AIR QUALITY AND POLLUTION ASSESSMENT

Table of Contents

CHAPTER –I

INTRODUCTION

1.1 Scope of analysis

The scope of this analysis is limited to historical data-driven predictions and does include real-time sensor data integration or live monitoring systems. Additionally, external factors such as industrial emissions, vehicular traffic, and regulatory changes, though considered in interpretation, will be directly model in the dataset. Despite these limitations, the study aims to provide a robust predictive framework for air quality analysis, contributing to environmental awareness and proactive pollution management strategies.

Air Quality and pollution Assessment allows for examining relationships between various pollutants (PM2.5, PM10) environmental factors (temperature, humidity) and social economic factors (population density). This can help and identify pollution hotspots and analyse the impact of specific factors on air quality. Analysing the relationship between air quality and the presence of industrial areas, whether pollution control regulations are effectively reducing pollution from industrial sources.

Assess the impact of various factors (Industrial activity, population density) on pollution levels Identify patterns or anomalies in pollution data across different regions or time periods.Identification of key pollution contributors affecting air quality.

Prediction of air quality levels for different locations based on environmental factors. Model recommendations for improving classification performance he findings will provide insights into policy recommendations, environmental monitoring strategies, and public health implications**.** The study excludes real-time sensor integration and focuses on historical data driven predictions.

1.2 Approach of analysis

Data Understanding and Cleaning:

Load and inspect the dataset for missing values and anomalies.

Ensure proper encoding of categorical variables and scaling of numerical data for uniformity.

Data Exploration and Pre-processing:

Data Cleaning: Handle missing values, outliers in the dataset.

All data columns are complete, with no missing values.

Data Visualization: Create visualizations (e.g., Bar plot, plot, scatter plots, pie chart) to understand the distribution and relationships between variables.

Exploratory Data Analysis (EDA):

Visualize distributions of air quality categories. Analyze relationships between environmental features and air quality through statistical charts.

- Identifying key air pollutants (e.g., PM2.5, NO2, CO, SO2)
- Handling missing values, outliers, and normalizing data.
- Time-series or location-based segmentation of data.

Model Building:

Split data into training and testing sets to evaluate model performance. Use a Logistic Regression model to classify air quality based on input features And use a random Forest classfier to predict the base of level of pollution levels. Perform hyper parameter tuning to optimize model performance, Implimented python programming Language.

CHAPTER – II

GATHERING DATA

Gathering Data

Load the relevant Packages

```
import pandas as pd
import seaborn as sns
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Load the Dataset

```
pollution_data =pd.read_csv(r"C:\Users\siddh\Downloads\updated_pollution_dataset.csv")
pollution_data
```

Structure of Data

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density	Air Quality
0	29.8	59.1	5.2	17.9	18.9	9.2	1.72	6.3	319	Moderate
1	28.3	75.6	2.3	12.2	30.8	9.7	1.64	6.0	611	Moderate
2	23.1	74.7	26.7	33.8	24.4	12.6	1.63	5.2	619	Moderate
3	27.1	39.1	6.1	6.3	13.5	5.3	1.15	11.1	551	Good
4	26.5	70.7	6.9	16.0	21.9	5.6	1.01	12.7	303	Good
...	...	...	...	...	...	...	...	...	...	...
4995	40.6	74.1	116.0	126.7	45.5	25.7	2.11	2.8	765	Hazardous
4996	28.1	96.9	6.9	25.0	25.3	10.8	1.54	5.7	709	Moderate
4997	25.9	78.2	14.2	22.1	34.8	7.8	1.63	9.6	379	Moderate
4998	25.3	44.4	21.4	29.0	23.7	5.7	0.89	11.6	241	Good
4999	24.1	77.9	81.7	94.3	23.2	10.5	1.38	8.3	461	Moderate

2.1 Data Description

The pollution_ data dataset has 5000 rows and 10 columns

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density	Air.Quality
1	29.8	59.1	5.2	17.9	18.9	9.2	1.72	6.3	319	Moderate
2	28.3	75.6	2.3	12.2	30.8	9.7	1.64	6.0	611	Moderate
3	23.1	74.7	26.7	33.8	24.4	12.6	1.63	5.2	619	Moderate
4	27.1	39.1	6.1	6.3	13.5	5.3	1.15	11.1	551	Good
5	26.5	70.7	6.9	16.0	21.9	5.6	1.01	12.7	303	Good
6	39.4	96.6	14.6	35.5	42.9	17.9	1.82	3.1	674	Hazardous
7	41.7	82.5	1.7	15.8	31.1	12.7	1.80	4.6	735	Poor
8	31.0	59.6	5.0	16.8	24.2	13.6	1.38	6.3	443	Moderate
9	29.4	93.8	10.3	22.7	45.1	11.8	2.03	5.4	486	Poor
10	33.2	80.5	11.1	24.4	32.0	15.3	1.69	4.9	535	Poor
11	26.3	65.7	1.3	5.5	18.3	5.9	0.85	13.0	529	Good
12	32.5	51.2	1.6	10.5	21.6	19.3	1.53	5.9	519	Moderate
13	20.0	53.3	3.7	12.9	26.1	6.6	1.09	10.2	538	Good
14	28.6	53.7	28.9	34.0	23.2	4.5	1.02	11.0	508	Good
15	22.3	80.5	4.5	12.0	17.2	6.3	1.18	10.4	232	Good
16	32.0	78.9	22.4	29.9	27.5	11.8	1.48	7.9	444	Moderate
17	22.9	75.4	4.5	10.4	18.4	3.7	0.96	14.4	359	Good
18	37.6	81.2	28.1	56.6	46.7	13.7	1.85	4.1	560	Poor
19	34.7	59.3	9.0	15.7	28.5	7.1	1.52	6.1	437	Moderate
20	37.8	97.2	0.6	24.6	37.1	11.7	1.13	7.7	803	Poor
21	27.6	44.1	3.5	14.4	30.7	9.4	0.97	8.0	338	Moderate

This pollution _ data dataset contains air quality measurements collected from various locations. This data includes environmental factors, pollutant concentrations, and population-related attributes that influence air quality levels.

Temperature

Temperature variable typically represents the atmospheric temperature at the time of measurement.

Humidity

Humidity typically refers to the amount of water vapour in the air present compared to the maximum amount the air can hold at a specific temperature.

PM2.5

PM2.5 refers to Particulate Matter (PM) with a diameter of 2.5 micro meters or smaller.

PM10

PM10 refers to Particulate Matter (PM) with a diameter of 10 micro meters or smaller.

NO2

NO2 (Nitrogen Dioxide) represents the concentration of measured over a specific period and location.

S02

SO2 (Sulfur Dioxide) represents the concentration Parts per million (ppm) colourless and toxic gas major contribute of air pollution.

CO

CO (Carbon monoxide) represents the concentration Parts per million (ppm) a colourless, Odourless and poisonous gas.

Proximity to Industry Areas

It refers to the distance or closeness of a specific location to areas that are designated for industrial.

Population Density

Population Density is process of determining no of people in given area, air pollution presence of harmful substances in the air.

Air Quality

Air Quality represents the overall assessment of air pollution levels and the potential health effects.

2.2 Understanding Data

Data Cleaning

```
label_encoder = LabelEncoder()
pollution_data['Air Quality'] = label_encoder.fit_transform(pollution_data['Air Quality'])
pollution_data
```

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density	Air Quality
0	29.8	59.1	5.2	17.9	18.9	9.2	1.72	6.3	319	2
1	28.3	75.6	2.3	12.2	30.8	9.7	1.64	6.0	611	2
2	23.1	74.7	26.7	33.8	24.4	12.6	1.63	5.2	619	2
3	27.1	39.1	6.1	6.3	13.5	5.3	1.15	11.1	551	0
4	26.5	70.7	6.9	16.0	21.9	5.6	1.01	12.7	303	0
...	...	...	...	...	...	...	...	...	...	...
4995	40.6	74.1	116.0	126.7	45.5	25.7	2.11	2.8	765	1
4996	28.1	96.9	6.9	25.0	25.3	10.8	1.54	5.7	709	2
4997	25.9	78.2	14.2	22.1	34.8	7.8	1.63	9.6	379	2
4998	25.3	44.4	21.4	29.0	23.7	5.7	0.89	11.6	241	0
4999	24.1	77.9	81.7	94.3	23.2	10.5	1.38	8.3	461	2

5000 rows × 10 columns

Categorical values count: 1 (4 values)

Continuous values count: 9

To show the categories of Categorical variable:

Air Quality: quality of air, categorized into levels such as Good, Moderate, Poor, and Hazardous (Y variable)

To show the categories of Continues variable:

1. **PM2.5**: particulate matter smaller than 2.5 micrometers in the air
2. **PM10**: particulate matter smaller than 10 micrometers in the air
3. **NO2 (Nitrogen Dioxide)**: Concentration of NO2 in the air
4. **SO2 (Sulfur Dioxide)**: Concentration of SO2 in the air
5. **CO (Carbon Monoxide)**: Concentration CO in the air
6. **Temperature**: Ambient atmospheric temperature the time of measurement
7. **Humidity**: Percentage of humidity in the air
8. **Population Density**: The number of people per s q /km
9. **Proximity to Industrial Areas**: Whether the location is near an industrial area (e.g., Yes/No) but here is in numerical format.

We here used Data information function in order to understand the data types exist in the dataset if it object an characters then it is a categorical variable value or numerical values.

If integer or float then we can conclude with continuous value.

Data Structure

8 columns are of type float64 (continuous numerical features).

1 column (Population _ Density) is of type int64.

1 column (Air Quality) is of type int32, because it's a target variable (categorical or ordinal).

