PRACTICAL TEST - DATA SCIENTIST

PROBLEM 1

A) Present you exploratory data analysis.

Performing exploratory data analysis, we identified relevant insights from the dataset examined.

Initially, upon observing outliers and concentration levels, it was discovered that certain sensors, such as PT08.S2, PT08.S3, and PT08.S1, do not have defined concentration limits. However, after excluding missing data, the mean concentration of carbon monoxide (CO) adjusted to 2.182 mg/m³, a value within acceptable limits and aligned with the guidelines of the World Health Organization, which stipulates a maximum limit of 4 mg/m³ for CO levels.

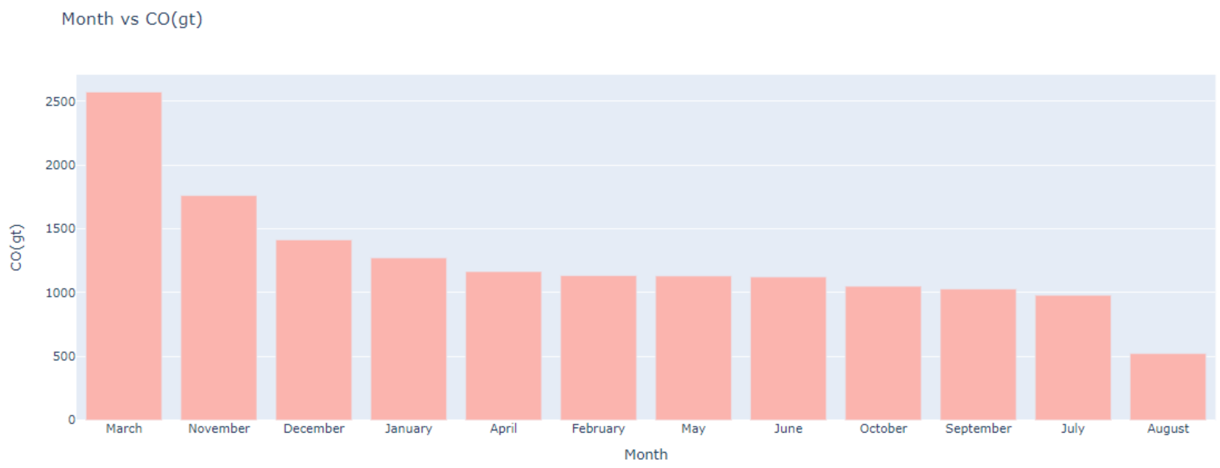


Figura 1 – Temporal Analisys

During the temporal analysis (Figure 1), a significant variation in CO levels throughout the year was noted, with March showing the highest indices and August the lowest, reflecting a pronounced difference attributable to seasonal changes. Analysis by days of the week revealed that CO levels tend to be higher during weekdays, with a notable reduction during weekends.

