Course Name: Data Analytics with Cognos

Project: Air Quality Analysis in Tamil Nadu

PROJECT OVERVIEW:

Our utilimate aim that analysis air quality from various location in Tamil Nadu. And these analysis data are depict in various visualization techniques. These data are take from monitoring stations of various locations in Tamil Nadu. From that analysis to gain insights of air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels. And creating a predictive model using Python and relevant libraries.

DESIGNING THINKING

The design thinking part our project is explained about so many aspect such as air quality trends, identify hotspot, steps to load dataset, Preprocess of data and visualization the dataset using visualization techniques. The air quality trends in Tamil Nadu which include about the current scenario summaries intends to reflect actual air quality and therefore includes concentrations that have been impacted by episodic events. Hotspots of pollution means a location where emission from specific or any sources may expose individuals population group to elevated risks adversy of health effect. The following steps to load dataset that firstLoad necessary packages into our working envirnoment. And load the air quality data set into your current session. Describe and Display the data. Summary of the data. Create a new data frame and remove missing values. Creation of factor (categorical) variables. Distribution of the data. And we finally visualized the data using Python codes.

DEVEOLPMENT PHASE:

In the development phase, we explained about how to analysis the air quality with the help of air quality index in Tamil Nadu using Python. Air quality is analysis based on chemical pollutant quantity. By using machine learning, we can AQ(Air quality). The dataset used in the development phases is https://tn.data.gov.in/resource/location-

wise-daily-ambient-air-quality-tamil-nadu-year-2014. We explained about how to load the, spillting and cleaning the dataset. And also it include about machine learning algorithm which used to build an predictive model for analysis process. And finally the depict visualize the final output which is easy values each onces.

ANALYSIS APPROACH:

According to the status report of air quality in Tamil Nadu in 2014, released by the Tamil Nadu Pollution Control Board (TNPCB), Chennai tops the list with five stations recording the highest concentration of Respirable Suspended Particulate Matter (RSPM) in 2014. The prescribed annual average standard for Respirable Suspended Particulate Matter (RSPM) was 60 micrograms/cubic metre. Tamil Nadu Pollution Control Board (TNPCB) had observed a decreasing trend over the years in the number of stations confirming to the standard in the presence of Respirable Suspended Particulate Matter (RSPM), with particle less than 10 micron size. Total Suspended Particulate Matter (TSPM) was also found high in Chennai. Thoothukodi continues to be a sulphur dioxide hotspot. The 3 stations in Thoothukudi are said to be moderately polluting and needs close monitoring. Thoothukudi ranks eighth in the SO2 levels in the country as per 2014 report of Tamil Nadu Pollution Control Board (TNPCB). The prescribed annual average standard for NO2 was 40 micrograms/cubic metre and presence of Nitrogen dioxide (NO2) in ambient air was within permissible limits in all stations in the state, except at SIDCO in Coimbatore. The trend of NO2 had indicated that its presence may exceed some stations in future. Then overall air quality in Tamil Nadu is 44.

DATA COLLECTION

In Data collection of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. It is an essential phase in all types of research, analysis, and decision-

making.The data collected is accurate data collection that not make false in our prediction and analysing part.

DATA VISUALIZATION

In this phase, we explained about data visualization process for using the apporiates python codes. More first data visualization means the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for peoples to present data to non-technical audiences without confusion. And we visualize the each component in the dataset using python code result are shown in various visualization methods.

PYTHON CODES:

