Forecasting Vehicle Smog Level using Machine Learning

Yifan Ren, Catherine Zheng, Jerry Wu, Zihao Shen

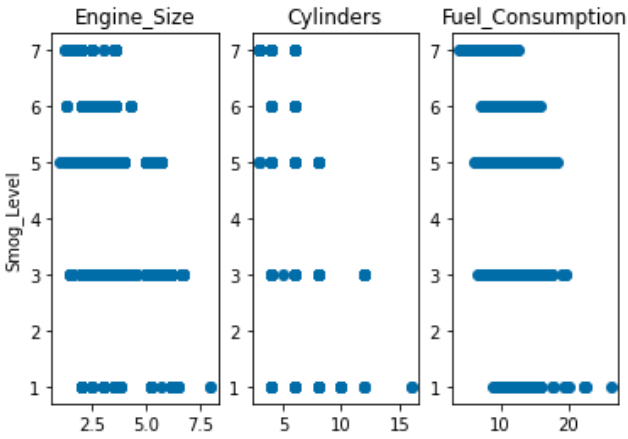
Introduction:

Due to the rapidly increasing demand for transportation methods including cars, trains, and airplanes, more contaminants are released into the air. In this study, we leverage a variety of statistics and five machine learning models to classify and forecast the cleanliness of vehicle emissions with features: engine size, number of cylinders, and combined fuel consumption. We optimize each model by finding the best parameters and then select the one with the highest accuracy.

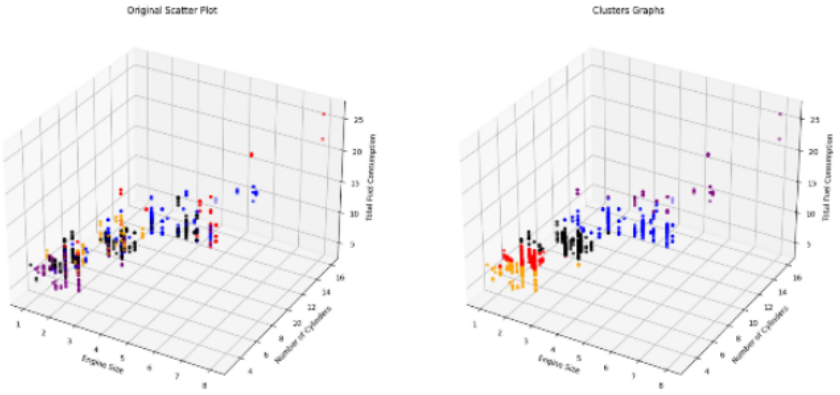
Data Exploration

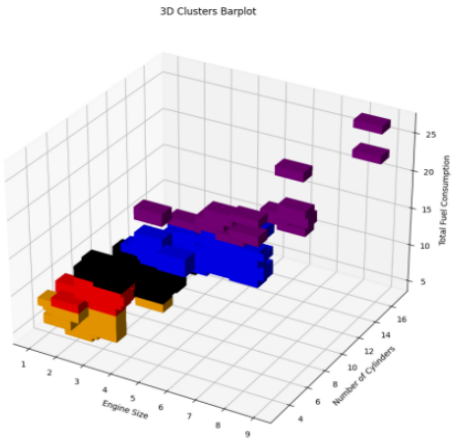
Our dataset, which we obtained from Kaggle, contains fundamental data on 935 vehicles from various brands. A car's smog level measures how much it contributes to air pollution, ranging from 1 to 10, with 10 being the cleanest. We transformed the smog level into a binary response to limit the variance: smog_stat = 0 for smog_level ≤ 4 and 1 for smog_level > 4. Following that, we divided the data into three sets: training, validation, and testing.

