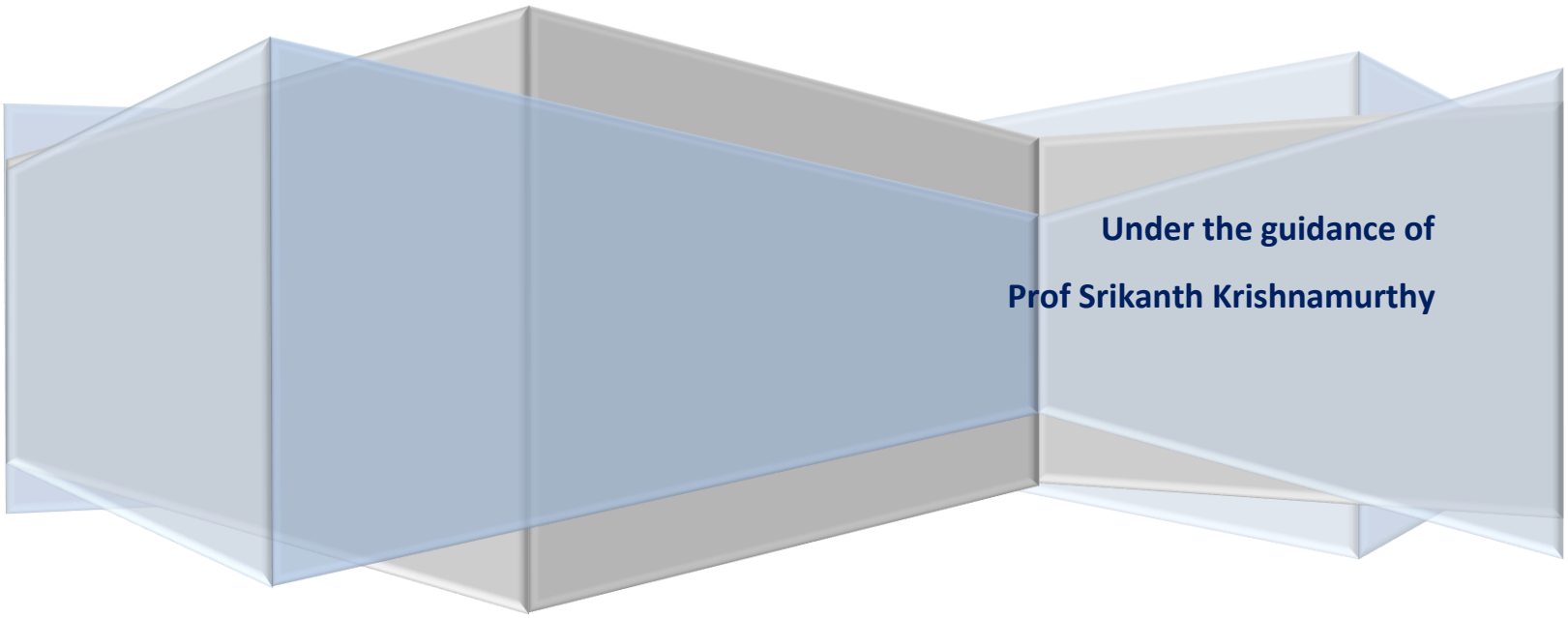
# Pollution Data Analysis and Machine learning Prediction

Team7: Snigdha Joshi & Vipra Shah



Under the guidance of

Prof Srikanth Krishnamurthy

## Part A

**Data Pipelining:**

Data pipelining is executed using luigi

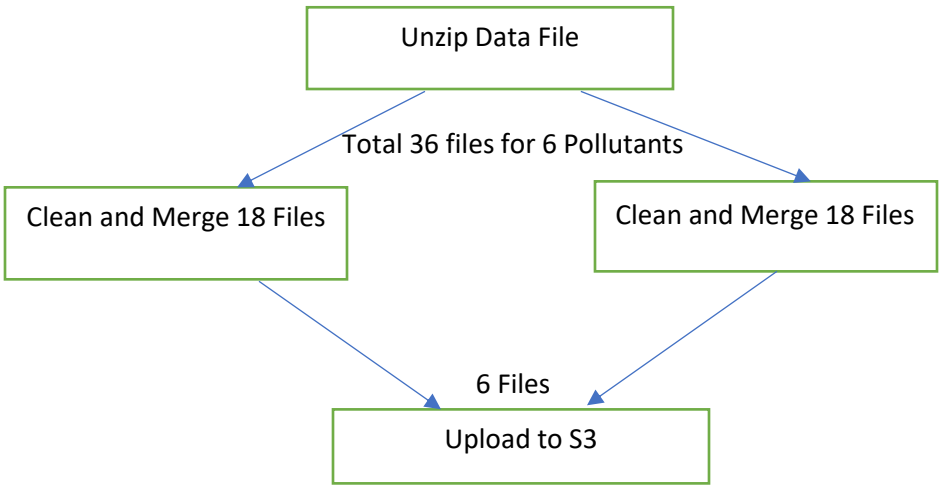
1. Unzip RawData.zip file(6 files for 6 pollutants)
2. Clean and Merge(Into 6 files)
3. Upload to S3 bucket

Run Following Command:

Docker pull joshisn/finalproject1:final

docker run -ti joshisn/finalproject1:final --module UploadData UploadAll --local-scheduler --id <AWS_access_id>  --key <AWS_secrate_key>

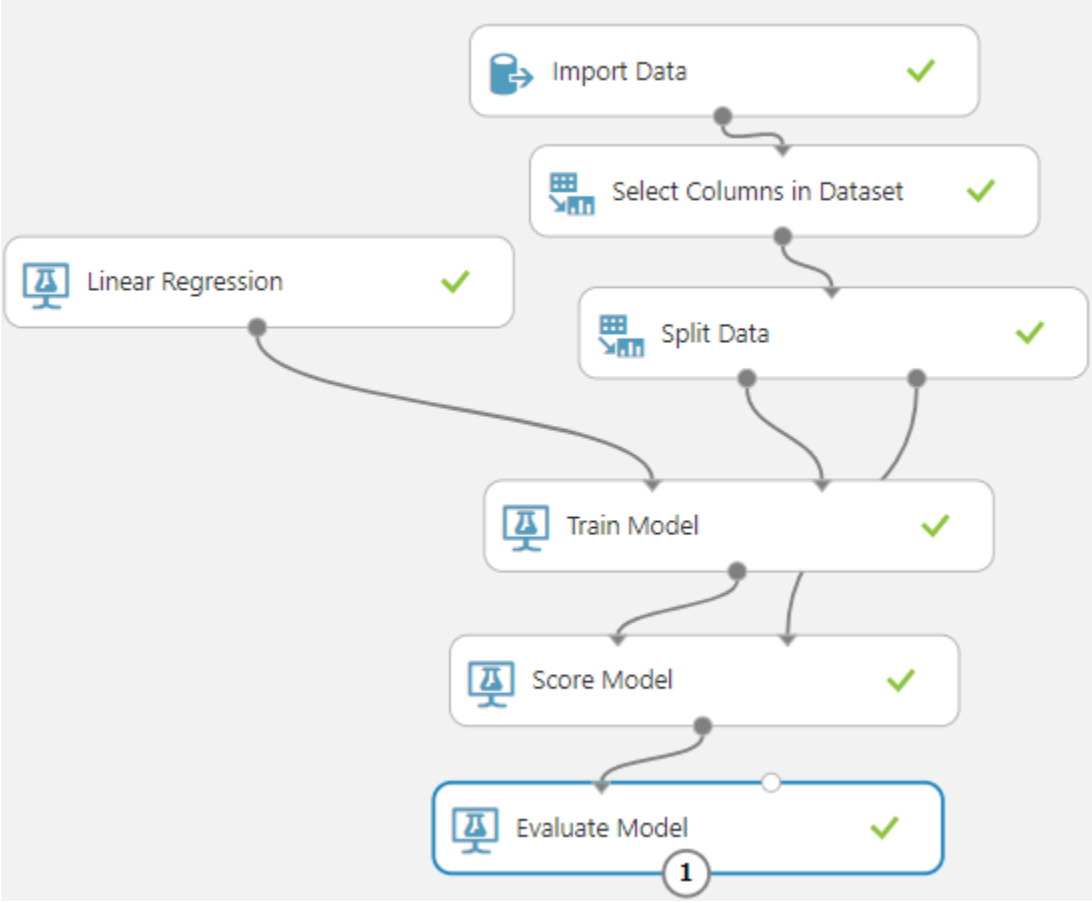
```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/tasks (master)
$ docker run -ti joshisn/finalproject1:final --module UploadData UploadAll --local-scheduler --id AKIAJ5D7X6D6F335AOQQ
--key PyYc9byj7w3TtiP+FWUXR/GTJkTNIv2R2zpDWJIC
DEBUG: Checking if UploadAll(id=AKIAJ5D7X6D6F335AOQQ, key=PyYc9byj7w3TtiP+FWUXR/GTJkTNIv2R2zpDWJIC) is complete
All tasks comlepted!
DEBUG: Checking if cleanAll() is complete
INFO: Informed scheduler that task   UploadAll_AKIAJ5D7X6D6F335_PyYc9byj7w3TtiP__aeecef1c0b   has status   PENDING
DEBUG: Checking if extractzip() is complete
All cleaning comlepted!
INFO: Informed scheduler that task   cleanAll__99914b932b   has status   PENDING
INFO: Informed scheduler that task   extractzip__99914b932b   has status   PENDING
INFO: Done scheduling tasks
INFO: Running Worker with 1 processes
DEBUG: Asking scheduler for work...
```




