**Iteration 4 BDAS ---- PM 2.5 and Weather Conditions**

Name: Haocheng Liu

UID: 917384044

UPI: hliu8

**Index**

[1. Situation Understanding 1](#_Toc25984)

[1.1 Situation objectives 1](#_Toc3773)

[1.2 Assess the situation 1](#_Toc7444)

[1.3 Data mining objectives 1](#_Toc7705)

[1.4 Project plan 2](#_Toc30813)

[2. Data Understanding 2](#_Toc27975)

[2.1 Collect initial data 2](#_Toc28758)

[2.2 Data Description 3](#_Toc15093)

[3. Data Preparation 5](#_Toc13352)

[3.1 Data selection 5](#_Toc12)

[3.2 Data Cleaning 5](#_Toc28119)

[3.3 Data Construction 7](#_Toc2110)

[3.4 Data Integration 7](#_Toc17700)

[3.5 Format the data 7](#_Toc979)

[4. Data Transformation 8](#_Toc23189)

[4.1 Data Reduction 8](#_Toc1296)

[4.2 Data Projection 9](#_Toc975)

[5. Data Mining Method Selection 10](#_Toc11696)

[5.1 Objective discussion and matching 10](#_Toc7072)

[5.2 Data mining method selection 10](#_Toc13637)

[6. Data Mining Algorithm Selection 10](#_Toc2919)

[6.1 Conduct Exploratory analysis and discuss 10](#_Toc30587)

[6.2 Data-mining Algorithms Selection 10](#_Toc8388)

[6.3 Model Building and Parameters Choices 10](#_Toc29071)

[7. Data Mining 12](#_Toc6574)

[7.1 Logical Test Design 12](#_Toc4224)

[7.2 Data Mining Conduction 12](#_Toc25192)

[7.3 Pattern Search and Model Outputs 13](#_Toc32196)

[8. Interpretation 16](#_Toc7147)

[8.1 Mined pattern discussion 16](#_Toc18040)

[8.2 Data Visualization and Interpretation 16](#_Toc30390)

[8.3 Assessment and Evaluation 21](#_Toc22904)

[8.4 Iteration 22](#_Toc3377)

# Situation Understanding

## Situation objectives

The situation objective of this study is first to figure out the main weather causes of the PM2.5 air pollution in major cities without disturbing normal city development and then raise a reasonable method to decrease the influence of PM2.5 in those cities.

## Assess the situation

Air pollution is one of the main problems damaging citizens’ health in cities of China. PM2.5 which are fine particulars in the air whose diameters are equal or smaller than 2.5 micron do the most harm among the air pollution. They are harmful for the reason that they are too small to be prohibited by the defence mechanism of the respiratory system. They can easily enter bronchus and harm the lung. In addition, they are able to combine with other negative substances such as noxious chemicals, viruses and bacteria, which further endanger humans’ health.

The source of PM2.5 is mainly the human activities such as automobile exhaust, industry waste gas and electricity production which are necessary activities in modern society. Hence, it is relatively difficult to deal with the problem from the source. Nevertheless, it is estimated that weather conditions have influence on the concentration of PM2.5 and diverse places may have diverse influencing factors. The main constraint in the study is that many values may be absent because of the late start of PM2.5 monitoring in China. Meanwhile, Some weather disasters may form contingencies in the study.

**Data source**: The data sets used in this experiment are obtained from the data used for a competition on Kaggle. The content on Kaggle contain the PM2.5 data and corresponding weather condition data in five main cities in China. Three of them will be used in this experiment.

**Hardware resource**: Personal PC, Access to Amazon Web Service

**Software resource**: Jupyter Notebook, Putty, GitHub, Packets for PySpark

## Data mining objectives

Due to the harms caused by PM2.5, the practical objective of this report is to reduce the harm of PM2.5 in major cities of China without disturbing normal production activities. There are two ways to achieve the practical objective which are reducing PM2.5 concentration and raising the alarm before the rise of PM2.5 concentration. They are both related to nature conditions especially the weather. Therefore, the research objective is to discover the relationship between weather conditions and the harm level of PM2.5 from the related data sets and find the factor which has the most influence on PM2.5. It is estimated that the level of PM2.5 can be reduced through simulating the influence. In addition, it is possible that predicting the appearance of high level of PM2.5 through observing weather conditions.

## Project plan

This iteration is the experiment performed with PySpark and Amazon Web Service. In the situation understanding phase, the specific conditions of each selected city are carefully examined. Then, in the data understanding phase, the content of the data sets should be reviewed carefully and the features of the data are examined in order to prepare for data cleaning. In the data preparation phase and data transformation, the data should be cleaned through and some data types should be altered to appropriate forms. Meanwhile, features are deleted and constructed to fit the requirements of the experiments. In the method selection phase, the data mining method which fits the objective will be selected. Then, according to the selected method, several algorithms will be tried and their models will be tested. Finally, the optimal will be selected to perform data mining. In the data mining phase, the model will be run and the pattern will be introduced. In the interpretation phase, the results obtained will be complemented.

Table 1-4-1 Project Timetable

|  |  |
| --- | --- |
| Phases | Possible Using Time |
| Situation Understanding | 1 hour |
| Data Understanding | 2 hours |
| Data Preparation | 4 hours |
| Data Transformation | 1 hour |
| Data-mining Method Selection | 1 hour |
| Data-mining Algorithm Selection | 4 hour |
| Data Mining | 4 hour |
| Interpretation | 4 hour |

# Data Understanding

## Collect initial data

The historical PM2.5 data with corresponding weather information of major cities from 2010/01/01/00:00 to 2015/12/31/23:00 are selected as the initial data in this study. The selected Chinese cities in this study are Shenyang, Chengdu and Guangzhou which are typical in their position and geological environment which represents the cities in north, middle and south China. The data are collected in three excel files “ShenyangPM20100101\_20151231”, “ChengduPM20100101\_20151231” and “GuangzhouPM20100101\_20151231”. All the three files are downloaded from the Kaggle. However, the issue related to the initial data is that some PM2.5 data must be absent due to the reason that the monitoring of PM2.5 of diverse places began at a diverse time.

## Data Description

In this iteration, PySpark is applied to assist the data analysis. Similar to data analysis in Python with Pandas, the data are treated in a data frame. While the data frame of PySpark is different from the data frame of Pandas. PySpark provides “printSchema()” function to obtain the data type of each feature and “describe()” to obtain abstract description of the data. This section only displays the data description of Shenyang data set. All the three data sets are in a similar condition.