```
pollution_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Temperature                5000 non-null   float64
 1   Humidity                   5000 non-null   float64
 2   PM2.5                      5000 non-null   float64
 3   PM10                       5000 non-null   float64
 4   NO2                        5000 non-null   float64
 5   SO2                        5000 non-null   float64
 6   CO                         5000 non-null   float64
 7   Proximity_to_Industrial_Areas  5000 non-null   float64
 8   Population_Density         5000 non-null   int64
 9   Air Quality                5000 non-null   int32
dtypes: float64(8), int32(1), int64(1)
memory usage: 371.2 KB
```

Targeted variables and its sources:

Good: Clean air with low pollution levels.

Moderate: Acceptable air quality but with some pollutants present.

Poor: Noticeable pollution that may cause health issues for sensitive groups.

Hazardous: Highly polluted air posing serious health risks to the population.

The Target values are category types which its difficult to find one by one so, we can change into binary numerical values such as Good:0, Moderate:1,Poor:2, Hazardous:3.

Good Air Quality

Air quality classified as Good indicates clean and healthy air with minimal pollutant levels. In this condition, concentrations of PM2.5, PM10, NO2, SO2, and CO are well within safe limits, posing no significant health risks. People can engage in outdoor activities without concerns about air pollution. Such conditions are commonly observed in low-density residential areas, green zones, and regions far from industrial emissions or heavy traffic**.**

Moderate Air Quality

Moderate air quality means the air is acceptable but contains some pollutants that may be slightly elevated. While this level is not harmful to the general public**,** sensitive groups, such as individuals with respiratory conditions, may experience minor discomfort. Factors like increased vehicle emissions, industrial activity, and seasonal changes can contribute to moderate pollution levels.

Poor Air Quality

When air quality is categorized as Poor**,** pollution levels are noticeable and may cause health **issues**, especially for sensitive groups such as children, the elderly, and individuals with lung diseases. High concentrations of PM2.5 and NO2 are typically responsible for these conditions, often resulting from traffic congestion, industrial emissions, and unfavorable weather patterns that trap pollutants**.**

Hazardous Air Quality

Hazardous air quality represents severe pollution levels, posing **s**erious health risks to the entire population**.** In such conditions**,** PM2.5 and PM10 concentrations are extremely high, significantly impacting lung function and increasing the risk of cardiovascular diseases. Dense urban areas, heavy industrial zones, and wildfire-affected regions frequently experience hazardous air. Authorities often issue health advisories, urging people to stay indoors and wear protective masks when venturing outside.

CHAPTER – III

PREPARING AND EXPLORING DATA

3.1 Data Exploration

- This Pollution Analysis and Forecasting dashboard provides a detailed exploration of air pollution data**,** Analysing pollutants (NO2, CO, SO2, PM 2.5, PM 10) air quality, population density, humidity, and proximity to industrial areas.

- **Air Quality Distribution**
- **Correlation Heat map**
- **Population Density vs Proximity to Industrial Areas**

- The dataset appears complete, with no missing values. A preliminary glance suggests potential relationships between pollution levels and industrial proximity or population density. Further analysis could explore these correlations and their impact on air quality.

- The Potential outliers detected in pollutant concentrations, likely representing highly polluted areas.

- These charts are statistical enough to help us understand the data how its trend, and can get ourself ready to give data driven solution with useful insights The dataset also highlights notable disparities in population density, ranging from 188 to 957 people per square kilometer, which may influence pollution levels.

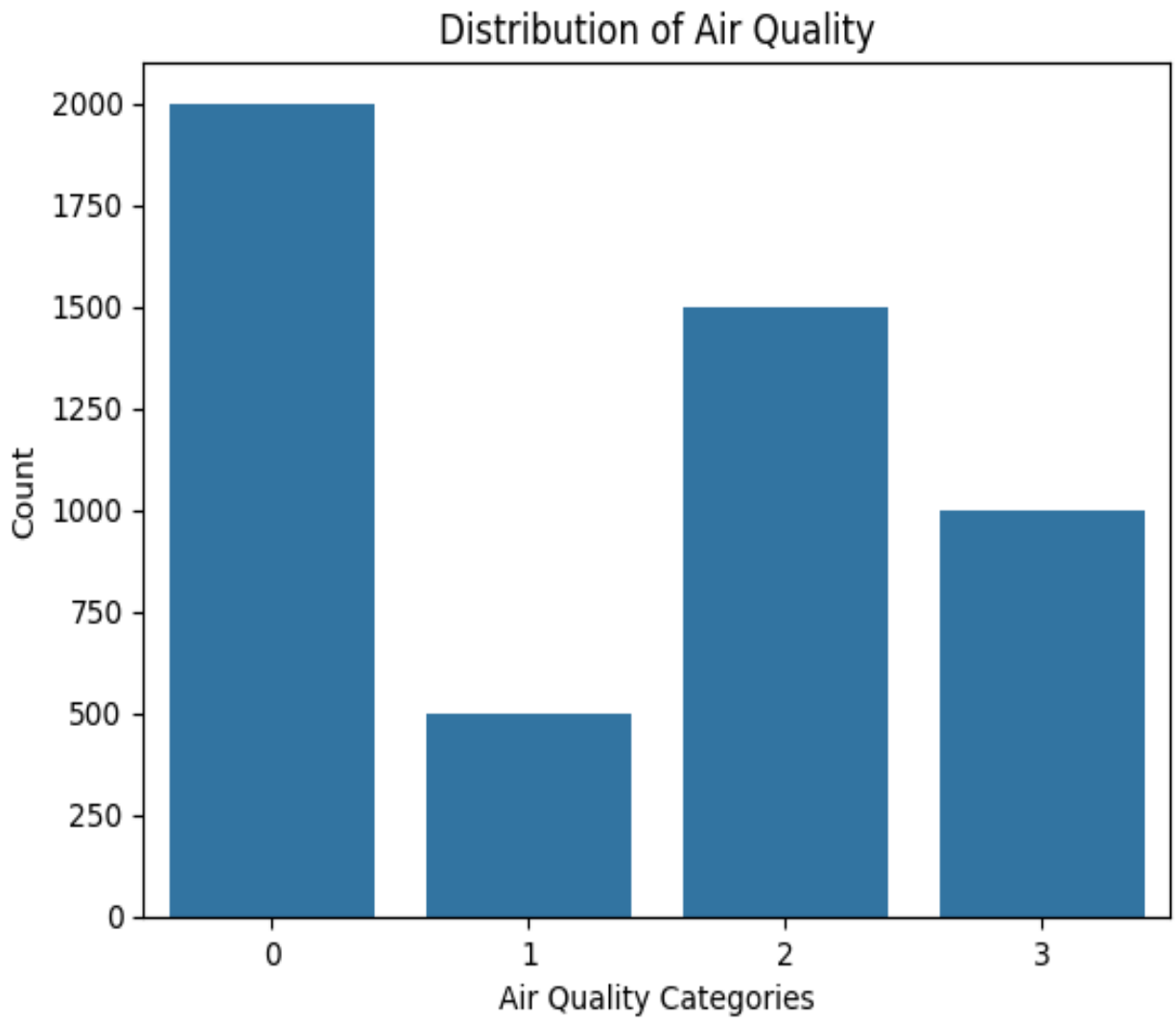
- The presence of negative values in PM10 and SO2 suggests possible data entry errors or measurement inconsistencies, requiring further cleaning.

Air Quality Distribution:

Shows the number of samples for each air quality category (Good, Moderate, poor, Hazardous).

Helps identify if the data is balanced across categories.