## Part B

**Data Models and Feature Selection using Azure:**

1. **Linear Regression:**




Metrics

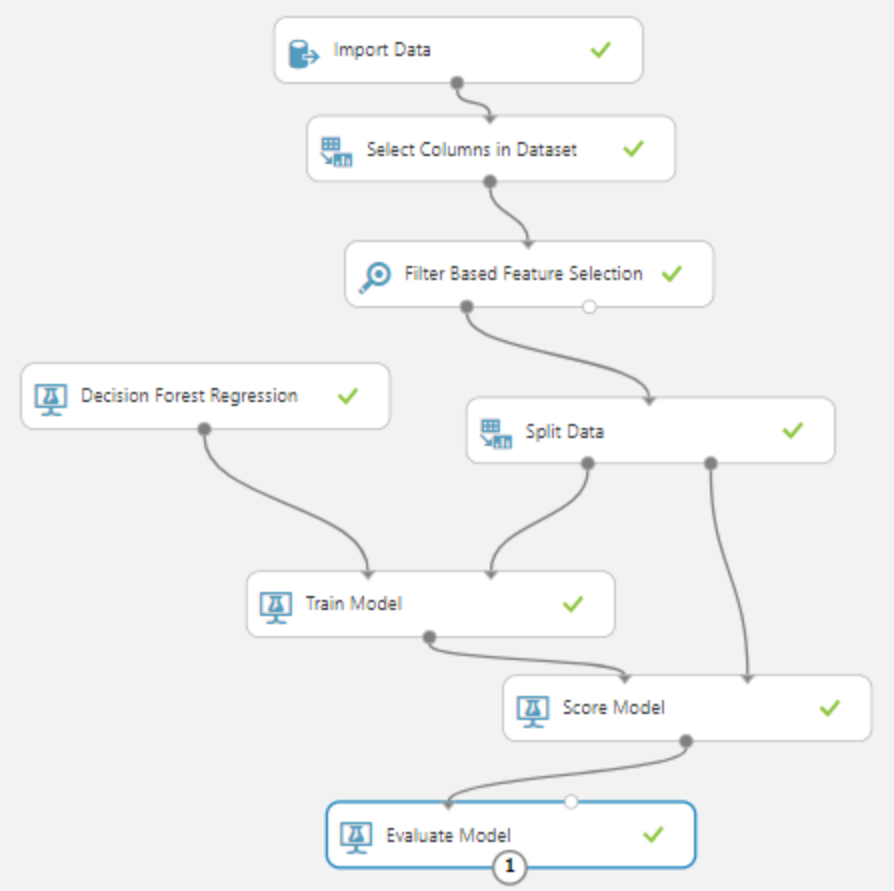
| | |
|---|---|
| Mean Absolute Error | 4.39086 |
| Root Mean Squared Error | 6.746139 |
| Relative Absolute Error | 0.358237 |
| Relative Squared Error | 0.130689 |
| Coefficient of Determination | 0.869311 |

As seen above, MAE and RMSE is too high for linear model. That is because AQI is non-linear itself.

2. **Decision Forest with Chi Feature Selection:**

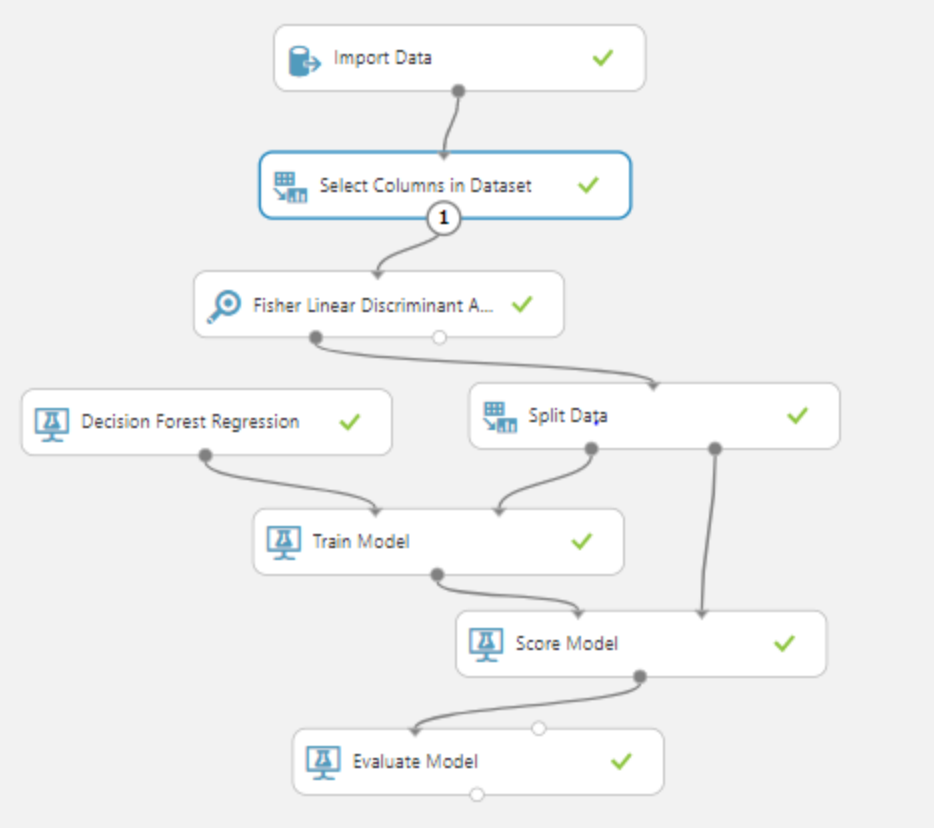
   Number of decision trees: 10

   Maximum depth of the Decision Trees: 15




| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 105089943553837.67 | 0.041034 | 1.285422 | 0.003348 | 0.004745 | 0.995255 |

