Prediction of Fine Dust by Weather

Team number 3

22000120 Kim Shin

22000350 Seo Jueun

22000532 Lee Seung Jae

22000690 Jung Yiju

- Problem description
    - Air pollution is measured by the atmospheric concentration of air pollutants such as sulfur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2), fine dust (PM10 and PM2.5), ozone (O3), lead, and benzene. Among them, fine dust has recently attracted the most attention due to its health effects. Fine dust (PM2.5) refers to dust with a diameter of 2.5 μm or less.
    - Korea has set certain standards (environmental standards) for major air pollutants such as sulfur dioxide, carbon monoxide, nitrogen dioxide, fine dust (PM10 and PM2.5), lead, and benzene.
    - Among the various air pollutants, fine dust is the most attractive attention in relation to national health. Fine dust (PM2.5), which has been a particular problem in recent years, is known to be more toxic because it has a small particle size and can go deep into the alveoli.
    - Internationally, Korea's fine dust pollution level is very high. According to OECD data, Korea's fine dust (PM2.5) concentration was 27.4$\mu g/m^3$ as of 2019, the lowest among the comparable countries. This is about twice as severe as the OECD average (13.9 $\mu g/m^3$). It is more than four times worse than Finland, which has the lowest level of fine dust pollution.

- Background and conventional approach
    - The method of predicting fine dust forecast is as follows. Recent fine dust forecast uses the air quality information and weather information. With this data, many models predict the fine dust concentration of each city. Many agencies use linear and nonlinear techniques using machine learning and time series models to predict fine dust. To use many weather data to predict exact fine dust, they usually use multiple

linear regression. Therefore, our team decided to create a model using Multiple Linear Regression and Random Forest to predict fine dust. By comparing the accuracy of the two models, we tried to determine which model is more efficient in predicting fine dust.

- Proposed method
  - Data collection and preparation
    - We need data on the monthly air pollution level and factors affecting the concentration of fine dust in Seoul. We imported air pollution level data from Seoul Open Data Square and weather element data from the Meteorological Agency's weather data opening portal. In the first data, there were 120 rows and 2 columns, and in the column, there were date and fine dust concentration data. In the second data, there were 120 rows and 13 columns, and in the column, there were data such as date, average wind speed, maximum wind speed, maximum wind direction, etc.

	일시	PM10
0	2012-01	60.0
1	2012-02	50.0
2	2012-03	47.0
3	2012-04	51.0
4	2012-05	52.0
...	...	...
115	2021-08	21.0
116	2021-09	15.0
117	2021-10	27.0
118	2021-11	46.0
119	2021-12	39.0

120 rows × 2 columns

Figure 1. the monthly fine dust concentration data

	일시	평균풍속(m/s)	최대풍속(m/s)	최대풍속풍향(deg)	평균기온(℃)	평균최고기온(℃)	평균최저기온(℃)	강수량(mm)	평균습도(%rh)	최저습도(%rh)	일조합(hr)	일조율(%)	일사합(MJ/m2)
0	2012-01	2.5	7.5	320	-2.8	1.3	-6.3	6.7	49.0	12.0	190.5	62.05	234.50
1	2012-02	2.9	8.1	290	-2.0	3.0	-6.0	0.8	43.0	11.0	224.9	71.74	325.25
2	2012-03	3.5	10.2	290	5.1	9.5	1.5	47.4	52.0	9.0	191.8	51.63	367.51
3	2012-04	3.4	12.0	290	12.3	17.9	7.8	157.0	54.0	9.0	212.5	53.72	448.48
4	2012-05	2.7	8.6	250	19.7	25.1	15.4	8.2	48.0	11.0	251.3	57.04	553.40
...	...	...	...	...	...	...	...	...	...	...	...	...	...
115	2021-08	2.1	8.3	320	25.9	29.7	22.8	211.2	74.0	39.0	127.3	30.24	428.88
116	2021-09	2.3	7.1	320	22.6	26.9	18.8	131.0	71.0	38.0	182.0	48.82	474.13
117	2021-10	2.1	10.6	290	15.6	20.5	11.6	57.0	70.0	25.0	168.4	48.38	353.62
118	2021-11	2.1	9.3	270	8.2	13.1	4.0	62.4	68.0	27.0	163.6	53.57	271.64
119	2021-12	2.3	8.1	290	0.6	5.1	-3.5	7.9	62.0	26.0	185.4	62.07	257.25

120 rows × 13 columns

Figure 2. the monthly factors data affecting the concentration of fine dust

The data was so well organized that we only had to grasp two things. The first is to determine if the data has null values, and the second is to merge the two data on a date basis. Fortunately, all the data didn't have null value, so I was able to move on right away.

```
df.isnull().sum()
```

```
일시              0
PM10            0
평균풍속(m/s)       0
최대풍속(m/s)       0
최대풍속풍향(deg)     0
평균기온(℃)         0
평균최고기온(℃)       0
평균최저기온(℃)       0
강수량(mm)          0
평균습도(%rh)        0
최저습도(%rh)        0
일조합(hr)          0
일조율(%)           0
일사합(MJ/m2)       0
dtype: int64
```

