AIR QUALITY ANALYSIS AND PREDICTION IN TAMIL NADU

PROJECT OBJECTIVES:

Data Collection and Integration:

Gather historical air quality data from various monitoring stations across Tamil Nadu.

Integrate data sources to create a comprehensive dataset for analysis.

Data Preprocessing and Cleaning:

Clean and preprocess the collected data to handle missing values and outliers.

Standardize data formats and units to ensure consistency.

Exploratory Data Analysis (EDA):

Perform EDA to understand the spatial and temporal trends in air quality.

Identify key pollutants, sources, and hotspots.

Model Development:

Develop predictive models for air quality, considering factors like meteorological data, emission sources, and historical air quality.

Explore machine learning algorithms like regression, time series forecasting, or neural networks.

Real-Time Monitoring:

Integrate real-time data sources for ongoing monitoring of air quality.

Develop a dashboard or application for public access to current air quality information.

Spatial Analysis:

Analyze the spatial distribution of air quality, identifying areas with consistent poor air quality.

Investigate correlations with population density and industrial zones.

Temporal Analysis:

Examine seasonal and daily patterns in air quality.

Identify trends and anomalies over time.

Prediction and Early Warning System:

Develop a predictive model for short-term air quality forecasts.

Implement an early warning system to alert the public and authorities of potential air quality issues.

Health Impact Assessment:

Assess the health impact of poor air quality, considering the population at risk.

Provide recommendations for minimizing health risks.

Policy Recommendations:

Offer recommendations for policies and regulations to improve air quality.

Suggest mitigation strategies for common sources of pollution.

Community Engagement:

Engage with local communities to raise awareness about air quality issues.

Encourage public participation in monitoring and reporting air quality data.

Validation and Model Improvement:

Continuously validate the prediction models against actual air quality data.

Update and improve models as needed based on ongoing monitoring and feedback.

Documentation and Reporting:

Document the entire project process, including data sources, methodologies, and results.

Prepare reports and presentations for stakeholders, including government agencies, environmental organizations, and the public.

Sustainability and Long-Term Monitoring:

Ensure the project's sustainability by setting up a long-term monitoring system for ongoing air quality assessment.

Collaboration and Knowledge Sharing:

Collaborate with other research institutions and organizations working on air quality issues.

Share findings and insights with the broader scientific community.

These objectives provide a framework for an Air Quality Analysis and Prediction project in Tamil Nadu, with a focus on improving air quality, public health, and environmental sustainability in the region.

ANALYSIS APPROACH:

1. **Statistical Analysis**
2. **Time Series Analysis**
3. **Spatial Analysis**
4. **Machine Learning and Data Mining**
5. **Deep Learning and Neural Networks**
6. **Chemical Transport Models**
7. **Emission Source Apportionment**
8. **Quality Control and Quality Assurance (QA/QC)**
9. **Health Impact Assessment**
10. **Anomaly Detection**
11. **Real-Time Monitoring and Sensor Data Analysis**
12. **Cross-Disciplinary Collaboration**
13. **Policy and Regulation Analysis**
14. **Public Engagement and Social Science Research**
15. **Community Surveys and Feedback Analysis**

These analysis approaches are applied in a coordinated manner to provide a comprehensive understanding of air quality in Tamil Nadu, helping with prediction, mitigation, and informed decision-making to improve air quality and public health in the region.

Top of Form

VISUALIZE TECHNIQUE:

SCATTERPLOT

A scatterplot is a fundamental data visualization technique used in statistics and data analysis to represent the relationship between two variables. It is a graphical representation of data points in a two-dimensional coordinate system, with each data point plotted as a dot or a "scatter" on the graph. Scatterplots are especially useful for understanding the distribution of data, identifying patterns, and assessing the relationship between variables.

Here are some key components and characteristics of a scatterplot:

1. Axes: A scatterplot typically has two axes, often referred to as the x-axis and y-axis. Each axis represents one of the two variables being studied.
2. Data Points: Each data point is represented by a single dot or marker on the graph. The location of the dot is determined by the values of the two variables for that data point. For example, if you are studying the relationship between height and weight, each dot on the scatterplot would represent one individual, with their height on the x-axis and their weight on the y-axis.
3. Variables: The two variables being compared are often referred to as the independent variable (x-axis) and the dependent variable (y-axis). The independent variable is typically on the horizontal axis, while the dependent variable is on the vertical axis.
4. Patterns: The arrangement of data points on the scatterplot can reveal different patterns or relationships between the two variables. Some common patterns include:
   * Positive Relationship: When the data points tend to form an upward-sloping line, it suggests a positive correlation. This means that as one variable increases, the other variable tends to increase as well.
   * Negative Relationship: When the data points tend to form a downward-sloping line, it indicates a negative correlation. This means that as one variable increases, the other variable tends to decrease.
   * No Relationship: If the data points are scattered randomly and do not form any discernible pattern, it suggests that there is little to no correlation between the variables.
5. Outliers: Outliers are data points that are significantly different from the rest of the data and are usually located far away from the main cluster of points. Scatterplots make it easy to identify outliers, which can be important in understanding the data distribution.

Scatterplots are particularly useful for exploring data, identifying trends, and making initial assessments of correlation or causation between variables. They are commonly used in various fields, including statistics, economics, social sciences, and natural sciences, to visualize and analyze data relationships. Additionally, scatterplots can be enhanced with labels, colors, and other graphical elements to provide more information or context about the data being displayed.