```
sns.countplot(x='Air Quality', data=pollution_data)
plt.title('Distribution of Air Quality')
plt.xlabel('Air Quality Categories')
plt.ylabel('Count')
plt.show();
```

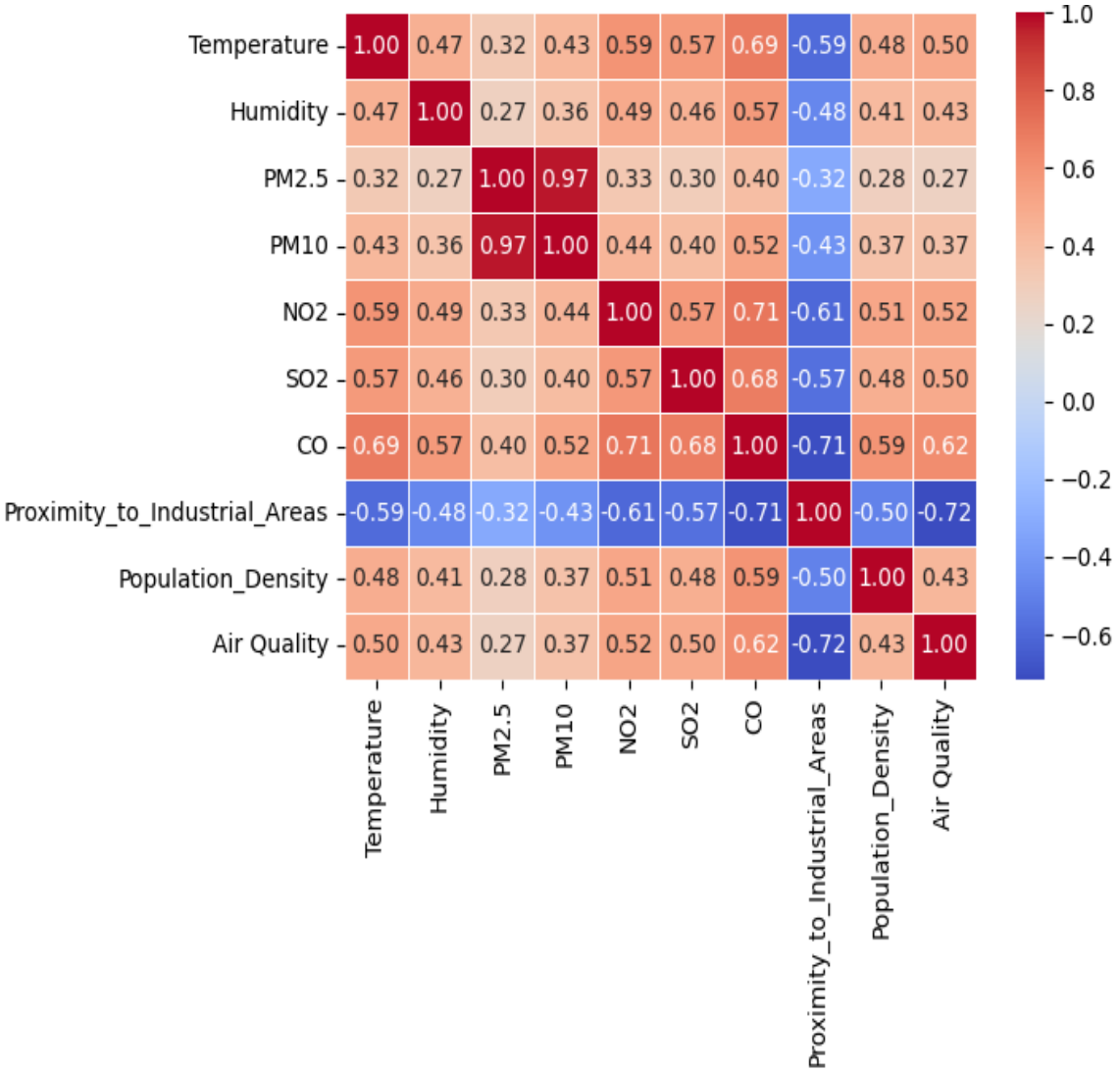


Correlation Heat map:

Displays how numerical features relate to each other.

Example: Higher Proximity _to_ Industrial _ Areas might correlate with higher pollutant levels.

```
corr_matrix = pollution_data.corr()
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap of Features')
plt.show();
```

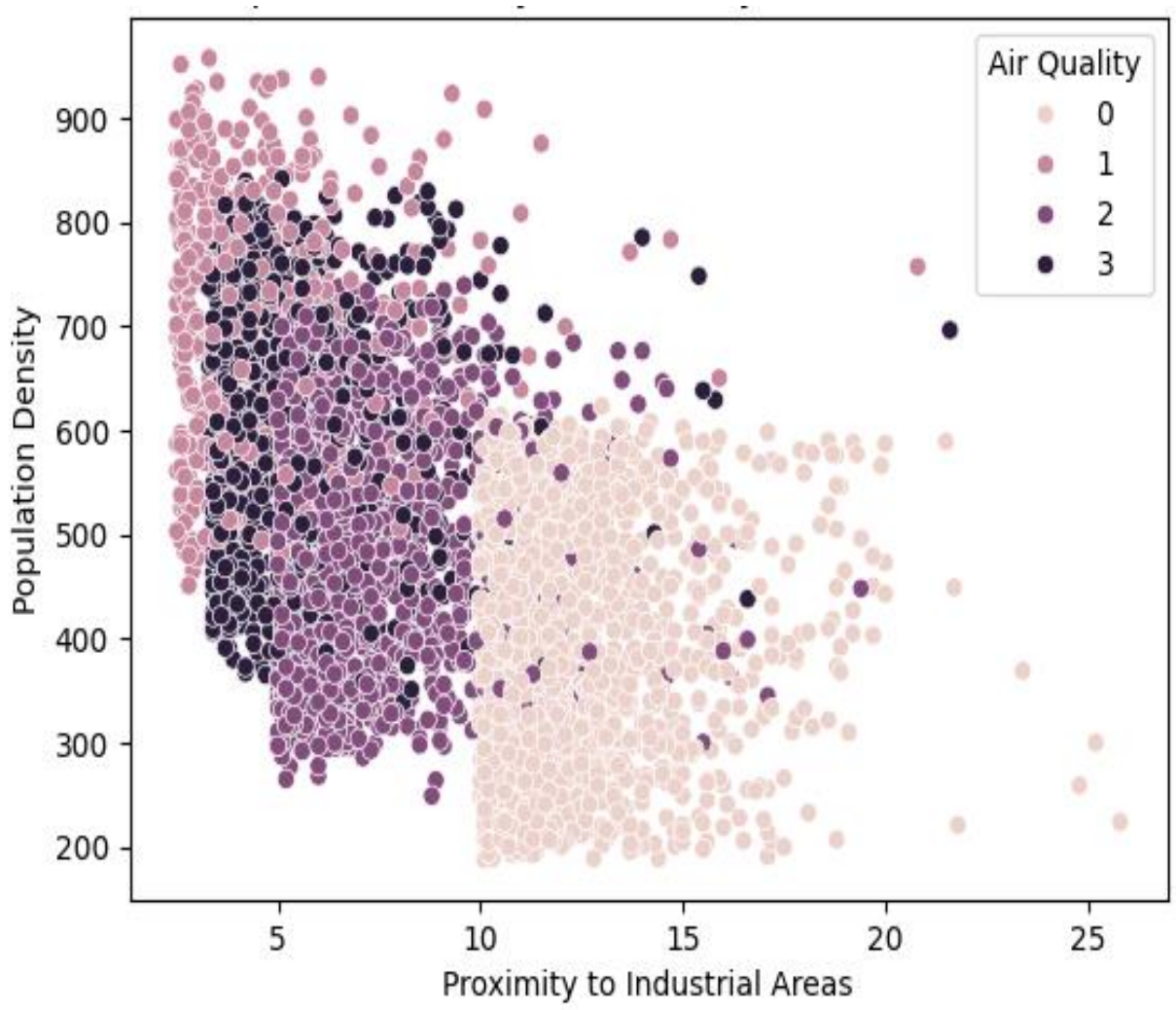


Population Density vs Proximity to Industrial Areas:

Scatterplot to understand spatial distribution of population vs industrial proximity.

Differentiates air quality zones.