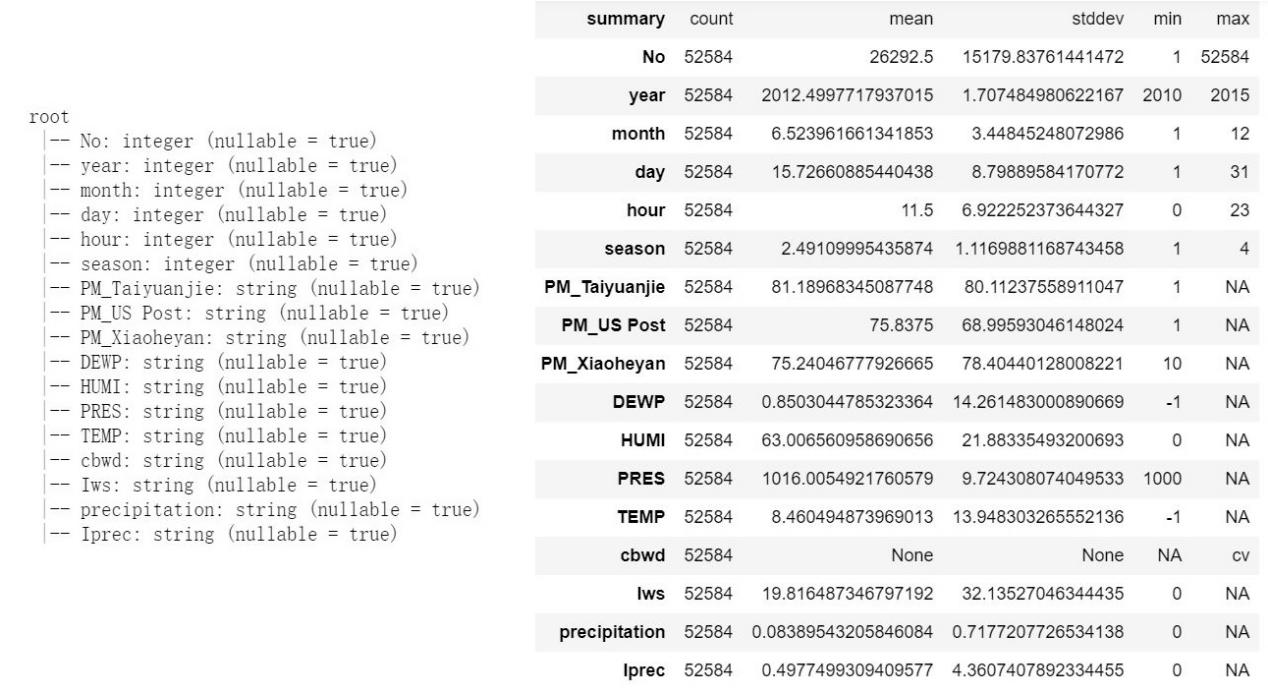


Figure 2-2-1 Schema and Abstract Description of Shenyang Data Set

Figure 2-2-1 displays the result of the two functions on the data set of Shenyang. It can be seen that the total number of instances in the data set of Shenyang is 52584 and the data set is consist of 17 features. There is a “No” feature to indicate ids and “year”, “month”, “day”, “hour”, “season” features to indicate time. Features describing weather conditions includes “DEWP” (dew point), “HUMI” (humidity), “PRES” (pressure), “TEMP” (temperature), “cbwd” (combined wind direction), “Iws” (cumulated wind speed), “precipitation” (hourly), and “Iprec” (cumulated precipitation). And Each data set has three kinds of PM2.5 features. In addition, the data type of many weather condition features whose data type should be numeric are string and many of the max value of features are “NA” which means “Not Available”, which greatly influence the data expression. Therefore, the first step of data cleaning should be “NA” elimination.

Table 2-2-1 Description of Features

|  |  |
| --- | --- |
| PM | PM2.5 concentration (ug/m^3) |
| DEWP | Dew Point (Celsius Degree) |
| HUMI | Humidity (%) |
| PRES | Pressure (Pa) |
| TEMP | Temperature (Celsius Degree) |
| cbwd | Combined wind direction |
| Iws | Cumulated wind speed (m/s) |
| precipitation | Hourly precipitation (mm) |
| Iprec | Cumulated precipitation (mm) |

* 1. **Data Exploration**

Due to the influence of “NA”, most weather conditions are treated with the incorrect data type which should be “double” instead of “string”. And the data with “string” type can not be used to construct graphs in Pyplot. In addition, only time features can be visualized, which is meaningless. Furthermore, the PySpark data frame restricts the visualization of “string” type data. The “string” type data will be lost during the transformation to Pandas data frame. Therefore, it is impossible to visualize the raw data before data cleaning and the correction of data type.

The data exploration can only be realized from abstract description, schema and content expression. Figure 2-3-1 displays the screenshot of the first five instances and the last five instances taken from the result of “toPandas()” function which has a better appearance than “show()” provided by PySpark. Observed from the figure 2-3-1, there are three conclusions can be achieved. First, the PM\_US Post is more average and can be selected as the representative. Second, the wind direction is related to the season. Third, “DEWP” follows “TEMP”.



Figure 2-3-1 Part of Instances in Shenyang Data Set

* 1. **Data Quality Verification**

Figure 2-2-1 shows the schema and the basic information of the data set and Figure 2-3-1 shows the part of the draft instances. These two figures indicate the nature of the data set of Shenyang. It is clear that there are many values absent which are represented as “NA” especially the PM2.5 records and it is essential to deal with them before further processing. In addition, due to the influence of “NA”, the data type of weather condition features are incorrect, which requires correction after the elimination of “NA” instances. All the three data sets are in the similar manner.

# Data Preparation

## Data selection

Due to the reason that not all the attributes in tables are useful in determining the final result, it is essential to select essential features and delete useless parts which might be obvious. Through observing the similarities of three tables, all the three tables have a kind of PM2.5 value column named “PM\_US Post” whose values are representative. Meanwhile, “PM\_US Post” records more data than other types of PM.5 records and their differences are not huge and acceptable. In order to simplify the structure and increase the universality of PM2.5 value in the city, only “PM\_US Post” is selected as the study object in this study.

In addition, according to the objectives of this study, features demonstrating information about No, year, month, day, hour and season should be removed because they are not considered as weather factors in influencing PM2.5 concentration in this study. And the effect of season can be represented by pressure and temperature features.

The “drop()” function is applied to remove previously mentioned columns.

## Data Cleaning

As mentioned in the previous step, the instances with “NA” values are required to be removed. Figure 3-2-1 shows the code to realize the procedure through the “filter()” and Figure 3-2-2 shows the description of the data set after the elimination of “NA”. However, it can be seen that the max value of “PM\_US Post” is 99 which is different from the real value. This issue results from the incorrect data type. Figure 3-2-3 shows the code to correct the data type of weather condition features manually by “cast()” and “withColumn()” and Figure 3-2-4 displays the effect. All the three data sets are handled in the same manner.