	Model_Year	Engine_Size	Cylinders	Fuel_Consumption_in_City(L/100 km)	Fuel_Consumption_in_City_Hwy(L/100 km)	Fuel_Consumption_comb(L/100km)	CO2_Emissions	Smog_Level	Smog_stat
count	935.0	935.000000	935.000000	935.000000	935.000000	935.000000	935.000000	935.000000	935.000000
mean	2021.0	3.214866	5.716578	12.498610	9.306203	11.060214	258.529412	4.726203	0.681283
std	0.0	1.388513	1.977359	3.487271	2.215819	2.867028	64.442768	1.712127	0.466228
min	2021.0	1.000000	3.000000	4.000000	3.900000	4.000000	94.000000	1.000000	0.000000
25%	2021.0	2.000000	4.000000	10.100000	7.700000	9.100000	213.000000	3.000000	0.000000
50%	2021.0	3.000000	6.000000	12.000000	9.000000	10.700000	255.000000	5.000000	1.000000
75%	2021.0	4.000000	8.000000	14.800000	10.800000	13.100000	303.500000	6.000000	1.000000
max	2021.0	8.000000	16.000000	30.300000	20.900000	26.100000	608.000000	7.000000	1.000000



We visualized the original and the clustered 3D scatter plot to better understand that if the labels were missing, clustering has a higher performance in forecasting the smog level at the centroid of the dataset, but lower performance elsewhere.





Model Selection

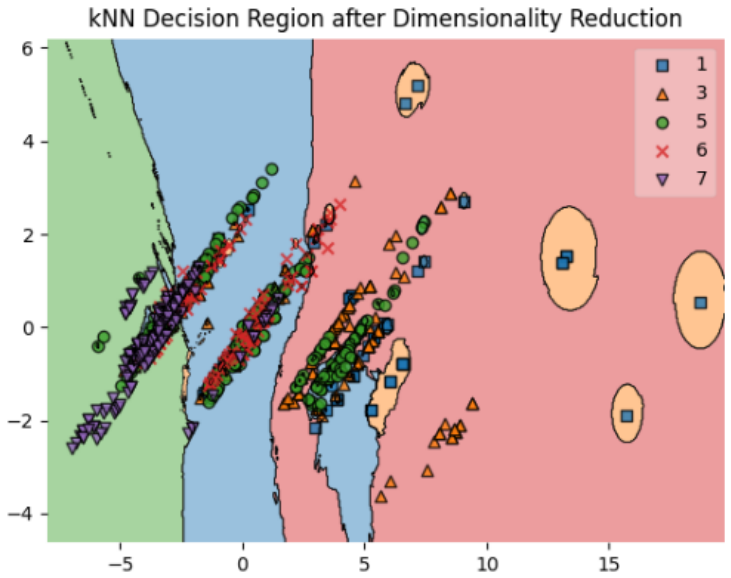
I: KNN

We leverage the weighted KNN method by

$$\frac{1}{d_i}\left(\sum_{i=1}^{N}|a_i - b_i|^p\right)^{\frac{1}{p}}$$

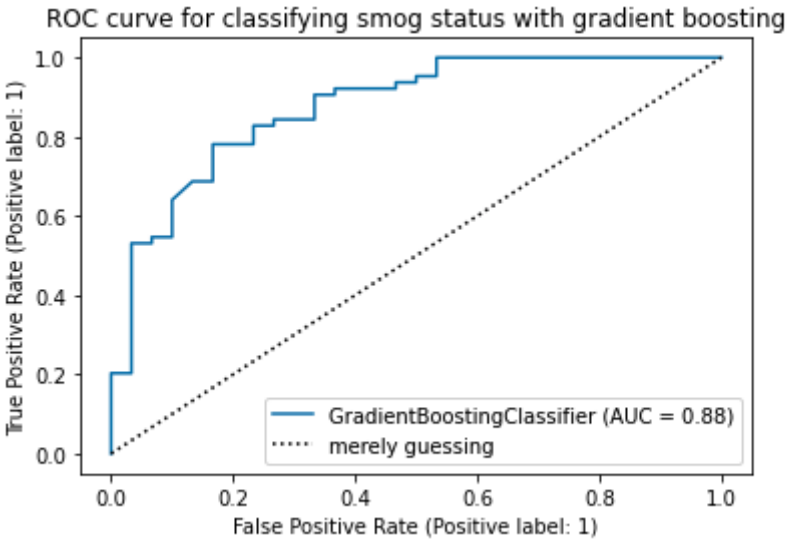
as the distance of each new point to the nearest neighbor. When using the kNN approach to deal with multi-class labels, the classification accuracy is unsatisfactory. We utilize PCA dimensionality reduction and standardization to reduce overfitting. Based on the optimal hyperparameters: n_neighbors = 94, p=2, weights = distance.

	kNN	PCA	Normalization
Train Accuracy	0.817	0.823	0.823
Test Accuracy	0.585	0.628	0.637



