# Data classification for the analysis of car-caused pollution

Vasco Miguel Guerreiro Gomes
Computer Engineering Student
Instituto Politécnico de Beja
Beja, Portugal
19921@stu.ipbeja.pt

Ricardo Jorge Rodrigues Sequeira
Computer Engineering Student
Instituto Politécnico de Beja
Beja, Portugal
21905@stu.ipbeja.pt

Cleyde Swailla Figueiredo Varela
Computer Engineering Student
Instituto Politécnico de Beja
Beja, Portugal
21684@stu.ipbeja.pt

*Abstract* — **This electronic document is a technical article that aims to describe the process of data classification for the analysis of car-caused pollution based on a dataset.**

***Keywords - car; pollution; machine learning; algorithms.***

## I. Introduction

Car emissions are a major source of pollution. Cars and trucks emit twenty percent of all greenhouse gases emitted in the United States [1], which is the world's second-largest greenhouse gas emitter. In this article, we aim to classify and analyze data related to car pollution from a chosen dataset. To support the study this article is divided into the following sections:

i. Introduction

ii. Data set and objectives

iii. Methodology

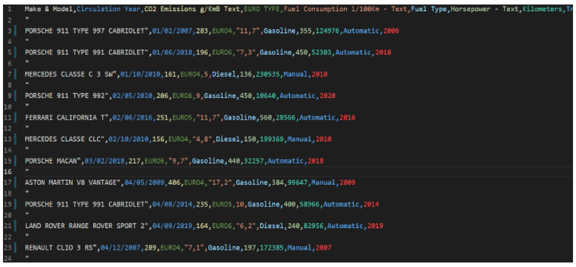
iv. Results

v. Conclusion

## II. Data set and objectives

### A. Data set

The data set used was taken from *Kaggle* [2], it is represented in tabular form where the rows are the records of events and the columns are the characteristics of these events. It has data related to cars, such as their model and type, fuel type, euro type, fuel consumption, horsepower, circulation year, CO2 emissions, kilometers, transmission, and year.



```
Make & Model,Circulation Year,CO2 Emissions g/KmB Text,EURO TYPE,Fuel Consumption 1/100Km - Text,Fuel Type,Horsepower - Text,Kilometers,Tr
"
PORSCHE 911 TYPE 997 CABRIOLET",01/02/2007,283,EURO4,"11,7",Gasoline,355,124976,Automatic,2006
"
PORSCHE 911 TYPE 991 CABRIOLET",01/06/2018,196,EURO6,"7,3",Gasoline,450,52303,Automatic,2018
"
MERCEDES CLASSE C 3 SW",01/10/2010,161,EURO4,5,Diesel,136,230535,Manual,2010
"
PORSCHE 911 TYPE 992",02/05/2020,206,EURO6,9,Gasoline,450,10640,Automatic,2020
"
FERRARI CALIFORNIA T",02/06/2016,251,EURO5,"11,7",Gasoline,560,28566,Automatic,2016
"
MERCEDES CLASSE CLC",02/10/2010,156,EURO4,"4,8",Diesel,150,199369,Manual,2010
"
PORSCHE MACAN",03/02/2018,217,EURO6,"9,7",Gasoline,440,32257,Automatic,2018
"
ASTON MARTIN V8 VANTAGE",04/05/2009,406,EURO4,"17,2",Gasoline,384,99647,Manual,2009
"
PORSCHE 911 TYPE 991 CABRIOLET",04/08/2014,235,EURO5,10,Gasoline,400,58966,Automatic,2014
"
LAND ROVER RANGE ROVER SPORT 2",04/09/2019,164,EURO6,"6,2",Diesel,240,82956,Automatic,2019
"
RENAULT CLIO 3 RS",04/12/2007,209,EURO4,"7,1",Gasoline,197,172385,Manual,2007
"
```

*Figure 1 – Dataset open in Visual Studio Code*

The ETL process (Extract Transform Load) was performed using *Microsoft PowerBI.*

### B. Objectives

The process described in this article is one of the required components for the conclusion of the Information Systems curricular unit in our bachelor, where we are prompted to apply some of the concepts of Data Mining and algorithms of Machine Learning to the chosen dataset, using a given platform, *Orange*.