RANDOM FOREST

SAMPLE CODE:

```
# importing Randomforest
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import RandomForestRegressor

# creating model
m1 = RandomForestRegressor()

# separating class label and other attributes
train1 = train.drop(['air_quality_index'], axis=1)
target = train['air_quality_index']

# Fitting the model
m1.fit(train1, target)
'''RandomForestRegressor(bootstrap=True,
ccp_alpha=0.0, criterion='mse',
                       max_depth=None,
max_features='auto', max_leaf_nodes=None,
                       max_samples=None,
min_impurity_decrease=0.0,
                       min_impurity_split=None,
min_samples_leaf=1,
                       min_samples_split=2,
min_weight_fraction_leaf=0.0,
                       n_estimators=100, n_jobs=None,
oob_score=False,
                       random_state=None, verbose=0,
warm_start=False)'''
```

```
# calculating the score and the score
is  97.96360799890066%
m1.score(train1, target) * 100

# predicting the model with other values (testing the
data)
# so AQI is 123.71
m1.predict([[123, 45, 67, 34, 5, 0, 23]])

# Adaboost model
# importing module
 # defining model
m2 = AdaBoostRegressor()
# Fitting the model
m2.fit(train1, target)

'''AdaBoostRegressor(base_estimator=None,
learning_rate=1.0, loss='linear',
                   n_estimators=50,
random_state=None)'''

# calculating the score and the score
is  96.15377360010211%
m2.score(train1, target)*100

# predicting the model with other values (testing the
data)
# so AQI is 94.42105263
m2.predict([[123, 45, 67, 34, 5, 0, 23]])
```

Load the dataset

```
# Load the dataset
df = pd.read_csv('cpcb_dly_aq_tamil_nadu-2014.csv')
print(df.head())
```

```
   Stn Code Sampling Date       State City/Town/Village/Area  \
0        38      01-02-14  Tamil Nadu                Chennai
1        38      01-07-14  Tamil Nadu                Chennai
2        38      21-01-14  Tamil Nadu                Chennai
3        38      23-01-14  Tamil Nadu                Chennai
4        38      28-01-14  Tamil Nadu                Chennai

                    Location of Monitoring Station  \
0  Kathivakkam, Municipal Kalyana Mandapam, Chennai
1  Kathivakkam, Municipal Kalyana Mandapam, Chennai
2  Kathivakkam, Municipal Kalyana Mandapam, Chennai
3  Kathivakkam, Municipal Kalyana Mandapam, Chennai
4  Kathivakkam, Municipal Kalyana Mandapam, Chennai

                                  Agency Type of Location   SO2   NO2  \
0  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0
1  Tamilnadu State Pollution Control Board  Industrial Area  13.0  17.0
2  Tamilnadu State Pollution Control Board  Industrial Area  12.0  18.0
3  Tamilnadu State Pollution Control Board  Industrial Area  15.0  16.0
4  Tamilnadu State Pollution Control Board  Industrial Area  13.0  14.0

   RSPM/PM10  PM 2.5
0       55.0     NaN
1       45.0     NaN
2       50.0     NaN
3       46.0     NaN
4       42.0     NaN
```

DATA PREPROCESSING

```
# Load the dataset
df = pd.read_csv('cpcb_dly_aq_tamil_nadu-2014.csv')
print(df.head())
```

```
   Stn Code Sampling Date       State City/Town/Village/Area  \
0        38      01-02-14  Tamil Nadu                Chennai
1        38      01-07-14  Tamil Nadu                Chennai
2        38      21-01-14  Tamil Nadu                Chennai
3        38      23-01-14  Tamil Nadu                Chennai
4        38      28-01-14  Tamil Nadu                Chennai

                    Location of Monitoring Station  \
0  Kathivakkam, Municipal Kalyana Mandapam, Chennai
1  Kathivakkam, Municipal Kalyana Mandapam, Chennai
2  Kathivakkam, Municipal Kalyana Mandapam, Chennai
3  Kathivakkam, Municipal Kalyana Mandapam, Chennai
4  Kathivakkam, Municipal Kalyana Mandapam, Chennai

                                  Agency Type of Location   SO2   NO2  \
0  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0
1  Tamilnadu State Pollution Control Board  Industrial Area  13.0  17.0
2  Tamilnadu State Pollution Control Board  Industrial Area  12.0  18.0
3  Tamilnadu State Pollution Control Board  Industrial Area  15.0  16.0
4  Tamilnadu State Pollution Control Board  Industrial Area  13.0  14.0

   RSPM/PM10  PM 2.5
0       55.0     NaN
1       45.0     NaN
2       50.0     NaN
3       46.0     NaN
4       42.0     NaN
```