Figure 3-2-1 Code to eliminate “NA”

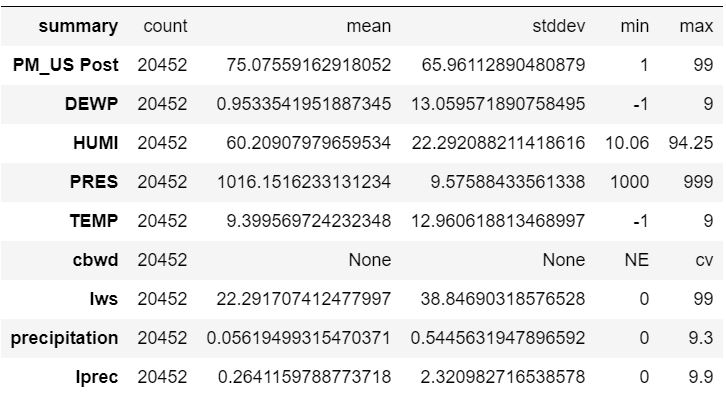


Figure 3-2-2 Description of Shenyang Data Set after the Elimination of “NA”

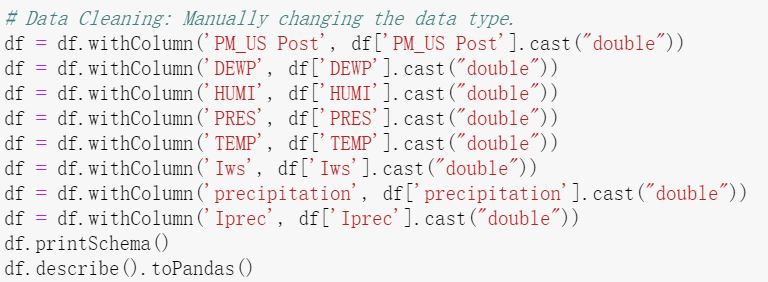


Figure 3-2-3 Code to correct the Data Type

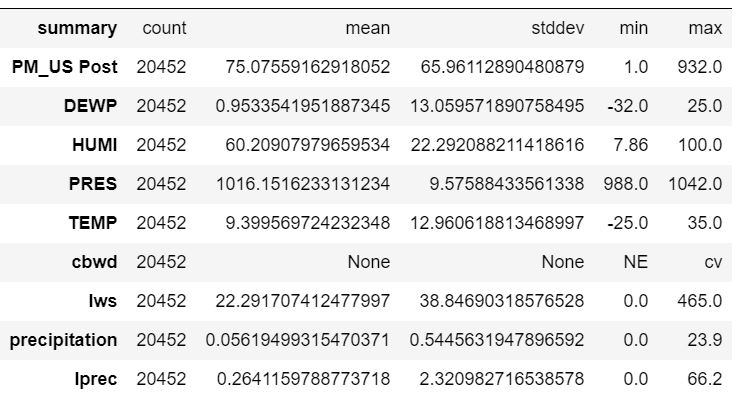


Figure 3-2-4 Description of Shenyang Data Set after the Data Type Correction

In addition, after the correction of data type, some abnormal values have been found. In Shenyang data set, there is a “DEWP” whose value is -97 which is far away from its neighbors. And in Guangzhou data set, there exists “DEWP” and “HUMI” whose value is -9999. These values must be abnormal values and should be deleted.

## Data Construction

In order to make the harm of PM2.5 concentration more clear, a new attribute describing the harm level of PM2.5 due to its concentration is constructed. The new attribute is called “Harm”. If “PM\_US Post” is less than 75, “Harm” will be zero which represents that air quality is relatively good; If “PM\_US Post” is equal or greater than 75 but less than 150, “Harm” will be one which means that the pollution may influence daily life; If “PM\_US Post” is greater than 150, “Harm” will be two which represents serious air pollution situation and people should take actions to protect themselves. This criteria is according to Chinese criteria towards PM2.5 in major cities. All the three tables are required to construct the “Harm” attribute in the same manner.

The construction of the new feature is realized by “udf”(User Defined Function) provided by PySpark and “withColumn()” function.(Figure 3-3-1)

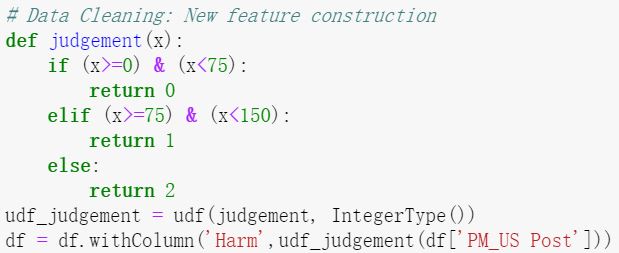


Figure 3-3-1 Construction of “Harm”

## Data Integration

Due to the difference in locations of sample cities, it is better to analyze tables independently. For the reason that weather data may interfere with each other if they are combined and analyzed together. Diverse cities may have diverse main cause of increasing the PM2.5 level. Therefore, the integration of the three data sets is abandoned.

## 3.5 Format the data

The structures of the three data sets are same. On the other hand, data sets have already been formatted through previous cleaning procedures which coerce outliners and do something to deal with undesired extremes. However, in order to check the correlation of weather condition features to the target feature “Harm”, it is essential to replace the value of “cbwd” with integer numbers for the reason that the correlation check function only supports numeric values. Figure 3-5-1 shows the code of the replacement. A function called “cbwd\_vc” has been created to change those value to numbers that “cv” is 0, “SW” is 1, “SE” is 2, “NE” is 3, and “NW” is 4. And the udf function is used to realize the application in PySpark.

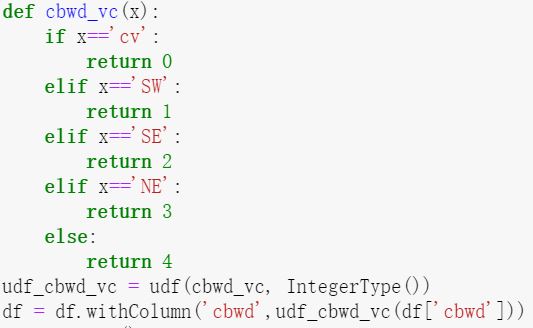


Figure 3-5-1 Value Replacement of “cbwd”

# Data Transformation

## Data Reduction

In this section, further reduction of features is required to be performed. For the reason that we just pay attention to the harm level which is “Harm”, “PM\_US Post” which is the source of the “Harm” should not be considered any more and ought to be removed.

