**Under the guidance of**

**Prof Srikanth Krishnamurthy**

**Pollution Data Analysis and Machine learning Prediction**

**Team7: Snigdha Joshi & Vipra Shah**

**Part A**

**Data Pipelining:**

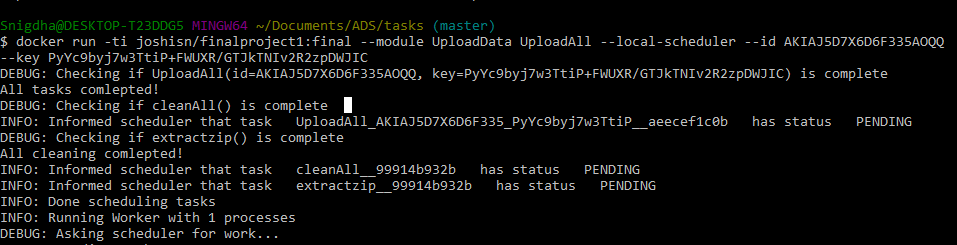
Data pipelining is executed using luigi

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>



Unzip Data File

Total 36 files for 6 Pollutants

Clean and Merge 18 Files

Clean and Merge 18 Files

18

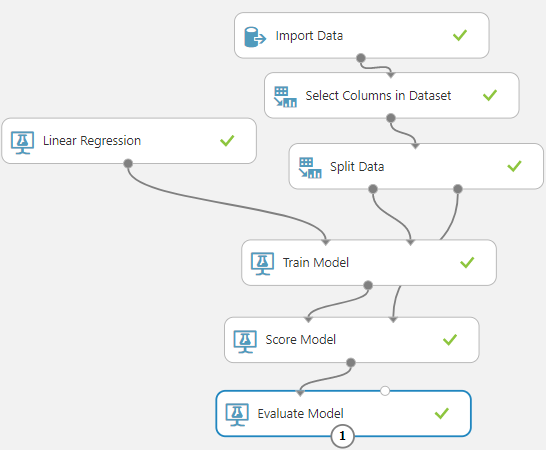
6 Files

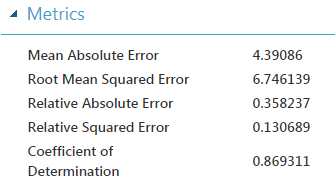
Upload to S3

**Part B**

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

1. **Linear Regression:**



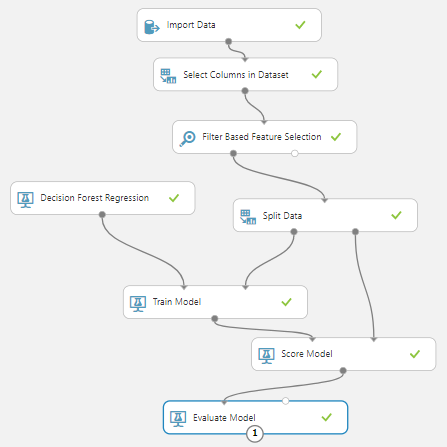


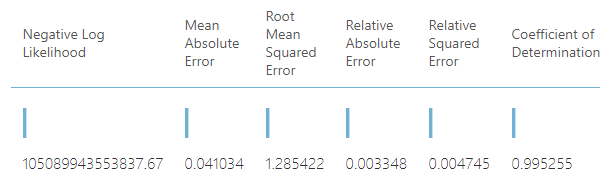
As seen above, MAE and RMSE is too high for linear model. That is because AQI is non-linear itself.