**Decision Forest with Fisher Linear Discriminant Feature Selection:**



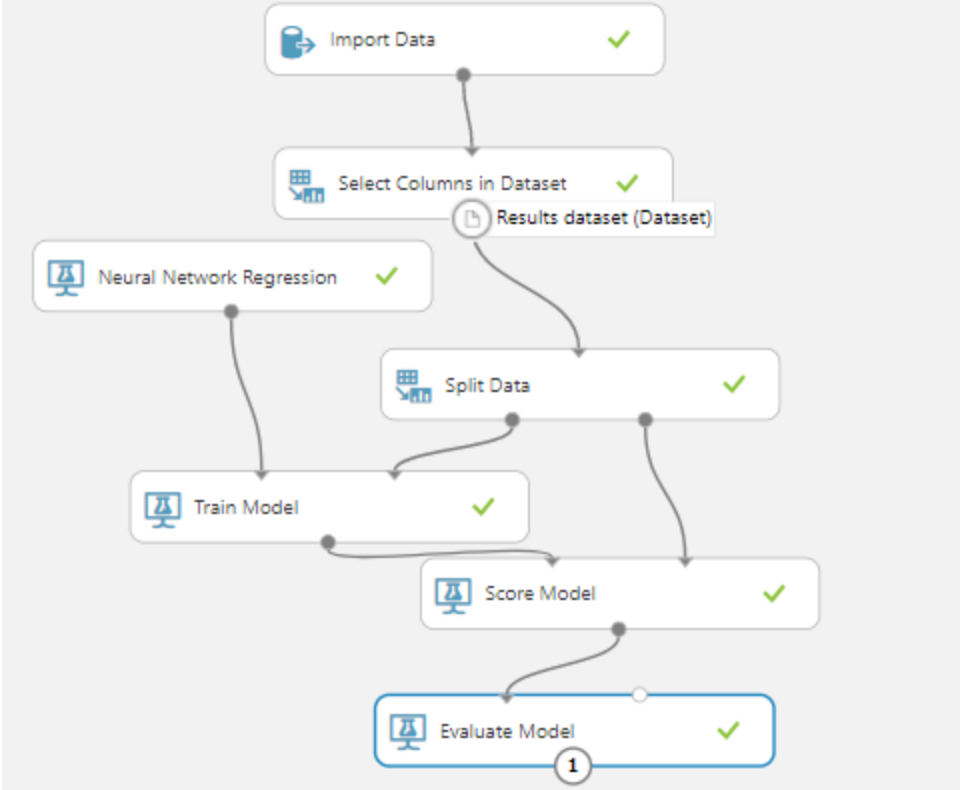

| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| -1095859.078411 | 0.021766 | 0.635584 | 0.001776 | 0.00116 | 0.99884 |

As seen above, Decision Forest gives us decent MAE and RMSE without taking long processing time.

Common features selected using Filter based(Chi) feature selection and Fisher Linear Discriminent Selection are:

- State Code(Contains city, county and site code)
- Arithmetic Mean
- First Max hour
- Date Local

3. **Neural Network:**




## Metrics

| | |
|---|---|
| Mean Absolute Error | 0.849674 |
| Root Mean Squared Error | 1.564987 |
| Relative Absolute Error | 0.069322 |
| Relative Squared Error | 0.007033 |
| Coefficient of Determination | 0.992967 |

Neural Network also gives us good results still less accurate than Decision Forest.

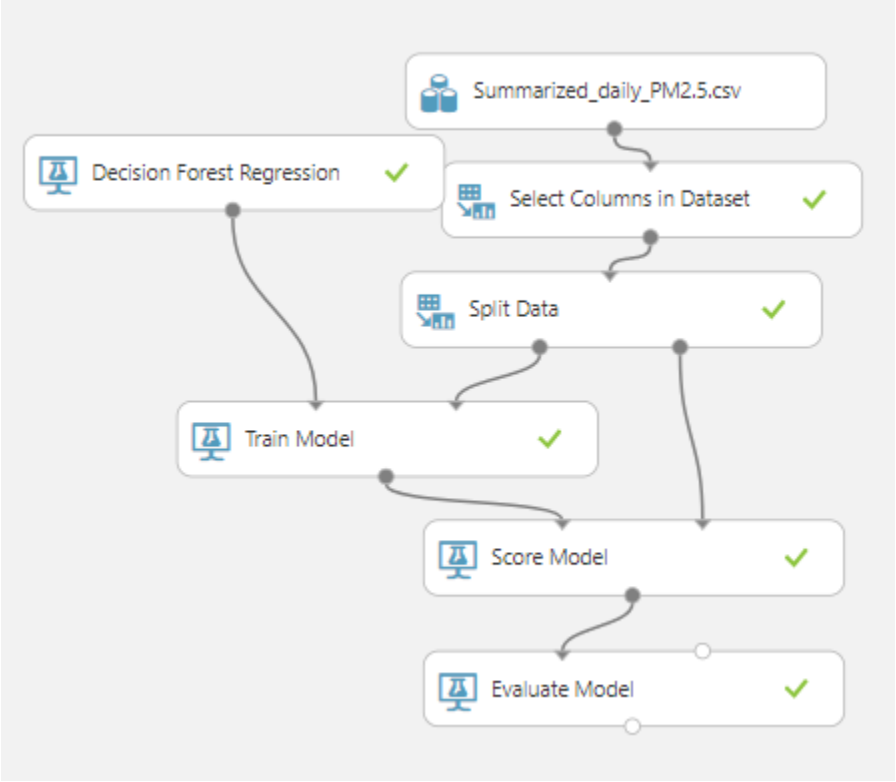
Also, it takes more processing time due to increase in number of inner nodes to 100 for obtaining above results.

**Conclusion:**

From above experiments we have decided to use Decision Forest algorithms for further analysis due to :

i. Good Accuracy
ii. Less Processing Time

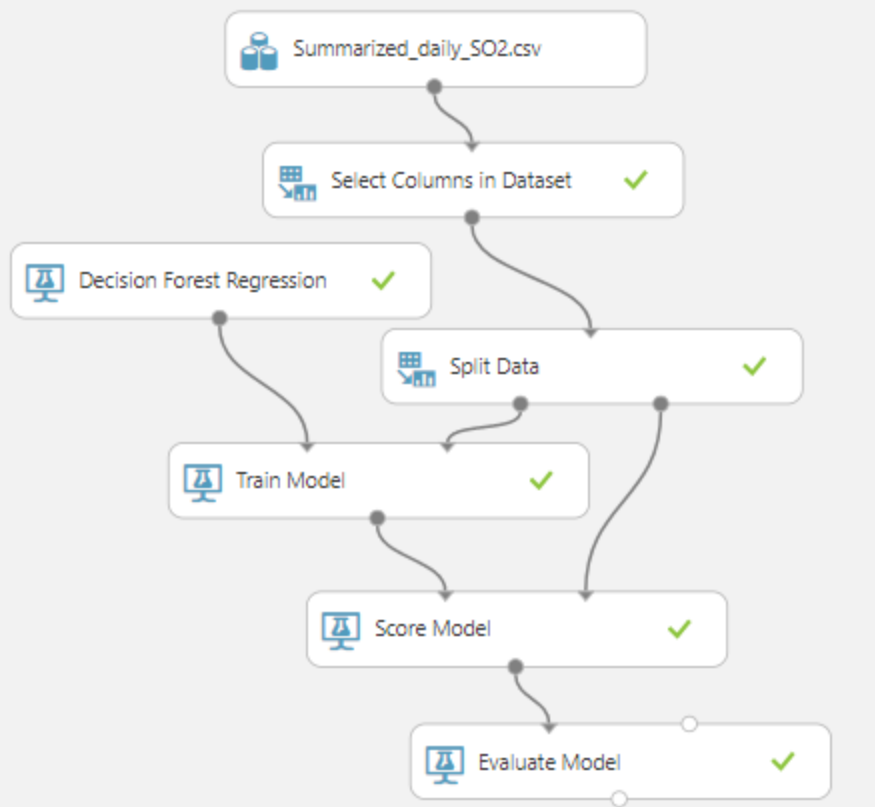
- **For PM2.5**




view as

| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 78366992623590.73 | 0.061886 | 0.911294 | 0.003697 | 0.001798 | 0.998202 |