## III. Methodology

Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values [3].

There are several algorithms that support machine learning, upon analyzing our targets the following appeared to be the most suitable ones: Decision Tree, Random Forest, and KNN (K Nearest Neighbor). Note that all the mentioned algorithms were applied using *Orange*.

### A. Pre-processing

In the pre-processing stage, the targets were set. For a better understanding and to ensure a great analysis of the data two targets were chosen, EURO_TYPE (European exhaust emission standards, which are the legal requirements related to air pollution into the atmosphere, note that the higher the Euro type is the less the car pollutes) and FUEL_TYPE (The standards vehicle's fuel types).

In the same way that 2 Select Columns widgets were used in our Orange model below, 2 Data Sampler widgets were also used, because it was necessary to subdivide the data that was going to be used to carry out the training of the model from the data that was going to be effectively classified, test data. To carry out this subdivision, we used a fixed portion of data and divided the training data and test data into percentages of 70% for training and 30% for testing. Next, we loaded Sample data so that it would perform the division of the dataset.

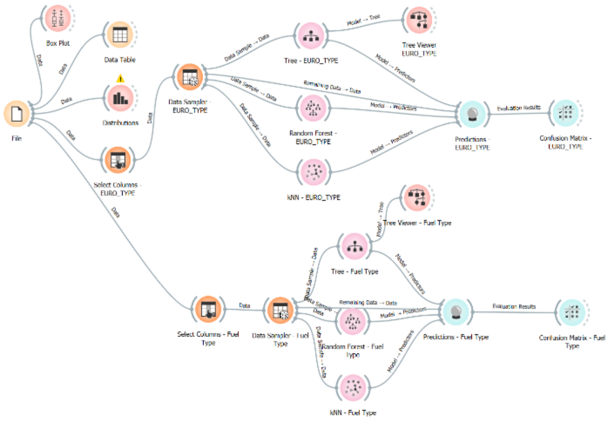


*Figure 2 – Model in Orange*

## B. Algorithms

Machine learning algorithms are ordered sequences of rules, commands, and instructions, that help to explore, analyze and find meaning in datasets complexes. Their goal is to establish or discover patterns that people can use make predictions or categorize information.

Since two targets were used in our model and the data was divided, the chosen algorithms had to be applied twice, for each subgroup of data.

- The *Decision Tree* is a supervised learning technique that according to the target defined in the Select Columns widget, in this case, is the EURO_TYPE for the first and FUEL_TYPE for the second, represents the data in a leaf node, and the decision rules also known as the features of our dataset are the branches [4].
- The *Random Forest* is also a supervised learning technique that can be used to make a regression approach or a classification approach which is what we are doing by classifying the cars in our dataset [5].
- The kNN or K Nearest Neighbor is one of the simplest algorithms based on supervised learning, this algorithm stores the data and makes a classification based on similarity. This algorithm can be used in Regression or in Classification as we are implementing. The way that it is implemented is based on the similarity of the features provided it makes a classification [6].

The application of the above algorithms can be visualized below:

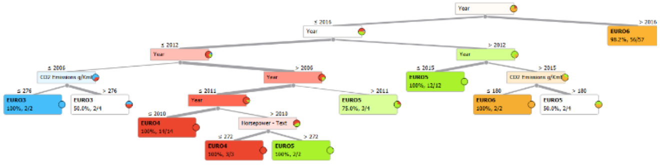


*Figure 3 – Decision tree with EURO_TYPE target*

This Decision Tree uses the Select Columns widget and selects the target as well as the features necessary to make the classification such as Year, CO2 Emissions and ultimately the Horsepower, removing the unnecessary features like Kilometers and Transmissions Type that make no sense here since the number of Kilometers and the type of Transmission does not determine a car's EURO, only determines if an automatic car or a manual car has less or more kilometers, that being that it is more durable. The feature that has more relevance here is the year because according to that we can see with type of EURO Norm a specific vehicle is equipped and using other features like the Horsepower and CO2 Emissions to make the final decisions that the feature year cannot determine.

