**Symbiosis Skills and Professional University**

**Kiwale, Pune**

PROJECT REPORT

On

India’s Air Quality Index

Submitted By

Tanuja Patil.

Ashatai Jadhav.

Meghana Jadhav.

Swapnali sangale.

REGISERED BATCH: ML7

UNDER THE GUIDENCE OF

TRAINER: AMRITA AJOTIKAR MA’AM.

**STUDENT DECLARATION AND ATTESTATION BY TRAINER**

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. If any part of the report is found to be plagiarized, I shall take full responsibility for it.

**Name Of Student Signature**

Tanuja Patil

Ashatai Jadhav.

Meghana Jadhav.

Swapnali Sangle.

**Signature of trainer**

**Amrita ma’am**

**CERTIFICATE**

**This is to certify that,**

**Tanuja Patill,**

**Ashatai Jadhav.**

**Meghana Jadhav.**

**Swapnali Sangle.**

**Has completed and submitted the project-entitled,  
“India’s Air Quality Index” under the guidance of Amrita ma’am, to**

**Symbiosis Skills and professional University, Pune, Maharashtra, India, is a record of Bonafide project work carried out by them and is worthy of consideration for the completion course in “Machine Learning”.**

**Date:**

**Signature Supervisor**

**Miss Amrita ma’am**

**INDEX**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SR.No | INDEX | | | | | | REMARK | Page .No | | |
|  | Acknowledgement | | | | | |  |  | | |
| 1 | Introduction | | | | | |  |  | | |
|  | 1.1 | | | | | Overview |  |  | | |
|  | 1.2 | | | | | Motivation |  |  | | |
|  | 1.3 | | | | | Problem statement |  |  | | |
|  | 1.4 | | | | | Objective |  |  | | |
|  | 1.5 | | | | | Scope |  |  | | |
| 2 | Methodologies of solving problem | | | | | |  |  | | |
|  | 2.1 | | | | | Algorithm |  |  | | |
|  | 2.2 | | | | | Hypothesis |  |  | | |
|  | 2.3 | | | | | Hypothesis Testing |  |  | | |
|  | 2.4 | | | | | Time series analysis |  |  | | |
|  | 2.5 | | | | | Python libraries |  |  | | |
| 3 | Software requirement specification | | | | | |  |  | | |
|  | 3.1 | | Assumptions And Dependences | | | |  |  | | |
|  | 3.2 | | Performance Requirements | | | |  |  | | |
|  | 3.3 | | System Requirements: | | | |  |  | | |
|  | Project Implementation | | | | | |  |  | | |
| 4 | 4.1 | | | Overview Of Project Module | | |  |  | | |
|  | 1: Data Flow Diagram | | | | | |  |  | | |
|  | | 1.1 | | | DFD Level 0 | |  | |  |
|  | | 1.2 | | | DFD Level 1 | |  | |  |
|  | | 1.3 | | | DFD Level 2 | |  | |  |
|  | | 1.4 | | | DFD Level 3 | |  | |  |
|  | | Conclusion | | | | |  | |  |

**ACKNOWLEDGEMENT**

We are profoundly grateful to trainer Amrita ma’am for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to express our deepest appreciation towards SYMBIOSIS SKILLS & PROFESSIONAL UNIVERSITY, Kiwale. Prof. Sumeet Sir and Prof. Baliram Sir whose invaluable guidance supported us in completing this project and also JP Morgan for funding and giving us this opportunity.

At last, we must express our sincere heartfelt gratitude to all staff members of Computer Science & Engineering Department who helped us directly or indirectly during this course of work.

Thanking you all!!

**ABSTRACT**

The air quality index(AQI) is an index for reporting air quality on a daily basis. A low AQI indicates good air quality and low levels of pollution while a higher AQI suggests increased concentration of pollutants in the air which is extremely detrimental to human health.

Air pollution is one of the major environmental issues. It can cause adverse health effects such as cancer, cardiovascular diseases and high mortality rates. High population density is a huge contributory factor of air pollution in cities and urbanized areas. The main air pollutants of concern are suspended particles and poly aromatic hydrocarbons. Other sources of air pollution are transport, local heating and possibly a pollution transfer from a industries. This thesis deals mainly with long term time series, including air pollutants (PM10, PM2.5, NO2, NO2,SO2,CO,O3,NH3), and meteorological variables. The purpose of this project is to compare the analysis before and after corona Air Quality index.

PM 10(smaller particles) can irritate your eyes, nose, throat. These particles can get cause into lungs and also some may even get bloodstream.