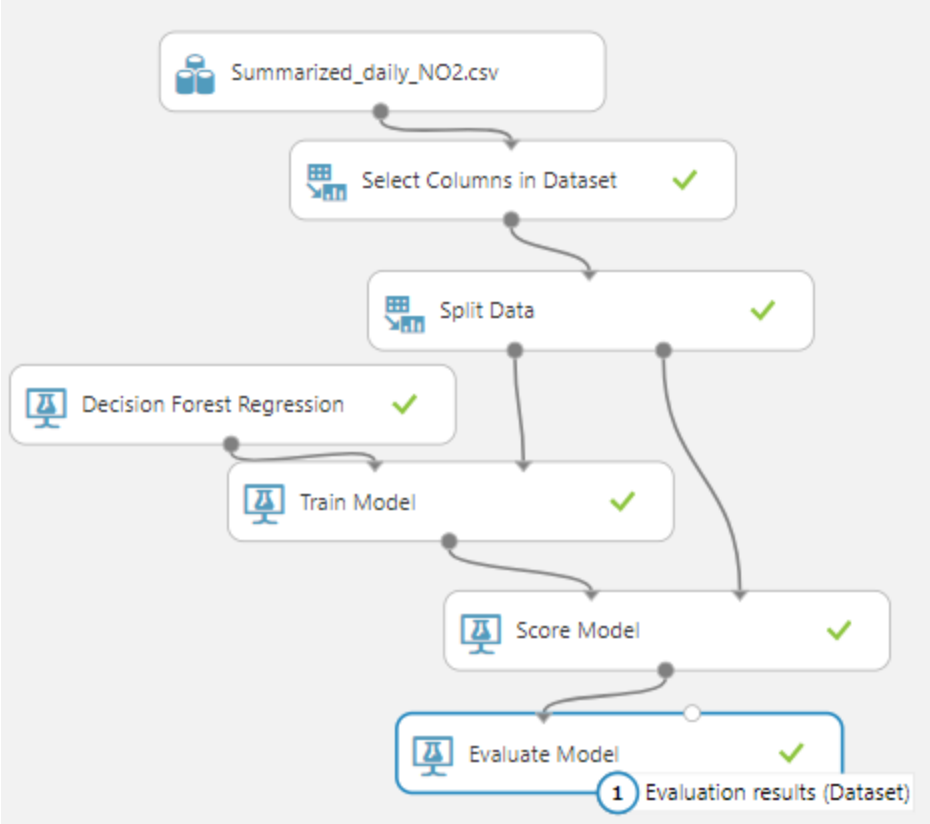
- **For SO2**




rows 1 columns 6

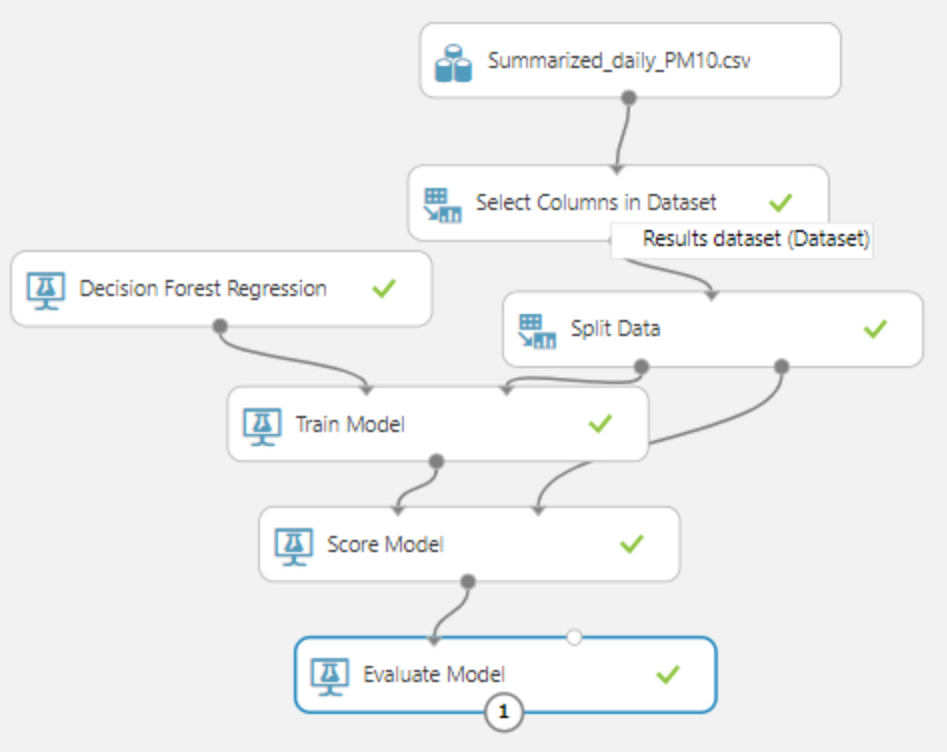
| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | 14765514928040.361 | 0.239601 | 2.745836 | 0.034238 | 0.029002 | 0.970998 |

- **For NO2**



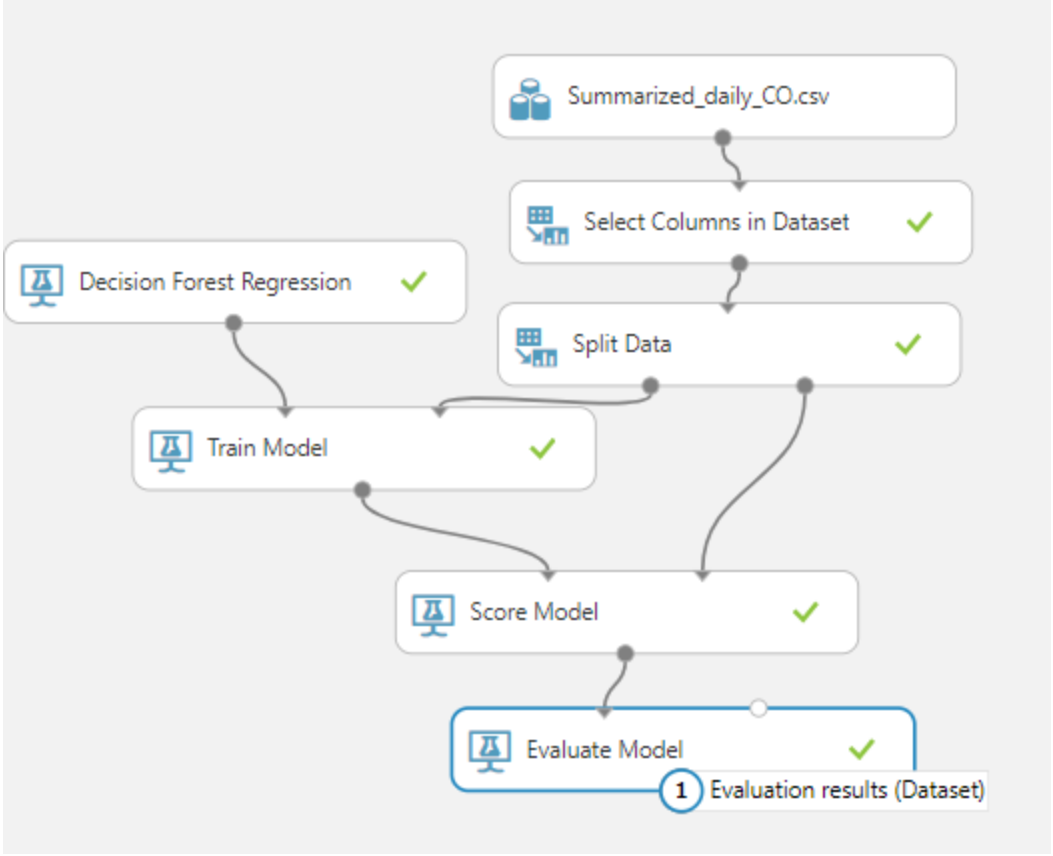

| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 868933990300.9795 | 0.001491 | 0.133265 | 0.000146 | 0.000114 | 0.999886 |