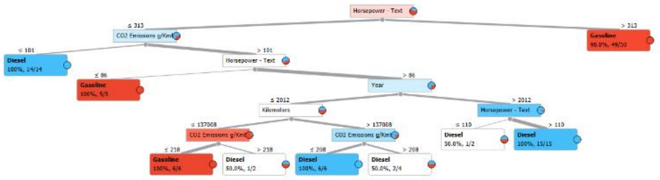


*Figure 4 – Decision tree with FUEL_TYPE target*

This Decision Tree uses the Select Columns widget and selects the target as the Fuel Type and the features necessary to make the classification such as Year, CO2 Emissions, Horsepower and Kilometers, removing the unnecessary features like Euro Type and Transmissions Type that do not change the outcome of the Decision Tree. The most important feature in this decision tree is the CO2 Emissions, due to the fact that if a car has a determined amount of horsepower and yet it is not conclusive its Fuel Type, then we can look at the CO2 Emissions because the Diesel cars emit fewer amounts than the Gasoline counterpart since the Diesel cars have a higher engine capacity and most of the times the help of a turbo, meanwhile, the Gasoline cars with less Horsepower emit more CO2 because they have to Rev the engine to higher RPM's to be able to catch the Diesel car. The car's height also plays a big role in the emissions.

Another interesting feature is also used to determine the Fuel Type of a car, the Kilometers, due to the fact that usually, Diesel cars have a higher millage since Diesel was democratized as being a green fuel in the late '90s and early '00s, however, it is not to say the Gasoline cars cannot accumulate many miles or kilometers like 600.000 Miles or 1.000.000 Kilometers, they can but they were sold mainly with small engines to put in fuel-efficient city cars or in the other end of the spectrum as high performance cars since Diesel in a performance car like an Audi R8 with a V12 Diesel engine, or a VW Touareg equipped with a 2.5 TDI outputting 275 Horsepower was not yet heard of.

## C. Validations

The Predictions widget shows all of the 3 algorithms we talked about earlier in the article, the Decision Tree, the Random Forest, and the kNN, as well as the error of each individual classification with the target EURO TYPE and FUEL TYPE and all of the features with the metrics of:

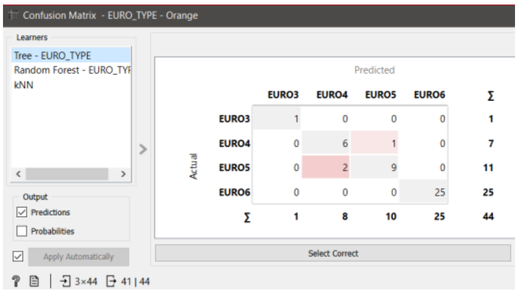
- *AUC* (Area under the ROC Curve) varies in between the values 0 to 1, this means if the previsions are 100% correct the AUC value is 1.0, and on the opposite, if they are totally incorrect the AUC value is 0.0 [7].
- *CA* stands for Classification Accuracy, and it is one metric for evaluating classification models, so we could say that having a 91% of score would be great, but it may not be the case, we have to check for True Negatives and False Positives, or the Precision and Recall values [8].
- *Precision* stands for the proportion of the positive identifications that are actually correct [9].
- *Recall* stands for True and False tax; basically what proportion of positives was identified correctly [9].
- *F1* stands for an evaluation metric that measures a model's accuracy, it uses precision and recall to be able to do so [10].

## IV. Results

For the first target used, Euro Type, from all of the parameters and all of the models, we can see that the kNN did not have the best classification due to having the lowest classifications of them all as we can see its Precision and Recall were very low compared with the other 2 algorithms impacting also the F1 value as the overall accuracy of the model and AUC. The Decision Tree appeared to be better than the Random Forest, and that is due to the F1, Precision and Recall values being higher than those in the Random Forest, so as we can see the Precision in the Decision Tree was higher than the Random Forest being 0.935 to 0.875 as well as the Recall value from 0.932 to 0.886 consequently, the F1 being higher also because the F1 is calculated with the Precisions and Recall values.

