****Project V1 –Business Understanding & Problem Discovery****

**Air Pollution Forecast**

**AIDI 1002 AI Algorithm**

**Course Facilitator: Marcos Bittencourt**

**Prepared By:**

Yali Huang -100788425

OCT 28, 2020

Contents

[**1.** **Problem Statement** 1](#_Toc54782282)

[**2. Rational Statement** 2](#_Toc54782283)

[**3. Approach** 2](#_Toc54782284)

[1) Dataset Preprocessing 2](#_Toc54782285)

[2) Building model 2](#_Toc54782286)

[3) Evaluation 2](#_Toc54782287)

[4) Suggestion 2](#_Toc54782288)

[**4. Data Source** 3](#_Toc54782289)

[1) Date Set 3](#_Toc54782290)

[2) Data Set Description: 3](#_Toc54782291)

[3) Attribute Information: 3](#_Toc54782292)

**Air Pollution Forecast**

1. **Problem Statement**

In recent years, air pollution has become a global concern, Especially in some large cities with large populations like Beijing, London, Mumbai. It refers to the release of pollutants into the air that are detrimental to human health and the planet as a whole. Car emissions, chemicals from factories, dust, and pollen and mold spores may be suspended as particles. Among all those gauges of air pollution, PM2.5 concentration is no doubt the most famous and effective. In order to deal with this serious problem,

Can people monitor the air pollution by predicting the value of PM2.5? With this question, we start this project.



## Rational Statement

In this project, we use a dataset from UCI Repository (with time series data from 2010 to 2014) to explore the factors influencing PM2.5 in Beijing, perform machine learning algorithms to predict the PM2.5 values. Suggestion on how to improve air quality in Beijing will be given at the end.

## Approach

## ****Dataset Preprocessing****

This process includes dataset loading, data normalization, plotting the graph for each column, and feature selection.

## ****Building model****

**Split the dataset to train and test set**

**Using different algorithms to build a model**

**Training our model using train dataset**

## ****Evaluation****

**Predicting the PM2.5 value on test dataset**

**Using different evaluation methods to evaluate our model.**

**Compare and select the most fitable algorithm**

## ****Suggestion****

**Based on the below analyze and prediction. Producing some suggestions to reduce air pollution.**

## Data Source

## Date Set

https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

## Data Set Description:

This hourly data set contains the PM2.5 data of the US Embassy in Beijing, including 43824 instances and each instance has 13 attributes. The time period is between Jan 1st, 2010 to Dec 31st, 2014. Missing data are denoted as NAN.

## Attribute Information:

No: row number  
year: year of data in this row  
month: month of data in this row  
day: day of data in this row  
hour: hour of data in this row  
pm2.5: PM2.5 concentration (ug/m^3)  
DEWP: Dew Point (â„ƒ)  
TEMP: Temperature (â„ƒ)  
PRES: Pressure (hPa)  
cbwd: Combined wind direction  
Iws: Cumulated wind speed (m/s)  
Is: Cumulated hours of snow  
Ir: Cumulated hours of rain