STATISTICAL

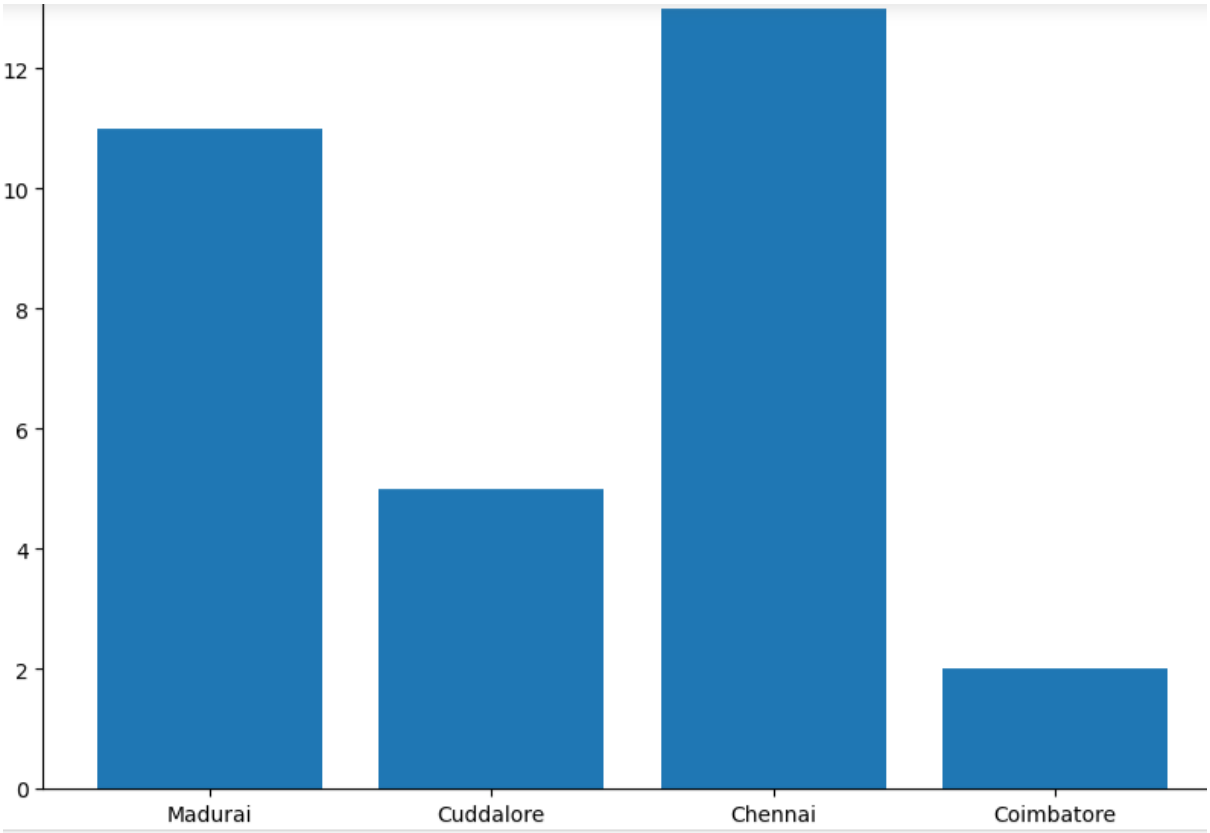
```
df.describe()
```

	Stn Code	SO2	NO2	RSPM/PM10	PM 2.5
count	2879.000000	2868.000000	2866.000000	2875.000000	0.0
mean	475.750261	11.503138	22.136776	62.494261	NaN
std	277.675577	5.051702	7.128694	31.368745	NaN
min	38.000000	2.000000	5.000000	12.000000	NaN
25%	238.000000	8.000000	17.000000	41.000000	NaN
50%	366.000000	12.000000	22.000000	55.000000	NaN
75%	764.000000	15.000000	25.000000	78.000000	NaN
max	773.000000	49.000000	71.000000	269.000000	NaN

BAR CHART

```
import pandas as pd
from matplotlib import pyplot as plt

# Read CSV into pandas
data = pd.read_excel("Book1.xlsx")
data.head()
df = pd.DataFrame(data)

City= df['City'].head(4)
SO2 = df['SO2'].head(4)

# Figure Size
fig = plt.figure(figsize =(10, 7))

# Horizontal Bar Plot
plt.bar(City[0:10], SO2[0:10])

# Show Plot
plt.show()
```

OUTPUT



BOXPLOT

```
import seaborn as sns
sns. boxplot(df['SO2'])
sns. despine()
```