- **For PM10**



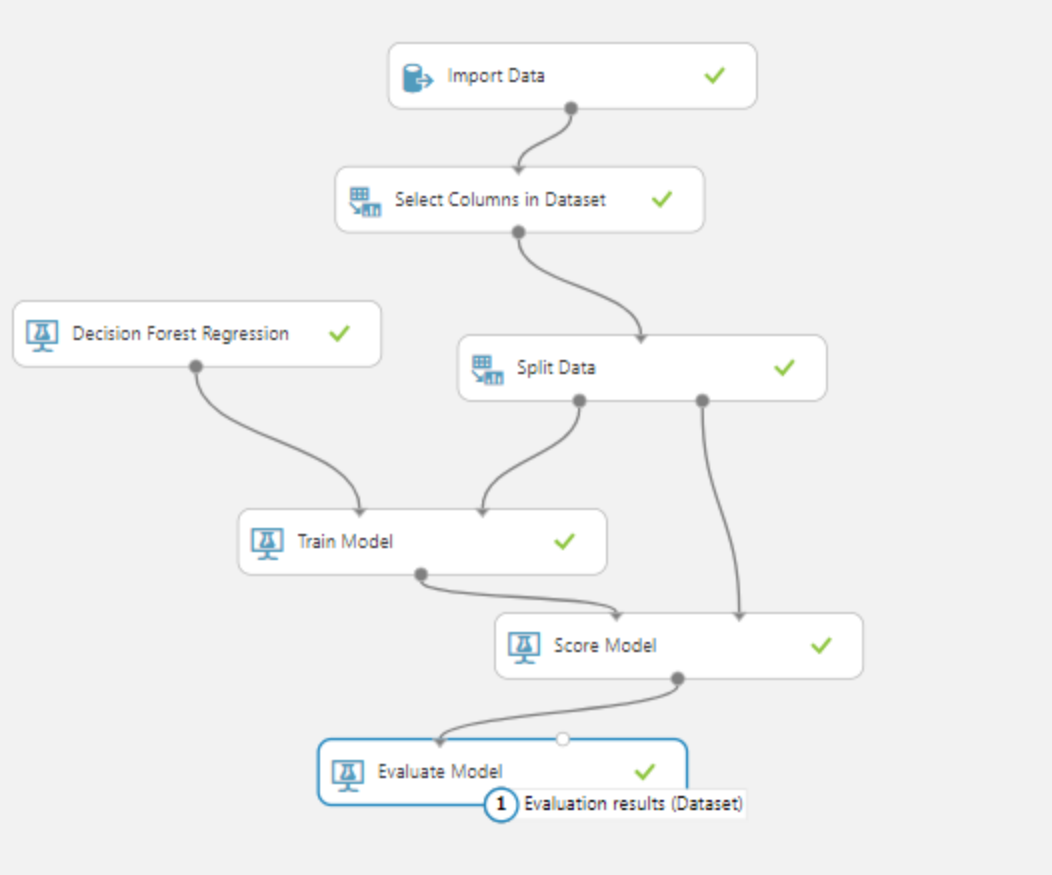

| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 331900740682 71.477 | 0.111268 | 3.606325 | 0.010415 | 0.02426 | 0.97574 |

- **For CO:**



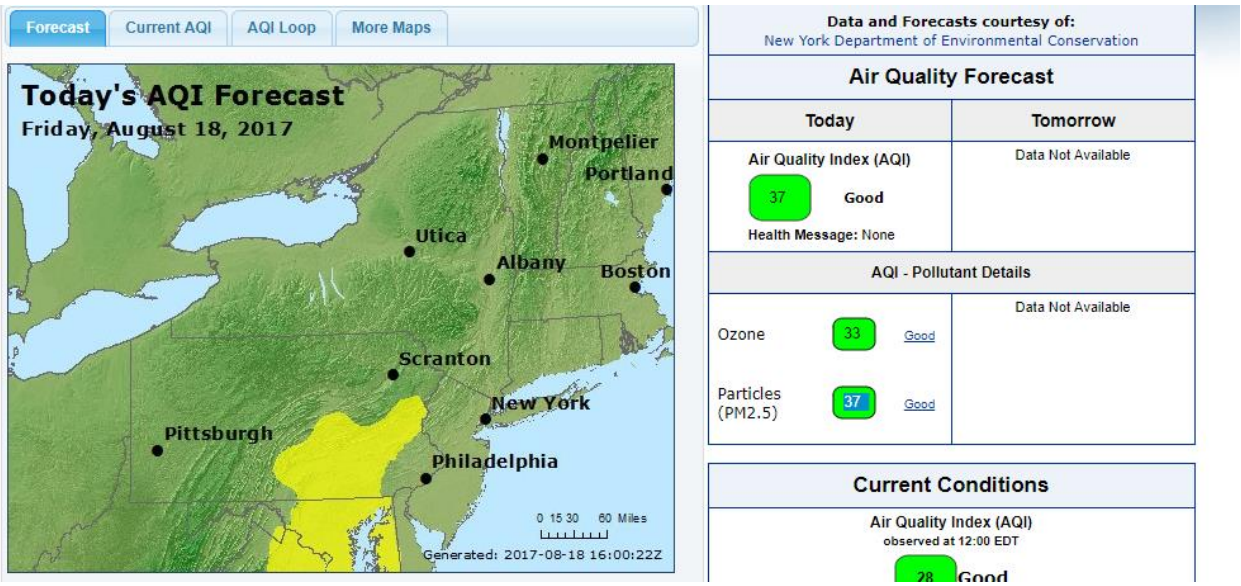

| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 211724085894.2526 | 0.0008 | 0.136094 | 0.000078 | 0.000119 | 0.999881 |

- **For Ozone:**

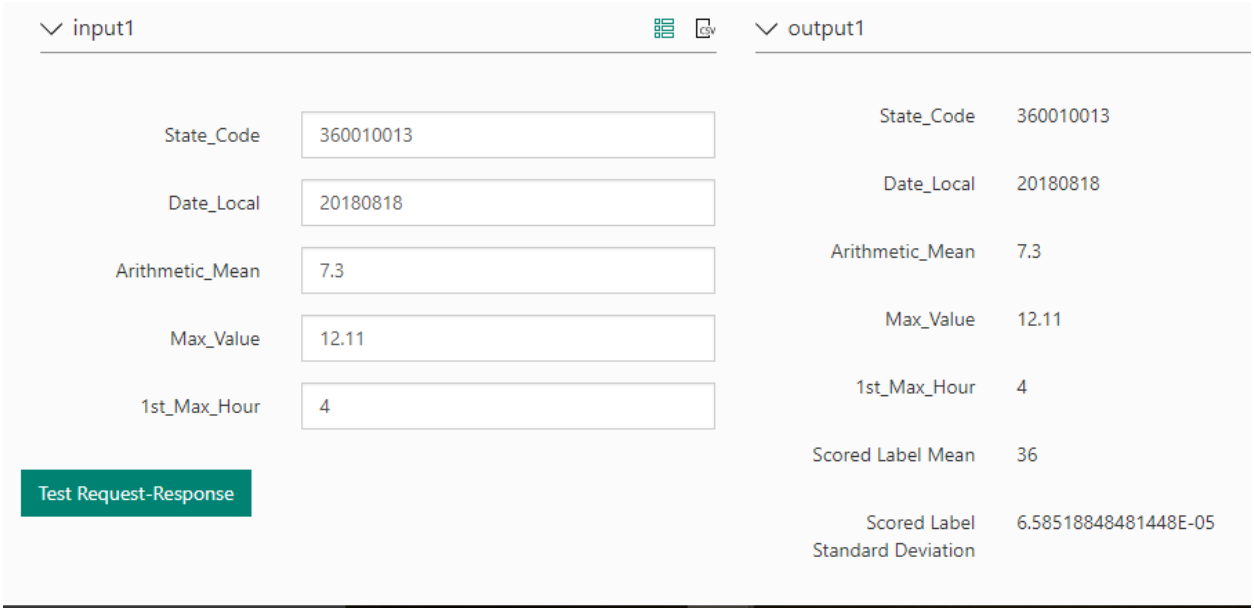



| Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|
| 139870352684838.9 | 0.040988 | 1.343999 | 0.003344 | 0.005187 | 0.994813 |

For New  York State – Alabany County


Forecast
Current AQI
AQI Loop
More Maps
Today's AQI Forecast
Friday, August 18, 2017
Montpelier
Portland
Utica
Albany
Boston
Scranton
New York
Pittsburgh
Philadelphia
0 15 30 60 Miles
Generated: 2017-08-18 16:00:22Z
Data and Forecasts courtesy of:
New York Department of Environmental Conservation
Air Quality Forecast
Today
Tomorrow
Air Quality Index (AQI)
37
Good
Health Message: None
Data Not Available
AQI - Pollutant Details
Ozone
33
Good
Particles (PM2.5)
37
Good
Data Not Available
Current Conditions
Air Quality Index (AQI)
observed at 12:00 EDT
28
Good

