**Study of Factors Contributing Air Quality**

Submitted to: **Yogendra Sir**

Submitted by: **Aditya Kumar**

Tables of Content:

* Certificate
* Acknowledgement
* Abstract
* List of figures
* Introduction
* Theory
* Methodology
* Data Pre-processing
* Regression Method Used
* Classification Method used
* Data visualization
* Result Analysis

**Certificate**

Date: 27/06/17

This is to certify that Mr. **Aditya Kumar**, student of 4th Year from Department of Information Technology, Poornima College of Engineering , Jaipur, has undergone a Project work from May 29, 2017 to July 05, 2017 in **Data Science & Machine Learning** titled “**Study of Factors Contributing Air Quality** ”

Project Incharge Seal

Acknowledgement

For making this project we would like to thank our course instructor, Yogendra Sir, he has helped us in learning the various concepts of the Machine Learning. Here we would also like to thank Kunal Sir and Rohit Sir for mentoring us throughout this project.

Abstract

Rapid growth of data comes with a challenge of sorting and analysing them, where raw data exists in graphical form, textual form or in images. Data science and machine learning finds its application in various fields like stock market, recommendation systems, image processing, aerial photography, military, weather forecasting etc.

This report is about our project on “Study of Factors Contributing to Air Quality in Jaipur” that addresses about data pre-processing and post processing which includes plotting, classification and prediction of air quality in real time and the ability of machine learning algorithms to deal with different set of data. In this project, we have tackled a regression problem of predicting the air quality factors for next day. We have tested and used Logistic Regression, Decision Tree and random forest to determine the results. In addition to this, we have also made use of several libraries to plot the data points.

Introduction

The rate at which urban air pollution has grown across India is alarming. A vast majority of cities are caught in the toxic web as air quality fails to meet health-based standards. Almost all cities are reeling under severe particulate pollution while newer pollutants like oxides of nitrogen and air toxics have begun to add to the public health challenge.

According to WHO, India ranks among the world’s most polluted countries. Out of the 20 most polluted cities in the world, 13 are in India. In which, Jaipur is one of the most polluted city in the world today so to know its reason is needed.

Exposure to particulate matter for a long time can lead to respiratory and cardiovascular diseases such as asthma, bronchitis, lung cancer and heart attack. Last year, the Global Burden of Disease study pinned outdoor air pollution as the fifth largest killer in India, after high blood pressure, indoor air pollution, tobacco smoking, and poor nutrition. In 2010, about 620,000 early deaths in India occurred from air pollution-related diseases. The Central Pollution Control Board (CPCB) sponsored the study that links the pollutants, pm 10 (particulate matter smaller than 10 microns), the cause of these diseases. The central regulatory authority recently regulated stricter norms for a number of air toxins and pollutants but omitted revision of the standard for pm 10.

Theory

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization. Data science – discovery of data insight This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviours, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions. For example: Netflix data mines movie-viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce Data science – development of data product A "data product" is a technical asset that: (1) utilizes data as input, and (2) processes that data to return algorithmically generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data. For example: Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you. Machine learning and statistics are part of data science. The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes or supervised clustering.

Supervised and unsupervised learning describe two ways in which machines algorithms can be set loose on a data set and expected to learn something useful from it. **Supervised:** If we are training our machine-learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training. Our learning algorithm seeks a function from inputs to the respective targets. If the targets are expressed in some classes, it is called *classification* problem. Alternatively, if the target space is continuous, it is called *regression* problem.

**Regression** analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

**Classification** model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset.

**Unsupervised:** If we are training our machine-learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. Most important unsupervised learning is *clustering*, which will create different cluster of inputs and will be able to put any new input in appropriate cluster.

**Cluster** analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

1. **Decision Trees:** A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility. From a business decision point of view, a decision tree is the minimum number of yes/no questions that one has to ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.

2. **Logistic Regression:** Logistic regression is a powerful statistical way of modelling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

**3. Linear Regression** It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b.

**4. KNN (K- Nearest Neighbors)** It is also a lazy algorithm. What this means is that it does not use the training data points to do any generalization. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing KNN modelling.

Methodology

To understand the methodology adopted, we first understand our dataset and the variables

**Dataset**

Our dataset consists of more than 6000 days of recorded values for SO2, NO2 and PM10. The target is to train machine-learning algorithm so that it can predict the next days values for same.

**Features:**

Sampling Date: the date when the same was recorded.

Location of Station: the place where it was recorded

SO2: value of so2

NO2: value of no2

PM10: value of pm10

Type of Area: either Residential, Industrial

Data Pre-processing

For pre-processing, we considered Python as our options for the project. After some experimentation, we found that while R was easier for statistics and analysis of the data, the lack of uniformity among the various ML packages made Python our preference. The ML algorithms provided by the scikit-learn package do not function if the input data has missing values. Hence, we either had to input data at the missing slots or remove the instances that had these missing fields. In our dataset we had dates in different format so we needed to make it in same format. Also we used fillna() method to fill NaN values with mean of the column data.

Libraries used

numpy

pandas

pickle

scikitlearn

o sklearn.cross\_validation

o sklearn.preprocessing

o sklearn.linear\_model

o sklearn.neighbors

o sklearn.ensemble

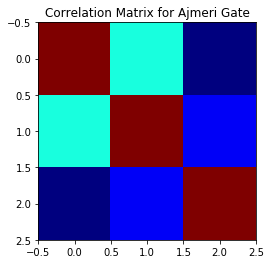
o sklearn.metrics

Regression Method Used

Here in this project we had used Logistic Regression. Here we had used Logistic Regression 2 times, first to train the data for Latitude part and another time for the Longitude part. However, as the data is very large and we had many attributes so the training part of the data took very long time. So, we had trained the data single time only and had saved the trained data in the pickle file, from where we can use that train data to make the predictions. Here we had used Logistic Regression only because we were very much familiar with this method, so it was easy for us to use it and our requirements of doing predictions completely met with the features of the Logistic Regression.

Data Visualization

For doing visualization of the data we had used library matplotlib. Through matplotlib we made first made the visualization for finding the correlation matrix between SO2, NO2, and PM10 for different locations.

****

Conclusion

We hope this article was an enriching experience for you and provided you enough insights about the factors that can lead to rise in air pollution. so, watch out before we contribute to air pollution, knowingly or unknowingly.