Answers for questions

What are the data types? (Only numeric and categorical)
A: only numeric

Are there missing values?
A: no

What are the likely distributions of the numeric variables?
A: Looking at the Q-Q plot, histogram, and distplots above, it appears that only a few independent variables are roughly following normal distribution, like temperature, humidity, Raw H2, and Raw Ethanol, others are not following a normal distribution, which means there are a tone of outliers

Which independent variables are useful to predict a target (dependent variable)? (Use at least three methods)
A: from logistic regression:

1. NC0.5
2. Humidity

from xgboost:

1. PM1.0
2. PM2.5
3. Pressure
4. TVOC

from random forest:

1. Pressure
2. TVOC
3. Raw Ethanol

Which independent variables have missing data? How much?
A: no missing data in this data set

Do the training and test sets have the same data?
A: no, the data set is split 90% into a training set and 10% into a test set

Is the predictor variables independent of all the other predictor variables?
A: No, they are not all independent of each other, we can know that from the correlation heap map.

1. features like PMs and NCs' that describe the particles in the air have a strong correlation with each other

2. TVOC has a strong correlation with Raw ethanol

Which predictor variables are the most important?
A: from logistic regression:

1. NC0.5
2. Humidity

from xgboost:

1. PM1.0
2. PM2.5
3. Pressure
4. TVOC

from random forest:

1. Pressure
2. TVOC
3. Raw Ethanol

What are the distributions of the predictor variables?
A: Looking at the Q-Q plot, histogram, and distplots above, it appears that only a few independent variables are roughly following normal distribution, like temperature, humidity, Raw H2, and Raw Ethanol, others are not following a normal distribution, which means there are a tone of outliers

Do the ranges of the predictor variables make sense?
A: After processing outliers, the range of predictor variables becomes more explainable.

1. the range of temperature is from -22 C to 59.9C, the max temperature is too high for our daily life, but it maybe occurs when the fire really happened, so the temperature could be so high
2. the humidity is between 38.98 and 61.8 which is make sense
3. the pressure stays between 930 to 939 which makes sense to our daily life
4. the PMs' and NCs' are very regular except NC0.5's max is 22.56 that may be due to the fire increase in the number of particles in the air
5. eco2 stay around 400 except in some situation it reaches 495, that may be the outliers or some extreme situation.
6. Row H2 and Raw Ethanol don't change sharply, so it makes sense.

Remove outliers and keep outliers (does if have an effect of the final predictive model)?
A: yes, removing outliers improve the accuracy from 87.8% to 88.9%, and the MSE improved from 0.12 to 0.11

Remove 1%, 5%, and 10% of your data randomly and impute the values back using at least 3 imputation methods. How well did the methods recover the missing values? That is remove some data, check the % error on residuals for numeric data and check for bias and variance of the error.

A:

no deleting: accuracy: 0.8893, MSE: 0.11

Method: fill with mean

deleting 1%: accuracy: 0.8746, MSE: 0.13

deleting 5%: accuracy: 0.8623, MSE: 0.14

deleting 10%: accuracy: 0.8427, MSE: 0.16

Method: drop the NaN

deleting 1%: accuracy: 0.8734, MSE: 0.13

deleting 5%: accuracy: 0.8926, MSE: 0.11

deleting 10%: accuracy: 0.8704, MSE: 0.13

Method: Linear interpolation

deleting 1%: accuracy: 0.8912, MSE: 0.11

deleting 5%: accuracy: 0.8866, MSE: 0.11

deleting 10%: accuracy: 0.8887, MSE: 0.11