In addition, PySpark has provided a “corr()” function to check the correlation of the features with the target feature which is “Harm” in this study. Hence, this function is applied to assist the reduction. Figure 4-1-1 displays the correlation of each weather condition to the target feature “Harm” in Shenyang data set. It can be seen that “Precipitation” and “Iprec” have the lowest correlation in this case. After checking the data of Chengdu and Guangzhou, it has been found that the correlation rates of “Precipitation” and “Iprec” are the lowest in all the three data set. Therefore, it is determined to remove the two features.

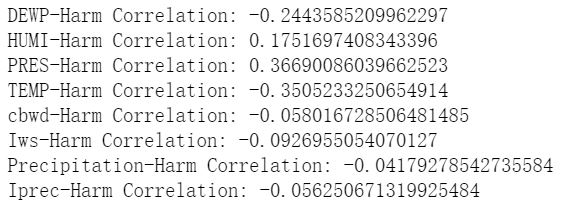


Figure 4-1-1 Correlation Rate Display in Shenyang Data Set

Figure 4-1-2 shows the code to realize the procedures in this section.

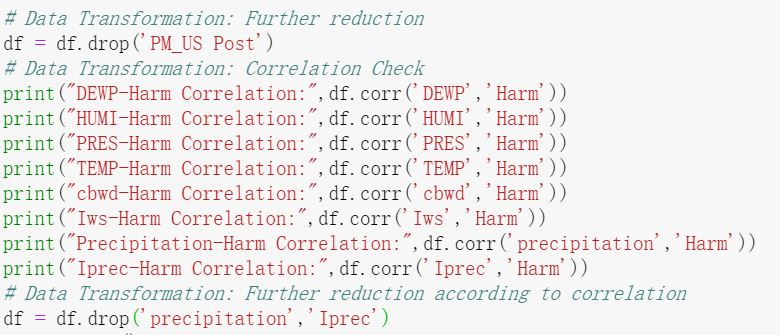


Figure 4-1-2 Code of Data Reduction

## Data Projection

In order to check the distribution status of the data, the data frame of PySpark is transformed to Pandas data frame in order to apply the “value\_counts()” function. Figure 4-2-1 shows the distribution of the target “Harm” in the three data sets. It can be seen that the distribution is unbalanced. So it is assumed that the accuracy of the result will be increased after the balancing of the data. The method taken to balance the data is oversampling the instances whose “Harm” is 1 or 2. The accuracy of the results generated by unbalanced data and balanced data should be compared in the step 7. Figure 4-2-2 shows the code to balance the data in Shenyang data set. The value of “frac” parameter is obtained through the ratio of the number of instances whose “Harm” is 0 to the number of instances whose “Harm” is 1 or 2.

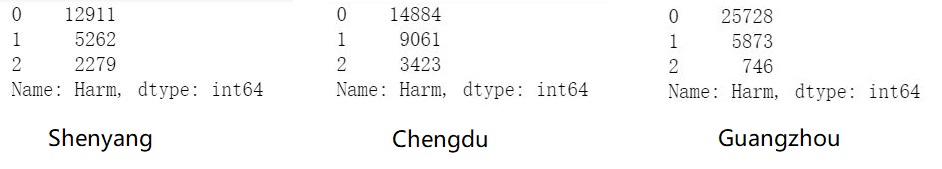


Figure 4-2-1 Distribution of Data in the three Data Sets

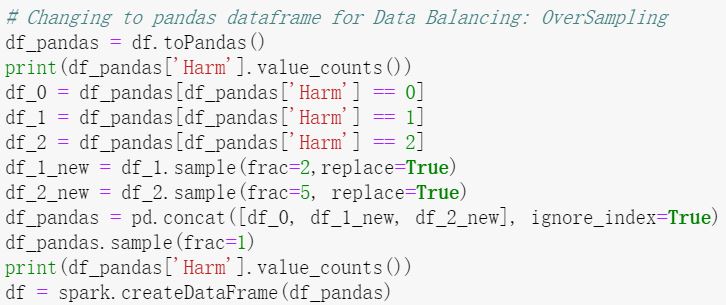


Figure 4-2-2 Code to balance the Shenyang Data Set

# Data Mining Method Selection

## Objective discussion and matching

The practical objective is to reduce the harm of PM2.5 in major cities of China without disturbing normal production activities. Meanwhile, the research objective is to discover the relationship between weather conditions and the harm level PM2.5 and find the main cause. In order to show the harm of PM2.5 intuitively, the concentration of PM2.5 is separated into three levels, which means that it is measured as nominal with three values. Due to this nature mentioned above, the classification method category is determined to be used.

## Data mining method selection

According to previous discussion, the classification method should be applied.

# Data Mining Algorithm Selection

## Conduct Exploratory analysis and discuss

According to the step five, the classification method has been selected as the data mining method in this study. Due to the classification method, PySpark has provided several classification algorithms such as Logical Regression, Decision tree, Random forest, [Gradient-boosted tree classifier](https://spark.apache.org/docs/2.1.0/ml-classification-regression.html" \l "gradient-boosted-tree-classifier), [Multilayer perceptron](https://spark.apache.org/docs/2.1.0/ml-classification-regression.html" \l "multilayer-perceptron-classifier), [One-vs-Rest](https://spark.apache.org/docs/2.1.0/ml-classification-regression.html" \l "one-vs-rest-classifier-aka-one-vs-all), and [Naive Bayes](https://spark.apache.org/docs/2.1.0/ml-classification-regression.html" \l "naive-bayes).

Due to the reason that the target feature “Harm” has three possible values, the classification is multiple class classification. Therefore, Logic Regression and One-vs-Rest is abandoned that its performance is not acceptable in multiple class classification. Multilayer perception is a kind of neural network algorithm and it will take relatively long time. In addition, according to the objective, the importance of weather condition features to the target feature is desired to be found. Hence, the model constructed can output feature importance is preferred. GBT classifier is the ensemble of decision trees, which takes much longer time than random forest.

## Data-mining Algorithms Selection

According to the previous statement, the random forest classifier and the decision tree classifier are selected as the algorithms to perform the data mining.

## Model Building and Parameters Choices

Although only two algorithms are chosen to perform the data mining, it is a problem to configure the parameters and conditions for each algorithm. Two steps are constructed to form the procedures.

The first step is parameter selection comparison. In this step, both decision tree classifier and random forest classifier are tested with different parameters. For decision tree classifier, all the parameters can be set to default. For random forest classifier, the “numTrees” will be set to 10, 20, 30 to test the performance under different number of trees because 20 is the default value and it is essential to find the influence of it on the result. In addition, the “maxDepth” parameter is set to 6. In this step, the data are not balanced by the method mentioned above.

The first step will generate the algorithm with possible optimal parameters. The second step is to compare the result of models with data balancing and without data balancing in order to see whether the data balancing can work.