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

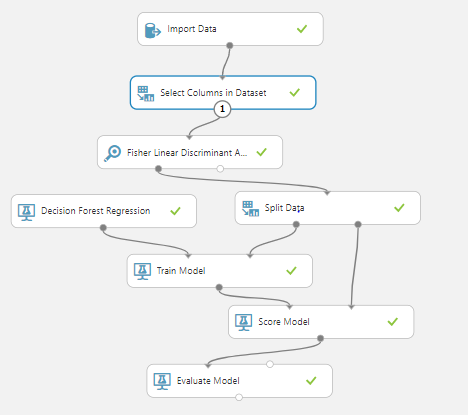
Number of decision trees: 10

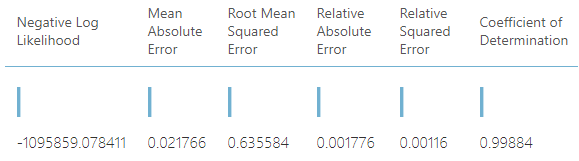
Maximum depth of the Decision Trees: 15





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



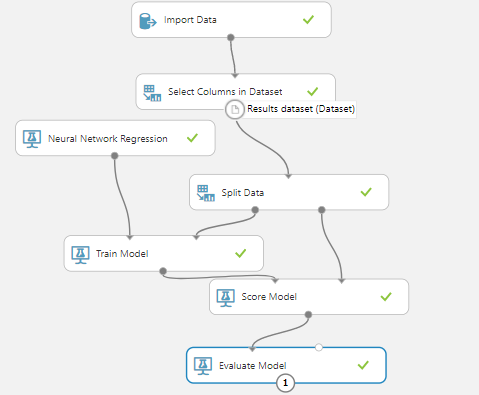


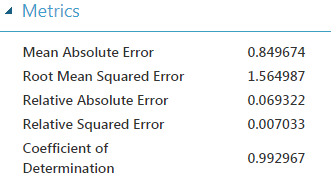
As seen above, Decision Forest gives us decent MAE and RMSE without taking long processing time.

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

1. **Neural Network:**





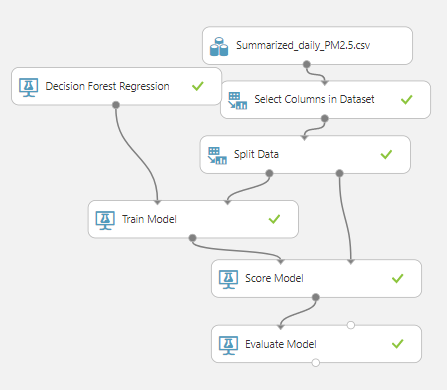
Neural Network also gives us good results still less accurate than Decision Forest.

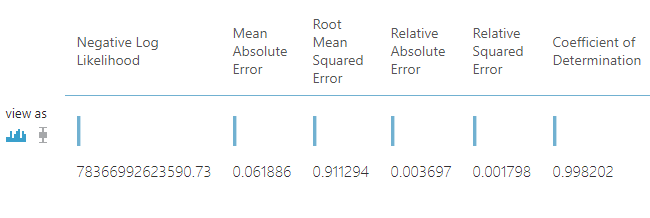
Also, it takes more processing time due to increase in number of inner nodes to 100 for obtaining above results.

**Conclusion:**

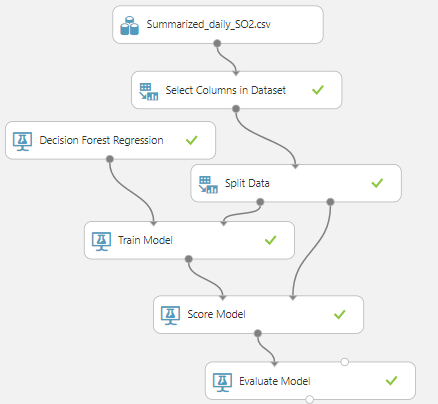
From above experiments we have decided to use Decision Forest algorithms for further analysis due to :

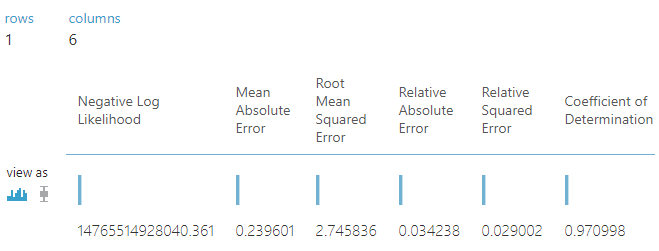
* + 1. Good Accuracy
    2. Less Processing Time
* **For PM2.5**