The Confusion Matrix widget shows all of the models as well as the Precisions and Recalls, as we can see, with the Decision Tree model there was 1 EURO3 predicted as well as 1 correct. In the EURO4 there were 7 vehicles and the Decision Tree Classified 8, 2 of them being Classified as EURO5 having a misclassification. In the EURO5 there were 11 vehicles equipped with that EURO_TYPE, and the model predicted 10 of them were in fact EURO5 and one of them was predicted as being a EURO4 vehicle. To Finish off the EURO6 had 25 vehicles equipped with this technology and there were predicted 25.

The Random Forest model had a very similar performance, although it made a bad classification in the EURO3, this can be due to the Random Forest containing multiple trees and the classification being different.



Confusion Matrix - EURO_TYPE - Orange

Learners
- Tree - EURO_TYPE
- Random Forest - EURO_TYP
- kNN

| Actual \ Predicted | EURO3 | EURO4 | EURO5 | EURO6 | Σ |
|---|---|---|---|---|---|
| EURO3 | 1 | 0 | 0 | 0 | 1 |
| EURO4 | 0 | 6 | 1 | 0 | 7 |
| EURO5 | 0 | 2 | 9 | 0 | 11 |
| EURO6 | 0 | 0 | 0 | 25 | 25 |
| Σ | 1 | 8 | 10 | 25 | 44 |

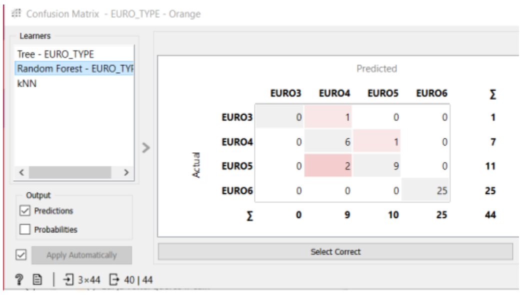
Output
☑ Predictions
☐ Probabilities

☑ Apply Automatically

Select Correct

3×44 41 | 44

*Figure 5 – Confusion matrix for the decision tree in Euro Type*



Confusion Matrix - EURO_TYPE - Orange

Learners
- Tree - EURO_TYPE
- Random Forest - EURO_TYP
- kNN

| Actual \ Predicted | EURO3 | EURO4 | EURO5 | EURO6 | Σ |
|---|---|---|---|---|---|
| EURO3 | 0 | 1 | 0 | 0 | 1 |
| EURO4 | 0 | 6 | 1 | 0 | 7 |
| EURO5 | 0 | 2 | 9 | 0 | 11 |
| EURO6 | 0 | 0 | 0 | 25 | 25 |
| Σ | 0 | 9 | 10 | 25 | 44 |

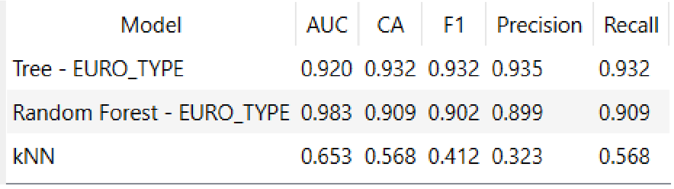
Output
☑ Predictions
☐ Probabilities

☑ Apply Automatically

Select Correct

3×44 40 | 44

*Figure 6 - Confusion matrix for the random forest in Euro Type*



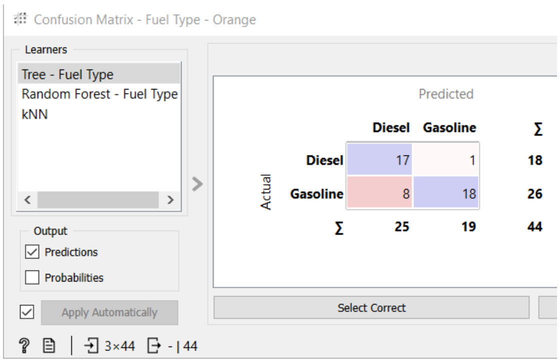
| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Tree - EURO_TYPE | 0.920 | 0.932 | 0.932 | 0.935 | 0.932 |
| Random Forest - EURO_TYPE | 0.983 | 0.909 | 0.902 | 0.899 | 0.909 |
| kNN | 0.653 | 0.568 | 0.412 | 0.323 | 0.568 |

