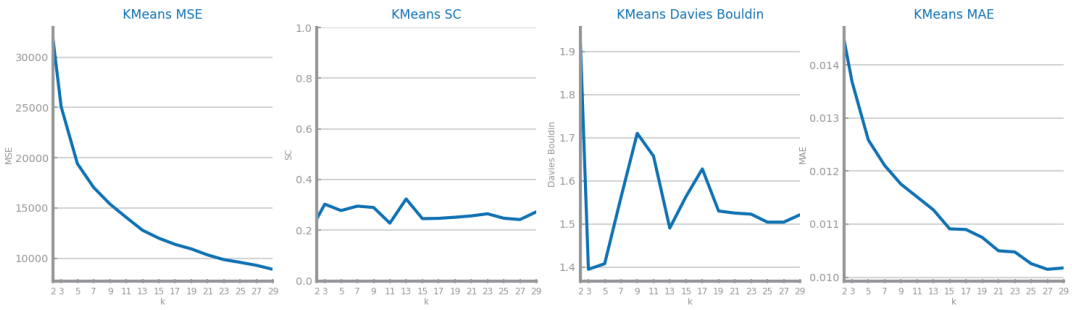

Figure 41 – NYC w/PCA - Metrics using KMeans, with number of clusters.

Regarding AQ we see some improvements after PCA, with a maximum Silhouette of around 0.9 being achieved by the DBSCAN with again an EPS=5, and a similarly good score with K=2 for Hierarchical, both considerably great clusterings. Varying the minimum sample size also resulted in equally good results. The same analysis of the non-PCA dataset DBSCAN results applies here again, only this time we have a considerably reduced number of variables. The improvements could originate from the reduced number of variables having a greater variance between themselves, and similar variables ending up closer together. The resulting clustering plot can be seen in ***Figure 46***.

As for NYC with PCA, we see a decrease of the overall Silhouette score, which might indicate that the number of components picked for PCA was too low, leading to a low variance between variables, and therefore very close variables, leading to unclear borders between clusters. Another possible explanation is that PCA actually has too many components, which would lead to a scenario that would approximate that of NYC before PCA, yet this can probably be discarded as the produced results are actually worse than those without PCA, where we see a maximum score of around 0.31 being achieved by both K-Means with K=13, and Hierarchical with K=15. We also see a

considerable drop in both MSE and MAE throughout the K increase, which is expectable as more clusters should always decrease the error, since their centres will always be overall closer to the data points the more clusters there are, so this metric can be discarded in favour of Silhouette. The resulting clustering plot can be seen in ***Figure 47***.

5 Association Rules

We can only apply Pattern Mining to symbolic data, so we discretized the numeric variables from both datasets. For this, we generated three new datasets from the previously scaled and feature-selected datasets: a dataset with no numeric variables (Only applicable to NYC), a discretized dataset with 5 bins of the same interval width (equal width), and a discretized dataset with 100 bins containing roughly the same number of elements but not necessarily bins of the same interval width (equal frequency). We then had to dummify all these new datasets for the apriori algorithm.

For the AQ dataset with equal width about 1020 patterns and 57,000 association rules are found in at least 20% (support) of the records, and when the support drops to 8.5%, the patterns jump to about 1520 and the rules to over 75000. Both average confidence and lift drop when either support or confidence decrease. The top rules we found are the proof of what we predicted previously: the levels of CO, PM2.5, and PM10 are highly correlated (be it the min, max, or mean variables) and predictive of the target variable. We present only the charts for 'equal width' in ***Figure 42*** for simplicity's sake.

For the AQ dataset with equal frequency, the number of patterns were significantly fewer than with 'equal width,' but when support dropped to 4%, we found as many patterns as before, which is expected - there are always many patterns when support is low. Lift differed substantially: while in 'equal width' lift stayed around the 1 mark for roughly all values of confidence, in 'equal frequency' the lift skyrocketed (top 10 rules with lift > 200 and top 25% rules with lift approaching 50) when confidence dropped below 1. The different discretization explains this phenomenon: in 'equal width' each variable is divided in, surprisingly, equal-size bins, so values are always within the bin limits. But in 'equal frequency,' the values are not bounded that rigorously, so values inside a bin can be very low or very high; hence the high discrepant lift values. Finally, we again arrived at the same rules and conclusions about CO, PM2.5, and PM10.

Figure 42: Pattern Mining - AQ (equal width): Confidence and Lift Charts

For the NYC dataset with no numeric variables, about 7000 patterns and 200,000 association rules are found with support lower than 2% and confidence below 50%. Those numbers plummet to 1400 and 55,000, respectively, when we increase the support to 10% and confidence to 90%. The average lift remains stable throughout support and confidence fluctuations at a little under 2. The top rules reflect common sense: bicyclists are most at risk when crossing the street and at intersections (likewise, drivers, and passengers obviously, crash at intersections, too); people complaining of pain or nausea are conscious. We present the confidence and lift charts in ***Figure 43***.

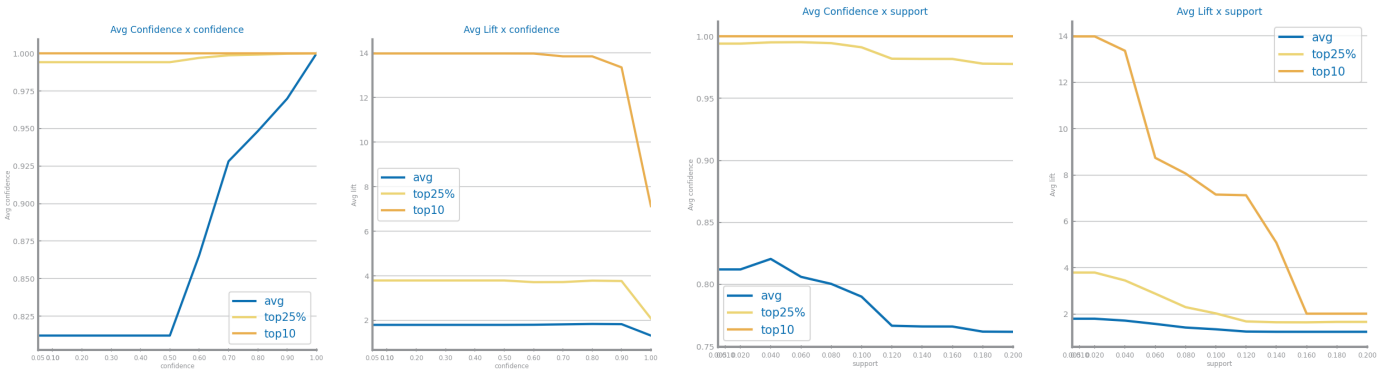

Figure 43: Pattern Mining - NYC: Confidence and Lift Charts

6 Critical Analysis

All around, our models presented positive results for both datasets, achieving high accuracy and recall, even in the presence of a highly unbalanced target class, though questionable decisions regarding the data preparation were made, in particular the dummification in NYC, which could have led to a loss of performance, as variables that could have been relevant were turned into various binary variables, losing its descriptive power. We also see that even if most models presented better results in accuracy in both the Feature Selection and PCA datasets, it came at the cost of a drop in either recall or specificity, which is undesirable considering the unbalance present in our target variables.