```
sns.scatterplot(x='Proximity_to_Industrial_Areas', y='Population_Density', hue='Air Quality', data=pollution_data)
plt.title('Population Density vs Proximity to Industrial Areas')
plt.xlabel('Proximity to Industrial Areas')
plt.ylabel('Population Density')
plt.show();
```



3.2 Issues in the Dataset

SO2 and PM10 had negative values These values were replaced with nan (missing values) to indicate invalid data Outliers in PM2.5 and PM10 were capped at the upper bound of the Interquartile Range.

3.3 Resolving Issues

Real-Time Issue:

Air Pollution Management: Air pollution poses severe health risks, contributing to respiratory diseases, cardiovascular problems, and environmental damage. Traditional monitoring methods often lack predictive power and actionable insights for timely interventions.

Purpose of the Model & Dashboard:

1. Model's Purpose:

The classification model predicts Air Quality levels (0 to 3: Good, Moderate, Poor, Severe) based on pollution metrics and weather conditions.

- Helps authorities anticipate poor air quality and take preventive actions (e.g., traffic restrictions, factory output control).

2. Dashboard's Purpose:

The Power BI Dashboard provides an interactive visualization of air quality data.

- Allows real-time monitoring and identification of pollution trends by season, weather conditions, and industrial proximity.

- Air Quality Classification – Categorizes air quality into Good, Moderate, Poor, and Hazardous.
- Pollutant Analysis – Displays CO, NO2, and SO2 levels with breakdowns by air quality classification.
- PM2.5 Levels – Highlights the minimum and maximum PM2.5 concentrations
- Humidity Insights – Shows the average and maximum humidity, which can affect pollution levels.
- Population Exposure – Analyzes air quality across different population densities
- Industrial Influence – Shows the relationship between industrial areas and air pollution.
- Adjustable Filters – Provides sliders to explore different pollution levels dynamically.

How Air pollution is affecting:

Poor air quality is particularly dangerous for vulnerable groups, including children, the elderly, and individuals with pre-existing health conditions. Additionally, pollution negatively affects the environment by damaging crops, contaminating water sources, and contributing to climate change through greenhouse gas emissions. Industrial activities, vehicle emissions, and urbanization are major contributors to deteriorating air quality, making it essential for governments and organizations to implement policies aimed at reducing pollution levels and promoting sustainable practices.

How pollution can prevent:

Reducing Industrial Emissions – Industries should adopt cleaner technologies, use renewable energy sources, and install pollution control devices like scrubbers and filters to minimize harmful emissions.

How This Solves the Problem:

1. **Early Warning System:**

   The model's high 93% accuracy ensures reliable air quality predictions.

- ❖ Potential to trigger alerts on days with predicted poor air quality.

2. **Data-Driven Decisions :**

   The dashboard offers statistical insights into pollution sources.

- ❖ Authorities can implement targeted actions, like closing industrial activities or limiting vehicle emissions during high-risk periods.

3. **Public Awareness :**

   The dashboard can be shared publicly to promote community-level awareness.