On an average (overall all cities), more than 24% decrease has been observed for all the AQI of pollutant. Poor air quality has a drastic effect on human repository system because of the small particles.

**INTRODUCTION**

**1.1OVERVIEW:**

Air is a pollution worldwide is a growing threat to human health and the natural environment. Air pollution may be described as contamination of the atmosphere by gaseous, liquid, or solid wastes or by-products that can endanger human health and welfare of plants and animals, attack materials, well as indoor. Mobile, stationary, area, and natural sources all emit pollution into the air.

An increase or decrease in the composition of these gases is harmful to survival. This imbalance in the gaseous composition has resulted in an increase in earth’s temperature, which is known as global warming.

Air pollution refers to any physical, chemical or biological change in the air. It is the contamination of air by harmful gases, dust and smoke which affects plants, animals and humans drastically. There is a certain percentage of gases present in the atmosphere.

**1.2 MOTIVATION:**

Air pollution is all around us. Indoors, outdoors, in cities and in the countryside. It affects us all, whether we realize it or not. For the longest time, we have taken the air we breathe for granted. There was air, there were smells, there was cold wind, there was hot air.

Take Pune, for instance, we all have experienced what it feels like inhaling in the ‘deadly’ smog that remained for about a week, after Diwali. Citizens were advised not to leave their homes and were asked to wear masks whenever going outside. Low visibility, a high number of deaths, etc. were the effects of pollution.

**1.3 PROBLEM STATEMENT :**

Analysing Air Quality Index using Machine Learning.

**1.3 OBJECTIVES:**

1.Finding the most polluted cities in recent years and analysing the levels of pollutants here. Understanding the impact of covid-19 induced lockdowns on Air Quality in some of the major cities

2.Analysing which cities underwent the most drastic improvement in Air Quality and cities showed a spike in AQI levels in spite of a stringent lockdown.

3.We do a time series analysis of the data and fit a SARIMA model with computed orders to forecast India AQI.

4.Finding the most populated content.

**1.4 SCOPE:**

This analysis will help us to understand the pattern of Air Pollution.

Find the pattern in Air pollution that can signifying how we can improve air quality.

Forecasting the future Air Quality In

**METHODOLOGIES OF PROBLEM SOLVING**

**2.1 Algorithm:**

**(a) Decision Tree:** Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision trees are composed of three main parts—decision nodes (denoting choice), chance nodes (denoting probability), and end nodes (denoting outcomes). Decision trees can be used to deal with complex datasets, and can be pruned if necessary to avoid overfitting.

**(b) Random Forest:** Random Forest is solid choice for nearly any prediction problem (even non-linear ones). It's a relatively new machine learning strategy (it came out of Bell Labs in the 90s) and it can be used for just about anything. It belongs to a larger class of machine learning algorithms called ensemble methods.

**(c) SVM:**  Support Vector Machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. In Support Vector Regression, the straight line that is required to fit the data is referred to as **hyperplane**.

**2.2 Hypothesis Testing:** The Pearson’s Chi-Square statistical hypothesis is a test for independence between categorical variables. In this article, we will perform the test using a mathematical approach and then using Python’s SciPy module. We start by defining the null hypothesis (H0) which states that there is no relation between the variables. An alternate hypothesis would state that there is a significant relation between the two. If our calculated value of chi-square is less or equal to the tabular (also called critical) value of chi-square, then H0 holds true.

**2.3TIME SERIES ANALYSIS:**

**a) Augmented Dicky Fuller Test:**

The test the Null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey-Fuller test for a larger and more complicated set of time series models.

The testing procedure for the ADF test is the same as for the [Dickey–Fuller test](https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test) but it is applied to the model:

Δyt = α + βt +γt-1 +δ1Δyt-1+…Δyt-p+1+€t,

Where,

Α is a constant, β is the coefficient on a time trend and the p the lag order of the autoregressive process. Imposing the constraints α=0 and β=0 corresponds to modelling a random walk and using the constraint β =0 corresponds to modelling a random walk with a drift.

**b) SARIMA:**

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

**2.4LIBRARIES USED:**

**❖ Pandas:** It provides fast, expressive, and flexible data structures to easily (and intuitively) work with structured (tabular, multidimensional, potentially heterogeneous)

**❖ NumPy:** It has advanced math functions and a rudimentary scientific computing package. NumPy is a popular array – processing package of Python. It provides good support for different dimensional array objects as well as for matrices.

**❖ Matplotlib:** Matplotlib helps with data analysing, and is a numerical plotting library. Matplotlib can create such quality figures that are really good for publication. Figures you create with Matplotlib are available in hardcopy formats across different interactive platforms.