II: Gradient Boosting

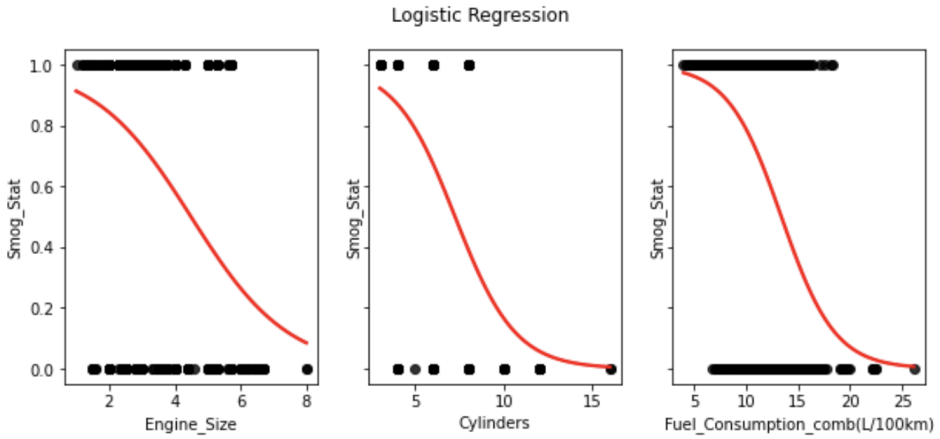
We select Gradient Boosting Classification as it iteratively creates a new model to correct the previous model's errors and the final ensemble model combines all models. Using GridSearch, we find the best parameters: α=0.25, max_depth = 8, max_features = "sqrt". Gradient Boosting is known to be good at reducing bias and variance to prevent underfitting and handling large datasets, however, it could overfit the data and take a longer time to train.



III: Logistic regression

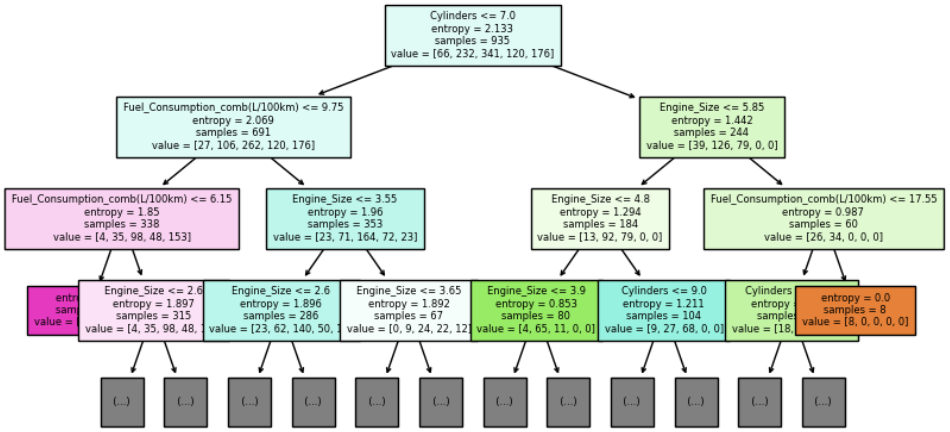
We also choose to fit the Logistic Regression model to examine the correlation between the car's features and smog_status. To explore the potential interactions between the car's features and improve the model's performance, Polynomial Features is also applied.

Since the Logistic Regression model is difficult to perform different dimensions, we treat each feature of the car as an independent predictor when we are visualizing it. The coefficients given by our multiclass Logistic Regression model are all negative, inferring the smog_status decreases as the car's features increase, as shown in the curves.