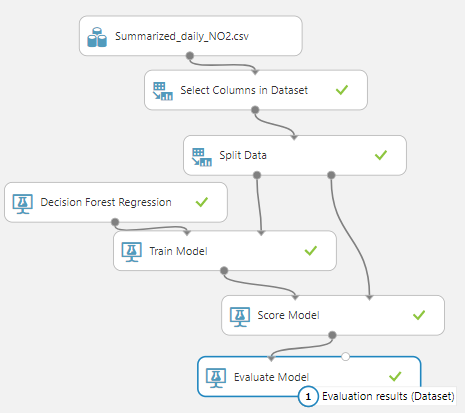


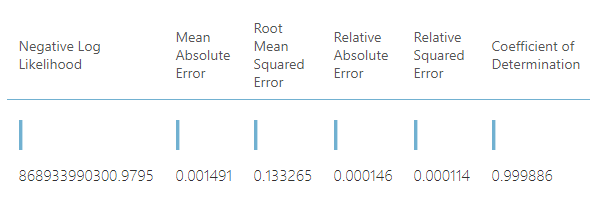
* **For SO2**



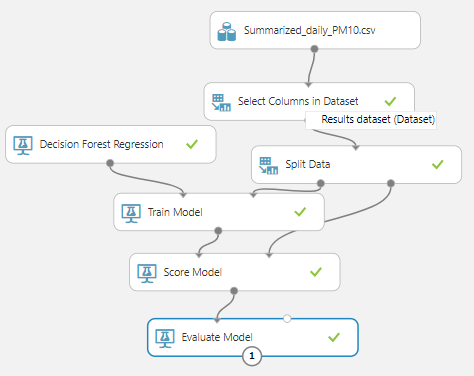


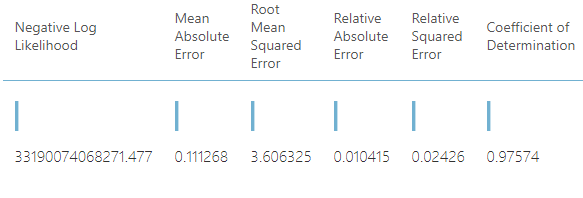
* **For NO2**



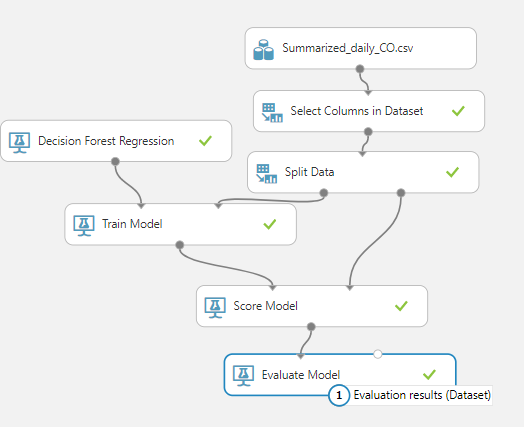


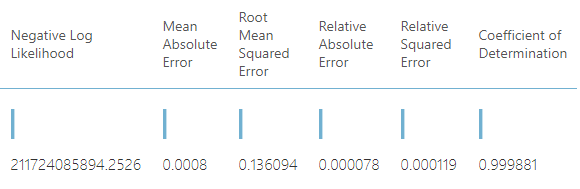
* **For PM10**