**❖ Seaborn:** It provides a high-level interface for drawing attractive and informative statistical graphics.

**❖SK-learn:** It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

**❖ SciPy:** The SciPy library, a collection of numerical algorithms and domainspecific toolboxes, including signal processing, optimization, statistics, and much more. Matplotlib, a mature and popular plotting package that provides publication-quality 2-D plotting, as well as rudimentary 3-D plotting.

**SOFTWARE REQUREMENT SPECIFICATIONS**

**3.1 Assumptions And Dependences:**

**3.1.1 Assumptions:**

I. The end user device should be a laptop.

ii. Additionally, the end user has an active internet connection in his/her laptop.

**3.1.2 Dependencies:**

i. The system browser is dependent on the end user device.

ii. The prediction and analysis purpose are dependent on the types of algorithms used.

**3.2 Performance Requirements:**

**i. Accuracy:** The system can predict with varying accuracy between 50 to 60% using one of the Algorithms c which gives maximum accuracy right now, but later on, as the number of responses will increase the accuracy will also increase.

**ii. Privacy:** Data will be totally secured and will not be leak as no personal details are asked.

**3.3 System Requirements:**

**3.3.1 Database Requirement:**

• Ms-Excel

**3.3.2 Software Requirement:**

• Jupyter Notebook

• Programming Language- Python.

**3.3.3 Hardware Requirement:**

• Any Device With Brower Support

**PROJECT IMPLEMENTATION**

**4.1 Overview Of Project Module:**

**The project is comprised of following steps:**

• Collecting of data

• Importing data into python

• Data cleaning

• Performing exploratory data analysis

• Hypothesis testing

• Feature engineering

• Implementation of algorithms

• Comparison of algorithms via accuracy

**System Design:**

**1: Data Flow Diagram:**

A data flow diagram (DFD) is a graphical representation of the “flow” of data through an information system. DFD is a preliminary step to create overview of system.

* 1. **DFD Level 0:**

In DFD level 0 shows that get the input from user then display output.

Procedure

**❖ Collection of data:**

Data collected from website Kaggle dataset.

**❖ Importing the data and cleaning the data:**

All the implementation is done in python programming language on Jupiter notebook. The dataset was imported using pandas library. Head, tail and shape of data was checked. Columns of the dataset were checked. Times. Hence we fill the missing values.

**❖ Performing exploratory data analysis:**

Before doing data exploration, we converting the categorical data set into numeric data set. After doing this, we performed data visualization techniques. Plotted each graph of every question to find the spread of the data and to check measure of tendency. We also plotted boxplots and density plots too. Group by was also done check which gives more clear picture.

* 1. **DFD Level 1:**

In DFD level 1 show that data cleaning and EDA perform and display output.

Output

Input

❖ **Feature engineering:**

After knowing which were the independent and dependent variables, feature engineering techniques were used for X and y. it is important step as the accuracy will be dependent on X and y variable. It was try and error method because to get the highest accuracy we needed to check the dependency.

**❖ Implementation of algorithms:**

Before implementing the algorithms, the data was spitted for train and test purpose. 80% of the data was for train and remaining 20% was for test. We also plotted heatmap and pair plot to get a better idea. Algorithms we used in this were Linear Regression, Naïve Bayes Algorithm, KNearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine. In all algorithms we plotted confusion matrix and overviewed classification report. On bases of y-test and y-pred we calculated accuracy.

* 1. **DFD Level 2:**

In DFD level-2 show that Give dataset as Input, data cleaning and EDA perform, Algorithm apply and display output in accuracy result.

Algorithm

EDA

**Cleaning**

**❖**

**Comparison of algorithms via accuracy:**

After implementing all the algorithms, we created a data frame of all accuracy with respect to the algorithms. So that we can clearly overlook which has highest accuracy. Mostly accuracy

were under 50% due to less number of data. But as the data will increase accuracy will also increase in upcoming days.

Δ��=�+��+���−1+�1Δ��−1+⋯+��−1Δ��−�+1+��,**Conclusion:**

1.Vehicular pollution contents are more related to air quality index.

2.Delhi is the most polluted city in terms of vehicular pollution contents.

3.Ahmadabad is the most polluted city in terms of industrial pollution content.

4.After COVID19 pandemic there is gradual decrease in vehicular pollution contents, industrial pollution content.

5.Extra Gradient Boost classifier 100% accurately classify the target variable.

Δ��=�+��+���−1+�1Δ��−1+⋯+��−1Δ��−�+1+��,