BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Work Integrated Learning Programmes Division
M.Tech (Data Science and Engineering)
Machine Learning
DSECLZG565
Second Semester, 2022 -23

Assignment 1 – PS10
Global Air Pollution

Weightage: 20 Marks

General Instructions :

1. Inside each of the submission documents, you are required to mention Group details - group number, group members.
2. Organize your code (PART B) in separate sections for each task. Add comments to make the code readable.
3. Notebooks without output shall not be considered for evaluation.
4. **For a given dataset, where class labels may not be found, choose the right target variable and discretize the same for classification.**

Submission guidelines:

1. Each group should upload in CANVAS in respective locations under ASSIGNMENT Tab. Assignment submitted via means other than through CANVAS will not be graded.

2. Upload your submission documents for PART A in .pdf format and PART B in .ipynb and .pdf format (for PART-B both formats are mandatory)

Dataset :

Download the dataset assigned to your group from the below link :
https://drive.google.com/file/d/1iZPFsGnr7CQ8_MDrvhYIQ3KHfHErELW1/view?usp=share_link

PART A: (5-marks) Research

Select the research paper of your choice. **Attach the chosen paper along with the assignment submission**. Write a synopsis and find below pointers:

3. Paper Contribution
4. Data Pre-processing
5. Machine Learning Activity
6. Result analysis with metrics used from paper
7. Exploratory Data Analysis / Visualization

PART B: (15 – marks) Dataset-based Implementation

Refer to the dataset mapped against your group. Use python based APIs and perform the following three classes of activities.

EDA

1. Perform Exploratory Data Analysis to gather insight from the dataset. Write your inference about the analysis learned from visualizations (minimum 3) [3]

Classification

CLASSIFICATION (any of the Logistic Regression / SVM / Decision Tree/ Naïve Bayes/KNN/ANN). Justify your design choices at each step: Write as a markdown cell in jupyter notebook at the beginning of each subsection.

1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]
2. Perform the Machine Learning activity. Explain the choice of target attribute, classification type, model selected with reason [1.5]
3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]
4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]

Regression

Any of the Linear Regression (any of Gradient / Stochastic / MiniBatch)/linear basis models/KNN/Locally weighted regression/ any of the regularization techniques). Justify your design choices at each step: Write as a markdown cell in jupyter notebook at the beginning of each subsection.

1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]
2. Perform the Machine Learning activity. Explain Attributes of interest, Regularization type with reason, model selected with reason [1.5]
3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]
4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]

Ensemble ML

Justify your design choices at each step: Write as a markdown cell in jupyter notebook at the beginning of each subsection.

1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]
2. Perform the Machine Learning activity. Explain Attributes of interest, base classifier chosen with reason, model selected with reason [1.5]
3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]
4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]