# Beijing Air Quality

## Final Presentation

21 NOVEMBER 2023

PROJECT OVERVIEW

# Problem Statement

The aim of this project is to develop predictive models for two critical air pollutants, PM2.5 pollutants, PM2.5 and PM10, using a comprehensive dataset encompassing hourly air hourly air quality measurements and meteorological data from 12 nationally-controlled air-controlled air-quality monitoring sites in Beijing. The goal is to build a regression model for regression model for PM2.5 prediction and a classification model to assess the level of air level of air quality based on PM10 concentrations. The project seeks to address the the challenges of pollution monitoring and forecasting, contributing to improved public improved public health, urban planning, and environmental policy-making.

# Project pipeline


The goal for our final presentation will be to train models and evaluate their effectiveness. Our models will be based upon the seasonality hypothesis we presented hypothesis we presented in our last deadline as well as applying feature selection

# Data Collection

The data includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017

PROJECT OVERVIEW

# Expected Hypothesis

Based on the initial exploratory data analysis, we can see that data has temporal temporal relationships. We have observed that yearly predictions exhibit inconsistency inconsistency across different years. However, when month-wise analysis was conducted, conducted, it identified recurring patterns in pollutant and chemical conditions, even conditions, even during anomalous years like 2017 with sudden increases in these factors. these factors. Therefore, our hypothesis is that by extending our analysis to longer time longer time periods, such as seasons, we can better capture and predict repetitive repetitive meteorological patterns that influence air quality.

# Proving Cyclical Nature of Pollutants

01
THEORY

Based on a few of the hypothesis tests previously presented, the pollutants **NO2, NO2, SO2 and CO had a positive** correlation with our our target variable whereas **O3 O3 had a negative** correlation correlation

02
CLUSTERING METHOD

The modelling process consisted in a consisted in a grid search **hyper-hyper-parameter tuning** of the **features** selected defining the input input data points and the **number of number of clusters**.

03

CLUSTERING PARAMETERS

The hyper-parameters were defined based on the correlation values with the target variable. A list of **eight combinations of features** and a list of **four numbers of clusters**, resulting in $8 * 4 = 32$ models.

04
HYPER-PARAMETERS

Set of combinations of features:

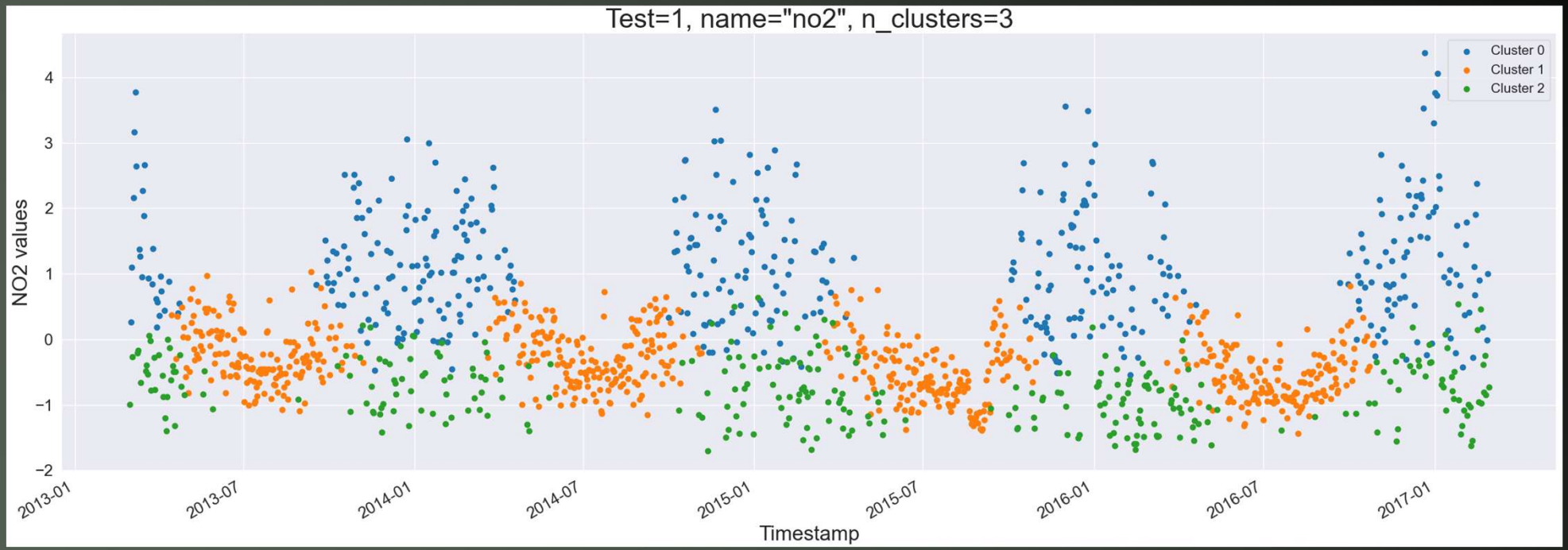
- "['Plltnt', 'TEMP', 'WSPM']".
- "['Plltnt', 'TEMP', 'PRES', 'WSPM']".
- "['Plltnt', 'TEMP', 'DEWP', 'WSPM']".
- "['Plltnt', 'TEMP', 'RAIN', 'WSPM']".
- "['Plltnt', 'TEMP', 'PRES', 'DEWP', 'WSPM']".
- "['Plltnt', 'TEMP', 'PRES', 'RAIN', 'WSPM']".
- "['Plltnt', 'TEMP', 'DEWP', 'RAIN', 'WSPM']".
- "['Plltnt', 'TEMP', 'PRES', 'DEWP', 'RAIN', 'WSPM']".

Set of number of clusters: "[2, 3, 4, 5]".

CLUSTERING RESULTS

# NO2

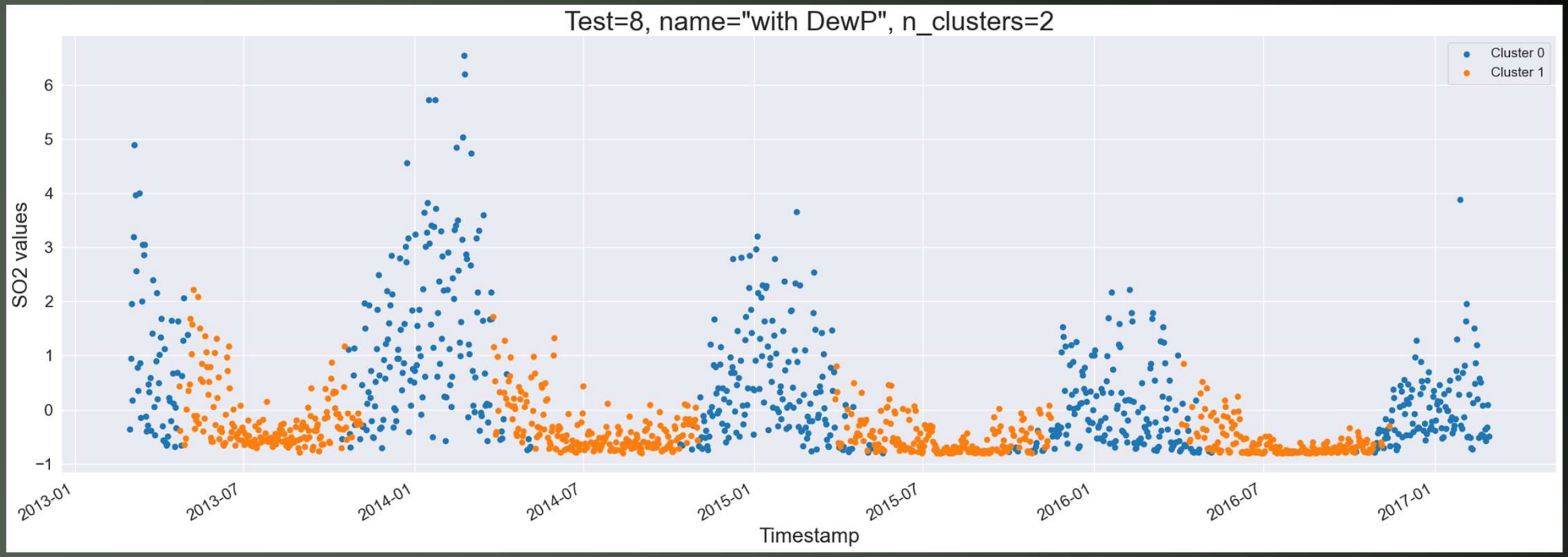
The pollutant NO2 had 5 such clusters out of 32 that had silhouette scores of ~0.45