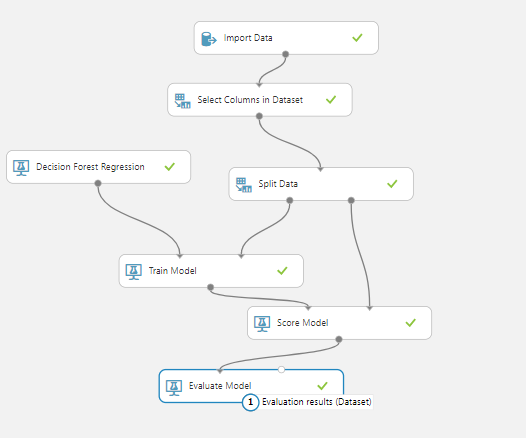


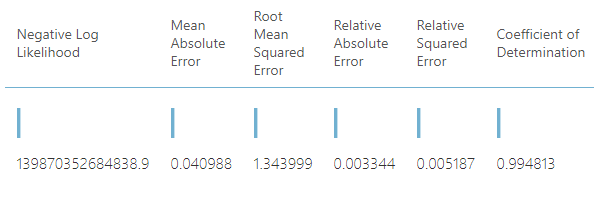
* **For CO:**



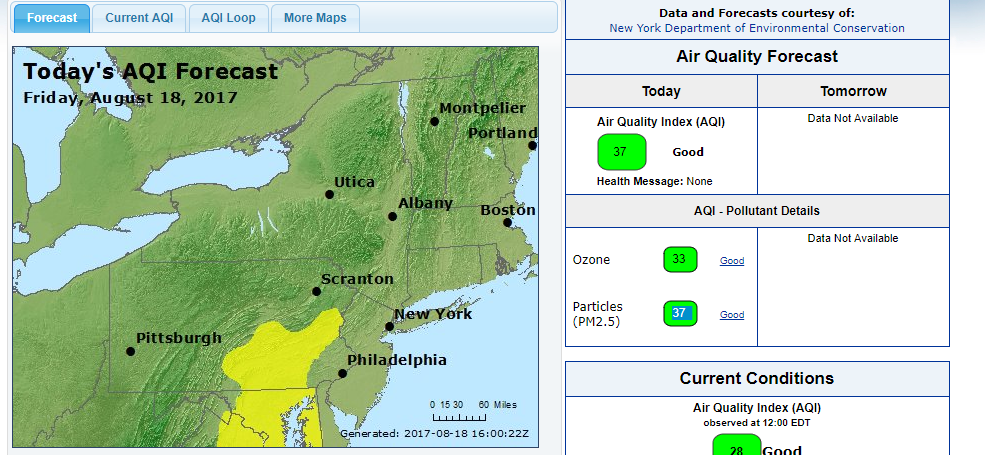


* **For Ozone:**

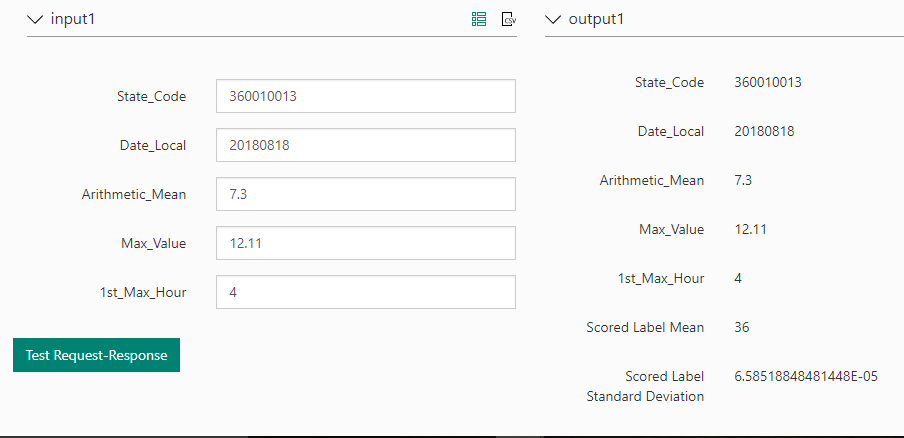




For New York State – Alabany County

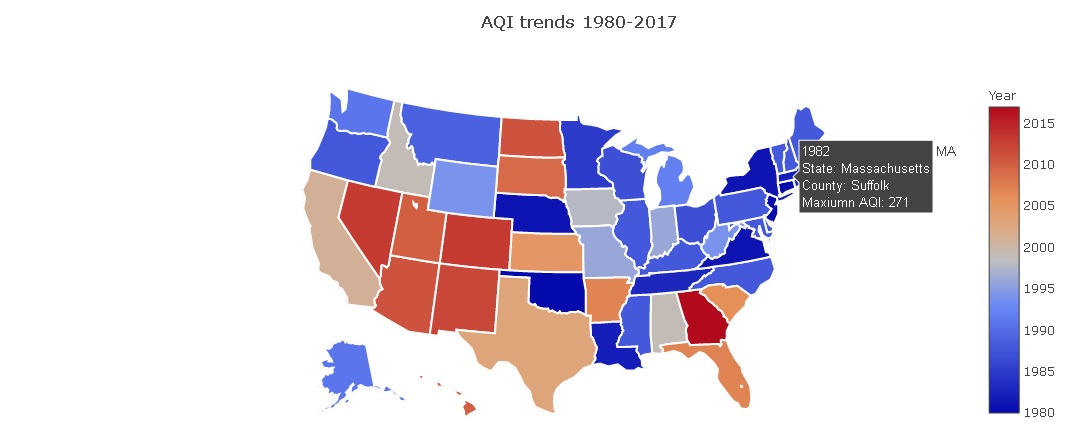


PM2.5

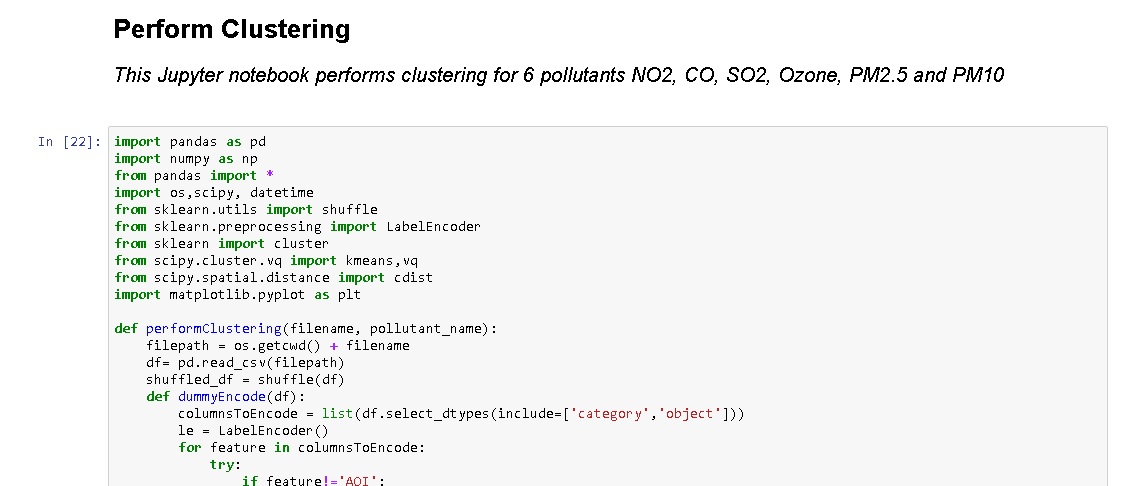