PM2.5

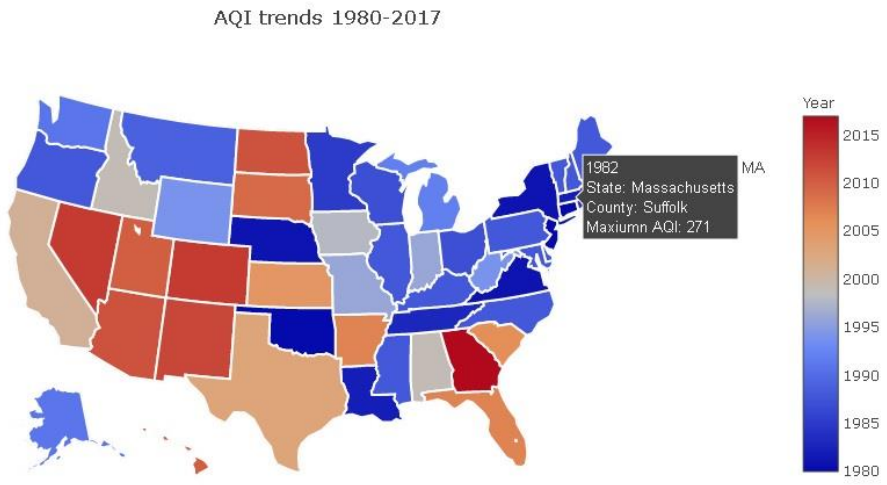

input1
State_Code 360010013
Date_Local 20180818
Arithmetic_Mean 7.3
Max_Value 12.11
1st_Max_Hour 4
Test Request-Response
output1
State_Code 360010013
Date_Local 20180818
Arithmetic_Mean 7.3
Max_Value 12.11
1st_Max_Hour 4
Scored Label Mean 36
Scored Label Standard Deviation 6.58518848481448E-05

# AQI Trends over 37 years

## Maximum AQI for every state over 37 years

*This Jupyter notebook shows state and county details where maximum AQI recorded from year 1980 to 2017*

In [49]:
```
import plotly.plotly as py
import pandas as pd
import plotly, os
plotly.offline.init_notebook_mode()

filepath = os.getcwd() + "/Statewise_AQI.csv"
df = pd.read_csv(filepath)

for col in df.columns:
    df[col] = df[col].astype(str)

scl = [[0.0, 'rgb(242,240,247)'],[0.2, 'rgb(218,218,235)'],[0.4, 'rgb(188,189,220)']]

df['text'] = 'State: ' + df['State'] +'<br>' + 'County: ' + df['County'] + '<br>' +'Maxiumn AQI: '+df['Max AQI']

data = [ dict(
        type='choropleth',
        colorscale = scl,
        autocolorscale = False,
        locations = df['Code'],
```



# Classification:

For classification we have used the combined Annual AQI files for each state for 37 years and split the train and test data as 70-30%.

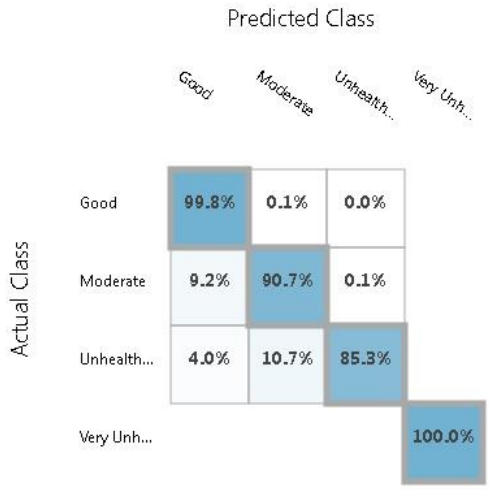
## Logistic Regression

Logistic Regression Model › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| Overall accuracy | 0.9894 |
| Average accuracy | 0.9947 |
| Micro-averaged precision | 0.9894 |
| Macro-averaged precision | 0.973822 |
| Micro-averaged recall | 0.9894 |
| Macro-averaged recall | 0.939751 |

Logistic Regression Model › Evaluate Model › Evaluation results

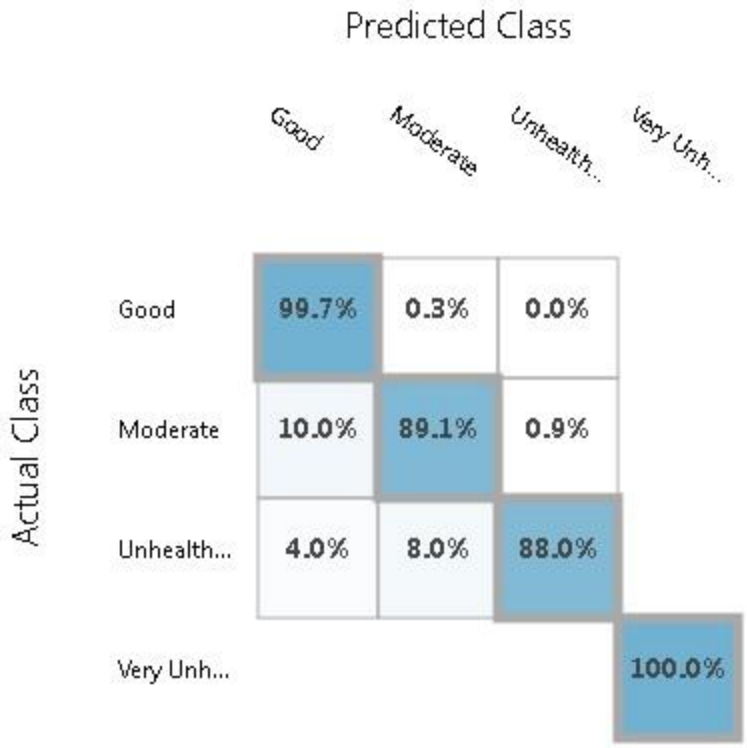



## 2. Decision Forest

Desicion Forest Model › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| Overall accuracy | 0.986679 |
| Average accuracy | 0.99334 |
| Micro-averaged precision | 0.986679 |
| Macro-averaged precision | 0.954207 |
| Micro-averaged recall | 0.986679 |
| Macro-averaged recall | 0.941962 |

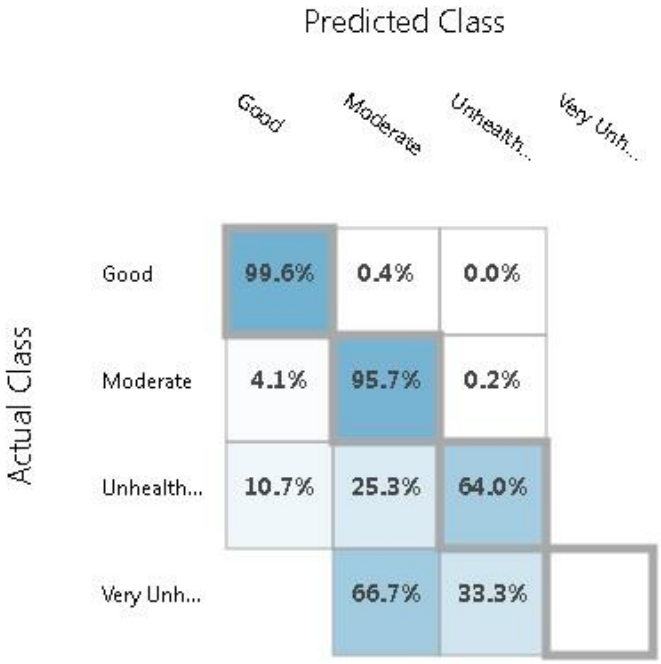



# 3. Neural Network

Neural Network Model › Evaluate Model › Evaluation results

## Metrics

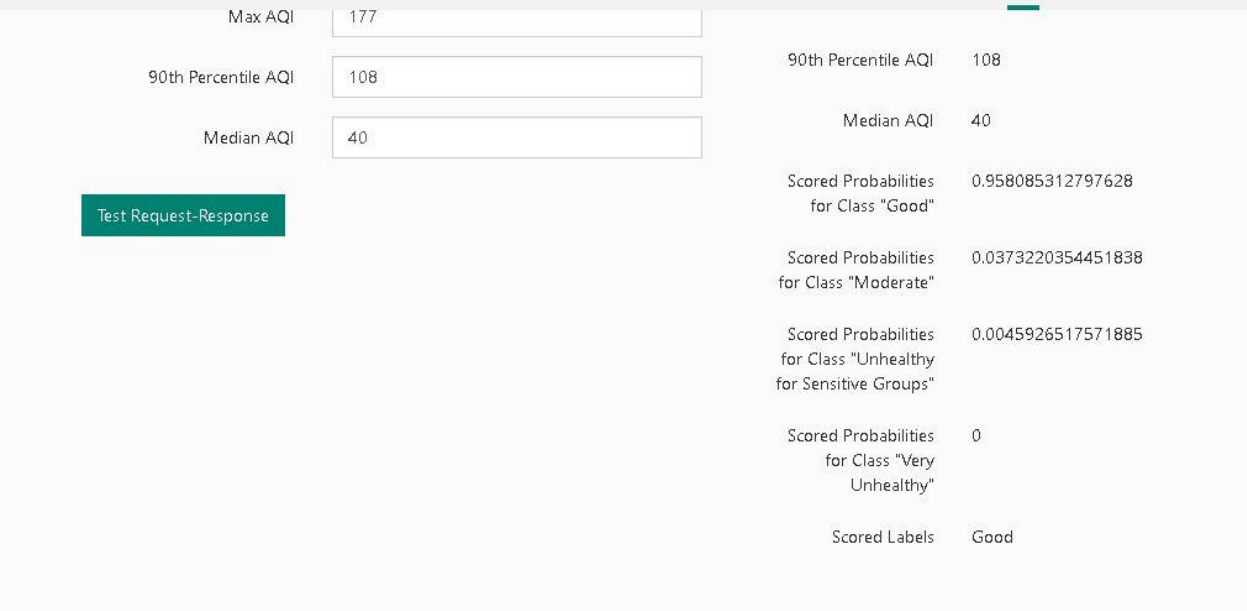
| | |
|---|---|
| **Overall accuracy** | **0.989493** |
| Average accuracy | 0.994747 |
| **Micro-averaged precision** | **0.989493** |
| **Macro-averaged precision** | **NaN** |
| **Micro-averaged recall** | **0.989493** |
| **Macro-averaged recall** | **0.648137** |



