Assignment 2
CSCI 6515 – Fall 2021
Posting date: Oct 25, 2021
Due Date: Nov 8, 2021, 11.30 PM (Halifax Time)

Background Information

We all are responsible for taking care of our air. Nova Scotia Environment (NSE) strives to monitor and protect our outdoor air quality through regulations and programs to reduce pollutants that lead to issues like smog, acid rain, climate change, and the thinning ozone layer. Poor air quality can affect our health, lead to increased health care costs, and also affect natural resources. Motivated by this fact, we try to integrate machine learning research with air pollution epidemiology to support our environment. We mainly focus on the pollutant 'carbon monoxide (CO)', monitored and measured by the air quality station 'Halifax'. We explore the Nova Scotia provincial ambient carbon monoxide (CO) hourly data. Carbon monoxide (CO) is a highly toxic gas produced when fuels burn incompletely. Simultaneously, we investigate if the traffic flow impacts the carbon monoxide levels of the roadside, and hence, the air quality. Finally, we decided to do a few experiments. Initially, we apply linear regression for the prediction of the average CO levels. However, we want to investigate some patterns in the data. So, we decided to apply clustering analysis to identify groups where traffic regions and CO levels are correlated. Finally, we build two different models, one using the decision tree classifier and another using Naïve Bayes to classify the dataset according to the labels provided by the clustering method.

To access and collect the hourly ambient CO data follow the link below:
https://data.novascotia.ca/Environment-and-Energy/Nova-Scotia-Provincial-Ambient-Carbon-Monoxide-CO-/8tvc-9ah2

You will find a clean csv file named "cleaned_traffic_data.csv" with assignment 1. Use this data to solve the assignment along with the CO data. To understand traffic flow data, you can access following link:
https://data.novascotia.ca/Roads-Driving-and-Transport/Traffic-Volumes-Provincial-Highway-System/8524-ec3n

Note that you may use open-source libraries such as SK-learn, NumPy, and Pandas. If you are interested in using any other available library, you can consult with your instructor.

Your Tasks:

[1] Your first task is doing some research on the dataset. You pre-process the CO data (Suggested year: 2019), and traffic data (csv file). Feature selection can also be done. In this step, you do a descriptive analysis of your data to better understand it. This step creates a dataset to you be able to work on. You also have to include one summary visualization of the data.

[2] In this task, you have to apply linear regression on the dataset in order to predict the average of CO levels. Use previous observations as training set and remaining as test set, because that is how we collect information in real-world. Example: jan-jul as training set, and aug-dec as test set. Select an evaluation metric and explain the reason for that choice. Show the results and provide an analysis and conclusions on that. You can also include visual representation of your results.
i. Describe how the linear regression algorithm works.
ii. What is the most relevant features for the prediction?
iii. Give an explanation about the evaluation metric and what this information says about your dataset.
iv. Provide some insights and conclusions.

[3] In this task, you have to apply K-means method to obtain classes for your dataset. For this task, include the average CO levels as features. In order to evaluate your clusters, you have to apply silhouette measure. Select the number of groups (k) that provides the highest silhouette for the dataset. Show the results and provide an analysis and conclusions on that. You also have to include visual representation of your results, such as the silhouette measures. PCA can be used to illustrate the clusters in a 2D dimension.
i. Describe how the k-means algorithm works.
ii. What the clusters represent in this dataset? What are the similarities?
iii. Give an explanation about the silhouette measure and what this information says about your dataset.

[4] As the clustering provided the labels, you have to apply Naïve bayes (NB) and Decision tree (DT) in your new labelled dataset. Show the results, compare the models and provide an analysis and conclusions on that.
i. what information each model can provide about the dataset?
ii. which evaluation metric you used and why?
iii. which model provide high score? Why do you think this happened? Perform statistical significance testing for both NB and DT.
iv. how you can interpret the features relationship? The patterns found by the supervised models are the same as in the one presented in the clustering?

How to Submit

Assignment 2 will be done in Jupyter Notebook. After you finished your coding and make sure it is working well, you run your code step by step in a Python notebook and print the notebook in a PDF file. We call code.PDF. Then you prepare a document and answer the questions in the document. You must add one page describing what you have done before your answers. You also add one page on the summary of your results at the end of the document. In the end, you add all the references you have used in this assignment. Therefore, this document includes the following:

1. Dataset description and preprocessing (Task 1)
2. Task description, questions i, ii, iii, and iv (Task 2)

3. Task description, questions i, ii, and iii (Task 3)
4. Task description, questions i, ii, iii, and iv (Task 4)
5. Summary of your results and conclusions
6. References

We call this file Report.PDF. Now you merge Report.PDF and code.PDF and create your final document to submit. Therefore, the final document (FinalReport_YourBannerID.PDF) includes the content from both report.pdf and code.pdf files.

Now you submit the FinalReport_YourBannerID.PDF to Brightspace.