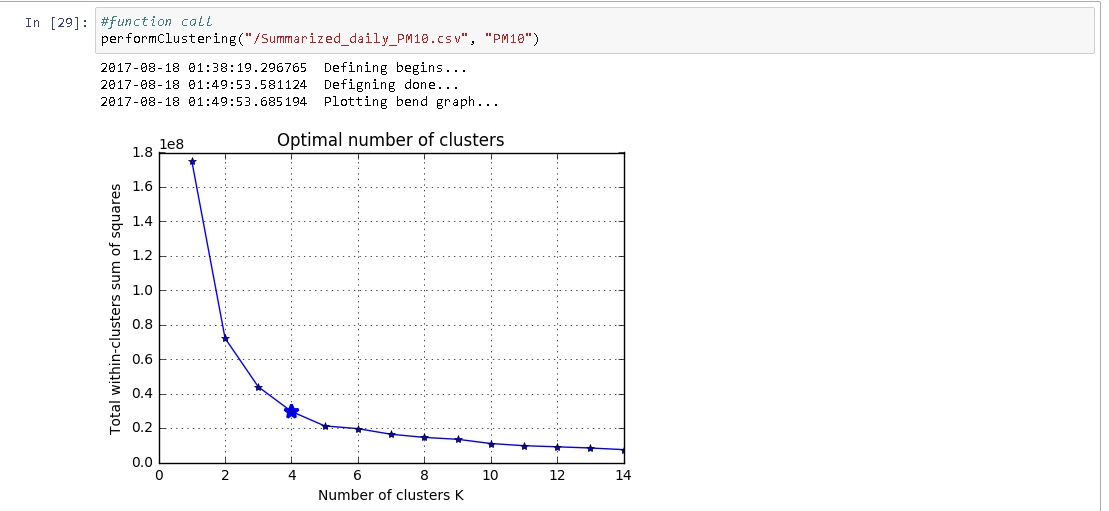
**AQI Trends over 37 years**

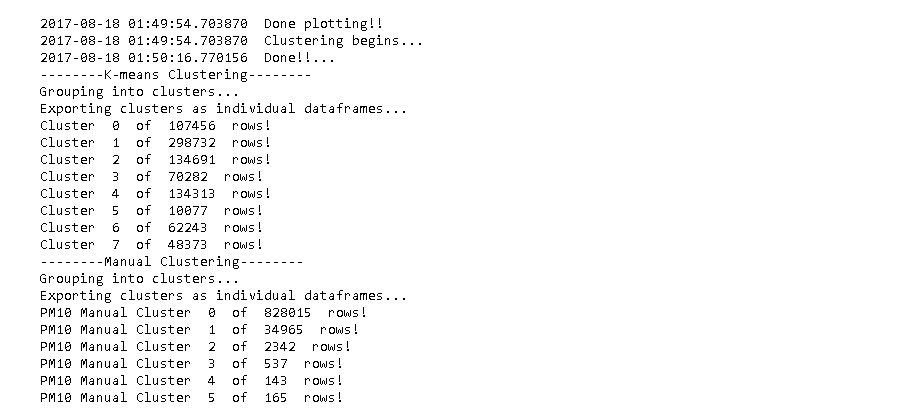
****

****

**Clustering:**

****

****

****

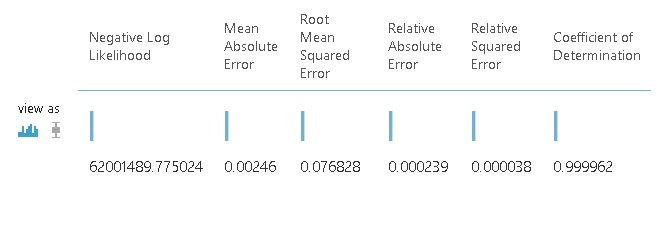
Technique1: K-Means Clustering

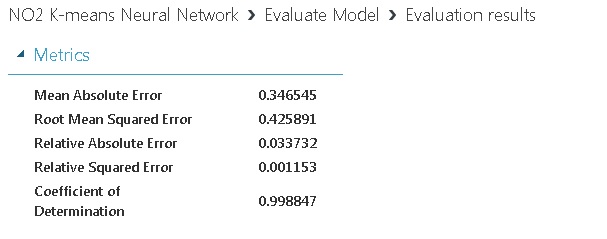
PM2.5 - Decision forest