Figure 3. There is no null value

	일시	PM10	평균풍속(m/s)	최대풍속(m/s)	최대풍속풍향(deg)	평균기온(℃)	평균최고기온(℃)	평균최저기온(℃)	강수량(mm)	평균습도(%rh)	최저습도(%rh)	일조합(hr)	일조율(%)	일사합(MJ/m2)
0	2012-01	60.0	2.5	7.5	320	-2.8	1.3	-6.3	6.7	49.0	12.0	190.5	62.05	234.50
1	2012-02	50.0	2.9	8.1	290	-2.0	3.0	-6.0	0.8	43.0	11.0	224.9	71.74	325.25
2	2012-03	47.0	3.5	10.2	290	5.1	9.5	1.5	47.4	52.0	9.0	191.8	51.63	367.51
3	2012-04	51.0	3.4	12.0	290	12.3	17.9	7.8	157.0	54.0	9.0	212.5	53.72	448.48
4	2012-05	52.0	2.7	8.6	250	19.7	25.1	15.4	8.2	48.0	11.0	251.3	57.04	553.40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
115	2021-08	21.0	2.1	8.3	320	25.9	29.7	22.8	211.2	74.0	39.0	127.3	30.24	428.88
116	2021-09	15.0	2.3	7.1	320	22.6	26.9	18.8	131.0	71.0	38.0	182.0	48.82	474.13
117	2021-10	27.0	2.1	10.6	290	15.6	20.5	11.6	57.0	70.0	25.0	168.4	48.38	353.62
118	2021-11	46.0	2.1	9.3	270	8.2	13.1	4.0	62.4	68.0	27.0	163.6	53.57	271.64
119	2021-12	39.0	2.3	8.1	290	0.6	5.1	-3.5	7.9	62.0	26.0	185.4	62.07	257.25

120 rows × 14 columns

Figure 4. Final data with two data merged

- AI approach
  - Multiple linear regression
    - The difference between multiple linear regression and linear regression is that multiple linear regression has multiple independent variables. There are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The second advantage is the ability to identify outliers, or anomalies.

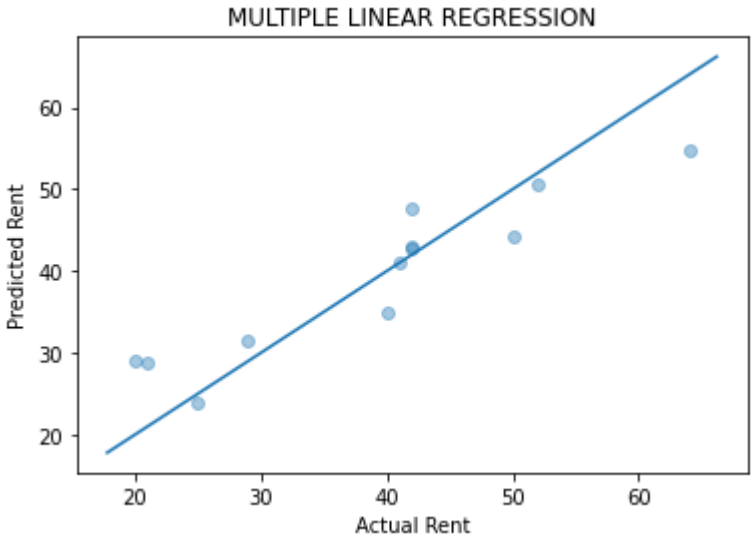

Figure 5. A graph of a multiple linear regression model that we learned with the data we prepared.

  - random forest
    - Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

- Evaluation criteria
  - Accuracy
    - Accuracy is the most intuitive indicator of the performance of the model.

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

Figure 7. Accuracy expression

- Results
  - First, the accuracy of multiple linear regression model is about 83 percent.

```
accuracy = mlr.score(x_test, y_test)
print("{:.2f}%".format(accuracy * 100))
```

```
82.89%
```

  - Second, the accuracy of the random forest model is about 50 percent.

```
In [11]: #기본적인 randomforest모형

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score # 정확도 함수

clf = RandomForestClassifier(n_estimators=50, max_depth=50,random_state=0)
clf.fit(train_x,train_y)

predict1 = clf.predict(test_x)
print(accuracy_score(test_y,predict1))
```

```
0.5
```

Figure 8. The accuracy of multiple linear regression model

- Conclusion
  - Multiple linear regression model's prediction accuracy is about 70 to 80 percent. so Linear regression techniques are more effective than random forest techniques.
  - Despite the small number of data, the use of multiple linear regression models enables some accurate prediction. If we secure a large amount of data, we will be able to increase the accuracy.
- Division of work
  - Kim Shin :
    - implementing random forest , Write a report
  - Seo Jueun
    - Research the background of the project, Creating presentation materials, Write a report
  - Lee Seungjae :
    - implementing data preprocessing, multiple linear regression, Write a report
  - Jung Yiju :
    - Research the data and background of project, Presentation, Write a report

- Discussion (per each member)
    - 22000690 Jung Yiju: The small size of the dataset was a big problem we could not solve for this project. We tried to use daily data because the monthly data was too small, but it was difficult to use because there were indiscriminately many null values. While doing the research for this project, I looked up various information about the existing fine dust prediction technology and I can see how it goes. Through this project, I found that several values are used to predict one value. In addition, I learned that in order to obtain accurate results, I had to be able to think about and prepare for many variables which have an influence on the results. And I can know the exact concept of Multi Linear Regression.

- Reference

    - https://hleecaster.com/ml-multiple-linear-regression-example/

      (Multiple Linear Regression)
    - https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do

      (the Meteorological Administration)
    - https://todayisbetterthanyesterday.tistory.com/51

      (random forest)
    - https://data.seoul.go.kr/dataList/OA-15526/S/1/datasetView.do

      (Seoul Air Pollution Measurement Information)