CLUSTERING RESULTS

# SO2

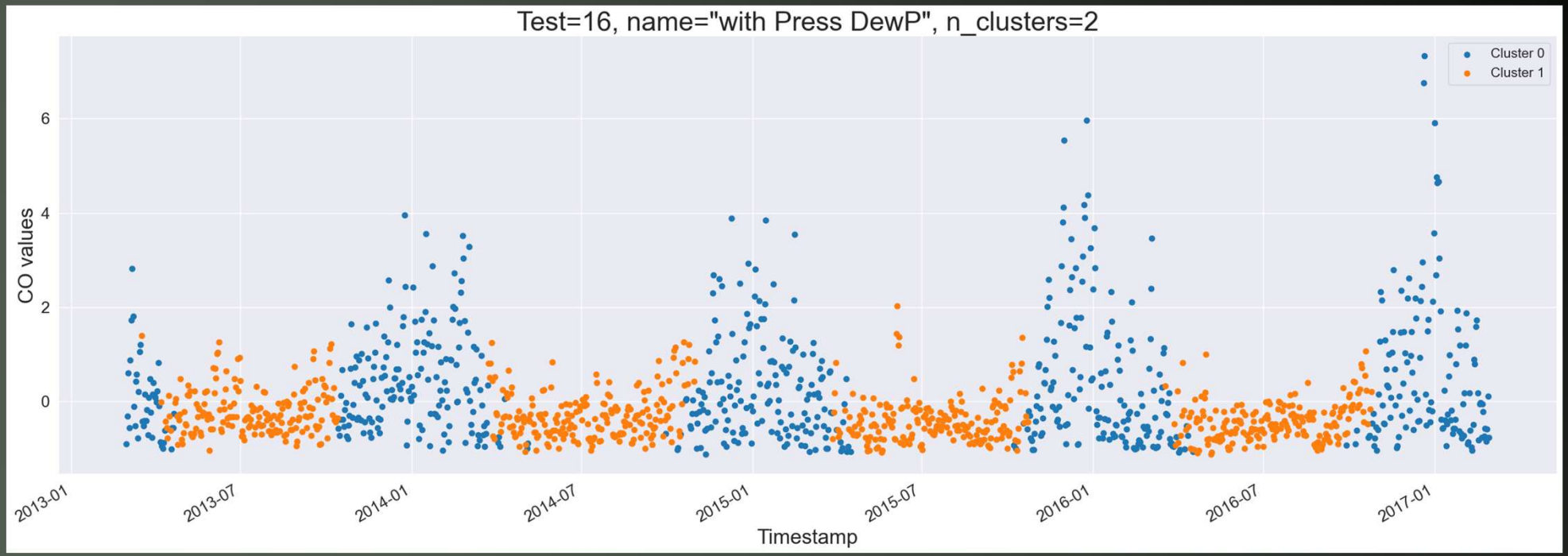
The pollutant SO2 had 4 such clusters out of 32 that had silhouette scores of ~0.45