PM2.5 – Neural Network

**Pollutant NO2:**

1) Decision Forest

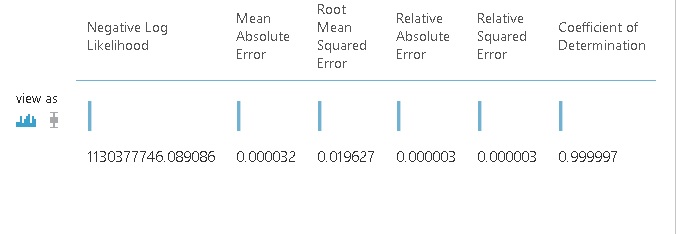




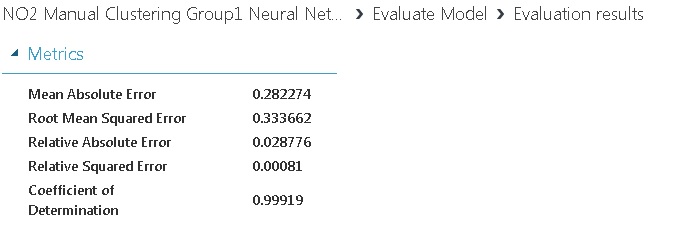
Technique2: Manual clustering

**Group1:**

Decision Forest

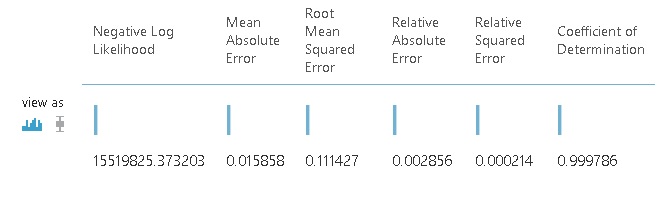


Neural Network

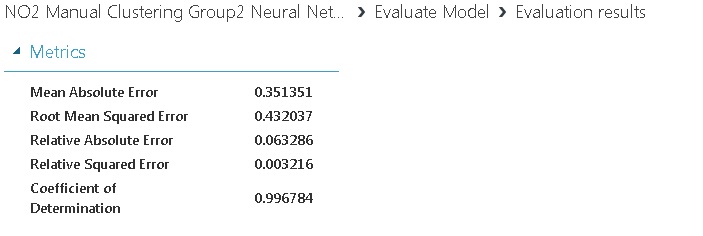


