BTIBM505

Major Project Synopsis

on

AIR QUALITY INDEX PREDICTOR

In partial fulfilment of requirements for the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

MANAY RAWAL [20100BTCSDSI07277]

RAHUL CHOUHAN [20100BTCSDSI07287]

DIVYANSH LASHKARI [20100BTCSDSI07269]

Under the guidance of

PROF. OM KANT SHARMA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SIIRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

SIIRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE

JUL.-DEC-2022

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

CONTENT

1. INTRODUCTION

Predicting Air Quality Index using Machine Learning

AQI: The air quality index is an index for reporting air quality on a daily basis. In other words, it is a measure of how air pollution affects one's health within a short time period. The AQI is calculated based on the average concentration of a particular pollutant measured over a standard time interval. Generally, the time interval is 24 hours for most pollutants, 8 hours for carbon monoxide and ozone.

AQI is calculated based on chemical pollutant quantity. By using machine learning, we can predict the AQI.

AQI Level	AQI Range
Good	0 – 50
Moderate	51 – 100
Unhealthy	101 – 150
Unhealthy for Strong People	151 – 200
Hazardous	201+

Description	AQI	PM10 µg/m³ 24 hr avg	PM2.5 µg/m³ 24 hr avg	CO ppm 8 hr avg	O3 ppb 8 hr avg	NO2 ppb 24 hr avg
Good + Satisfactory	0-100	0-100	0-60	0-1.7	0-50	0-43
Moderate	101-200	101-250	61-90	1.8-8.7	51-84	44-96
Poor	201-300	251-350	91-120	8.8-14.8	85-104	97-149
Very Poor	301-400	351-430	121-250	14.9-29.7	105-374	150-213
Severe	401-500	431-550	251-350	29.8-40	375-450	214-750

Data Set Description

It contains 8 attributes, of which 7 are chemical pollution quantities and one is Air Quality Index. PM2.5-AVG, PM10-AVG, NO2-AVG, NH3-AVG, SO2-AG, OZONE-AVG are independent attributes. air_quality_index is a dependent attribute. Since air_quality_index is calculated based on the 7 attributes.

As the data is numeric and there are no missing values in the data, so no pre-processing is required. Our goal is to predict the AQI, so this task is either Classification or regression. So as our class label is continuous, regression technique is required.

Regression is supervised learning technique that fits the data in a given range. Example Regression techniques in machine learning:

- Random Forest Regressor
- Random Forest Classifier
- Decision Tree Classifier
- Decision Tree Regression
- Linear Regression etc.
- Multi-Linear Regression
- Logistic Regression
- K-nearest Neighors Classification
- K-nearest Neighors Regression

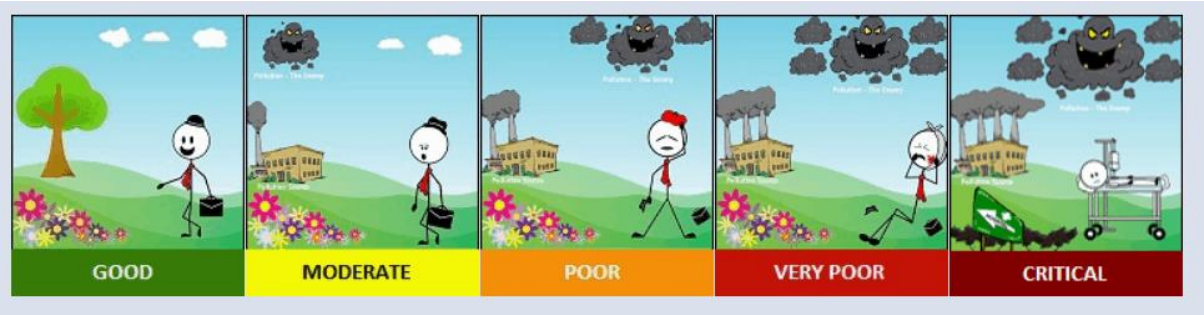
2. PROBLEM DOMAIN

The Air Quality Index (AQI) is used for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health affects you may experience within a few hours or days after breathing polluted air.

Why is it important to monitor the air quality?

Air is essential to life. Poor air quality threatens the health of all living things from humans to plants. There are many types of air pollution, and each have a different effect on human health. The two most common types of air pollution in the United States are ozone and particle pollution.

Understanding the Air Quality Index is important because it gives people vital information about the conditions of the air in their location and how the quality of the air in their city can impact their health.



3. SOLUTION DOMAIN

We are using following machine learning models on supervised learning for prediction of air quality index :-

- **Random Forest Regressor**

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

- **Random Forest Classifier**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

- **Decision Tree Classifier**

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Classification trees (Yes/No types)

What we've seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.

- **Decision Tree Regression**

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

Regression trees (Continuous data types)

Here the decision or the outcome variable is Continuous, e.g. a number like 123. Working Now that we know what a Decision Tree is, we'll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called as ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3. Before discussing the ID3 algorithm, we'll go through few definitions.

- **Simple Linear Regression**

Simple linear regression is a statistical method for establishing the relationship between two variables using a straight line. The line is drawn by finding the slope and intercept, which define the line and minimize regression errors.

