README: Dimensionality Reduction on Air Pollution Dataset

Table of Content

Demo

Link:https://colab.research.google.com/drive/12U-T3Le58Nd244Gq_bcP9RtmkSm21X81?usp=sharing

Context

This dataset deals with air pollution measurement information in Seoul, South Korea.Seoul Metropolitan Government provides much public data, including air pollution information, through the 'Open Data Plaza'. Used Dataset is structured by collecting and adjusting various air pollution related datasets provided by the Seoul Metropolitan Government.

This data provides average values for six pollutants (SO2, NO2, CO, O3, PM10, PM2.5).

- Data was measured every hour between 2017 and 2019.
- Data was measured for 25 districts in Seoul.
- This dataset is divided into four files.

1. Measurement info: Air pollution measurement information
    - 1 hour average measurement is provided after calibration
    - Instrument status:
        - 0: Normal, 1: Need for calibration, 2: Abnormal
        - 4: Power cut off, 8: Under repair, 9: abnormal data

2. Measurement item info: Information on air pollution measurement items
3. Measurement station info: Information on air pollution instrument stations
4. Measurement summary: A condensed dataset based on the above three data.

Dataset Link: https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul

Overview

This project deals with the dimensionality reduction algorithms on the given Air Pollution Dataset of Seoul. It is a unlabelled dataset so, initially dimensionality reduction algorithms like PCA and Correlation method have been used but just to make it little more explorative we have labeled the data afterwards and have applied algorithm like LDA which allows dimensionality reduction for labeled dataset. All algorithms have given reduced remaining dimension and good performance which has been further calculated with the help of classification report. Overall this project deals with dimensionality reduction through different algorithms and their performances which have been calculated.

Technical Aspect

This project is divided into three parts:

1) Application of different dimensionality reduction Algorithm on the Air pollution dataset.
2) Calculating their accuracy with the help of classification report.
   2.1) For calculation of accuracy data has been labeled through **K-means.**
   2.2) Appropriate number of clusters has been calculated through **WCSS.**
3) Unlabelled data has been changed to labeled data in order to apply LDA for exploratory analysis of the Dimensionality reduction Algorithm.

Installation

The Code is written in Python 3.7. If you don't have Python installed you can find it here. If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip. To install the required packages and libraries.

Below commands install all the required libraries to run this program.

```
!pip install pyspark
!pip install seaborn
!pip install pandas
!pip install scikit-learn
!pip install matplotlib
```

Run

STEP 1: Open this code on google colab/ Jupyter notebook/ Spyder. In case it has been opened in a Jupyter notebook or Spyder then file paths need to be changed and no other change is required.

STEP 2: Run the code cell by cell and it will show the output, no other user interventions needed.

Acknowledgements

Original Data is provided from here.

- https://data.seoul.go.kr/dataList/OA-15526/S/1/datasetView.do
- https://data.seoul.go.kr/dataList/OA-15516/S/1/datasetView.do
- https://data.seoul.go.kr/dataList/OA-15515/S/1/datasetView.do