CODE:

#import libraries

import pandas

#Dataset reading and activities

pandas.set\_option("display.max.rows",None)

pandas.set\_option("display.max.columns",None)

file\_data = pandas.read\_csv(r"Air quality analysis dataset.csv")

print(file\_data)

print(file\_data.head(200))

print(file\_data.tail(100))

print(file\_data.describe())

print(file\_data.info())

#Dataset cleaning

from sklearn.preprocessing import LabelEncoder

my\_le=LabelEncoder()

#Train and Testing

from sklearn.datasets import make\_classification

value1, y = make\_classification(

n\_features=6,

n\_classes=2,

n\_samples=800,

n\_informative=2,

random\_state=66,

n\_clusters\_per\_class=1,

)

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

value1, y, test\_size=0.33, random\_state=125

)

from sklearn.naive\_bayes import GaussianNB

model = GaussianNB()

model.fit(X\_train, y\_train)

predicted = model.predict([X\_test[6]])

print("Actual Value:", y\_test[6])

print("Predicted Value:", predicted[0])

from sklearn.metrics import (

accuracy\_score,

confusion\_matrix,

ConfusionMatrixDisplay,

f1\_score,

)

#Accuracy prediction

y\_pred = model.predict(X\_test)

accuray = accuracy\_score(y\_pred, y\_test)

print("Accuracy:", accuray)

import matplotlib.pyplot as plt

#visuailzation

plt.scatter(value1[:, 0], value1[:,1], c=y, marker="\*")

plt.show()

Output:

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

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2879 entries, 0 to 2878

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Stn Code 2879 non-null int64

1 Sampling Date 2879 non-null object

2 State 2879 non-null object

3 City/Town/Village/Area 2879 non-null object

4 Location of Monitoring Station 2879 non-null object

5 Agency 2879 non-null object

6 Type of Location 2879 non-null object

7 SO2 2868 non-null float64

8 NO2 2866 non-null float64

9 RSPM/PM10 2875 non-null float64

10 PM 2.5 0 non-null float64

dtypes: float64(4), int64(1), object(6)

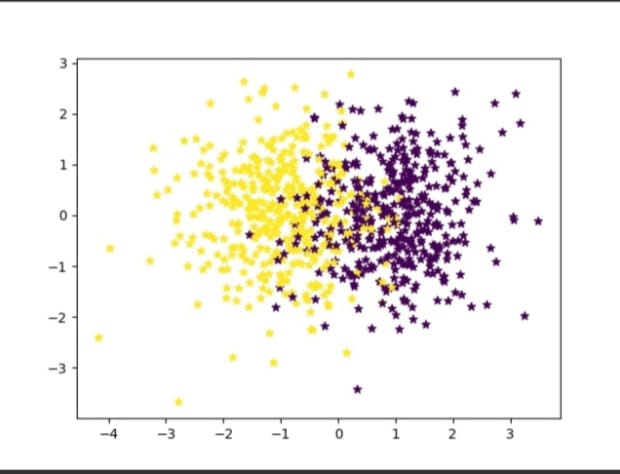
memory usage: 247.5+ KB

None

Actual Value: 1

Predicted Value: 1

Accuracy: 0.8977272727272727



***AIR POLLUTION TRENDS AND POLLUTION LEVELS IN TAMIL NADU***

To analyze air pollution trends and pollution levels in Tamil Nadu, several key steps and methods can be employed:

**Data Collection:**

Collect historical air quality data from various monitoring stations across Tamil Nadu. This data includes measurements of pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). Additionally, meteorological data should be gathered to assess its impact on air quality.

**Data Preprocessing:**

Clean and process the collected data to remove outliers, missing values, and inconsistencies. This ensures that the analysis is based on reliable information.

**Trend Analysis:**

Use statistical tools and techniques to identify long-term air pollution trends in Tamil Nadu. This can involve plotting time series data, calculating moving averages, and detecting seasonal patterns to understand how pollution levels have changed over the years.

**Spatial Analysis:**

Evaluate the spatial distribution of pollution levels by mapping air quality data across different regions of Tamil Nadu. This can help identify areas with consistently high or low pollution levels.

**Source Attribution:**

Employ source apportionment techniques, such as chemical transport modeling, to identify the major sources of air pollution in Tamil Nadu. This can help policymakers target specific sectors for mitigation efforts.

**Health Impacts Assessment:**

Assess the health impacts of air pollution by correlating pollution levels with public health data. This can highlight the consequences of poor air quality on the population's health.

**Policy Evaluation:**

Analyze the effectiveness of existing air quality regulations and policies in Tamil Nadu and recommend changes or improvements based on the insights gathered from the analysis.

**Future Projections:**

Use predictive modeling to forecast air pollution levels based on different scenarios, considering factors like population growth, industrial expansion, and climate change. This helps in planning for the future.

**Public Awareness:**

Communicate the findings and insights to the public through various channels, such as reports, websites, and public meetings, to raise awareness and encourage action to address air pollution.

By following these steps, an analysis can provide valuable insights into air pollution trends and pollution levels in Tamil Nadu, helping authorities and stakeholders make informed decisions to mitigate the impact of air pollution on public health and the environment.