# Classify condition:

Results:

1. Decision Forest




2. Neural Network


Based on the above results, we have chosen neural network, since it has best accuracy for prediction compared to decision forest.

# Clustering:

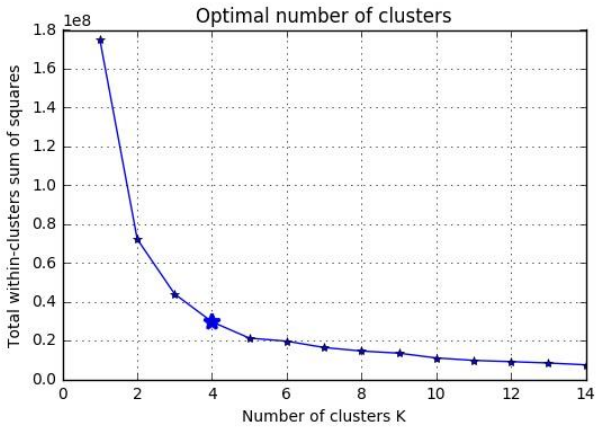
## Perform Clustering

*This Jupyter notebook performs clustering for 6 pollutants NO2, CO, SO2, Ozone, PM2.5 and PM10*

In [22]:
```
import pandas as pd
import numpy as np
from pandas import *
import os,scipy, datetime
from sklearn.utils import shuffle
from sklearn.preprocessing import LabelEncoder
from sklearn import cluster
from scipy.cluster.vq import kmeans,vq
from scipy.spatial.distance import cdist
import matplotlib.pyplot as plt

def performClustering(filename, pollutant_name):
    filepath = os.getcwd() + filename
    df= pd.read_csv(filepath)
    shuffled_df = shuffle(df)
    def dummyEncode(df):
        columnsToEncode = list(df.select_dtypes(include=['category','object']))
        le = LabelEncoder()
        for feature in columnsToEncode:
            try:
                if feature!='AOI':
```

In [29]:
```
#function call
performClustering("/Summarized_daily_PM10.csv", "PM10")
```

```
2017-08-18 01:38:19.296765  Defining begins...
2017-08-18 01:49:53.581124  Defigning done...
2017-08-18 01:49:53.685194  Plotting bend graph...
```



```
2017-08-18 01:49:54.703870  Done plotting!!
2017-08-18 01:49:54.703870  Clustering begins...
2017-08-18 01:50:16.770156  Done!!...
--------K-means Clustering--------
Grouping into clusters...
Exporting clusters as individual dataframes...
Cluster  0  of  107456  rows!
Cluster  1  of  298732  rows!
Cluster  2  of  134691  rows!
Cluster  3  of  70282  rows!
Cluster  4  of  134313  rows!
Cluster  5  of  10077  rows!
Cluster  6  of  62243  rows!
Cluster  7  of  48373  rows!
--------Manual Clustering--------
Grouping into clusters...
Exporting clusters as individual dataframes...
PM10 Manual Cluster  0  of  828015  rows!
PM10 Manual Cluster  1  of  34965  rows!
PM10 Manual Cluster  2  of  2342  rows!
PM10 Manual Cluster  3  of  537  rows!
PM10 Manual Cluster  4  of  143  rows!
PM10 Manual Cluster  5  of  165  rows!
```

Technique1: K-Means Clustering

PM2.5 - Decision forest

| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | | | | | | |
| | 179896384024.24814 | 0.030445 | 0.439232 | 0.003021 | 0.001334 | 0.998666 |

PM2.5 – Neural Network

Metrics

| | |
|---|---|
| **Mean Absolute Error** | **0.276125** |
| **Root Mean Squared Error** | **0.53404** |
| **Relative Absolute Error** | **0.027401** |
| **Relative Squared Error** | **0.001972** |
| **Coefficient of Determination** | **0.998028** |

**Pollutant NO2:**

1) Decision Forest

| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | 62001489.775024 | 0.00246 | 0.076828 | 0.000239 | 0.000038 | 0.999962 |

NO2 K-means Neural Network › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| **Mean Absolute Error** | **0.346545** |
| **Root Mean Squared Error** | **0.425891** |
| **Relative Absolute Error** | **0.033732** |
| **Relative Squared Error** | **0.001153** |
| **Coefficient of Determination** | **0.998847** |

Technique2: Manual clustering

**Group1:**

Decision Forest

| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | 1130377746.089086 | 0.000032 | 0.019627 | 0.000003 | 0.000003 | 0.999997 |

Neural Network

NO2 Manual Clustering Group1 Neural Net... › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| **Mean Absolute Error** | **0.282274** |
| **Root Mean Squared Error** | **0.333662** |
| **Relative Absolute Error** | **0.028776** |
| **Relative Squared Error** | **0.00081** |
| **Coefficient of Determination** | **0.99919** |

**Group2:**

Decision Forest

| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | | | | | | |
| | 15519825.373203 | 0.015858 | 0.111427 | 0.002856 | 0.000214 | 0.999786 |

Neural Network

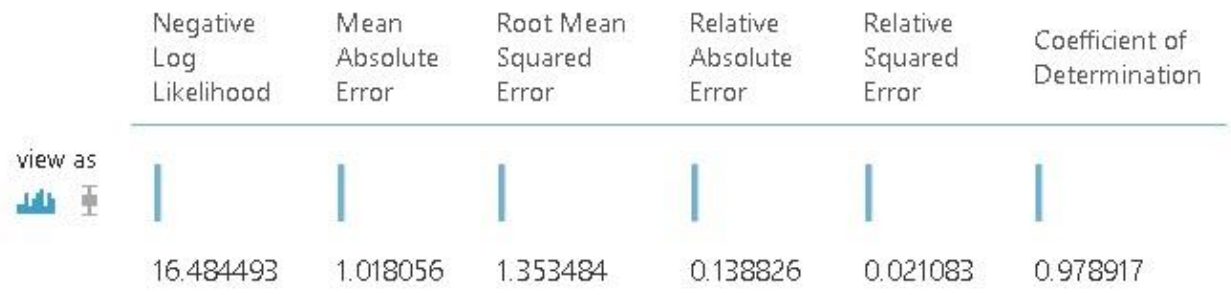
NO2 Manual Clustering Group2 Neural Net... › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| **Mean Absolute Error** | **0.351351** |
| **Root Mean Squared Error** | **0.432037** |
| **Relative Absolute Error** | **0.063286** |
| **Relative Squared Error** | **0.003216** |
| **Coefficient of Determination** | **0.996784** |