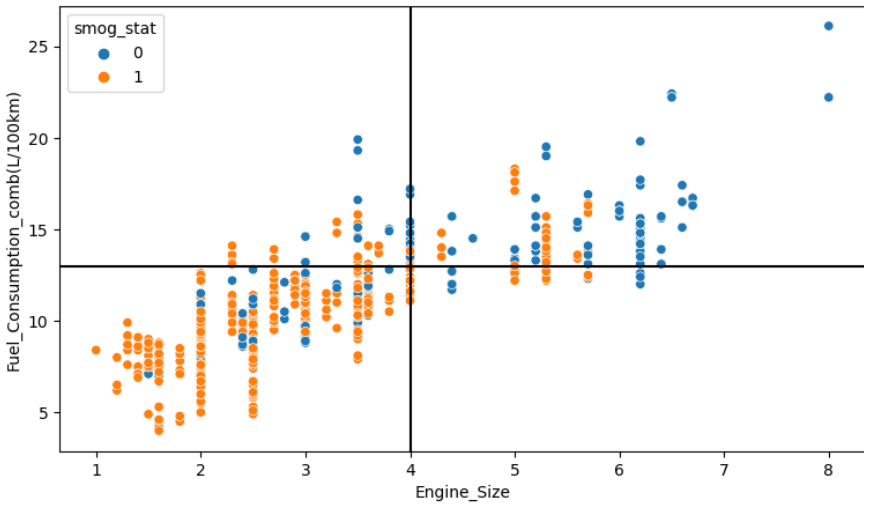


IV: Decision tree

We choose to do a decision tree classification. By visualizing those three features through a decision tree graph with max depth = 3, each class is colored by its importance, and breaks down to subsets for decision-making and increases readability.



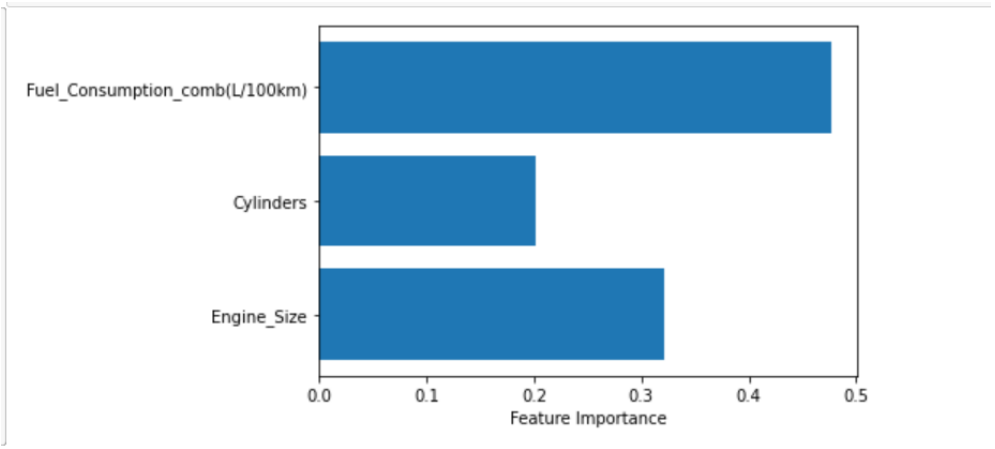
Through another visualization between engine size and fuel consumption combination, we expect that cars with engine size ≤ 4 with fuel consumption combination ≤ 13 have less pollution to the environment compared to others.



V: Random Forest

For decision tree model, overfitting occurs when many samples are fitted within training data, leading to more noise occurrence. However, randomforest model builds multiple individual trees and averages the result to lower the variance and prediction error to prevent overfitting. The accuracy rises after applying randomforest model with parameters Max depth = 10 and base estimators = 100.

Random forest regressor model also provides a built-in feature selection function to help us visualize the importance of each feature and the combined fuel consumption is the key component to predict smog_level of a car.



Results:

	Model	Accuracy	Precision	Recall
0	k-Nearest Neighbor	0.637	0.806	0.770
1	Gradient Boosting	0.819	0.831	0.922
2	Logstic Regression	0.840	0.820	0.920
3	Decesion Tree	0.817	0.850	0.910
4	Random Forest	0.837	0.861	0.932

With the results above, we find that the Logistic Regression using Polynomial Features at degree 3 obtains the highest testing accuracy. By applying the Grid Search, we optimize the model with C = 0.1, and max_iter = 1000. To avoid overfitting the data, L2 regularization is applied to the Logistic regression model.

Conclusion:

Based on the results from the Logistic regression with polynomial features, we believe that with the increases in engine size, cylinders, and combined fuel consumption, the vehicle tends to be dirtier and more likely to pollute the environment. In the future, we hope more data will be available for all smog levels to improve our model for multi-class classifications, also combining categorical features. We hope this can help car manufacturers to design cleaner cars.

Contributions:

Each member has contributed to a model and finished the report together.