Hypothesis:

By using a machine learning algorithm using the classification methods it will be possible to use the data of previous years to encounter in the current year which columns are most likely to cause problems, even with the columns encoded as they are. Once the most problematic components are identified it is possible to compare them with the original dataset to uncover what is the limit value that a component can reach before causing a malfunction.

Results:

- Can we reduce our expenses with this type of maintenance using AI techniques?

Yes it is possible to reduce expenses by observing the thresholds obtained via interpretation of the graphics created from the machine learning algorithm. Following are a few examples obtained from interpreting the graphics:

Ag_001

- High number of confirmed failure cases indicates it is one of the main responsible components

- Values between 0 and 30 are more likely to result in air system failures

- Values above 150 have a low failure rate; trucks with air systems having this component above this threshold are in a safe zone.

Cn_000

- High number of confirmed failure cases suggests it is one of the main responsible components

- Values between 0 and 10,000 are highly likely to cause air system failures

- Values above 100,000 have a low failure rate; trucks with air systems having this component above this value are in a safe zone

Ag_002

- High number of confirmed failure cases suggests it is one of the main responsible components

- Values between 0 and 0.0625 are highly likely to cause air system failures

- Values above 0.8 have a low failure rate; trucks with air systems having this component above this value are in a safe zone

- Can you present to me the main factors that point to a possible failure in this system?

Each component from the air filtration system has its own threshold values that indicate a possible failure on the system and subsequently having to send a vehicle to repairs. The analysis performed indicates that a handful of components have a much higher probability of causing issues to the air systems, by evaluating the graphics of these components thresholds of safety can be determined. By keeping a close eye on these values it is possible to detect a malfunction before it even happens and provide adequate maintenance, vastly reducing the costs of repair.

1. What steps would you take to solve this problem? Please describe as completely and clearly as possible all the steps that you see as essential for solving the problem.

By using a machine learning algorithm using the classification methods it will be possible to use the data of previous years to encounter in the current year which columns are most likely to cause problems, even with the columns encoded as they are. Once the most problematic components are identified it is possible to compare them with the original dataset to uncover what is the limit value that a component can reach before causing a malfunction.

The resulting predictions obtained from the algorithm had an accuracy of 97% and were subjected to outlier removal to further increase the likelihood of finding meaningful patterns on the data.

After the proper treatment was applied to the data frame, graphics of the most relevant components can be plotted and further analyzed and interpreted.

2. Which **technical** data science metric would you use to solve this challenge? Ex: absolute error, rmse, etc.

   The technical data used to solve this challenge was accuracy and recall. Both used to guarantee that the classification of data was as precise as possible in order to find which components were the most likely to cause issues to the vehicles

3. Which business metric ***would*** you use to solve the challenge?

   Forecasting and Risk Assessment are business metrics useful to analyze and explain the results obtained by the machine learning algorithm, providing solid data to be used.

4. How do technical metrics relate to the business metrics?

   Both Forecasting and Risk Assessment rely on the precision Accuracy and Recall produce. Therefore by guaranteeing that the data is being precisely classified it guarantees that the prediction of metrics and the detection of risks are as correct as they can be.

5. What types of analyzes would you like to perform on the customer database?

   Prediction to find which components are the most likely to cause failures, Classification in order to correctly label failures on the air system, KNN prediction to fill in the 'na' data based on their nearest values and outlier removal to "clean" the data for better graphic visualization.

6. What techniques would you use to reduce the dimensionality of the problem?

   PCA is a fairly common method of reducing dimensionality for Classification algorithms, however it can be considered optional depending on the methods used for the Machine Learning algorithm

7. What techniques would you use to select variables for your predictive model?

   Feature Importances was used to identify the most likely columns that contribute for a 'pos' result

8. What predictive models would you use or test for this problem? Please indicate at least 3.

Random Forest, KNN and Logic Regression

9. How would you rate which of the trained models is the best?

   The model was chosen by evaluating its higher accuracy, precision, recall, and F1 score, indicating a best performing model.

10. How would you explain the result of your model? Is it possible to know which variables are most important?

    The results display the correctly classified labels that were predicted by using the trained data, using these results it is possible to use Feature Importance to predict which variables are the most important.

11. How would you assess the financial impact of the proposed model?

    Since the model is able to classify the most likely culprits for causing failures in air systems it is also capable of predicting which values each component must be above or below in its threshold to indicate possible malfunctions. Knowing of said malfunctions beforehand drastically reduces the costs of repair since they can be done preventively.

12. What techniques would you use to perform the hyperparameter optimization of the chosen model?

    Random Search could be useful to optimize the algorithm, given that it relies on a fixed number of hyperparameter combinations from the specified ranges. This approach can find good hyperparameters more efficiently than grid search, which relies on brute force.

13. What risks or precautions would you present to the customer before putting this model into production?

    The model is currently experimental and not fully optimized, not only that it requires plenty of quality of life adjustments for ease of use before it can be put into production.

14. If your predictive model is approved, how would you put it into production?

The model should be Sterilysed and Refined before being placed into a Cloud Deployment Environment such as AWS. An API could then be used to access and send new data to the model for prediction and analysis

15. If the model is in production, how would you monitor it?

Classification Algorithms don’t necessarily need monitoring, however the use of monitoring tools to track the model’s performance and health in production is highly recommended. Tools like Prometheus and Grafana are good choices for that task

16. If the model is in production, how would you know when to retrain it?

It is possible to know the appropriate time to retain the model via the monitoring of its performance. By monitoring it’s precision, accuracy, recall, F1-score, MAE and MSE and establishing thresholds for their ideal values it is possible to be alerted to when it’s performance is less than optimal with ease, once these values fall below healthy numbers the model should be retained, retrained. retested and updated.