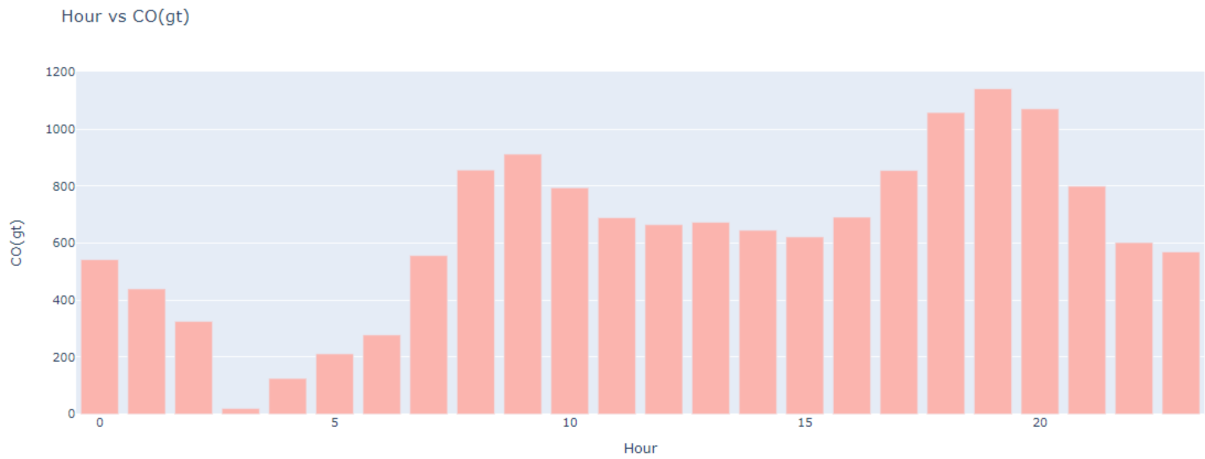


Figura 2 - Hour vs CO

Regarding peak CO concentration times (Figure 2), specific periods were identified in the early morning and late afternoon, possibly correlated with peak human activity times, such as commuting to schools and workplaces. The lowest CO concentrations were recorded in the early hours of the morning, suggesting a decrease in activities.

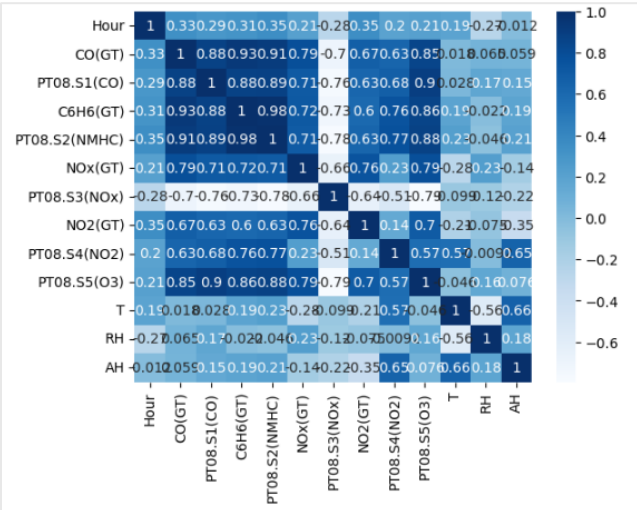


Figura 3 - Corr

The analysis also revealed a significant correlation between relative humidity and environmental factors, with a notable relationship between temperature and relative humidity. This finding indicates that warmer periods may exhibit lower relative humidity, which has important implications for environmental management, especially in urban areas affected by the heat island effect. Furthermore, the analysis suggested an interaction between relative humidity and nitrogen oxide levels, indicating potential complex interactions between air pollution and meteorological conditions.

B) Estimate Relative Humidity behavior based on its answer to other parameters.

To estimate the behavior of Relative Humidity (RH) concerning other key parameters, some important data are observed:

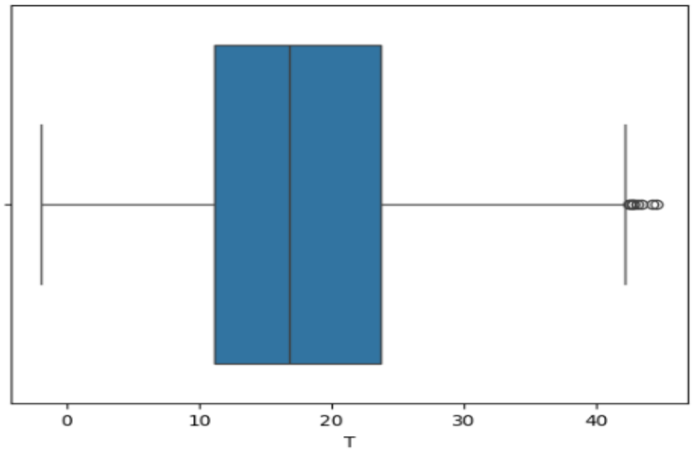


Figura 4 - BoxPlot Temperature

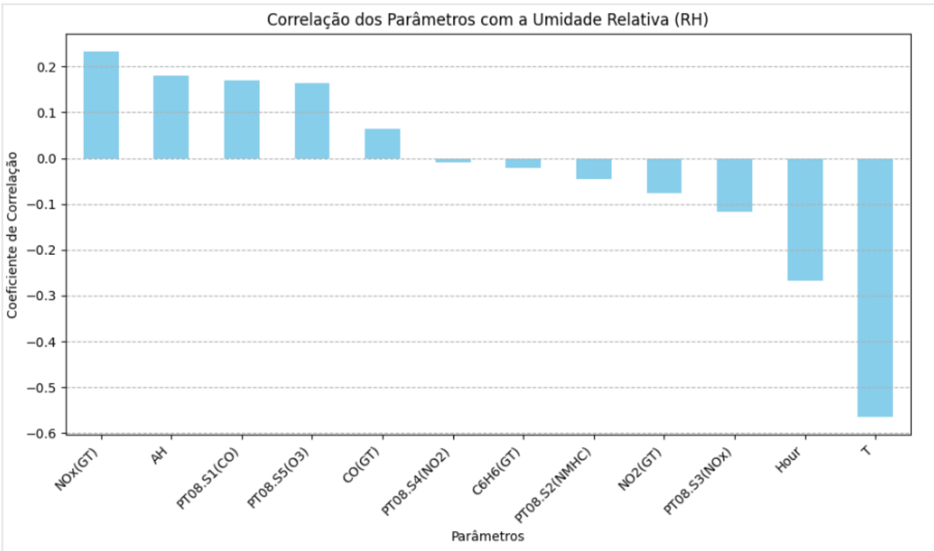


Figura 5 - Pearson Corr

First, Temperature (T) (Figure 4) showed a large variation, ranging from cold to hot (-1.9°C to 44.6°C), with an average of 17.8°C. We found that temperature has a strong inverse relationship with RH. This means that as temperature increases, relative humidity generally decreases. Therefore, on hotter days, we can expect the air to be drier.

Relative Humidity (RH), in turn, varies considerably (9.2% to 88.7%), with an average of 48.9%. This variation indicates that humidity can change significantly depending on the weather conditions and time of day.

Regarding pollutants, the concentration of Carbon Monoxide (CO) varies from 0.1 to 11.9 mg/m³, with an average close to 2.18 mg/m³. Through this exploration, we note the relationship of RH with nitrogen oxides (NOx), where we found a moderate positive correlation (0.23). This suggests that in environments with higher humidity, the concentration of NOx tends to be higher. We also observed slight positive correlations between RH and CO and ozone (O3) sensors, indicating that humidity may slightly affect the presence of these pollutants in the air.

Furthermore, the time of day influences RH, with a tendency for lower relative humidity during the hottest hours, reinforcing the relationship between temperature and humidity.

Regarding analytical models, Gradient Boosting stood out with excellent performance (98), showing that it is a reliable tool for predicting RH from these parameters. The removal of outliers did not significantly change the results, indicating that our observations are stable and reliable.

It is concluded that temperature is a crucial factor that inversely affects RH. The presence of certain pollutants such as NOx also increases with humidity. These findings help us better understand how relative humidity behaves and is influenced by different environmental conditions and pollutants.

With outliers:

```python
results_sorted = results.sort_values(by='Test R2', ascending=False)

# Exibir os melhores valores ordenados
print(results_sorted)
```

```
               Method  Training MSE  Training R2    Test MSE   Test R2
3   Gradient Boosting      2.951906     0.990239    3.458273  0.988831
0   Linear Regression     36.374550     0.879725   34.080248  0.889936
1    Huber Regression     38.935391     0.871258   35.434060  0.885564
4    Gaussian Process     37.591669     0.875701   35.520956  0.885283
6           Ada Boost     36.187602     0.880344   39.166687  0.873509
9                 MLP     68.065265     0.774938   67.539955  0.781877
5         K-Neighbors     64.184545     0.787770   96.915838  0.687006
2       Random Forest     97.705572     0.676931  102.727439  0.668237
8       Decision Tree    110.316721     0.635231  116.418195  0.624022
7                 SVR    193.345707     0.360691  195.634082  0.368190
```

Figura 6 - Methods resume.