We also observe the emergence of some patterns regarding the most relevant variables considered during the Decision Trees, Random Forests, Gradient Boosting and Feature Selection: PM2.5 and PM10 means for AQ; EMOTIONAL_STATUS and COMPLAINT for NYC. The conclusions we take from these patterns are: PM2.5 and PM10 are highly correlated to, highly predictive of, and responsible for the air quality (high quantities of these gases mean decreased air quality), seemingly superseding any other gas, as predicted during Data Profiling; in NYC, a person can only register a complaint if she is conscious, otherwise she is either dead or unconscious, which is correlated with the target variable, yet again as predicted by Data Profiling.

Finally, both the Multi-layer Perceptron and the Gradient Boosting models presented the best results for the AQ dataset; decision Trees were the best model for NYC. Both of these models had exceedingly good performance across all metrics. Their real life usage recommendation is, however, dependent on the impact of the task at hand, as it would be unwise to use such models in life-threatening tasks such as critical air quality classification, as these models were conceived using a linear selection process to reduce a combinatorial explosion of possible models. As a result, with a more thorough search, as well as various KDD iterations, better models and better data preparation transformations are almost guaranteed to be found. The testing was also done only taking into consideration four measures, which serve just a general indication of a model's quality. Yet, as simple classifiers, these models prove more than capable of executing the task at hand.

