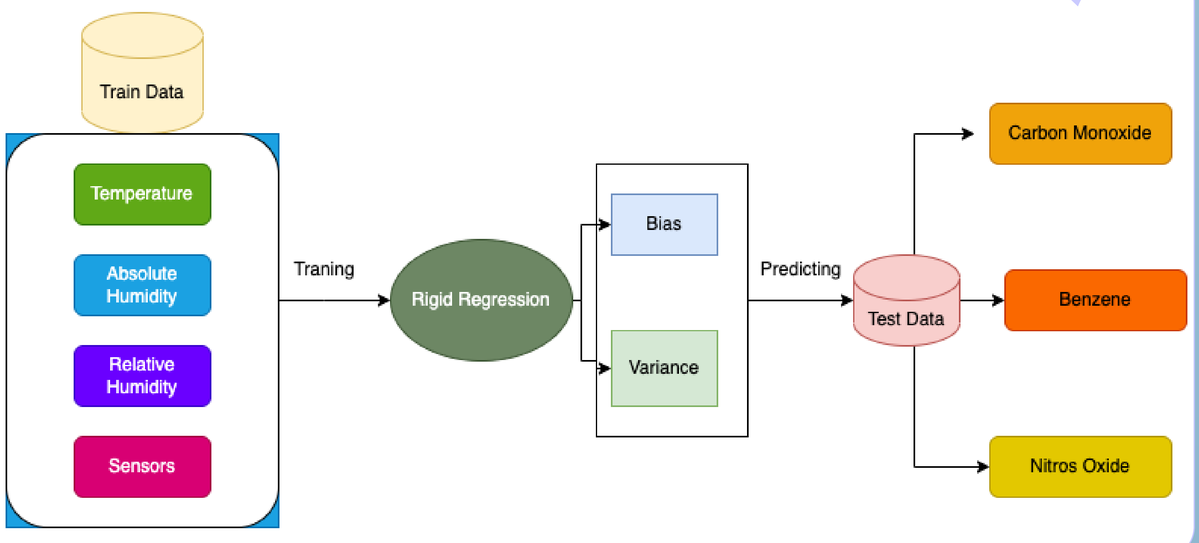
Predicting Air Pollution compositions

Shanika Iroshi Nanayakkara
Deakin University, Australia

Introduction

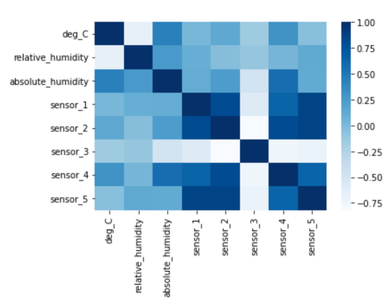
Air pollution is significant problem which arises due to many reasons. Among them technological evolution takes high priority. However, people need to move on with the technical world while keeping sustainable environment for the future generation. For that, almost all industrialization production processes need to concern on their emissions into the environment. For example, many industries tend to left different types of air particles into the atmosphere as a result of their production process. Though it does not seems to have quick threat on environment and the human life, it has long life impact. Consequently, people tend to find the solutions to mitigate these pollution in order to enhance the best balance between the industrialization and the environment protection. Initially, data is the major asset for everything in the world. Because, every application had data driven structure to analyse the current trends in order to predict the future impacts or enhancements. Machine learning and Artificial intelligent play vital role in the data driven applications. Therefore, in such studies present prediction and detection models based on the existing data.
In this study we reveals the high performing prediction model, using supervised learning model called regression, to predict the air pollution. For that, we use temperature values, relative and absolute humidity, and 5 sensor values as predictor variable for several months. The given target variables are compositions of carbon monoxide, benzene, and nitros oxide.

Proposed Framework



Multicollinearity and Rigid Regression

We propose the *Rigid Regression* algorithm to solve the research problem of *Air pollution prediction*. The *Rigid Regression* algorithm uses to hinder the impact of Multicollinearity

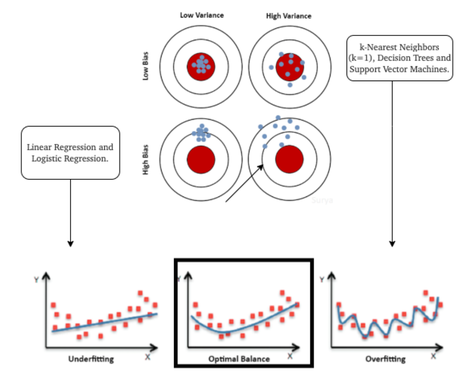


Multicollinearity .Multicollinearity is a aspect which occurs when two or more predictor variables are correlated. As a consequent, coefficients error may increase. Then the each variables behave insignificantly, though they should be significant.

Rigid Regression This is a technique used to implement multicollinearity of predictor variables which are highly correlated each other [5]. Generally, rigid regression is same like linear regression unless it add penalty values to reduce the loss or error of linear regression. This error can be happen due to bias

Preliminaries

Making prediction cannot be always trustworthy to have 100

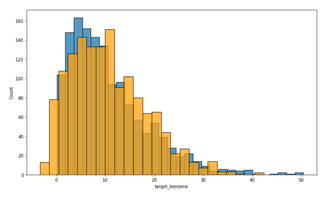


Bias Variance and determination of coefficient Bias and Variance are prediction errors. These errors are called Bias and Variance. We use this bias and variance values in the models to state accuracy levels of proposed models in our evaluation process. Bias is a training data error while variance is the testing data error.

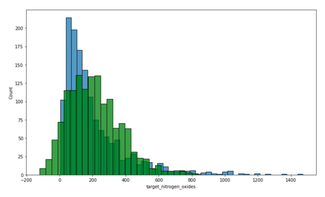
Experiment and Evaluations

The test data histogram is properly fitted with train data histogram. Further, in above However, we use coefficient determination factor to show how well our proposed models fits the observed data.

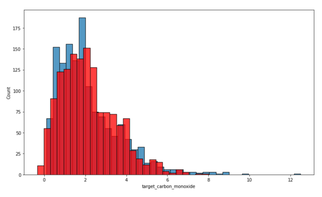
Method	CM	Benzene	NO
Bias	0.38	3.379	12916.6
Variance	0.001	0.009	30.591
CD Training	0.82	0.95	0.68
CD Testing	0.83	0.94	0.66



Benzene Prediction



Nitrous oxide Prediction



Carbon monoxide Prediction

All three models have overall best performance efficiency. For example, carbon monoxide prediction model hold 0.82 (82%) coefficient determination, which indicates high efficiency of the model. And such, benzene model holds 0.95 (95%) coefficient determination, which indicates highest efficiency. Moreover, compare to the other two models, the model for nitrous oxide has less efficiency since it has 0.6 coefficient determination.

Conclusion

Problem Definition we propose prediction model to predict air pollution components which are specified as carbon monoxide, benzene and nitrous oxide.

Algorithm Consequently, we use *rigid linear regression* instead general *linear regression* method.

Strategies We uses multiple predictors such as temperature, absolute and relative humidity and five sensor data. We identify that these predictors are co among each other. Therefore, we reveals the need of handli We clearly show the performance efficiency of pr nation of coefficient, bias and variance

Recommendations Evaluate how well behave on adding more noisy data. How to affect predi by noisy data. Increase the amount of predictors to predict air pol tion.

Acknowledgement
• Kaggle Project and FLIP00 team