Finally, after the comparison of the accuracy, the optimal one will be selected and further mining will be based on it. Due to the reason that the natures of the three data sets are similar, the selected model from one data set should be acceptable by other two data sets.



Figure 6-3-1 Construction of Models

Figure 6-3-1 shows the procedure to build the model in decision tree classifier and random forest classifier. It can be seen that the category feature “cbwd” has been transformed to the vector by “OneHotEncoder” in order to make it suitable for data mining. All the features used to predict are gathered into the assembler. The target feature “Harm” also has been transformed by “StringIndexer” and given the title called “label”. And all the factors required in building the model are assembled in the pipeline. Then, the PySpark data frame fits to the pipeline and a new data frame is constructed. Due to the reason that the features required have been gathered in the “feature”, only the “feature” is required to be remained. The “label” and “feature” will be used in the algorithms.

# Data Mining

## Logical Test Design

In order to check the acceptance of the model, the proportion of training and testing data should be determined. The accuracy is used to check the correctness of the model on the test data set. In order to obtain a higher accuracy rate, 80 percent of the data is used as the training data and 20 percent of the data is used to test the model, which is showed in Figure 6-3-1.

As mentioned above, the accuracy rate is selected as the criterion to examine the quality of the model. Figure 6-3-1 shows that the result of the model on the test data is stored in “predictions”. Figure 7-1-1 shows the code to compare the value predicted and the real value and generate the accuracy rate. The code can be used on both classifiers.

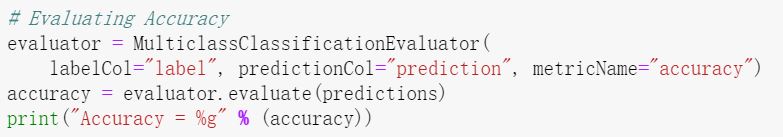


Figure 7-1-1 Code of Accuracy Evaluation

## Data Mining Conduction

As the plan states in the step 6, the first step is to check the models with diverse parameters on unbalanced data. This process is conducted on the Shenyang data set. First, the decision tree classifier using “gini” and “entropy” is examined. Figure 7-2-1 displays the result of them. It can be seen that the decision tree algorithm applying “gini” has a higher accuracy. Hence, “gini” is chosen for further processing.



Figure 7-2-1 Accuracy of Decision Tree Classifier

Second, the random forest classifiers with different “numTrees” parameter are examined. Figure 7-2-2 shows the results with “numTrees” whose values are 10, 20, and 30. It is obvious that when the “numTrees” is set to 30, it has the highest accuracy rate. Therefore, the random forest classifier whose “numTrees” is 30 is selected for further processing.

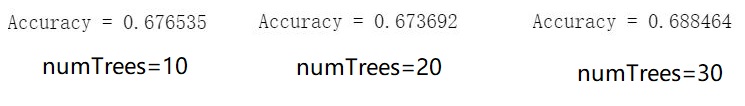


Figure 7-2-2 Accuracy of Random Forest Classifier

The second step is to compare the result of the classifier obtained in the first step with balanced and unbalanced data. Figure 7-2-3 displays the results of the four model on the test data. Two of them are retrieved from the first step in which data are unbalanced. The other two of them are processed on balanced data. It can be seen that the model constructed by random forest classifier with “numTrees=30” on unbalanced data has the highest accuracy rate. The result displayed is interesting that the model trained on unbalanced data has higher accuracy than the model trained on balanced data. And the process to balance data is exactly effective in iteration 3 and increase the accuracy by 20 percent. This situation is probably because the same algorithm used in PySpark and Python may have diverse internal working mechanism. Therefore, the model constructed by random forest with “numTrees=30” on unbalanced data is selected to assist the data mining.

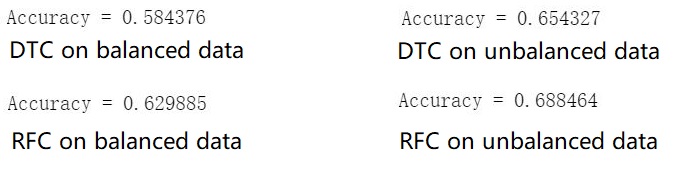


Figure 7-2-3 Accuracy of Balanced and Unbalanced Data Model

Due to the similarity of the nature of the data in the three data sets, also can be selected as the optimal algorithm for Chengdu and Guangzhou data sets. Figure 7-2-4 displays the prediction accuracy rate for Shenyang, Chengdu and Guangzhou.

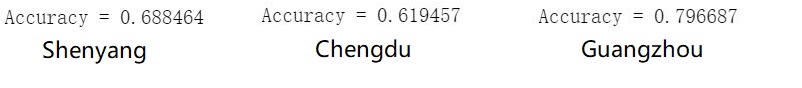


Figure 7-2-4 Accuracy of the RFC on Three Cities

## Pattern Search and Model Outputs

The objective is to figure out the relationship between weather conditions and the harmful level of PM2.5. Hence, it is critical to make the importance of each weather condition features towards the level of harm clear. The model constructed by random forest classifier has the function called “featureImportances” which generates the importance of each feature in a vector. Figure 7-3-1 shows the output of the “featureImportances” of Shenyang. The order of the features depends on the order of the features in the “assembler”. It can be seen that there are 9 features, while only 6 features are used to determined the target feature. This situation is because the application of “OneHotEncoder” which expresses the category type feature as four binary features which are included in the “cbwdVec” and occupy index 3 to 6.

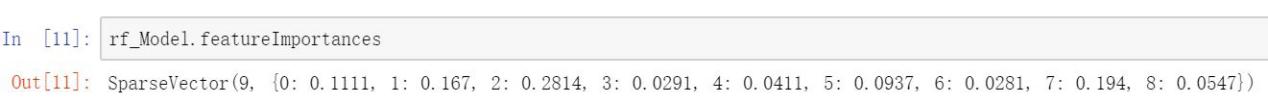


Figure 7-3-1 Output of Feature Importance

In order to have a better understanding, Visualization of the feature importance is necessary. Figure 7-3-2 displays code to visualize the feature importance through pyplot.

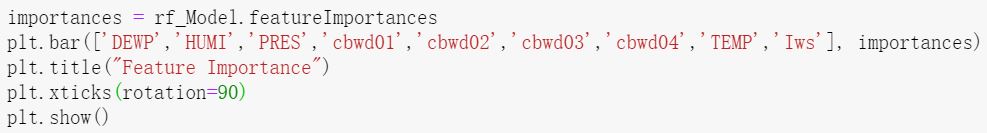


Figure 7-3-2 Code of Feature Importance Visualization

Figure 7-3-3 shows the importance of features in Shenyang. It can be seen that the pressure occupies the heaviest weight. The importance of the temperature and the humidity features is also relatively high.