*Figure 7 – Metrics for Euro Type Target*

For the second target, Fuel Type, we can see in the Predictions widget all of the models and their classifications, as expected the kNN was the worst algorithm, we can see it looking for the classifications.

Surprisingly the Random Forest algorithm performed better than the Decision Tree but using this new target and the Random Forest being made of multiple Decision Trees got an advantage, as we can see for the F1, Precision, and Recall values. The Precision (proportion of positive identifications) is greater in the Random Forest by almost 9%, the same can be said to the Recall value where the proportion of the actual positives were identified correctly, and one more time the Random Forest algorithm wipes the floor with the Decision Tree, meaning that had almost a 9% difference between the 2. With that being said since the F1 classification uses Precision and Recall, and it was higher in the Random Forest the F1 is also higher, meaning that the Random Forest is more accurate than the Decision Tree. We can also see that the Classification Accuracy of the Random Forest Algorithm is higher than in the Decision Tree, as well as the AUC values.

The Confusion Matrix widget, as we can see, with the Decision Tree model and the Random Forest where we can observe that there were classified more than 7 cars as having Diesel Engines to the actual Diesel Engine cars that exist in the dataset and in the Random Forest only 3, that being the False Positives or the Recall values. As we can see that the Random Forest has classified 23 cars as having Gasoline out of the 26 due to the false positives in the diesel fuel type, and the story is the same in the Decision Tree Model, we can observe 18 cars were classified as having a Gasoline engine out of the 26 that really have it. We can observe that in both models with was only one vehicle that was classified as having a Diesel Engine instead of a Gasoline one.



Confusion Matrix - Fuel Type - Orange

Learners
- Tree - Fuel Type
- Random Forest - Fuel Type
- kNN

Output
☑ Predictions
☐ Probabilities

☑ Apply Automatically

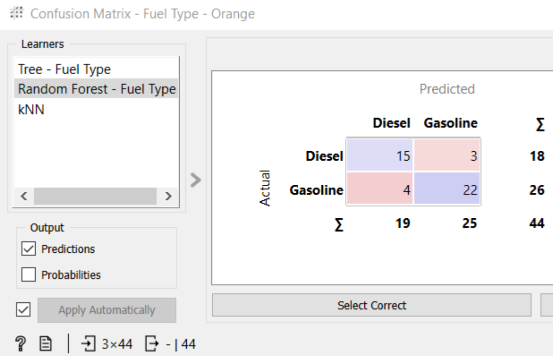
Predicted

| Actual | Diesel | Gasoline | ∑ |
|---|---|---|---|
| Diesel | 17 | 1 | 18 |
| Gasoline | 8 | 18 | 26 |
| ∑ | 25 | 19 | 44 |

Select Correct

3×44 | - | 44

*Figure 8 - Confusion matrix for the decision tree in Fuel Type*



Confusion Matrix - Fuel Type - Orange

Learners
- Tree - Fuel Type
- Random Forest - Fuel Type
- kNN

Output
☑ Predictions
☐ Probabilities

☑ Apply Automatically

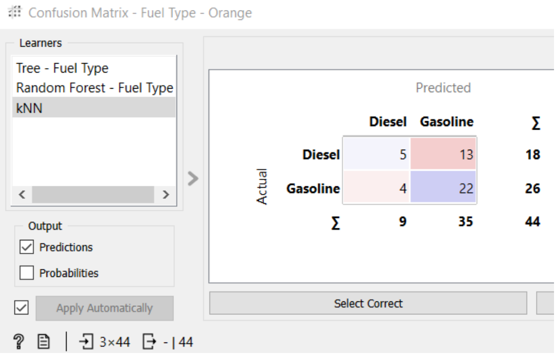
Predicted

| Actual | Diesel | Gasoline | ∑ |
|---|---|---|---|
| Diesel | 15 | 3 | 18 |
| Gasoline | 4 | 22 | 26 |
| ∑ | 19 | 25 | 44 |