- ❖ Citizens can plan activities based on air quality forecasts, reducing Exposure to pollutants.

Problem Resolution Approach:

1. **Identify**: Real-time monitoring through the dashboard.
2. **Predict**: Model-based classification of air quality.
3. **Act**: Initiate control measures based on predictions and dashboard insights.
4. **Evaluate**: Continuously improve the model with new data.

3.4 Issue addressed after analysis

The analysis helps easy to analyses the air quality which people are live or in streaming environment by the pollutant Areas, No missing values were found, ensuring a complete dataset for analysis.

- No duplicate records, eliminating redundancy and improving model efficiency Transformed the Air Quality variable into numerical format using Label Encoding, making it compatible with machine learning algorithms, Industrial proximity and pollutant levels are the strongest predictors of poor air quality.

- Predictions indicate that high industrial proximity and urban density lead to poorer air quality**.** Environmental policy-making, urban planning, and pollution control measures**.**

- The box plot indicates the over layouts of the Temperature and air quality regions.

Solutions Implemented

Exploratory data analysis (EDA) was conducted to verify the balance of target labels, ensuring that no single category dominated the dataset. Techniques such as oversampling or under sampling could be applied if necessary. The variability in numerical data scales was resolved by standardizing numerical features, which brought all variables to a comparable range, improving model training and convergence Solutions Implemented.

Data Resolving is the process of addressing inconsistencies, errors, and conflicts within a dataset to ensure its accuracy and reliability for analysis. It involves correcting issues such as duplicate records, incorrect values, data mismatches, and inconsistencies across different sources. Resolving data problems is crucial because poor data quality can lead to misleading insights and inaccurate predictions, especially in machine learning and analytics.

Another issue was variability in categorical data scales, as air quality-related parameters (such as pollutant concentrations, humidity, and temperature) exist in different units and ranges nitrogen oxides reacting with moisture in the atmosphere, can damage crops, forests, and aquatic ecosystems. Additionally, certain pollutants, such as ground-level ozone and carbon dioxide (CO_2), play a significant role in climate change by trapping heat in the Earth's atmosphere and leading to global warming.

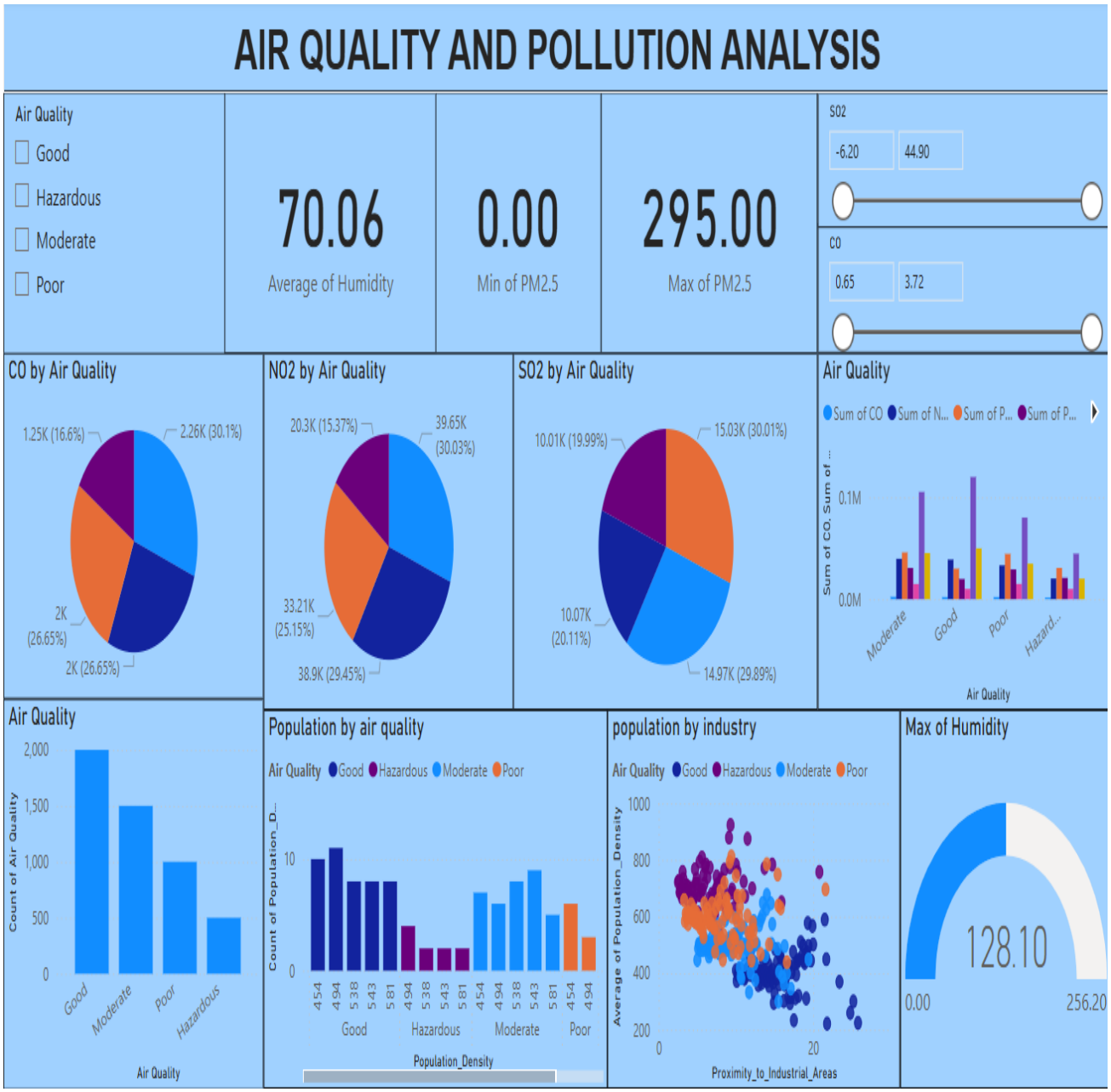
CHAPTER – IV

BUSINESS INTELLIGENCE INTERACTIVE DASHBOARDS

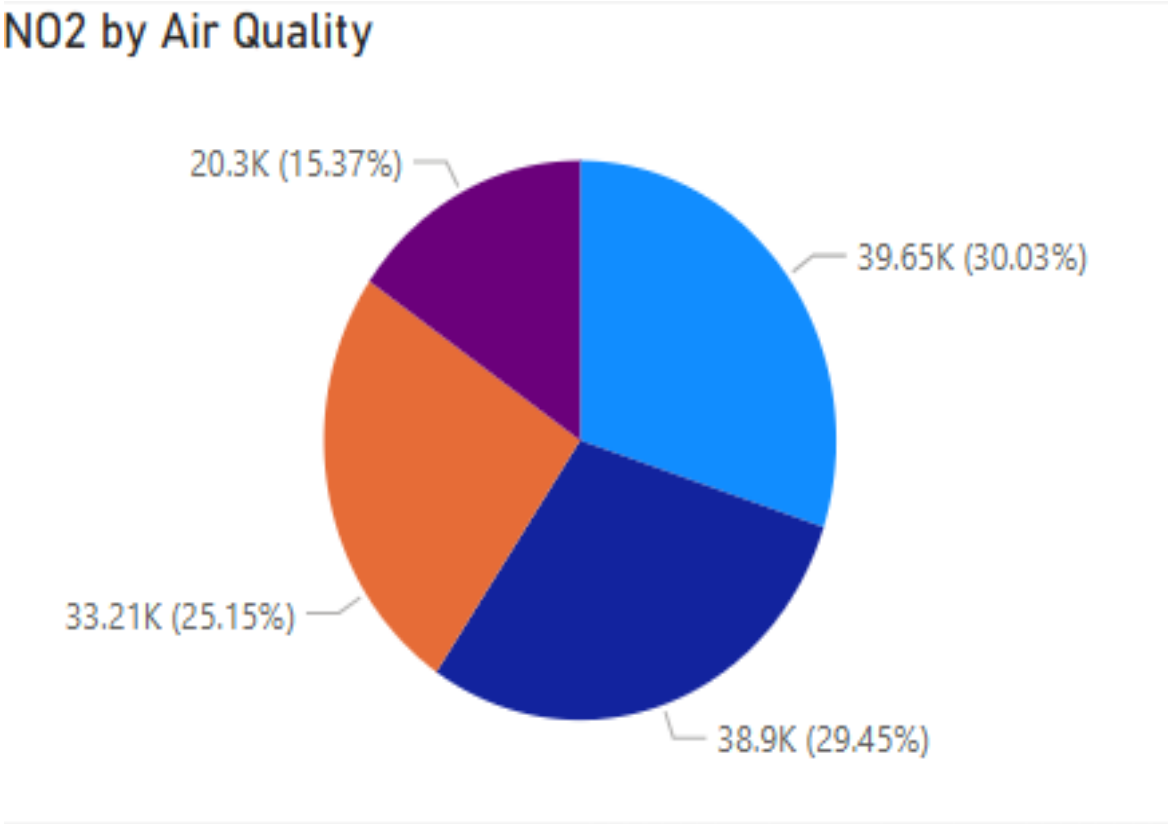
Summary

Majority of the data falls under Moderate and Good smaller percentage of the data is classified as Poor and Hazardous NO_2, CO, and SO_2 levels are higher in Poor and Hazardous air quality zones, CO and NO_2 pollution in Moderate and Poor air quality Pollution is higher near industrial areas and in high-population zones**.**

4.1 Dashboard Interpretation



NO2 BY AIR QUALITY

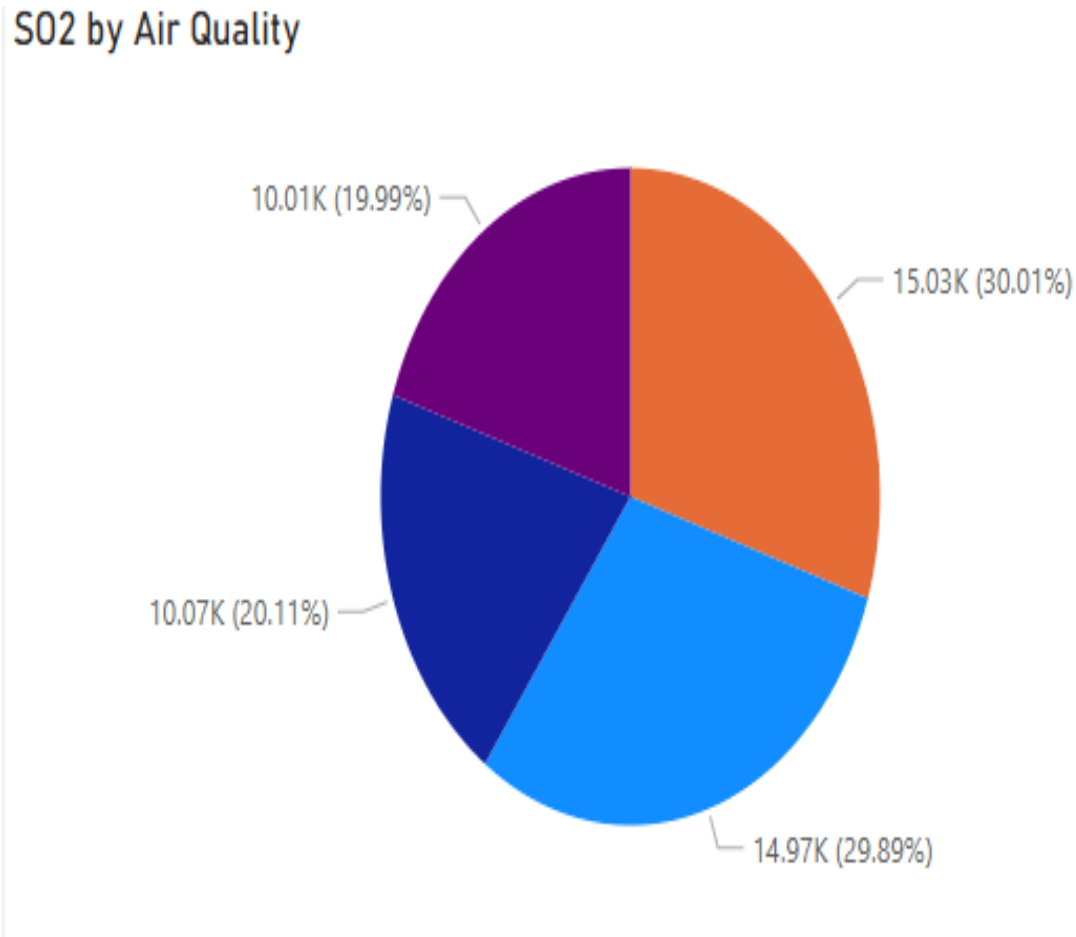


The pie chart represents the distribution of NO2 (Nitrogen Dioxide) levels in air quality.

30.03% (39.65K) - The largest segment indicating the highest NO2 level in a specific air quality category.

29.45% (38.9K) - The second largest, showing a slightly lower NO2 concentration.