CLUSTERING RESULTS

# CO

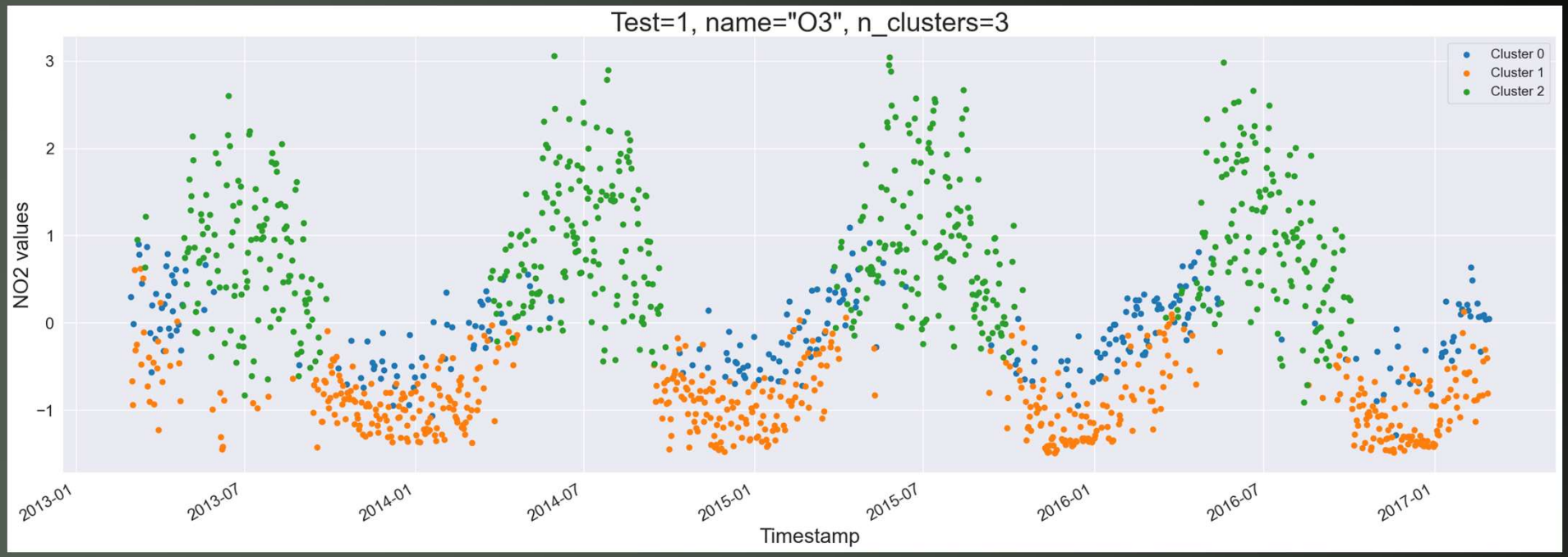
The pollutant CO2 had 6 such clusters out of 32 that had silhouette scores of ~0.45