Without outliers:

```python
results_sorted = results.sort_values(by='Test R2', ascending=False)

# Exibir os melhores valores ordenados
print(results_sorted)
```

```
               Method  Training MSE  Training R2    Test MSE   Test R2
3   Gradient Boosting      3.025345     0.990143    3.869986  0.986630
9                 MLP     21.689596     0.929331   22.210800  0.923266
0   Linear Regression     33.839091     0.889745   31.831988  0.890027
4    Gaussian Process     35.152506     0.885466   32.920851  0.886265
1    Huber Regression     39.870559     0.870094   37.697269  0.869763
6           Ada Boost     35.994745     0.882722   37.975108  0.868803
5         K-Neighbors     62.771240     0.795478   96.954307  0.665042
2       Random Forest     99.762547     0.674953  101.698491  0.648651
8       Decision Tree    112.437955     0.633654  117.110192  0.595407
7                 SVR    197.237813     0.357359  184.398454  0.362939
```

Figura 7 - Methods Resume (without outliers)

PROBLEM 2

A) Provide some insights on the data such as shape, distribution, and cross-category comparisons (data exploration)

When evaluating the dataset composed of 1728 rows and 7 columns, a detailed insight into the determinative features for automobile classification was obtained. The dataset is complete, without any missing values or duplicates.

Among the various insights extracted, the following stand out:

	frequencia	Porcentagem (%)
unacc	1210	0.700231
acc	384	0.222222
good	69	0.039931
vgood	65	0.037616

Figura 8 - Class Grouped

The 'Unacceptable' category is prevalent: Approximately 70% (Figure 8) of cars were classified as 'Unacceptable'. This demonstrates a clear trend, indicating that many cars do not meet certain desired criteria.