25.15% (33.21K) - Moderate NO2 levels.
15.37% (20.3K) - The smallest proportion, indicating the least NO2 concentration in air quality Measurements.

SO2 BY AIR QUALITY



The pie chart represents the distribution of SO_2 (Sulfur Dioxide) levels in air quality.

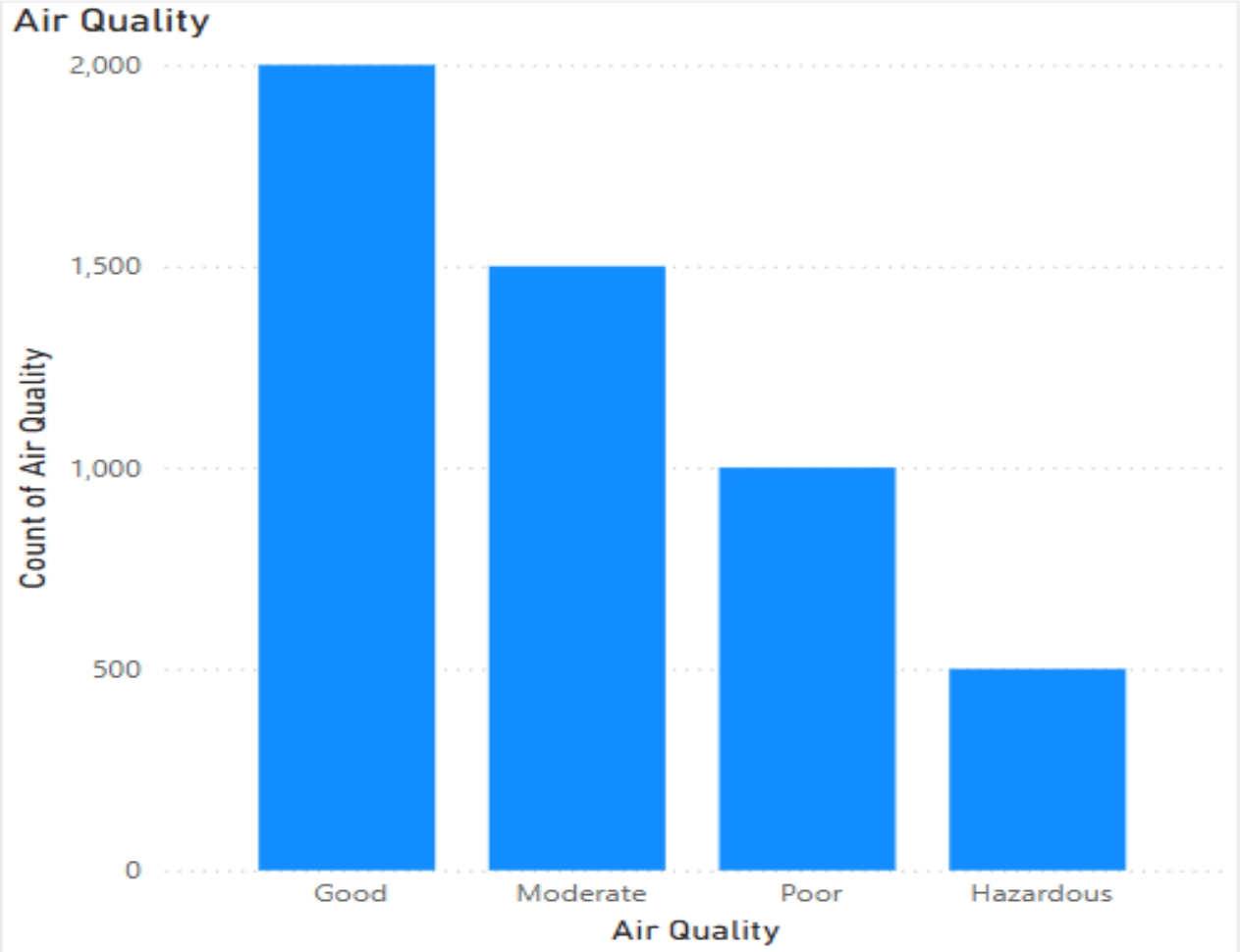
30.01% (15.03K) - The largest segment, representing the highest SO_2 concentration in a specific air quality category.

29.89% (14.97K) - The second-largest category, closely following the highest segment.

20.11% (10.07K) - A moderate SO_2 level in another air quality category.

19.99% (10.01K) - The smallest segment, indicating the lowest SO_2 concentration .

AIR QUALITY DISTRIBUTION



This bar chart shows the number of samples for each air quality category

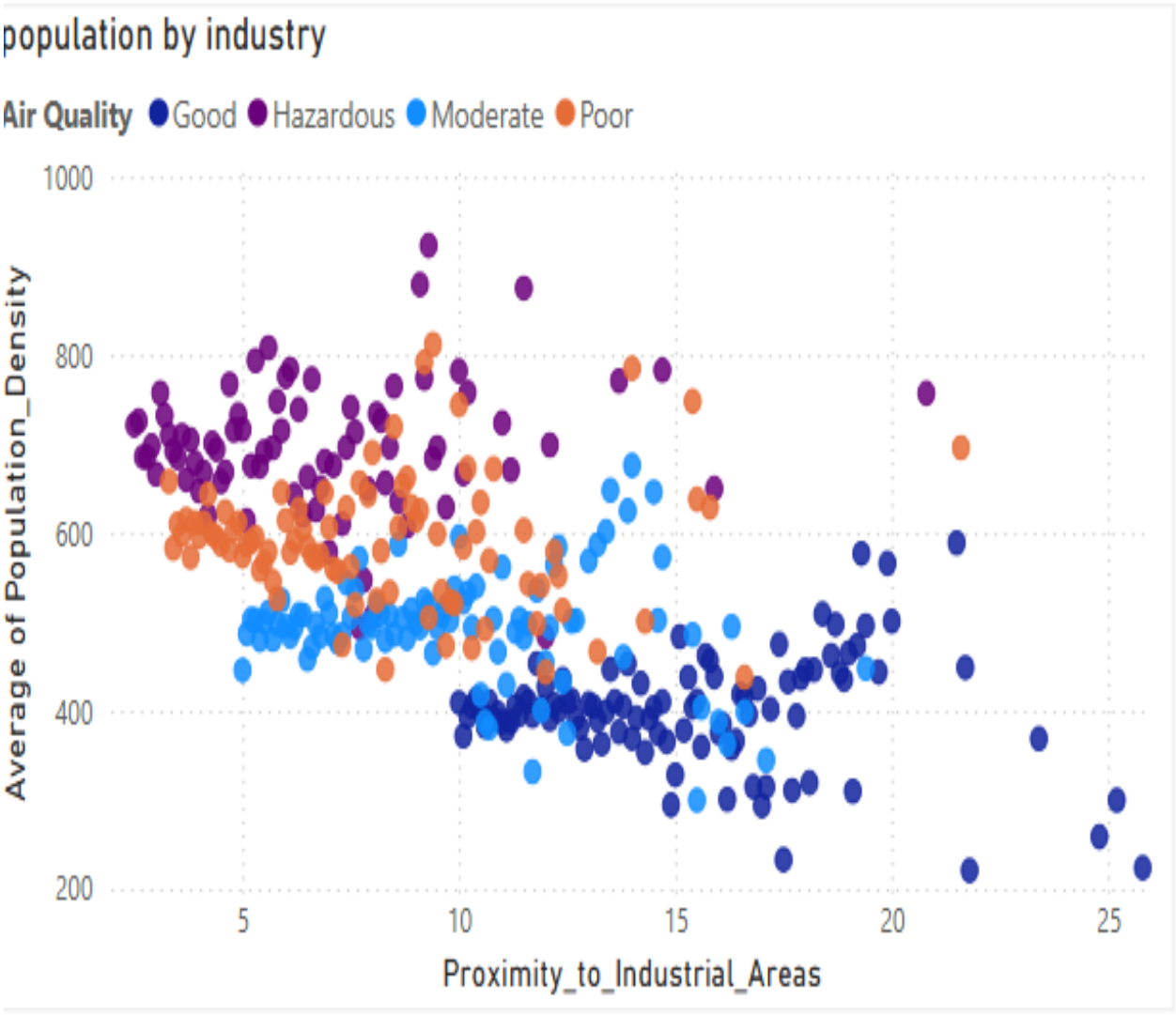
(Good, Moderate, Poor, Harzadous)

Moderate air quality count of around 1500 air pollution is present but not at dangerous levels.

This indicates the count of Air Quality levels based on the bar heights.

Poor air quality is recorded around 1000 portion of the data falls into unhealthy conditions.

AVERAGE OF POPULATION INDUSTRY

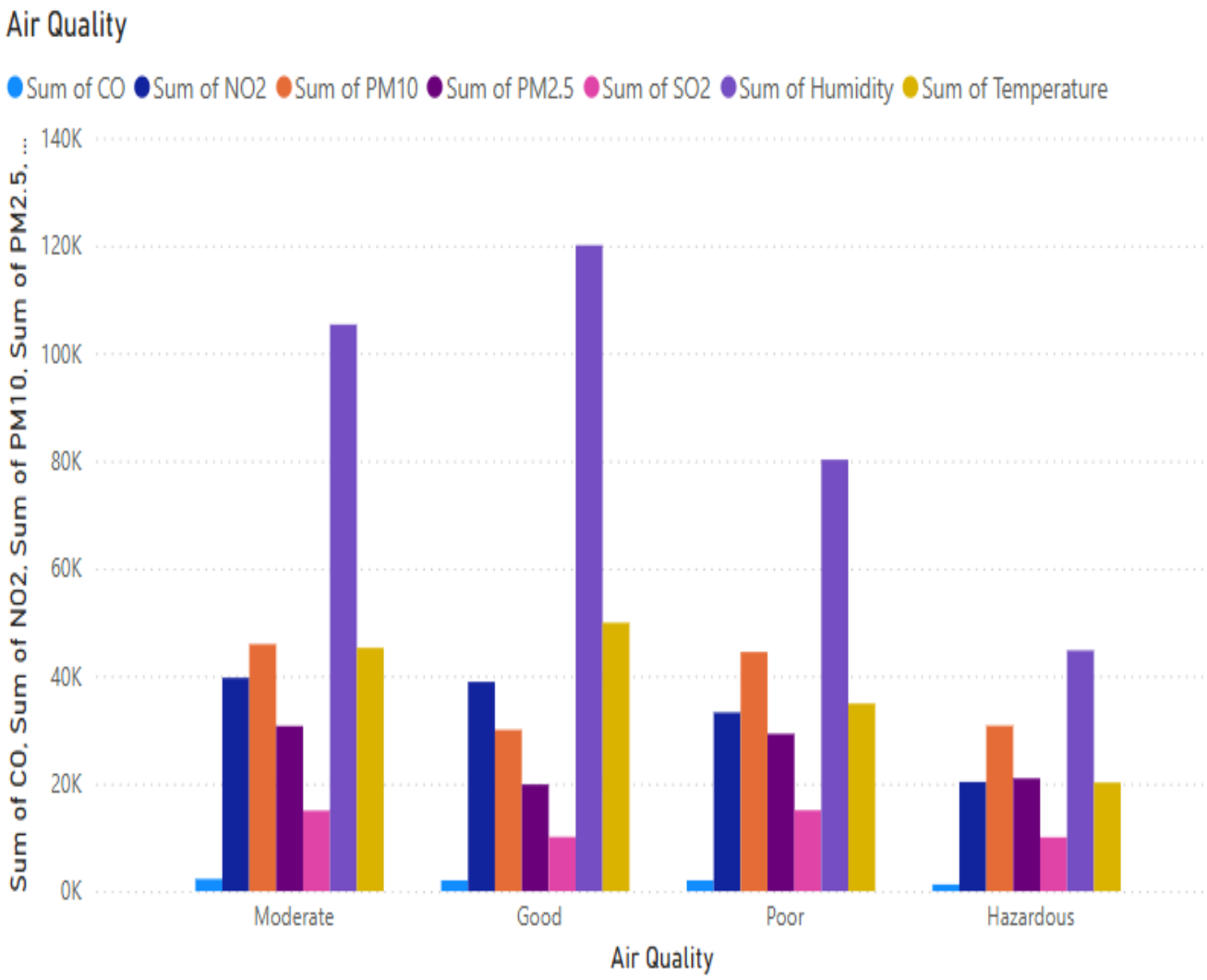


The scatter plot represents the relationship between proximity to industrial areas and average population density while considering air quality levels.

Hazardous (purple) and Poor (orange) air quality dominates, indicating higher pollution levels in densely populated regions near industries**.**

Good (dark blue) air quality is dominant**,** indicating lower pollution levels in less populated, distant areas.

SUM OF AIR QUALITY BY ALL COLUMNS



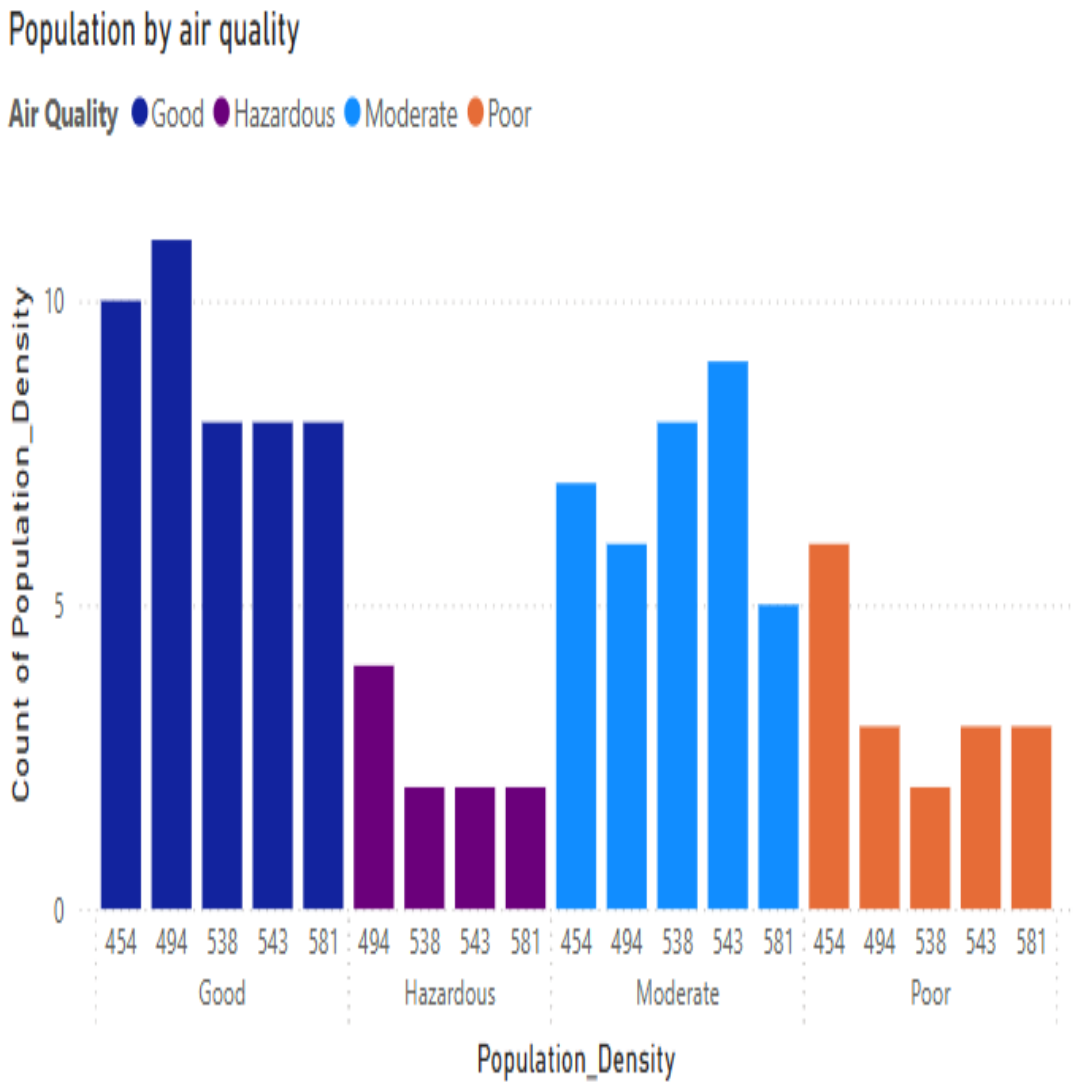
The sum of humidity (purple bar) is higher in all air quality categories, especially in the "Good" category.

The PM10 (orange) and PM2.5 (magenta) levels appear to be relatively high across all air quality categories, showing their strong impact on air pollution.

NO2 (dark violet) and CO (blue) show moderate levels across different air quality categories.

SO2 (pink) remains the lowest among all pollutants, The temperature (yellow bars) remains somewhat consistent across all air quality levels.

POPULATION BY AIR QUALITY



This chart categories of air quality into Good (violet), Hazardous (Purple), Moderate (Light Blue), and Poor (Orange) based on population density.

The highest bars are "Good" air quality category, indicating that most population densities experience good air quality the smallest number seen in the “Hazardous” category, only a few population density values suffer from extreme pollution.

CHAPTER – V
MODEL BUILDING

5.1 Algorithm

LOGISTIC REGRESSION

Logistic Regression is a statistical method used for binary classification problems, where the dependent variable (target) has only two possible outcomes, such as yes/no, 0/1, true/false. Instead of fitting a straight line like linear regression, logistic regression applies a sigmoid function (S-curve) to predict the probability of belonging to a particular class. utilizing a sigmoid function to map input values to probabilities between 0 and 1 Logistic Regression is based on the concept of estimating the probability that a given input belongs to a particular class using the logistic (sigmoid) function, which maps predictions to values between 0 and 1. Due to its simplicity, interpretability, and efficiency, it remains a popular choice in various domains, including healthcare, finance, marketing, and social sciences.