7 APPENDIX

Data Profiling

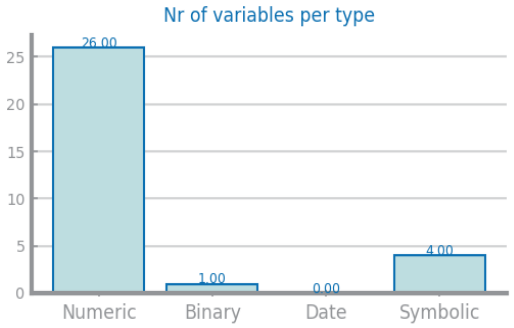

***Figure 1**: Nº of variables per type - AQ*

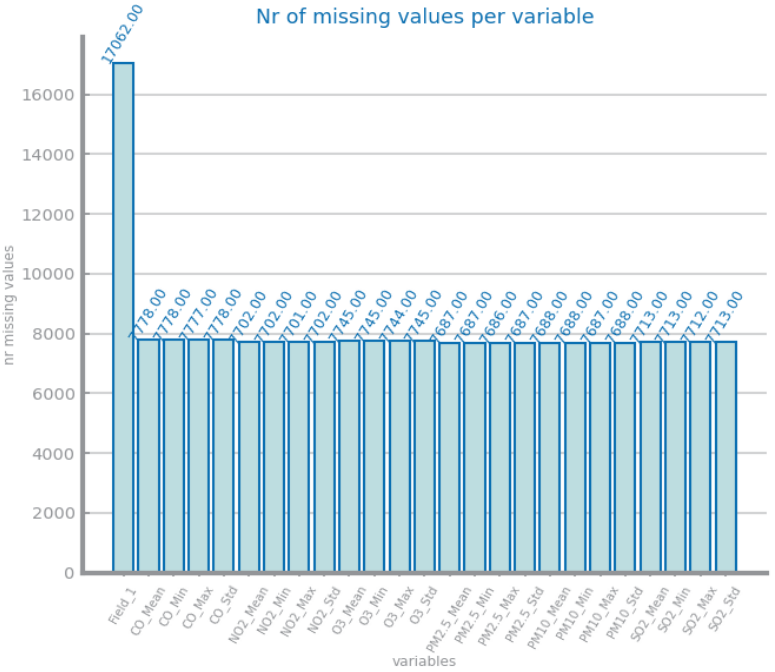

***Figure 2**: Nº of missing values per variables - AQ*

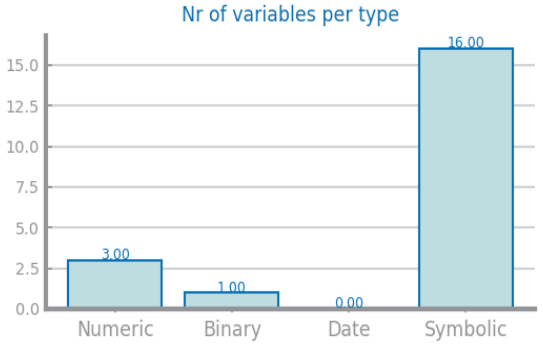

***Figure 3**: Nº of variables per type - NYC*

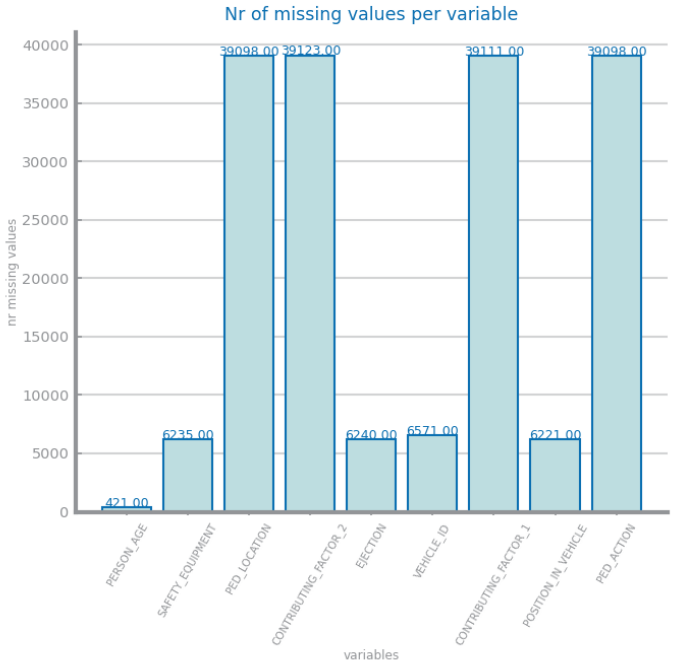

***Figure 4**: Nº of missing values per variables - NYC*

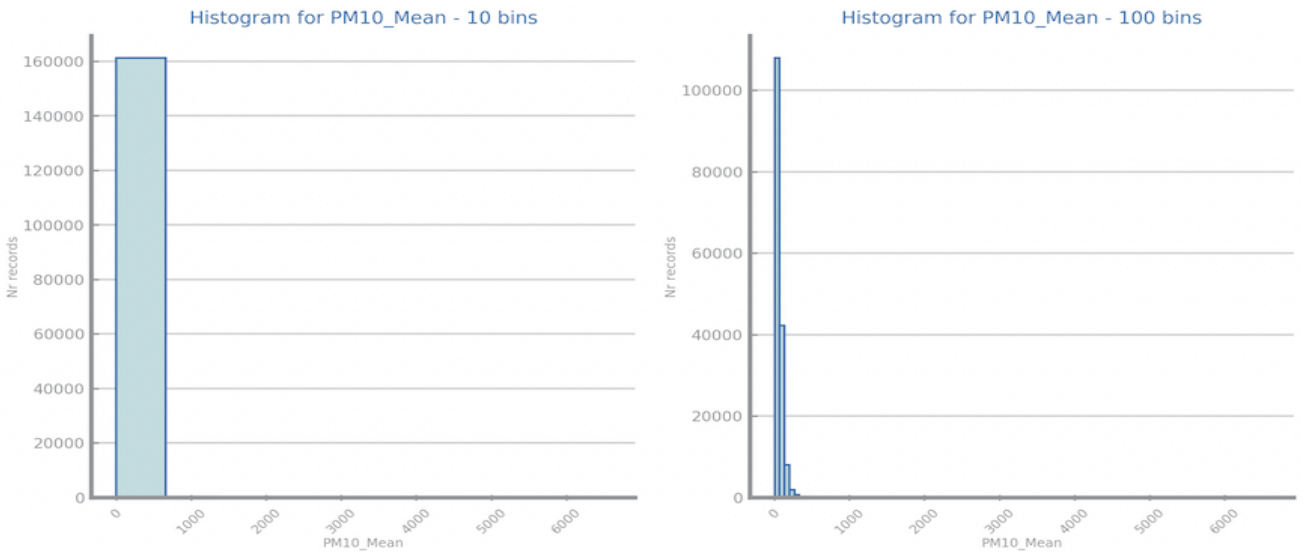