CLUSTERING RESULTS

# O3

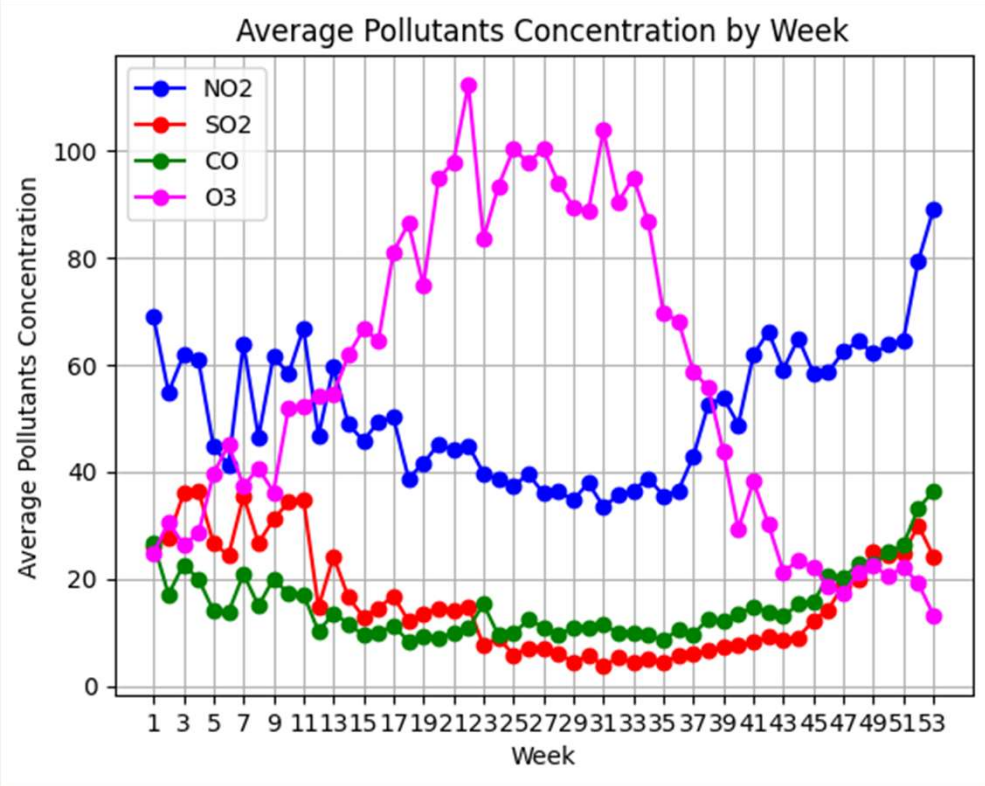
The pollutant O3 had 4 such clusters out of 32 that had silhouette scores of ~0.45



21 NOVEMBER 2023

# Choosing Our Seasons

Based on the Clusters formed by the pollutants, we decided to choose 3 seasons. The time periods of which were decided based on the gradients the pollutant levels followed in different weeks of the year



# Data Preprocessing for ML model application

01
APPLYING SEASONALITY

The 'Week' column was modified to to reflect the trends we saw in our our seasonal hypothesis.
0-11 Season 0
12-34 Season 1
35-52 Season 2

02
ONE HOT ENCODING

One hot encoding was applied to discrete data columns. Feature reduction using PCA was also applied which impacted model performance so was removed

03
DEFINING CLASSIFICATION VARIABLES

We ought to build models for PM10 and PM2.5 variables so we split our Classification goal into 3 categories.
0-100 Class 0
101-400 Class 1
400+ class 3