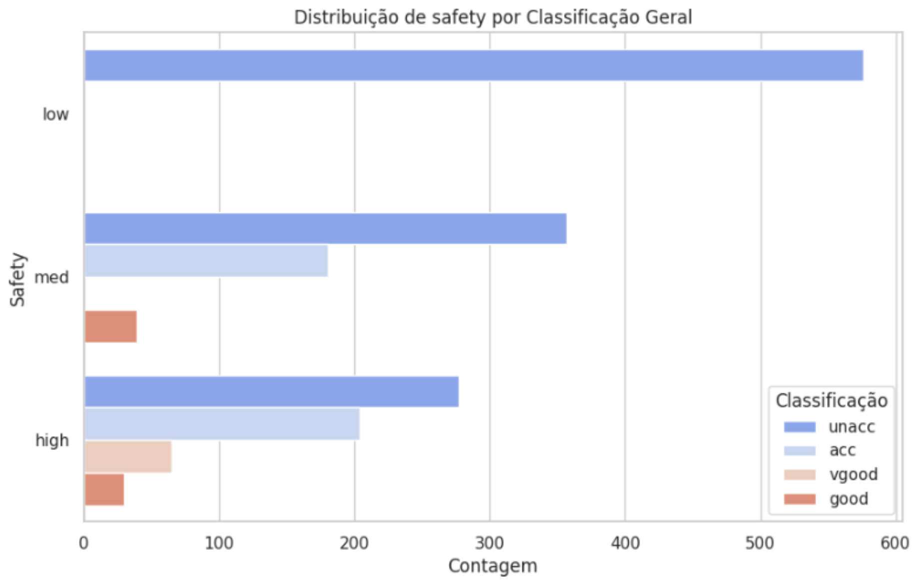


Figura 9 - Safety vs Class

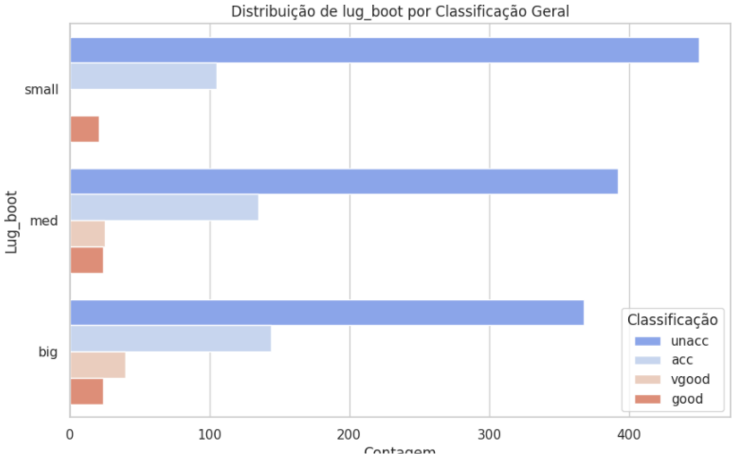


Figura 10 – Lug_boot vs Class

- Number of doors and evaluations: It was noticed that cars with 5 or more doors tend to receive slightly better ratings. This suggests that having more doors may be viewed positively, although the difference is not very significant.
- Safety and Classification: Cars classified as having low safety were all considered 'Unacceptable' (Figure 9). This highlights the importance of safety in evaluation criteria.
- • Trunk space: Vehicles with a large trunk space (Figure 10) received better ratings. This indicates that larger storage space is valued.
- Expensive cars are not always well-rated: Cars with higher prices did not necessarily receive better ratings, suggesting that consumers seek a good balance between cost and benefit.

Based on these insights, the recommendation for car manufacturers would be to focus on safety and cost-effectiveness, while also considering the importance of space. These are areas that appear to influence market expectations and, if improved upon, can lead to greater customer satisfaction.

B)   Given Logistic Regression, Random Forest Classifier and Decision Tree, which model performs better when predicting car class? Justify your answer with data.

During the comparative analysis of machine learning models, it was observed that the Random Forest model outperformed the Decision Tree and Logistic Regression models. Although the former models achieved accuracy rates of 81.9% and 75.72%, respectively, their macro f1-score scores were 39.98% and 35.14%, which may be attributed to the "unacc" class, negatively impacting the macro f1-score accuracy rate.

```
Classification accuracy RF = 96.82%

Taxa de acerto : 96.82
F1 Score Macro : 89.09
------------ Modelo RF -----------------
```

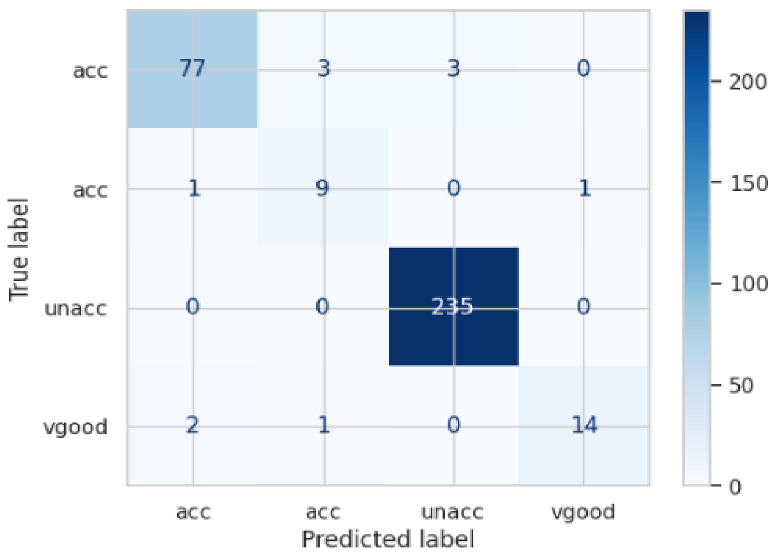


Figura 11 - RF

On the other hand, the Random Forest model achieved an accuracy rate of 96.82% (Figure 11), accompanied by a macro f1-score of 89%. This remarkable performance reflects the model's ability to make accurate predictions for the "unacc" class, thus elevating its macro f1-score score.

