Assignment 1

CSCI 6515 Machine Learning for Big Data - Fall 2023

Posting date: Sep 18, 2023, 11:00 AM

Due Date: Oct 2, 2023, 11:30 PM (Halifax Time)

Background Information

We all share the responsibility for taking care of our air. Nova Scotia Environment (NSE) strives to monitor and protect our outdoor air quality through regulations and programs aimed at reducing pollutants that contribute to issues such as smog, acid rain, climate change, and the depletion of the ozone layer. Poor air quality can have adverse effects on our health, lead to increased healthcare costs, as well as impact our natural resources. Motivated by these facts, we attempt to integrate machine learning research with air pollution epidemiology to support our environment. Our primary focus is on the pollutant 'ambient fine particulate matter (PM2.5),' which is monitored and measured by the air quality station 'Halifax.' We analyze the hourly data for ambient fine particulate matter, PM2.5 in Nova Scotia. PM2.5 is a common air pollutant that negatively affects public health, leading to heart and lung problems. Simultaneously, we investigate how traffic flow contributes to PM2.5 levels and, consequently, air quality. Finally, we build our model using a decision tree classifier to classify average PM2.5 levels. We classify the class label as 'High' when the PM2.5 level is greater than or equal to 0.5 (>=0.5) and 'Low' when the PM2.5 level is less than 0.5 (<0.5).

To access and collect the hourly ambient fine particulate matter (PM2.5) data follow the link below:
Nova Scotia Provincial Ambient Fine Particulate Matter (PM2.5) Hourly Data Halifax BAM/T640 | Open Data | Nova Scotia

To access and collect the traffic volume data follow the link below:
Traffic Volumes - Provincial Highway System | Open Data | Nova Scotia

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

<u>Your Task:</u>

(Points on the right side of each question indicates the weight of the individual question)

[1] Your first task is to conduct research to identify appropriate resources for obtaining such data and to understand their format.

[2] You will preprocess the PM2.5 data (recommended year: 2019) and the traffic data (in CSV format).

- **i.** The traffic data needs to be filtered to represent the Halifax region. 2
- **ii.** The PM2.5 data will serve as labels for the traffic dataset, so you should compute the daily averages. 2
- **iii.** Subsequently, you must normalize the PM2.5 levels and discretize them using a threshold of 0.5. During this step, perform a descriptive analysis of your data to gain a better understanding of it. This process will result in a dataset that you can work with. 2
- **iv.** Lastly, you should include one summary visualization of the data. 2

[3] Answer the following questions:

- **i.** Use the Information Gain (IG) as the decision criterion to select which attribute to split on. Show your calculations for the IG for the root node. 4
- **ii.** Repeat (i) using Gini Index criterion. 4
- **iii.**
  - ***a.*** Create a decision tree using IG with default parameters. 2
  - ***b.*** Create a decision tree using Gini index with default parameters. 2
  - ***c.*** Explain which splitting criterion works well for your data and model and why. Create confusion matrix and obtain accuracy, precision, recall, specificity, and f-measure. 2
  - ***d.*** Find optimal *max_depth*, *min_values_split*, or *min_values_leaf* for your model with 5-fold cross validation. 3
- **iv.** Fit a Random Forest to your data. Evaluate and compare the results with (iii). Describe which model gives a better performance and explain the reason. 5

Instructions:

1. Completing Assignment:

- Use **Jupyter Notebook** for Assignment 1.
- Follow the provided *Jupyter notebook* **template in Brightspace's** assignment section. You can write as many lines of code and markdown cells as necessary for answering the questions.
- Include a list of all references used in this assignment in the reference section. Follow *APA referencing format* for the references.

2. **Submission Guideline:**

   i. **Preparing File**: Your submission should be a **ZIP (.zip)** file containing two files**:**

   1. A Jupyter notebook file **(.ipynb)** with cell execution results**.**
   2. A **PDF** of the Jupyter notebook with cell execution results.

   ii. **Naming the Submissions:** As part of the final printout, name your files as follows:

   a. *Naming* ***Individual*** *Files*

   i. **Jupyter Notebook:** A1_<banner_id>.ipynb

   ii. **PDF file:** A1_<banner_id>.pdf

   b. *Naming the* ***ZIP*** *file*

   - A1_<your_name>_<your_banner_id>.zip

   **Here A1 means, Assignment-1.*

   ******Note[1]: Failure to follow the naming convention may result in a deduction of up to 2 points.***
   ****** Note[2]: Failing to provide both required files (.ipynb and .pdf) will result in a score of zero (0) for the assignment.***

   iii. **Submission Process:** **S**ubmissions should be made through Brightspace, adhering to the due date and time specified time.

3. **Late Submission Policy:** Late submission will follow the following policy-

- Day 1 (within first 24 hours) – 15% reduction of the total marks
- Day 2 (24-48 hours) – 20% reduction of the total marks
- Day 3 (48-72 hours) – 25% reduction of the total marks

****** Note[3]: No submissions will be accepted after the specified time; Later submission will result in zero (0) points for the assignment.***

******Note[4]: Markdown cheat sheet and APA reference generator website can be found in the reference section of the assignment template.***