- **Multi-Linear Regression**

Multiple linear regression is a statistical technique that uses multiple linear regression to model more complex relationships between two or more independent variables and one dependent variable. It is used when there are two or more x variables.

- **Logistic Regression**

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring.

- **K-nearest Neighors Classification**

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

- **K-nearest Neighors Regression**

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

4. SYSTEM DOMAIN

1. Tools

- **NumPy**

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- **Pandas**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license

- **Matplotlib**

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

- **Seaborn**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

1. Technology

- **Python 3.11.0**

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

2. Environment

- **Google collab**

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

3. Platform

- **MacOS, Windows, Linux**

macOS is a Unix operating system developed and marketed by Apple Inc. since 2001. It is the primary operating system for Apple's Mac computers. Within the market of desktop and laptop

computers it is the second most widely used desktop OS, after Microsoft Windows and ahead of ChromeOS.

Windows is a group of several proprietary graphical operating system families developed and marketed by Microsoft. Each family caters to a certain sector of the computing industry, for example, Windows NT for consumers, Windows Server for servers, and Windows IoT for embedded systems.

Linux is an open-source Unix-like operating system based on the Linux kernel, an operating system kernel first released on September 17, 1991, by Linus Torvalds. Linux is typically packaged as a Linux distribution.

4. Hardware/Software

- PC, laptop

A personal computer is a multi-purpose microcomputer whose size, capabilities, and price make it feasible for individual use. Personal computers are intended to be operated directly by an end user, rather than by a computer expert or technician.

A laptop, laptop computer, or notebook computer is a small, portable personal computer with a screen and alphanumeric keyboard.

5. APPLICATION DOMAIN

Background:

Information about local air quality is reported across the India using air quality alerts such as the Environmental Protection Agency's Air Quality Index. However, the role of such alerts in raising awareness of air quality is unknown. We conducted this study to evaluate associations between days with Air Quality Index ≥101, corresponding to a categorization of air quality as unhealthy for sensitive groups, unhealthy, very unhealthy, or hazardous, and air quality awareness among adults in the United States.

Methods:

Data from 12,396 respondents to the 2016–2018 Consumer Styles surveys were linked by geographic location and survey year to daily Air Quality Index data. We evaluated associations between the number of days in the past year with Air Quality Index ≥101 and responses to survey questions about awareness of air quality alerts, perception of air quality, and changes in behaviour to reduce air pollution exposure using logistic regression.

Results:

Awareness of air quality alerts (prevalence ratio [PR] = 1.23; 95% confidence interval [CI] = 1.15, 1.31), thinking/being informed air quality was bad (PR = 2.02; 95% CI = 1.81, 2.24), and changing behaviour (PR = 2.27; 95% CI = 1.94, 2.67) were higher among respondents living in counties with ≥15 days with Air Quality Index ≥101 than those in counties with zero days in the past year with Air Quality Index ≥101. Each aspect of air quality awareness was higher among adults with than without asthma, but no differences were observed by heart disease status. Across quintiles of the number of days with Air Quality Index ≥101, air quality awareness increased among those with and without selected respiratory and cardiovascular diseases.

6. EXPECTED OUTCOME

Confusion Matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one.

This is the confusion matrix which describes the excepted outcome of report.

```
print(lr.score(x_train,y_train))
print(lr.score(x_test, y_test))

from sklearn.metrics import confusion_matrix,classification_report, accuracy_score
print(confusion_matrix(y_pred_lr, y_test))
print(classification_report(y_pred_lr, y_test))
print(f'model_score- {lr.score(x_test,y_test)}')
print(f'accuracy_score- {accuracy_score(y_pred_lr, y_test)}')
```

```
[[ 32   1   0  11   0   0]
 [  3 464  71 248   3  16]
 [  0  61  45   2   4  35]
 [ 84  89   0 267   0   1]
 [  0   3   6   0  11   6]
 [  0  12  24   2   9  49]]
```

Classification Report

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives, and False Negatives are used to predict the metrics of a classification report.

This is the classification report based in confusion matrix created above

```
              precision    recall  f1-score   support

           0       0.27      0.73      0.39        44
           1       0.74      0.58      0.65       805
           2       0.31      0.31      0.31       147
           3       0.50      0.61      0.55       441
           4       0.41      0.42      0.42        26
           5       0.46      0.51      0.48        96

    accuracy                           0.56      1559
   macro avg       0.45      0.52      0.47      1559
weighted avg       0.59      0.56      0.57      1559

model_score- 0.5567671584348942
accuracy_score- 0.5567671584348942
```

7. REFERENCES

- **Kaggle –** https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?resource=download

- **Google Colab1** - https://colab.research.google.com/drive/1Hi1ur89p4DdpjSAWjd0gJV-LGjHSG7e-#scrollTo=-teNaxQN-B01

- **Google Colab2** - https://colab.research.google.com/drive/1GO0JBNjDj_WMdzNqXu0_kbI8TKzHRsyE

- **Geeks for Geeks -** https://www.geeksforgeeks.org/

- **Scikit learn** - https://scikit-learn.org/stable/

- **Python -** https://www.python.org/

- **Pandas -** https://pandas.pydata.org/

- **NumPy -** https://numpy.org/

https://drive.google.com/file/d/1X1uB7jXeZNVRL0J6TojhlYP-hrheoNU7/view