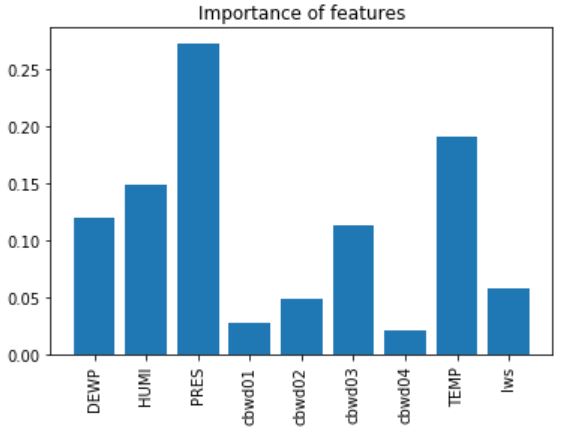


Figure 7-3-3 Feature Importance in Shenyang

Figure 7-3-4 shows the importance of features in the model of Chengdu. The temperature owns a extremely high importance. Then is the dew point whose value follows the trend of the temperature. Pressure, wind speed and humidity also have influence in affecting the harm level of PM2.5.

Figure 7-3-5 displays the importance of features in the model of Guangzhou. It can be observed that pressure, temperature, the dew point occupy the first three places of importance. While the temperature and the dew point have been observed a direct ratio relationship through data exploration. Hence, the importance of the wind speed and the humidity can not be ignored.

Due to the content mentioned above, the pattern can be concluded. In north China of which Shenyang is the representative city, the pressure, temperature and humidity are three main factors determining the harm level of PM2.5. In middle China of which Chengdu is the representative city, the temperature affects the harm level of PM2.5 a lot. Humidity, wind speed and pressure may have weak influence. In south China of which Guangzhou is the representative city, the pressure, temperature and wind speed may greatly affect the harm level of PM2.5. Meanwhile, the humidity may also have some influence on the harm level of PM2.5.

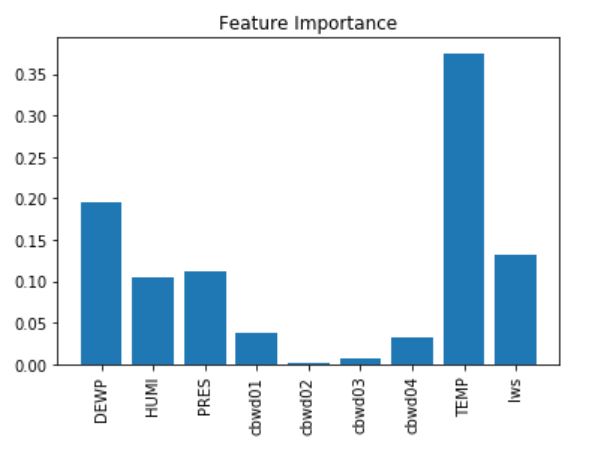


Figure 7-3-4 Feature Importance in Chengdu

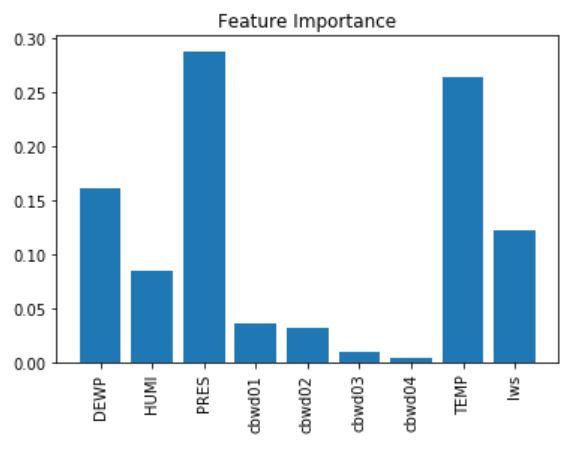


Figure 7-3-5 Feature Importance in Guangzhou

# Interpretation

## Mined pattern discussion

Towards the data used, the model based on unbalanced data has better performance than the model with balanced data, which is contrary to the hypothesis. This situation may be contributed by the difference of the internal working mechanism of the algorithm.

It can be seen that the accuracy of the models is not ideal. The highest one can not exceed 80%, which is different from applying SPSS and Python sklearn with the same procedures. There are three possible reasons. First is the difference in internal working mechanism. Second, the parameter selection can be improved such as the number of trees. Third, other data balancing method such as SMOTE may increase the accuracy of the prediction. Among the results of the three data sets, the model of Guangzhou has the highest accuracy rate and the model of Chengdu has a relatively low accuracy rate.

Observed from the mined pattern, temperature occupies a relatively important status in affecting the harmful level of PM2.5. The trend of it requires to be figured out. In addition, it is essential to find out how the combination of other weather conditions with relatively high importance affects the harmful level of PM2.5 and whether there are rules to follow for different cities.

The above issues will be solved through further visualizing and interpreting data. Due to the reason that cities in diverse places may have diverse situations, the visualization and interpretation will also be separated into three parts.

## Data Visualization and Interpretation

* + 1. Shenyang Data

Figure 8-2-1-1 shows the diagram of pressure and the harmful level of PM2.5 in Shenyang. Despite the generalization of the pressure, there is a rule which can be followed, which is displayed in the red outline. If the pressure is under 992Pa, the harmful level of PM2.5 will always be low. The possibility of higher level of PM2.5 harm will increase with the increase of the pressure. Therefore, pressure can be the first criterion in the tree and has the highest importance.

Figure 8-2-1-2 shows a scatter figure describing the relationship among humidity, temperature and the harmful level of PM2.5. The larger the circle, the higher the level of harm of PM2.5. The red line encloses the area with the highest density of large circles. It is clear that the occurrence of high level PM2.5 goes up when the humidity goes up and the temperature drops, especially when the humidity is over 50% and the temperature is under 10 centigrade.