***Figure 5**: Nr of records per PM10 mean value - AQ*

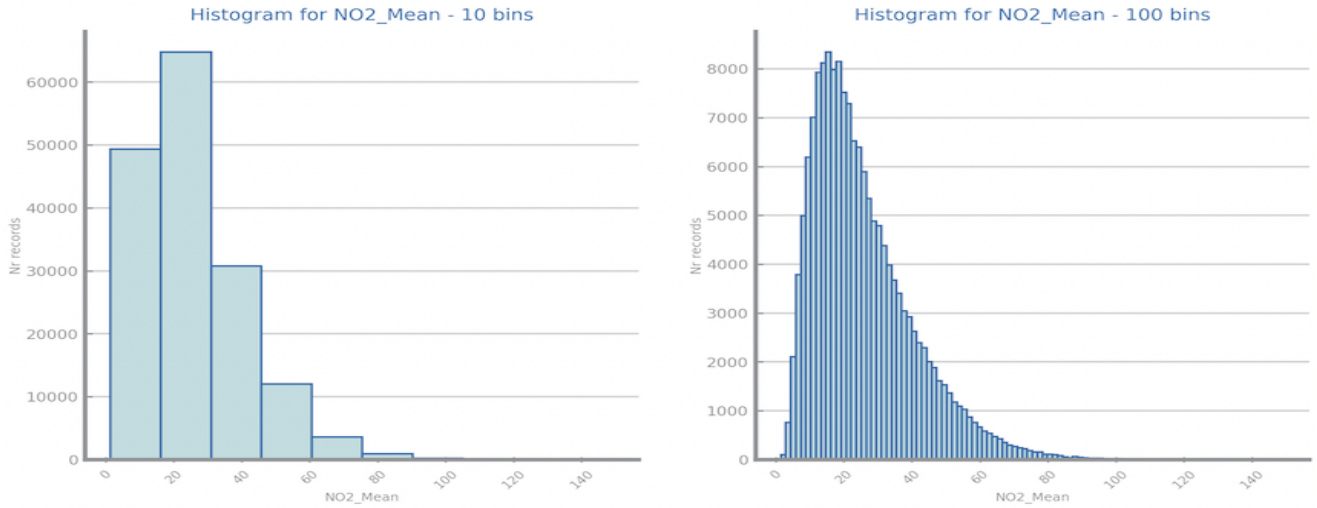

***Figure 6**: Nr of records per NO2 mean value - AQ*

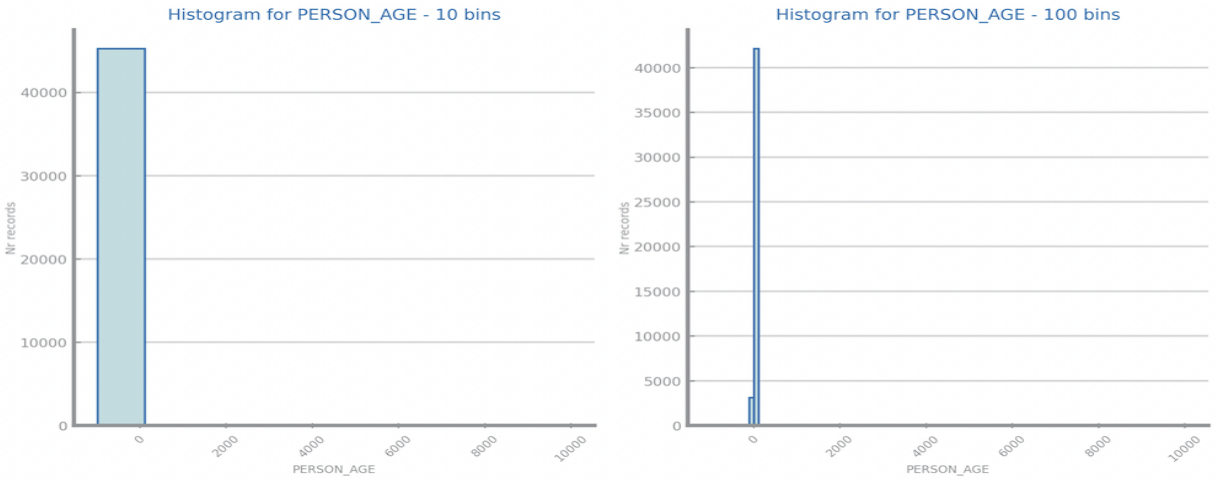

***Figure 7**: Nr of records per person age value - NYC*

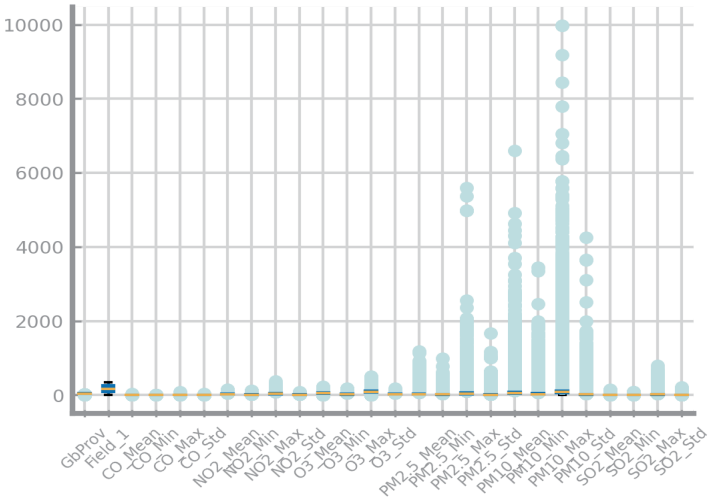

***Figure 8**: Boxplots for all the variables - AQ*

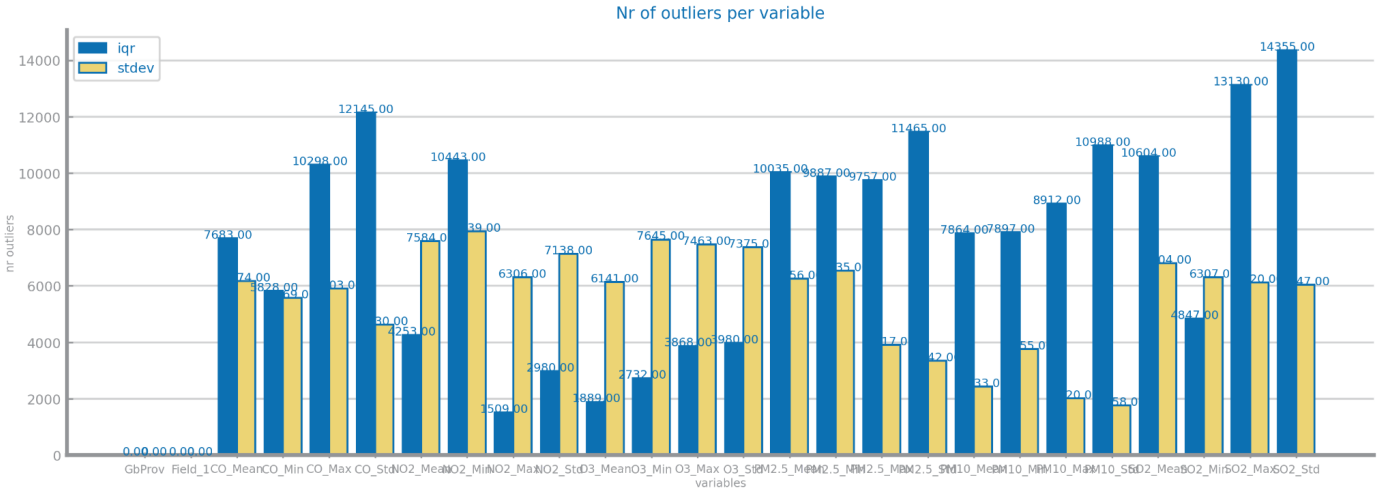