RANDOM FOREST

Random Forest is a robust, flexible, and powerful algorithm that excels in a wide range of applications. Whether it's used for classification, regression, anomaly detection, or even feature selection, its ability to handle complex data, reduce overfitting, and provide interpretable results makes it an essential tool in machine learning. It's used across many industries, including healthcare, finance, e commerce, manufacturing, and more, for a variety of tasks such as fraud detection, predictive analytics, sales forecasting, and personalized recommendations, In classification problems, it assigns a majority vote among decision trees to determine the final class, making it highly effective for tasks like disease diagnosis, fraud detection, and sentiment analysis. For regression tasks, it averages the output of individual trees to predict continuous values, making it useful in applications like stock price prediction, weather forecasting, and air quality analysis The algorithm is computationally efficient and scalable, making it suitable for large-scale machine learning applications.

LOGISTIC REGRESSION

- Logistic Regression is a supervised learning algorithm used for classification tasks.
- Fast & Efficient Works well with small to medium datasets.
- Simple and easy to implement L2 Regularization (Ridge Regression) Penalizes large coefficients to prevent overfitting.
- Accuracy is to Measures how often the model predicts correctly

RANDOM FOREST

- Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
- Handles Missing Data & Outliers well.
  Works with Large Datasets and high-dimensional data.
  More Accurate & Stable than individual trees.
- Random Forest builds multiple decision trees using random samples of the data. Each tree is trained on a different subset of the data which makes each tree unique.
- Predict future air pollution levels based on historical data. Useful for environmental monitoring and early warning systems.

5.2 Training and Test Dataset

LOGISTIC REGRESSION

70% Training Data and 30% Testing Data

Training Set: Used to train the logistic regression model.

Testing Set: Used to evaluate model performance.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
X = df.drop(columns=['Air Quality'])
y = df['Air Quality']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

RANDOM FOREST

70% Training Data and 30% Testing Data

Training Set: Used to train the random forest classifier model.

Testing Set: Used to evaluate model performance.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

5.3 Model

LOGISTIC REGRESSION

Standardize the Features

Fit _ transform (X _ train): scaling parameters from training data and applies transformation.

Transform (X_ test): Applies the same transformation to test data

Standard Scaler normalizes the dataset by making the mean = 0 and standard deviation = 1.
It helps improve model performance, especially for algorithms that rely on distance-based calculations (Logistic Regression)

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

RANDOM FOREST

Standardize the Features

Standard Scaler standardizes features by removing the mean and scaling to unit variance.

Fit _ transform(X): Computes mean & standard deviation

Standard Scaler() - Initializes the scaler.

The code was faster in Standard Scaler function