Select Correct

3×44 | - | 44

*Figure 9 - Confusion matrix for the random forest in Fuel Type*



Confusion Matrix - Fuel Type - Orange

Learners
- Tree - Fuel Type
- Random Forest - Fuel Type
- kNN

Output
☑ Predictions
☐ Probabilities

☑ Apply Automatically

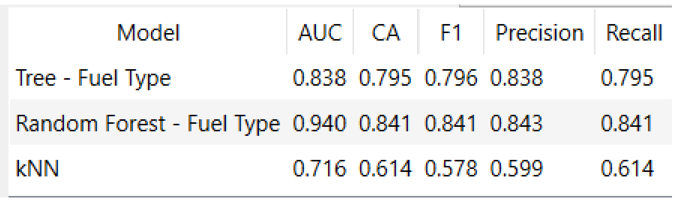
Predicted

| Actual | Diesel | Gasoline | ∑ |
|---|---|---|---|
| Diesel | 5 | 13 | 18 |
| Gasoline | 4 | 22 | 26 |
| ∑ | 9 | 35 | 44 |

Select Correct

3×44 | - | 44

*Figure 10 - Confusion matrix for the kNN in Fuel Type*



| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Tree - Fuel Type | 0.838 | 0.795 | 0.796 | 0.838 | 0.795 |
| Random Forest - Fuel Type | 0.940 | 0.841 | 0.841 | 0.843 | 0.841 |
| kNN | 0.716 | 0.614 | 0.578 | 0.599 | 0.614 |

*Figure 11 – Metrics for Fuel Type Target*

Note that we did not encounter any outliers during the analysis process.

## V. Conclusions

The main focus of this work was to learn and understand the machine learning algorithms by applying them to a dataset and using the Orange platform, a goal that we believe was successfully reached.

The conclusion we reached at the end of this work was that depending on the targets and features being used, the different algorithms can prove to be better or worse, as seen in the first classification with Euro Type target in which the decision tree algorithm had an accuracy higher than random forest and kNN, but the opposite was shown when the target changed to Fuel Type and the features changed, thus becoming the random forest to have a higher precision.

Additionally, from our knowledge this is the first classification (supervised analysis) work done in the chosen dataset. Related works in the same dataset include *Data mining on second hand French cars* [11] and *XG.Boost car price prediction* [12].

## References

[1] Lewis&Clark Law school, "Environmental, Natural Resources, & Energy Law Blog," 20 December 2020. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[2] GONZA, "French second hand cars," [Online]. Available: https://www.kaggle.com/datasets/spicemix/french-second-hand-car.

[3] E. Burns, "Machine Learning," [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[4] JavaTPoint, "Decision Tree Classification Algorithm," [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[5] JavaTPoint, "Random Forest Algorithm," [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm.

[6] JavaTPoint, "K-Nearest Neighbor(KNN) Algorithm for Machine Learning," [Online]. Available: javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[7] GoogleDevelopers, "Classification: ROC Curve and AUC," [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[8] GoogleDevelopers, "Classification: Accuracy," [Online]. Available: https://developers.google.com/machine-

learning/crash-course/classification/accuracy.

[9] GoogleDevelopers, "Classification: Precision and Recall," [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall.

[10] v7Labs, "F1 Score in Machine Learning: Intro & Calculation," [Online]. Available: https://www.v7labs.com/blog/f1-score-guide.

[11] GONZA, "Data mining on second hand cars data," [Online]. Available: https://www.kaggle.com/code/spicemix/data-mining-on-second-hand-cars-data.

[12] L3LLFF, "XGBoost. Car Price Prediction," [Online]. Available: https://www.kaggle.com/code/l3llff/xgboost-car-price-prediction.