***Figure 9**: Nr of outliers per variable - AQ*

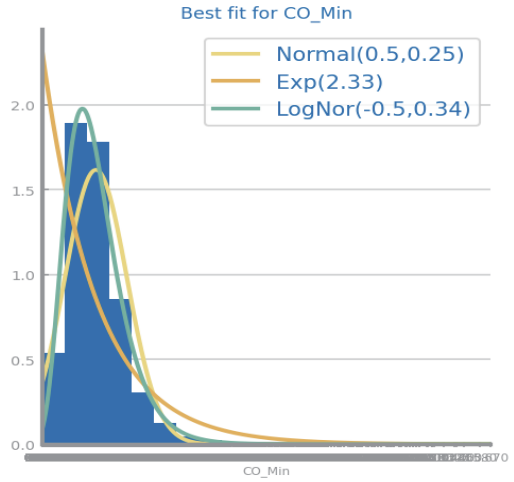

***Figure 10**: Trend histogram for CO_Min - AQ*

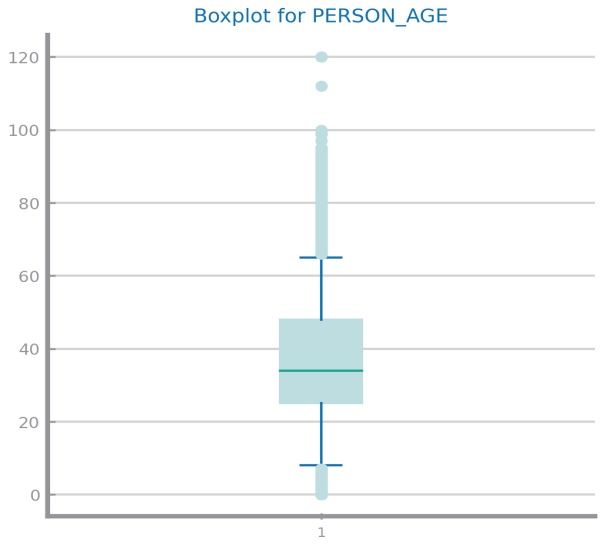

***Figure 11**: Boxplot for PERSON_AGE - NYC*

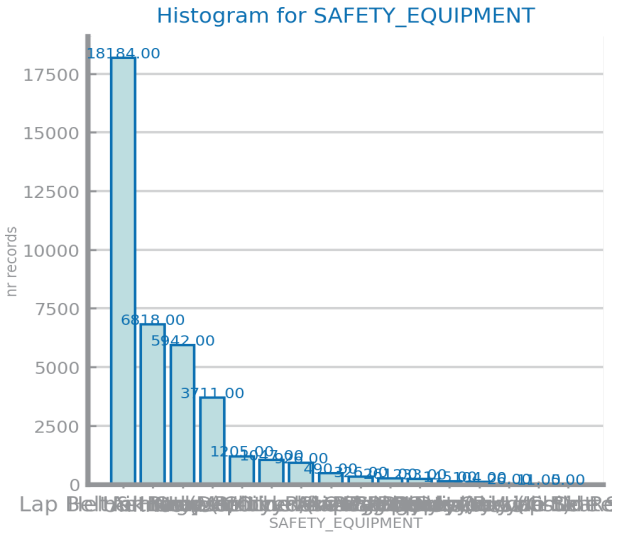

***Figure 12**: Histogram for SAFETY_EQUIPMENT - NYC*

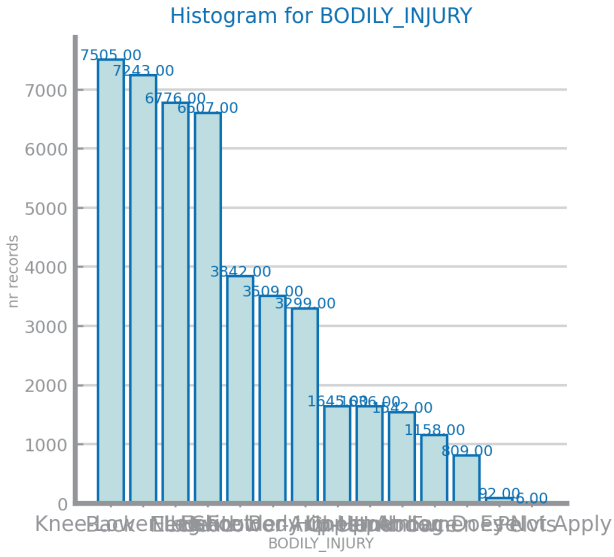

***Figure 13**: Histogram for BODILY_INJURY - NYC*

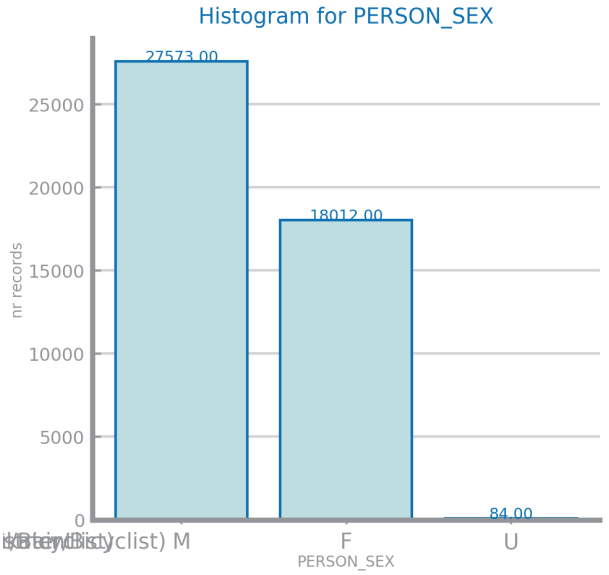