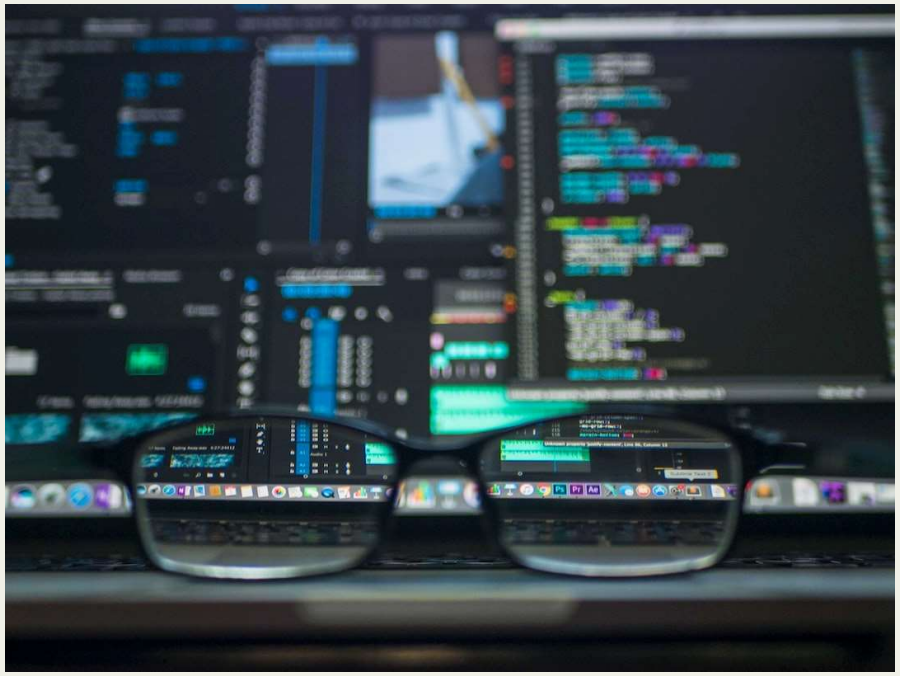
04
SCALING

Use of Sci-kit learn libraries to scale, divide into test-train splits

# Applying LazyRegressor

We used 1% of shuffled data and applied the lazypredict library to find the best model for the task of regression in predicting PM2.5 scores. The models with the best R2 scores are as follows:

| R2 SCORE | MODELS AND PERFORMANCE |
|---|---|
| 0.75 | **RANDOM FOREST REGRESSOR**<br>Mean Squared Error: 1540.2696018434438<br>Mean Absolute Error: 25.052746468453964 |
| 0.73 | **GRADIENT BOOSTING REGRESSOR**<br>Mean Squared Error: 1586.850455163867<br>Mean Absolute Error: 25.3810031552647 |
| 0.72 | **RIDGE CV**<br>Mean Squared Error: 1788.5467560336735<br>Mean Absolute Error: 29.287257183923515 |
| 0.72 | **LASSO CV**<br>Mean Squared Error: 1782.528811148812 Mean Absolute Error: 28.925176215801507 R2 Score: |



SUBHEADLINE

# Results

Post-using LazyRegressor, we identified that the best-performing model was the model was the **Random Forest Regressor**. The Random Forest Regressor is a Regressor is a meta-estimator that fits several decision trees on various data various data sub-samples and uses averaging to **improve** the predictive predictive **accuracy** and **control over-fitting**. This model was then fit on the on the entire dataset and the results obtained were as follows:

**1) Mean Squared Error**: 488.117

2) **Mean Absolute Error**: 13.003

3) **R2 Score**: 0.924

| Model | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | \ |
|---|---|---|---|---|---|
| LGBMClassifier | 0.86 | 0.84 | 0.84 | 0.86 | |
| RandomForestClassifier | 0.86 | 0.84 | 0.84 | 0.86 | |
| LinearSVC | 0.85 | 0.83 | 0.83 | 0.85 | |
| LogisticRegression | 0.85 | 0.83 | 0.83 | 0.85 | |
| XGBClassifier | 0.85 | 0.83 | 0.83 | 0.85 | |
| CalibratedClassifierCV | 0.85 | 0.83 | 0.83 | 0.85 | |
| AdaBoostClassifier | 0.85 | 0.82 | 0.82 | 0.84 | |
| BaggingClassifier | 0.84 | 0.82 | 0.82 | 0.84 | |
| SGDClassifier | 0.83 | 0.81 | 0.81 | 0.83 | |
| ExtraTreesClassifier | 0.84 | 0.80 | 0.80 | 0.84 | |
| SVC | 0.83 | 0.80 | 0.80 | 0.83 | |
| Perceptron | 0.81 | 0.80 | 0.80 | 0.81 | |
| LinearDiscriminantAnalysis | 0.82 | 0.80 | 0.80 | 0.82 | |
| NearestCentroid | 0.80 | 0.80 | 0.80 | 0.80 | |
| RidgeClassifierCV | 0.82 | 0.79 | 0.79 | 0.82 | |
| NuSVC | 0.83 | 0.79 | 0.79 | 0.82 | |
| RidgeClassifier | 0.82 | 0.79 | 0.79 | 0.82 | |
| PassiveAggressiveClassifier | 0.80 | 0.78 | 0.78 | 0.80 | |
| DecisionTreeClassifier | 0.79 | 0.77 | 0.77 | 0.79 | |
| BernoulliNB | 0.78 | 0.77 | 0.77 | 0.78 | |
| GaussianNB | 0.75 | 0.74 | 0.74 | 0.76 | |
| ExtraTreeClassifier | 0.74 | 0.71 | 0.71 | 0.74 | |
| KNeighborsClassifier | 0.75 | 0.70 | 0.70 | 0.74 | |