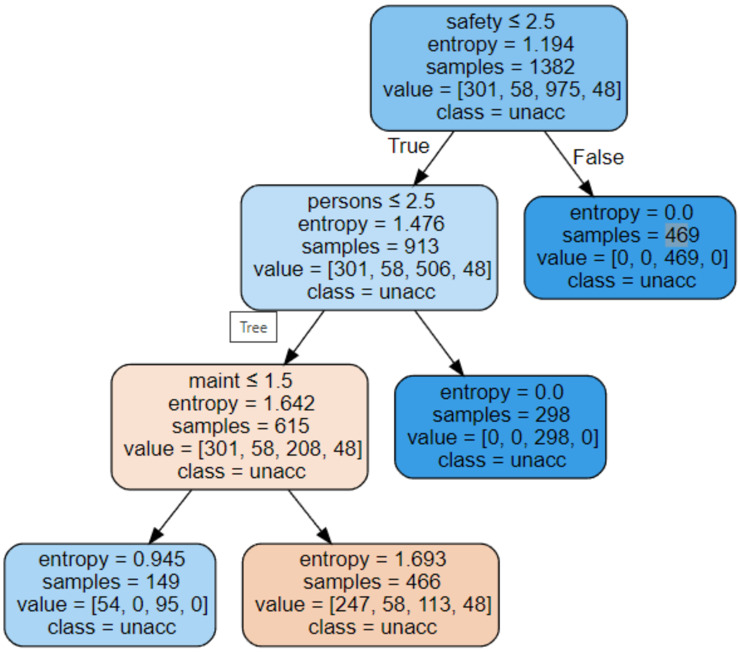
C)   Rank feature importance with respect to Random Forest Model and share your insights.

The ability of a vehicle to accommodate a certain number of passengers is not just a matter of space or convenience; it proves to be a key criterion in defining its classification, such as family car, sports car, or compact car. This factor directly influences aspects like safety, comfort, and functionality, which depend on the number of individuals the car can transport comfortably.

In vehicle data analysis, the RandomForest model assesses the importance of variables by observing how the splits made in the model's trees reduce node impurity. The variables that, when splitting the data, result in a greater reduction in impurity or error are seen as more important. Therefore, 'persons' and 'buying' are variables that stand out for significantly aiding in the differentiation and correct classification of training data.

It is important to note that the importance of variables depends on the specific dataset and how the model was trained. Changes in the data or training can alter which variables are considered more important. Furthermore, the importance of a variable does not indicate causality but rather that it is useful for the model to make predictions or classifications with the available data.

D)   Present a visualization of the Decision Tree and share your insights.



When analyzing the decision tree presented in the above figure, it is observed that the decisive criteria for classifying automobiles. The first and most notable division occurs in the safety attribute: vehicles with safety ratings less than or equal to 2.5 are promptly classified as 'unacceptable' (unacc). This criterion reflects the primacy of safety as an evaluation factor in the perception of car Qualy.

Following the structure of the tree, the second decisive factor reveals: the car's capacity in terms of the number of passengers. Vehicles with the capacity to accommodate 2.5 people or fewer equally receive the 'unacc' classification.

This strategic model appears to prioritize vehicle safety and capacity, with subsequent attention to maintenance, to fine-tune car categorization.

Furthermore, the model signals maintenance as another decisive factor. Cars with maintenance costs less than or equal to 1.5 remain in the 'acc' category, suggesting a correlation between lower maintenance costs and negative perceptions about the vehicle. As we proceed through the branches of the tree, the model meticulously minimizes entropy, i.e., disorder or uncertainty in decisions, leading to increasingly accurate and evident categorization.

The clear objective is to reduce entropy - a measure of uncertainty or impurity - thereby facilitating a more direct and less ambiguous classification. It is evident that the chosen factors are crucial for this analysis, demonstrating a methodical decision-making process to arrive at a consistent classification.