***Figure 14**: Histogram for PERSON_SEX - NYC*

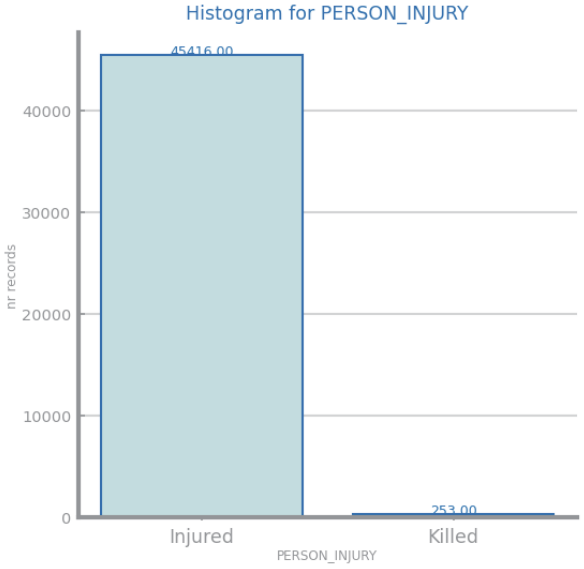

***Figure 15**: Histogram for PERSON_INJURY - NYC*

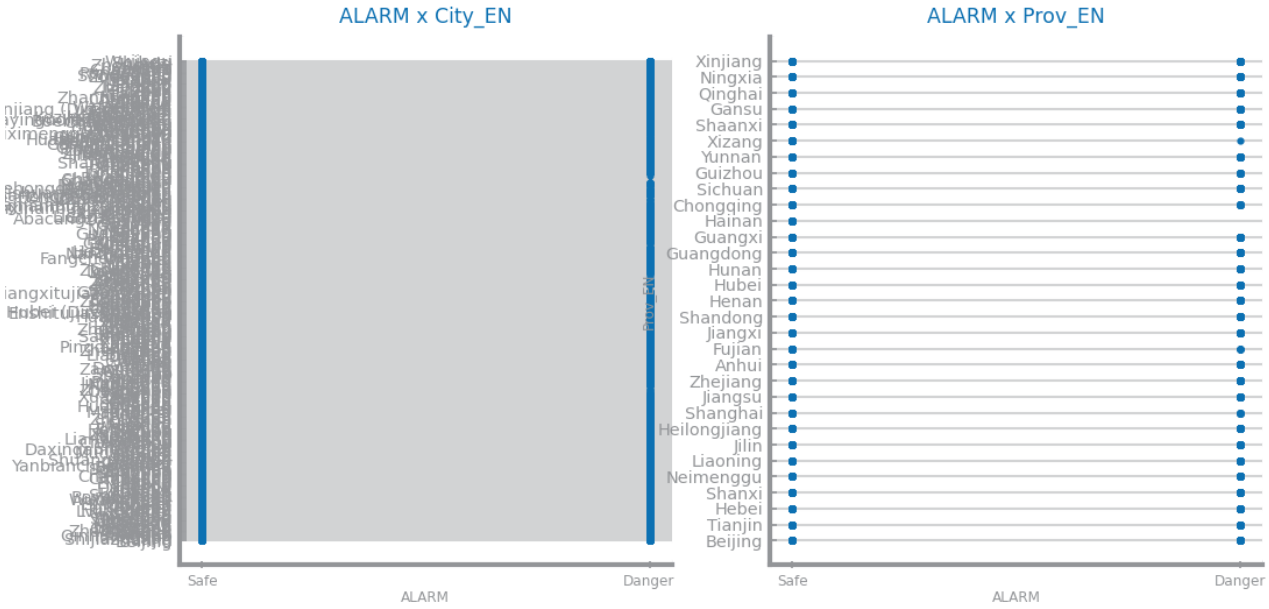

***Figure 16**: Sparsity for cities - AQ*

***Figure 17**: Sparsity for provinces - AQ*

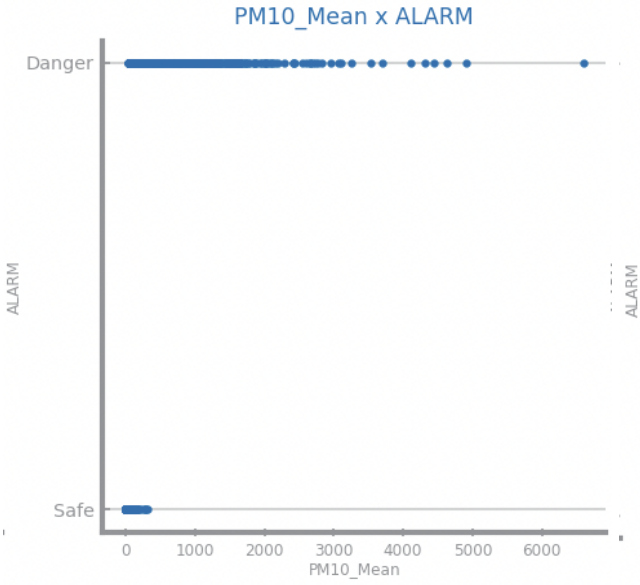

***Figure 18**: Scatter plot for PM10_Mean (w/ target) - AQ*

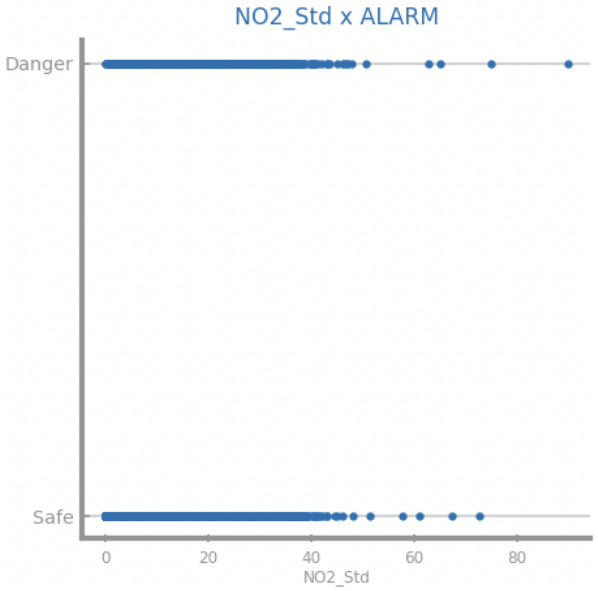