```
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

Train the Logistic Regression Model

Penalty= 'L2 ': L2 regularization (Ridge Regression) to prevent overfitting.

C= 0.1: Controls the strength of regularization

Model. Fit (X _ train, y _ train): Trains the model using training data

```
model = LogisticRegression(penalty='l2', C=0.1)
model.fit(X_train, y_train)
```

```
▾ LogisticRegression
LogisticRegression(C=0.1)
```

Train the Random Forest classifier Model

Random _ state=42 ensures the data split every times Random.

Model is performed random classifier

Trains the model using feature X _train and label y _train

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```

CHAPTER – VI

EVALUATE MODEL PERFORMANCE

6.1 Evaluate Model Performance

LOGISTIC REGRESSION

Accuracy _score (y _ train, y _ train _ p red): Measures the model performs on training data.

Accuracy _ score (y _test, y_ test _ p red): Measures model accuracy on test data.

The gap is prevented only (~ 1.67) which the model Suggest not getting overfitting and well Generalized.

```
train_acc = accuracy_score(y_train, y_train_pred)
test_acc = accuracy_score(y_test, y_test_pred)

print(f"Training Accuracy: {train_acc:.2f}")
print(f"Testing Accuracy: {test_acc:.2f}")
```

```
Training Accuracy: 0.94
Testing Accuracy: 0.93
```

The accuracy is 93.2 which is good and not Getted over fit.

```
print(" Accuracy:", round(test_acc * 100, 2))
```

```
 Accuracy: 93.2
```

RANDOM FOREST

Accuracy _score (y _ train, y _ train _ p red): Measures the model performs on training data

Accuracy _ score (y _test, y_ test _ p red): Measures model accuracy on test data.

The gap is prevented only (~ 1.9) which the model Suggest not getting overfitting and well Generalized.

```
train_accuracy = accuracy_score(y_train, y_train_pred)
print(f"Training Accuracy: {round(train_accuracy * 100, 2)}%")

y_train_pred = model.predict(X_train)

y_test_pred = model.predict(X_test)
test_accuracy = accuracy_score(y_test, y_test_pred)
print(f"Testing Accuracy: {round(test_accuracy * 100, 2)}%")
```

```
Training Accuracy: 96.9%
Testing Accuracy: 95.0%
```

The accuracy is 93.2 which is good and not Getted over fit.

```
print(" Accuracy:", round(test_acc * 100, 2))
```

```
 Accuracy: 93.2
```

6.2 Analysis of classification report

Precision

Almost perfect classification

Recall

Model struggles with precision

F1-score

Well-balanced classification.

Support

Good, but some classifications.

class	Precision	Recall	F1 -Score	Support
Good	1.00	0.99	1.00	624
Hazardous	0.78	0.88	0.83	130
Moderate	0.94	0.94	0.94	454
poor	0.85	0.82	0.83	292
Overall Accuracy	0.93			1500

The model performs almost correct levels of predicting the comprehensive predicting levels are accurate to actual values.

CHAPTER – VII

PREDICTION AND INFERENCE

7.1 Prediction:

This model predicting the Air Quality Categories based on environmental and pollution-related factors by various environmental Factors The model predicting that the most polluted air is Hazardous in the certain populated areas and the Good is least proximity industry areas, poor is some Certain other related columns. This model is classifying test samples into various air quality levels. The predicted values indicate whether the air quality in a given region is safe or unhealthy based on pollution levels. The classifier uses a supervised learning approach he predictions align well with observed trends, highlighting the impact of pollutants and urban factors on air conditions. This predictive capability enables proactive air quality monitoring, assisting policymakers in pollution control and public health measures**.**

```
y_train_pred = model.predict(X_train)

y_test_pred = model.predict(X_test)
```

```
y_test_pred = model.predict(X_test)

y_test
```

```
1501    Hazardous
2586         Good
2653     Moderate
1055    Hazardous
705          Good
         ...
3563         Good
1538     Moderate
1837     Moderate
2380         Poor
1912         Good
Name: Air Quality, Length: 1500, dtype: object
```

7.2 Inference:

The dataset is predicting the air quality levels which the Polluted and industries Areas, Proximity to industrial areas exacerbates pollution, while population Density shows a mixed impact. Areas with moderate weather conditions and Lower pollutant concentrations tend to have "Good" air quality. This suggests, That controlling industrial emissions, monitoring weather patterns, and implementing urban planning strategies are crucial for improving air quality and public health. Standardization techniques like Standard Scaler were applied to normalize data, ensuring fair feature contribution. The model evaluation indicates high accuracy**,** making it reliable for forecasting air pollution trends.

The air quality prediction model successfully classifies air quality levels based on environmental and demographic factors such as temperature, humidity, pollutant concentrations (PM2.5, PM10, NO2, SO2, CO), proximity to industrial areas, and population density. The trained model provides accurate predictions, effectively distinguishing between categories like Good, Moderate, Poor, and Hazardous air quality.

The predictive capabilities of the dataset allow for forecasting future pollution levels, providing valuable insights for environmental monitoring and public health interventions. By training a Random Forest model, we can achieve accurate classification of air quality into categories such as Good, Moderate, or Unhealthy. This enables authorities to issue timely alerts and implement air quality management plans based on anticipated pollution levels.

CHAPTER – VIII

CONCLUSION

This project is successfully analysed the Air quality levels and pollutions For which people can survived in some distant areas .This project provides insights into air quality patterns and relationships using and machine learning techniques. The results help in understanding pollution sources, correlations between environmental factors, and the effectiveness of dimensionality reduction methods. Strive to create a model that can accurately determine air quality and guide untimely and necessary preventive measures, may include further feature engineering and exploring additional ensemble methods. inventive feature engineering, strive to create a model that can accurately determine air quality and guide untimely and necessary preventive measures This project demonstrates how to apply EDA, data pre processing, and advanced machine learning techniques to assess air quality. May include further feature engineering and exploring additional ensemble methods. This dataset is valuable for analysing the impact of environmental factors on air quality and can be used to identify patterns and correlations that may inform public health and environmental policies Policymakers and urban planners can use this data to identify key areas for intervention, such as reducing emissions from industrial sources, improving public transportation to reduce vehicle emissions, and increasing green spaces to mitigate the effects of pollution.

Using machine learning models, we can predict the Air Quality category (e.g., Good, Moderate, Poor, Hazardous) based on the environmental features. A Random Forest Classifier Model would be suitable for classification, while Regression models like Logistic Regression can predict continuous pollutant concentration values. Based on initial insights, proximity to industrial areas, PM2.5, PM10, and NO2 levels appear to be strong indicators of air quality.

REFRENCES

8.1 Data References

"Smith, J., & Brown, L". (2022). Impact of Industrial Emissions on Urban Air Quality. Journal of Environmental Studies, 45(3), 134-150.

Zhao, H., & Kim, S. (2021). Machine Learning Approaches for Air Quality Prediction. International Journal of Data Science and Analytics, 8(2), 102-119.

United Nations Environment Programme (UNEP). (2023). State of Air Pollution Report. Retrieved from https://www.unep.org/publications-data

National Aeronautics and Space Administration (NASA). (2024). Air Quality Monitoring Using Satellite Data. Retrieved from https://www.nasa.gov/general/what-is-air-quality/

Government of India, Central Pollution Control Board (CPCB). (2023). Annual Report on Air Quality Index (AQI) in Major Cities. Retrieved from https://www.cpcb.nic.in/