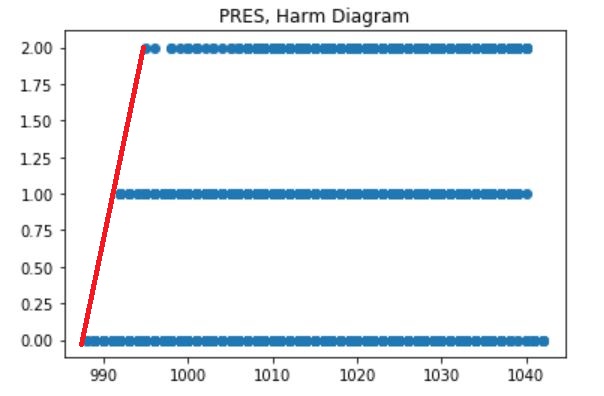


Figure 8-2-1-1 Pressure and Harm in Shenyang

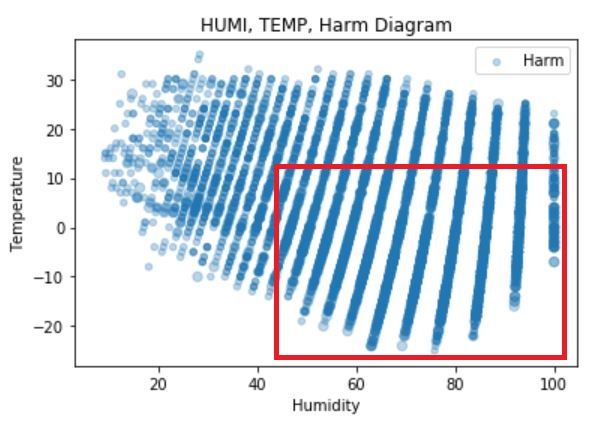


Figure 8-2-1-2 Humidity, Temperature and Harm in Shenyang

Due to the content discussed above, pressure plays an important roles in determining the harmful level of PM2.5 and it can be used to predict possible high level PM2.5 situation. On the other hand, when temperature is low and the humidity is high, the probability of occurring the high level PM2.5 situation is also high. Therefore, with high atmosphere pressure, low temperature and high humidity, it is essential to pay attention to the PM2.5 level and make the protection in advance in north China.

* + 1. Chengdu Data

Figure 8-2-2-1 shows the diagram of pressure and the harmful level of PM2.5 in Chengdu. It can be seen some rules in it. If the pressure is smaller than 992 Pa, the level of harm is always 1. However, the rule generated by pressure does not show a clear manner in Chengdu that there is a huge overlap among the data, which is outlined by the red line in the figure.

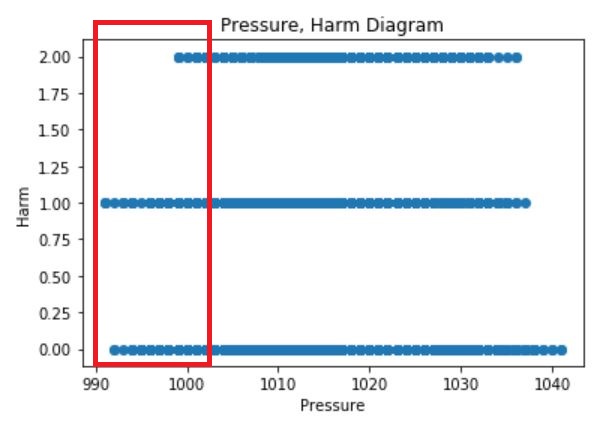


Figure 8-2-2-1 Pressure and Harm in Chengdu

Figure 8-2-2-2 displays the diagram of the wind speed and the harmful level of PM2.5 in Chengdu. The trend which is similar to the result of Shenyang can be discovered that the occurrence of high level of PM2.5 decreases with the increase of the wind speed. However, there are more noisy samples which are above the red line in the figure. Hence, the rule might be interfered by those samples.

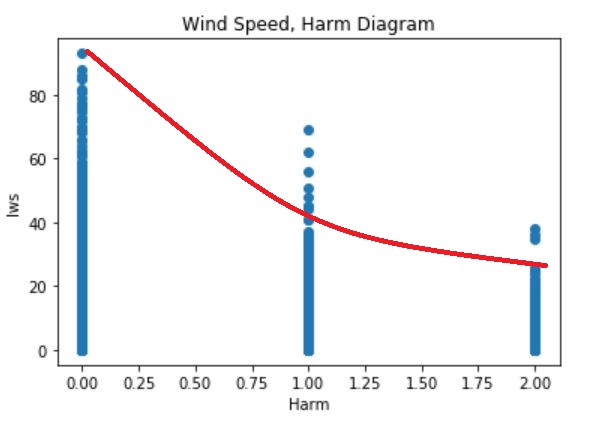


Figure 8-2-2-2 Wind Speed, Harm in Chengdu

Figure 8-2-2-3 displays the diagram among humidity, temperature and the harmful level of PM2.5 in Chengdu. The area between the two red line are the place with highest density of harm nodes with high harmful level. It can be seen that the occurrence of high harmful level of PM2.5 situation ascends with the increase of the humidity and enlarge the temperature period in which the harmful level of PM2.5 is high. In addition, when the humidity is fixed, high harmful level of PM2.5 will happen with lower temperature.

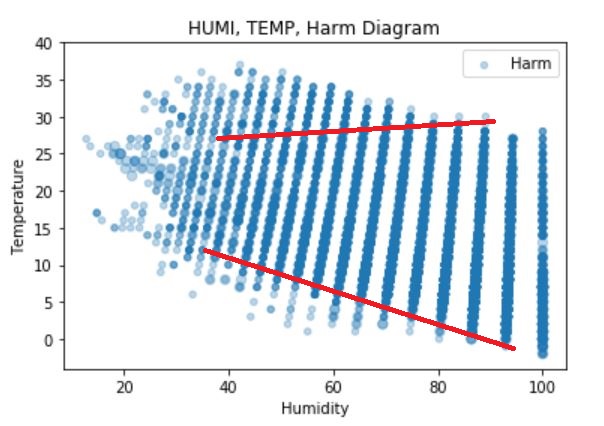


Figure 8-2-2-3 Humidity, Temperature and Harm in Chengdu

Due to the content mentioned above, the rule of the air pressure in Chengdu is not as clear as Shenyang, but it displays a similar trend. Meanwhile, the wind speed have a relatively strong effect on the harmful level of PM2.5. The higher the wind speed, the lower the possibility of the occurrence of the high PM2.5 harmful level. In addition, the temperature and humidity displays a bit different distribution but with the similar trend that low temperature and high humidity will cause PM2.5 situation with a high level. Therefore, the accuracy of the model in Chengdu is relatively not high. There might be other factors such as geography influencing the level of PM2.5 a lot.

* + 1. Guangzhou Data

Figure 8-2-3-1 shows the diagram of the pressure, the harmful level of PM2.5 in Guangzhou. It can be seen that the rule is much clearer that the pressure can separate the level of harm of PM2.5 when the atmosphere pressure is relatively low. When the pressure is under about 985 Pa, the harmful level of PM2.5 is exactly acceptable and the concentration of it is under 75.