# Running Classifiers on Original Data with 2 classes

We have used the Lazy Predictor Library to run different models on a best case 2 class system in PM10 classification, with the entire data.

The following slide shows the accuracies of the best models run once we only take into account the seasonality feature in the data, effectively dropping all other time series information thus significantly dropping the dimensions within data.

We also increased the number of classes to 3 to better indicate the level of severity in the PM10 levels in the city.
Despite effectively **reducing the number of features** to about half the original number and **increasing number of classes**, we don't see a significant change in **accuracy** for the best models

# Results on Applying Different Classifier Models

We used 1% of shuffled data and applied different classifier models to find the best model for the task of classification in predicting PM10 class. The models with the best accuracy scores are as follows:

| ACCURACY SCORE | MODELS AND PERFORMANCE |
|---|---|
| 0.845 | **RANDOM FOREST REGRESSOR**<br>Precision: 0.8438539529579784<br>Recall: 0.8442402154967517 |
| 0.83 | **GRADIENT BOOSTING CLASSIFIER**<br>Precision: 0.8316591463501383<br>Recall: 0.8334653779115829 |
| 0.82 | **SVC**<br>Precision: 0.8181931930383618<br>Recall: 0.819996830930122 |
| 0.81 | **LOGISTIC REGRESSION**<br>Precision: 0.8106972293417506<br>Recall: 0.8135002376802408 |

# Effect of Hypothesis Chosen Empirically

01

SEASONALITY HYPOTHESIS

Pollutant levels appeared to be cyclic in the data, and displayed patterns in longer time periods. Thus seasonal data(multiple weeks) should help capture this information.

02

RESULTS

The accuracy for both classification and regression.

R2 went from 0.824 to 0.924
Accuracy dropped only by 0.86 to 0.845 despite increase in classes as well.

03

WEEKEND HYPOTHESIS

Based on the EDA in the report and our hypothesis in the MidSem evaluation, weekends had significantly less pollution as compared to Weekdays.

04

RESULTS

We can switch from a day of the week week based feature to only having a a binary, weekend/ not weekend feature.

# Conclusion

In our project, we conducted an in-depth analysis of pollution data, which **was recorded hourly** and spanned an extensive time period. Through our investigations, we found that by strategically **sampling the data multiple times at intervals representing different seasons**, we achieved **comparable accuracy** while significantly **reducing the computational demands** associated with hourly measurements. Notably, our results demonstrated that the classification of PM10 levels exhibited minimal variance with this reduced sampling frequency.

Furthermore, our exploration into PM2.5 data revealed a noteworthy **improvement in predictive performance** when adopting a seasonality-based approach. The utilization of seasonal time periods contributed to a **higher R2 score**, underlining the efficacy of this strategy in capturing the **underlying patterns in the data**.

Our project highlighted the advantages of simplifying the temporal aspect by transforming the **day-of-the-week** information into a **binary feature**. This streamlined representation proved to be **more effective** in practice than the previous detailed day-wise approach, enhancing the overall model performance.

In conclusion, our findings suggest that strategic data sampling aligned with seasonal patterns and the adoption of binary features for temporal representation offer practical and efficient alternatives for pollution data analysis, providing valuable insights while mitigating computational complexities.

TEAM GRADE-IENT DESCENT

# Thank you

CSE558: DATA SCIENCE