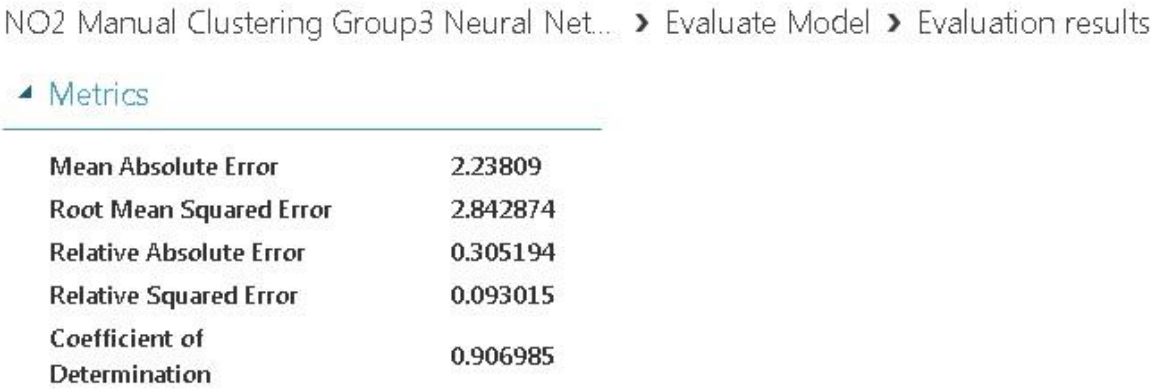
**Group3:**

Decision Forest



| | Negative Log Likelihood | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|---|---|---|---|---|---|---|
| view as | | | | | | |
| | 16.484493 | 1.018056 | 1.353484 | 0.138826 | 0.021083 | 0.978917 |

Neural Network




Bases on the all models, we choose Decision Forest Model as best model as it give us best accuracy and best values for MAE, RMSE, RAE, RSE and coefficient.

# Web Application:

## Predict

**state** New York

**state** NY-Bronx

**date** 08/19/2018

Get AQI

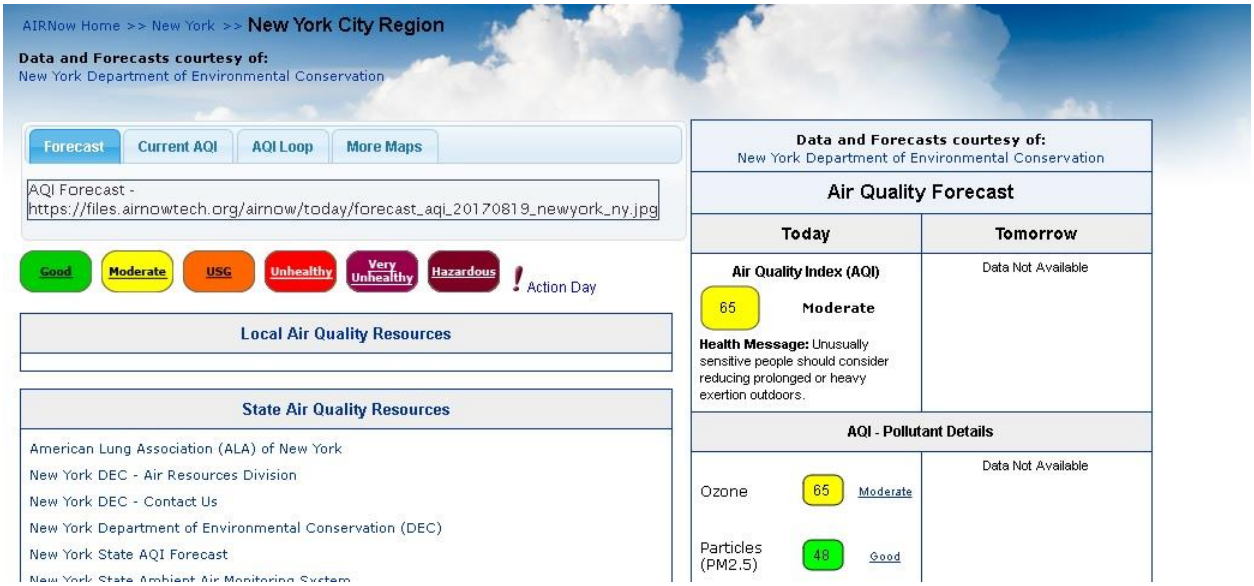
### PM2.5

Predicated AQI: 49.1792592592593
Message: Good: It's a great day to be active outside.

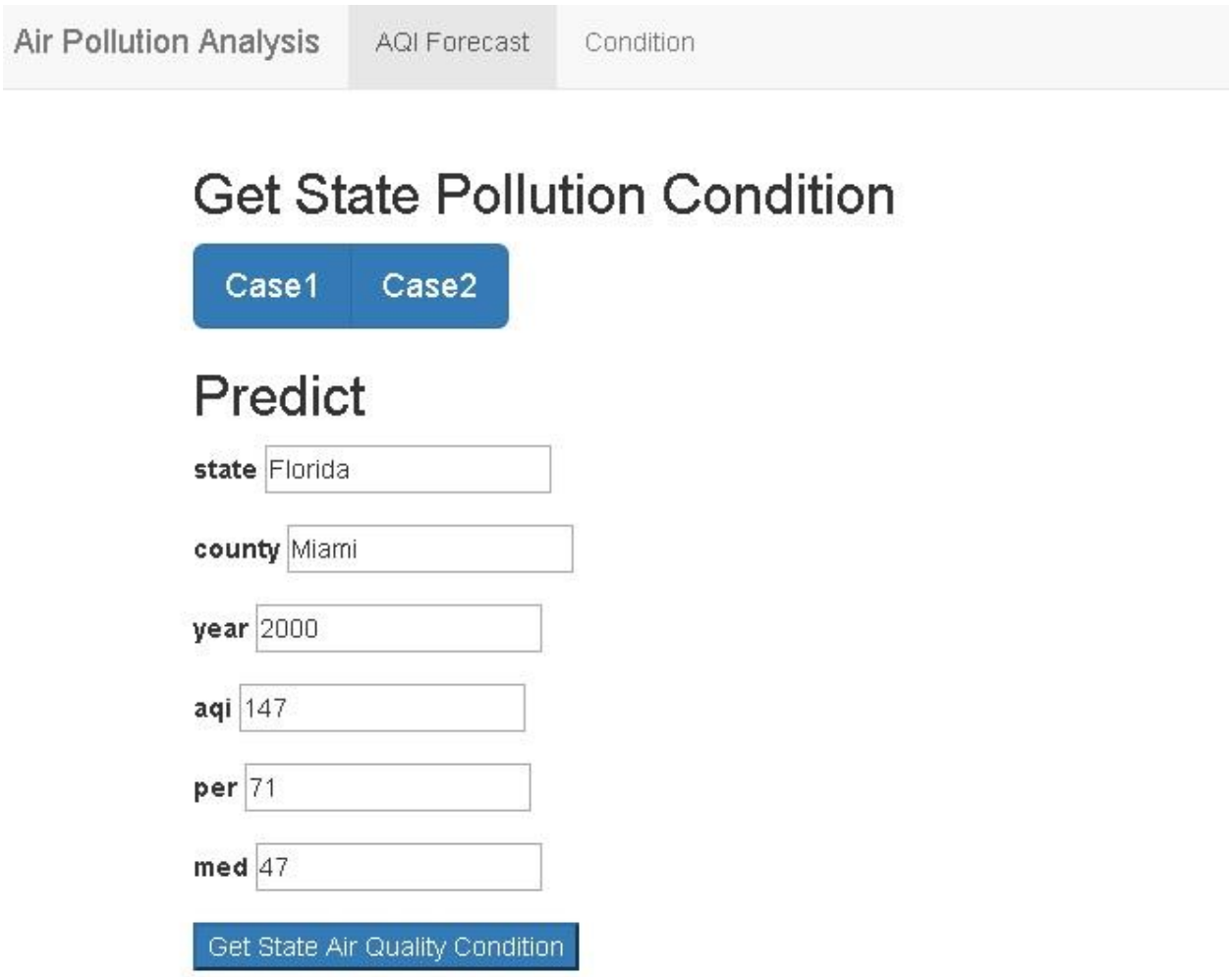
### Ozone

Predicated AQI: 67
Message: Unusually sensitive people: Consider reducing prolonged or heavy outdoor exertion. Watch for symptoms such as coughing or shortness of breath. These are signs to take it a little easier.Others: It's a good day to be active outside

Ozone and PM2.5 predicted AQI for next day.




Classification:




Visualization:

https://datastudio.google.com/open/0B0oq2j4ughAdWTFCaEQ3Q3d0cWM

Web Application:

https://morning-river-35298.herokuapp.com/

Demo Link:

https://www.youtube.com/watch?edit=vd&v=bUP-lwwh5HE