***Figure 19:** Scatter plot for NO2_Std (w/ target) - AQ*

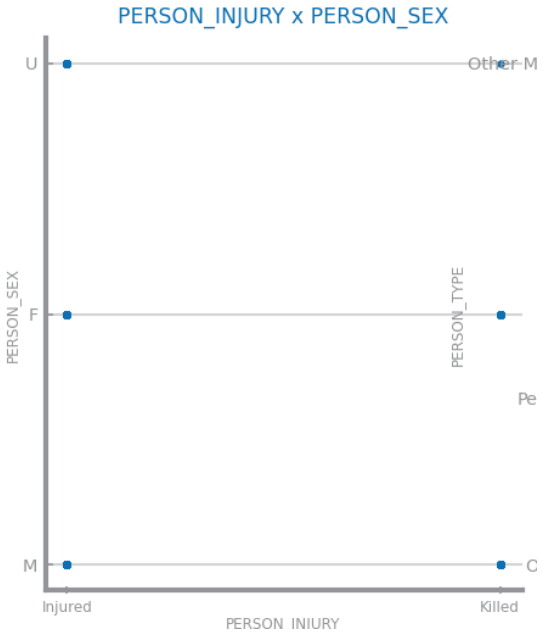

***Figure 20**: Scatter plot for PERSON_SEX (w/ target) - NYC*

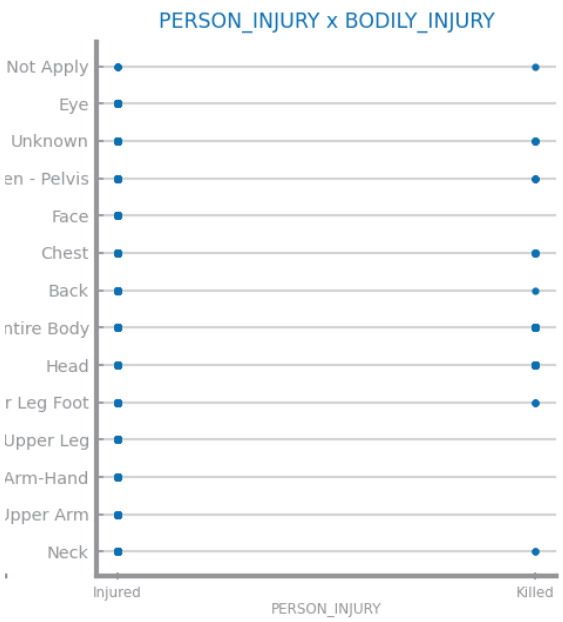

***Figure 21:** Scatter plot for BODILY_INJURY (w/ target) - NYC*

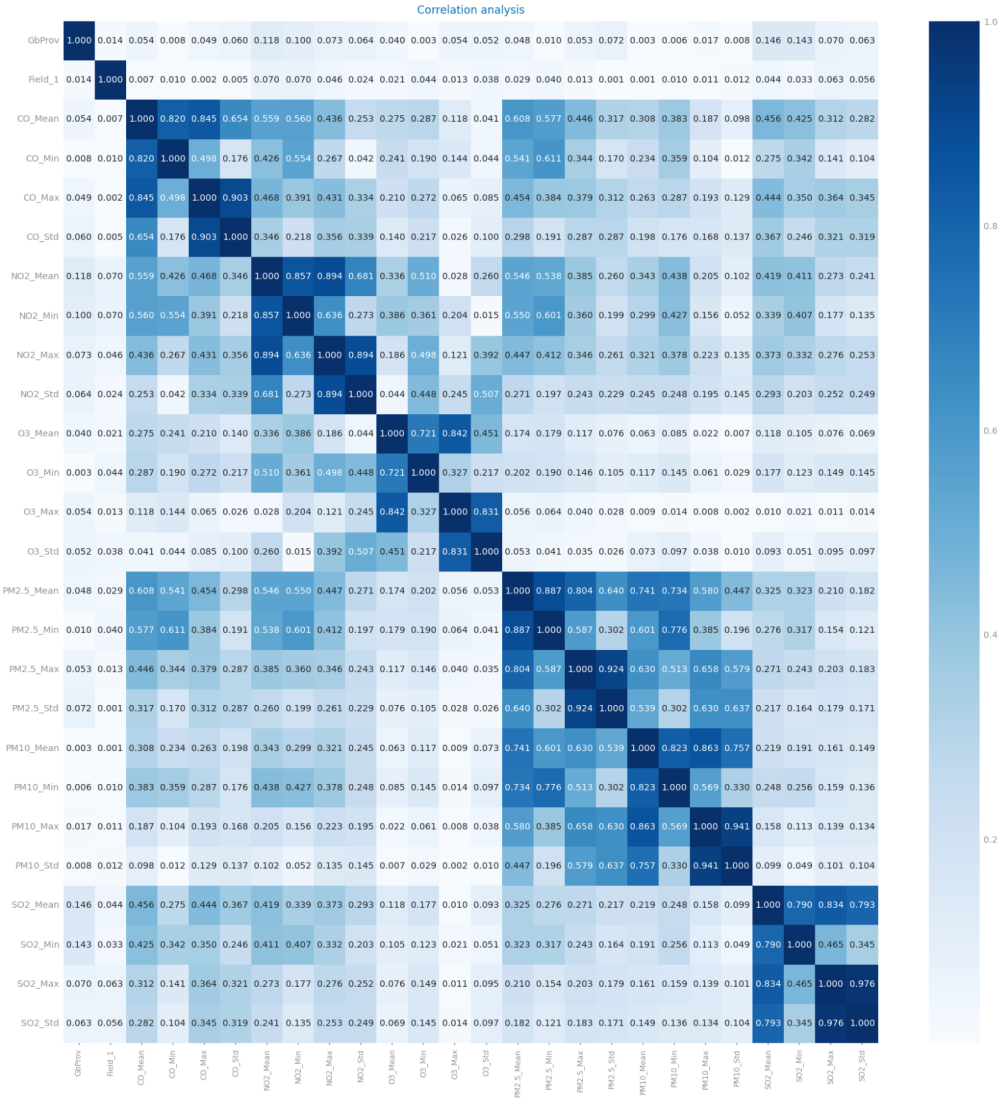

***Figure 22**: Correlation Matrix - AQ*

Clustering

***Figure 44**: AQ - DBSCAN with varying EPS - Scatter Plots*

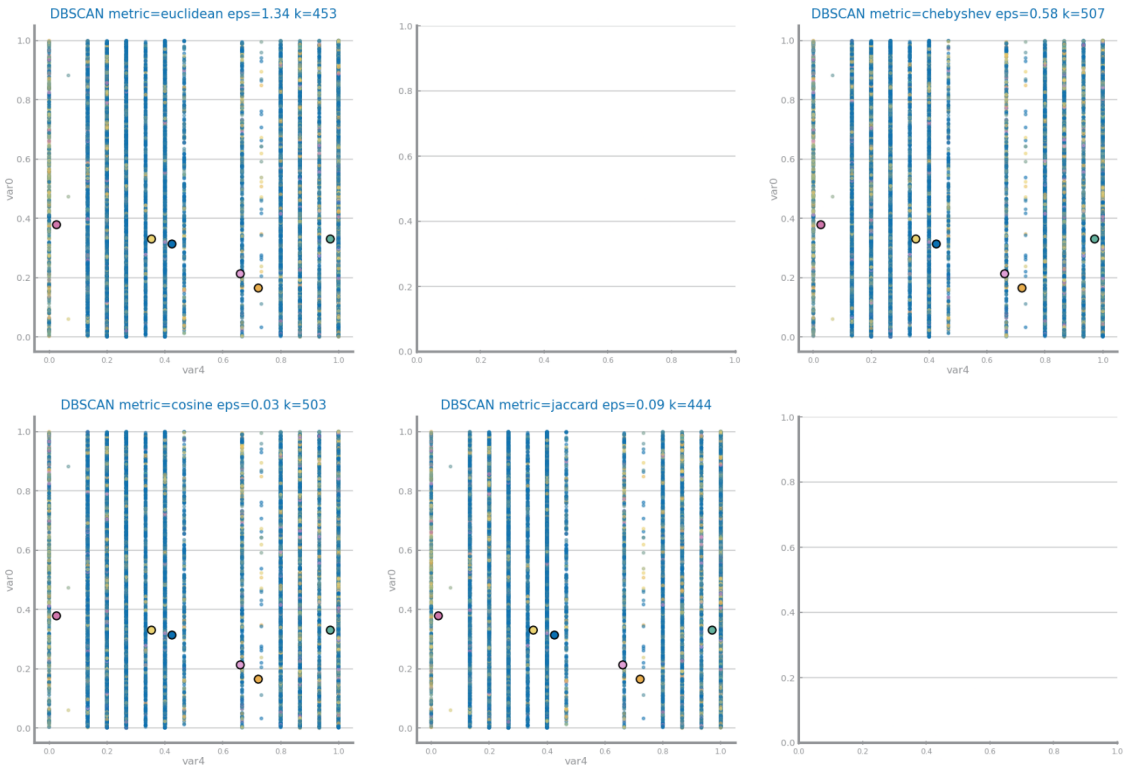