**Group2:**

Decision Forest

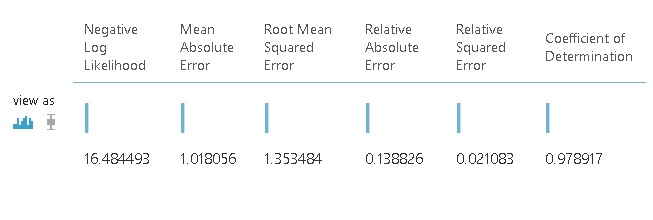
****

Neural Network

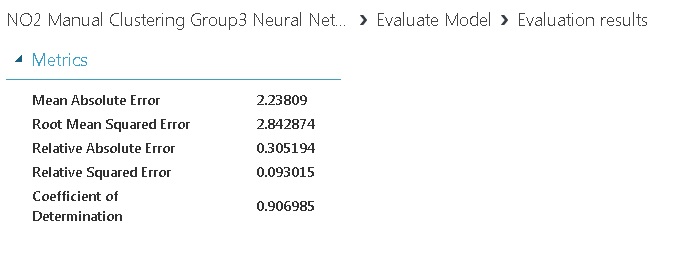


**Group3:**

Decision Forest

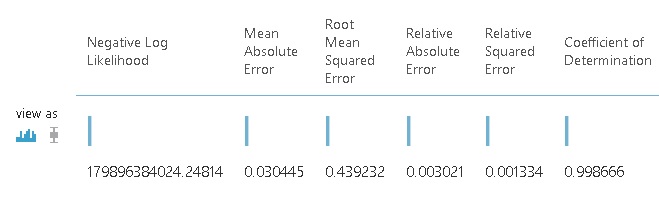


Neural Network



**Pollutant PM2.5**

Group1: Decision Forest



Group1: Neural Network