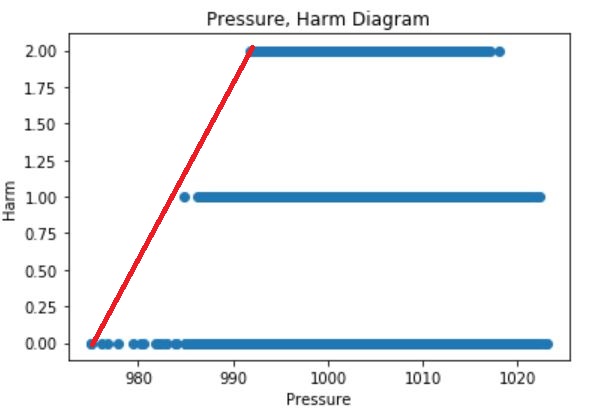


Figure 8-2-3-1 Pressure, Harm in Guangzhou

Figure 8-2-3-2 displays the diagram of the wind speed and the harmful level of PM2.5 in Guangzhou. In Guangzhou, the wind speed shows a great and clear effect on influencing the harmful level of PM2.5. When the cumulated wind speed is more than 100m/s, PM2.5 will not have any harmful influence. Nevertheless, when the cumulated wind speed is lower than 100m/s, its influence on PM2.5 is much weaker.

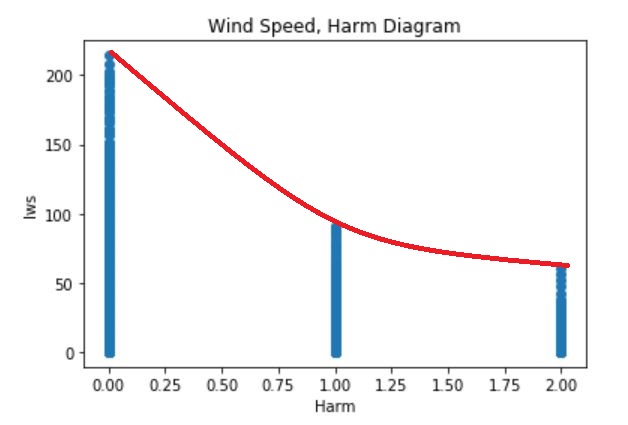


Figure 8-2-3-2 Wind Speed, Harm in Guangzhou

Figure 8-2-3-3 shows the diagram of humidity, temperature and the harmful level of PM2.5 in Guangzhou. It can be seen that the distribution of “Harm” nodes in Guangzhou diagram is different from the other cities. However, it follows the same trend that the occurrence of high level of PM2.5 situation increases with the increase of the humidity. What’s different is the influence of temperature. In Guangzhou, the high level of PM2.5 situation occurs between 15 and 30 centigrade. Meanwhile, the influence of humidity is stronger than temperature, which is observed from this diagram.

Due to the content mentioned above, the diagrams given by the Guangzhou data set is very clear, which might be the reason for the highest accuracy of the model displayed in step 7. It can conclude that when there is no wind with high atmosphere pressure and the humidity is high, it is essential to warn the citizens in south China to protect themselves from the PM2.5 pollution.

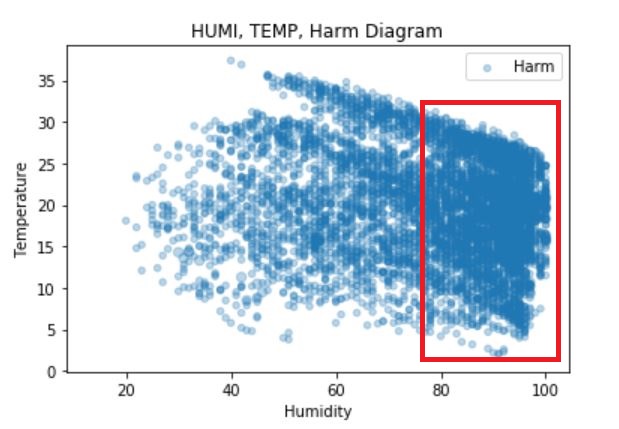


Figure 8-2-3-4 Humidity, Temperature and Harm in Guangzhou

* + 1. Conclusion

Through analyzing the change of atmosphere pressure, the wind speed, temperature and humidity, it is possible to predict the increase of the harmful level of PM2.5. However, there is no method to deal with the weather conditions to prevent the harm brought by PM2.5. Meanwhile, simulating low atmosphere and high wind speed indoor is also not possible due to forming a comfortable living environment. Nevertheless, It is found that humidity plays a relatively important role in all areas of China. Therefore, this idea can be applied that reducing the indoor humidity to reduce the influence of PM2.5 through simple methods such as applying dehumidifiers.

## Assessment and Evaluation

The evaluation of the models are realized through “MulticlassClassificationEvaluator” provided by PySpark due to the reason that the target feature has three possible values. And the accuracy which can mostly display the performance of the model has been chosen.

As displayed in Figure 7-2-4, the accuracy of the selected model on the test sets is not ideal. Three possible reasons have been concluded. First is the difference in internal working mechanism, which leads to the lower accuracy rate than SPSS and Python sklearn. Second, the parameter selection can be improved such as the number of trees. Third, other data balancing method such as SMOTE may increase the accuracy of the prediction. Among the results of the three data sets, the model of Guangzhou has the highest accuracy rate and the model of Chengdu has a relatively low accuracy rate.

Nevertheless, the pattern found is still meaningful. The way that weather conditions influence the harmful level of PM2.5 has been discovered roughly. An idea to avoid the harm brought by the PM2.5 indoor has also been given. Therefore, the objective of this study has been achieved. Although the model can not predict the harmful level of PM2.5 precisely.

## Iteration

There are several iterations taking place in the experiment. The first iteration is taken place in the algorithm selection step in order to obtain the optimal configuration. For decision tree classifier, the “impurity” has been set to “gini” and “entropy” and evaluate the results. For random forest, different “numTrees” are tested. It has been found that the more trees have been constructed the more accurate the model will be. Nevertheless, when “numTrees” reach a certain level, the increase is slight. Meanwhile, different “maxDepth” parameters have been tested. The test result is that the deeper the model the more accurate it will be. However, bigger “maxDepth” value means longer computing time, which leads to service block in AWS. Therefore, the “maxDepth” is finally set to 5.

The second iteration takes place to test the result of model on balanced data and unbalanced data in order to improve the accuracy obtained. Nonetheless, it is amazing that the balancing of data does not work. This issue may be corresponding to the internal working mechanism of the algorithm provided by the PySpark.

The third iteration should be trying other classifiers such as GBT classifier and multilayer perception classifier which requires long computing time. GBT classifier is the ensemble of decision trees and multilayer perception classifier is a kind of neural network algorithm. They may achieve a better accuracy than decision tree classifier and random forest classifier.

Due to the content mentioned above, many iterations are required such as parameter selection, data balancing method and algorithm selection in order to achieve an acceptable result.