***Figure 45**: NYC - DBSCAN with varying metrics - Scatter Plots*

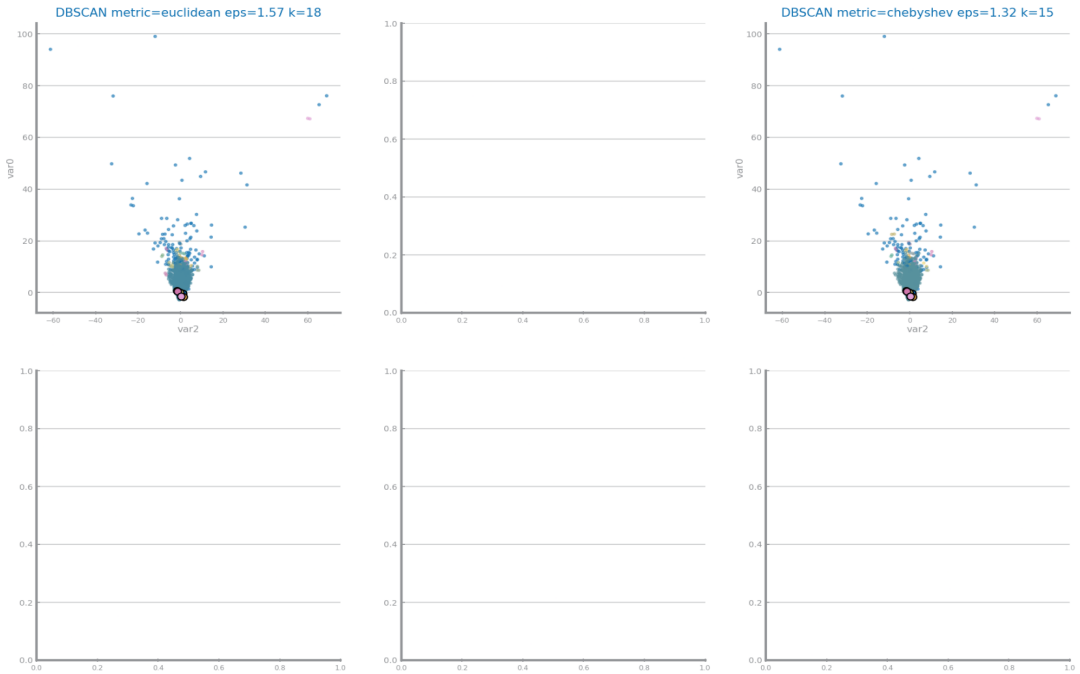

***Figure 46**: AQ w/PCA - DBSCAN with varying metrics- Scatter Plots*

***Figure 47**: NYC w/PCA - KMeans with varying K - Scatter Plots*

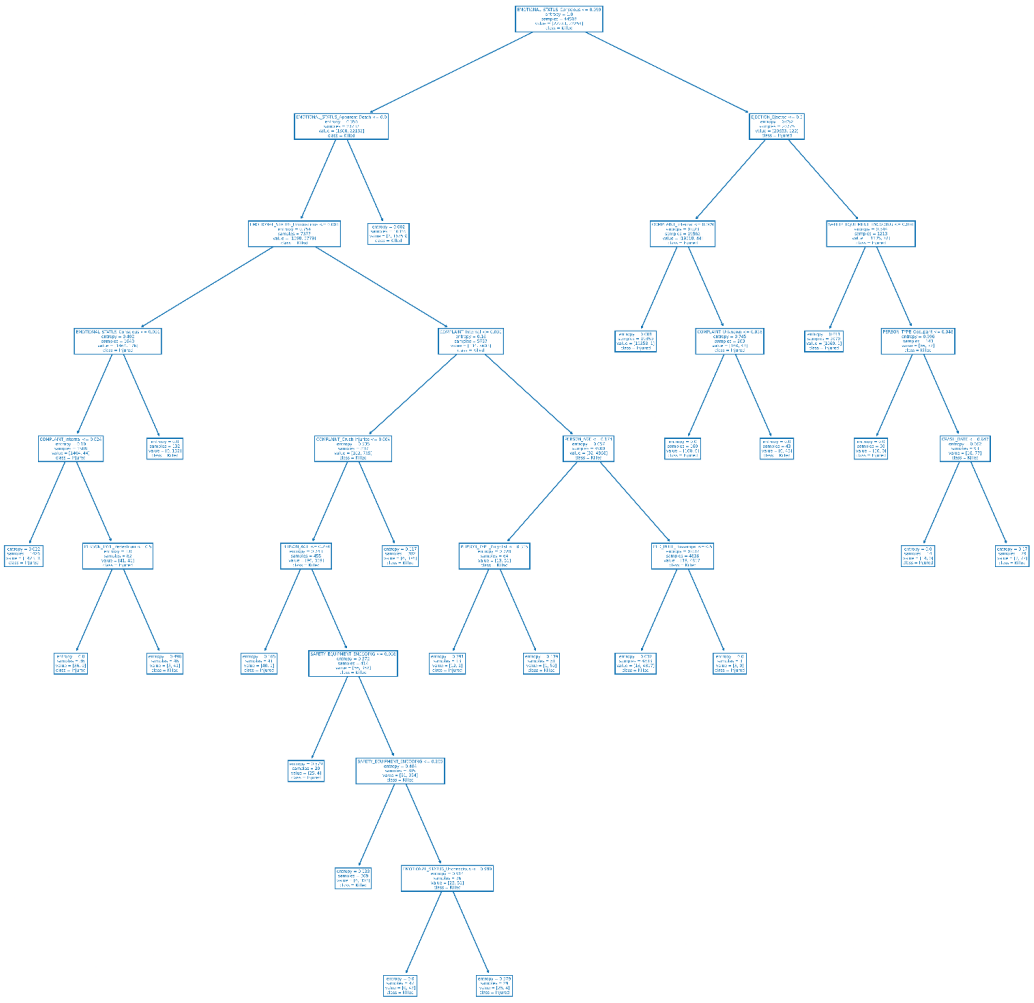

***Figure 48**: NYC – the best decision tree*

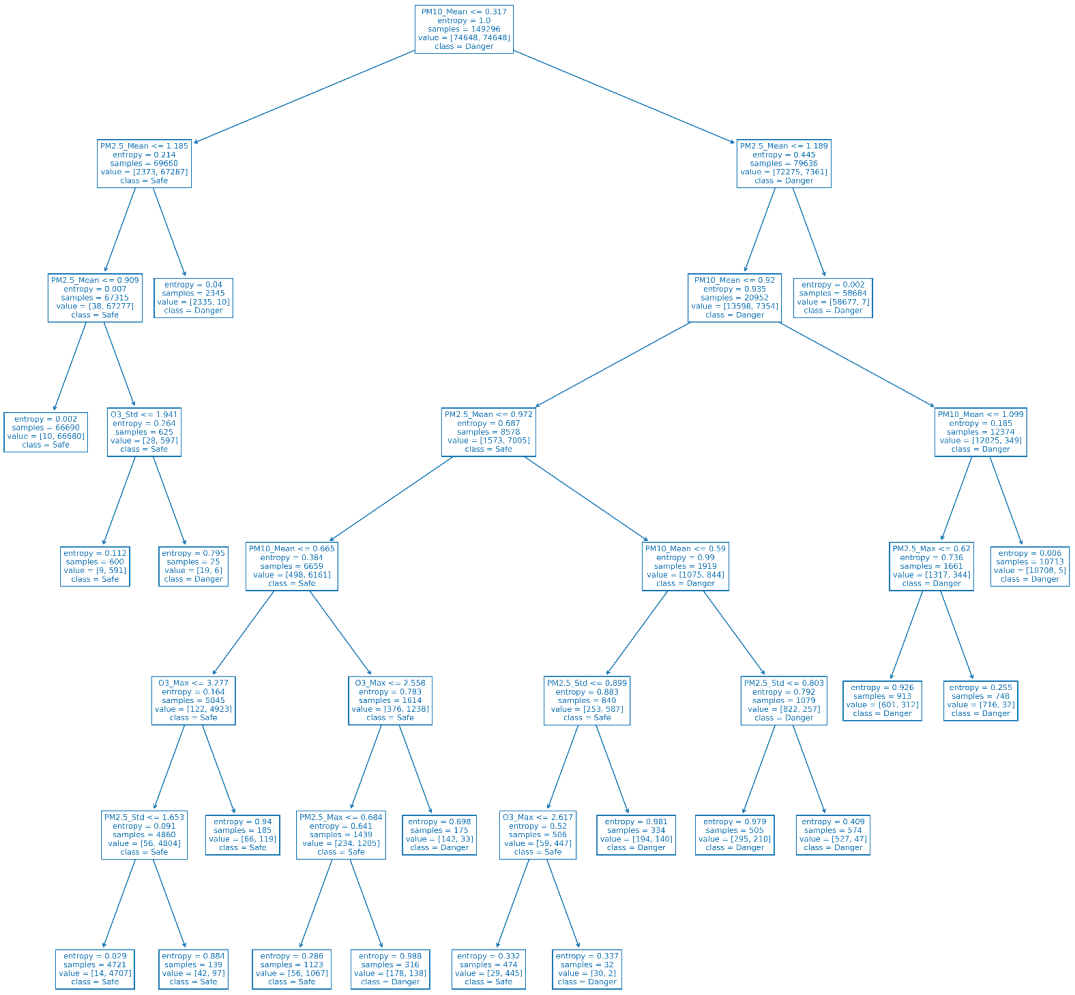

***Figure 49**: AQ – the best decision tree*