Chem Learning Challenge

Air pollution is a persistent problem of the world, which has been around for a while and will continue to exist in the future. It is a cause of a lot of problems and has extremely adverse effects. For curbing air pollution, its analysis is first necessary. For making predictions about the air quality, domain knowledge is essential, as are data analysis skills. This problem statement tests them all.

Problem Statement

Participating teams are invited to propose and implement a multiclass classification technique for a given data set of air quality measurements. We expect a well-structured report detailing the approach used for classification and its implementation. Teams would be provided with the training data set to be used. Teams are expected to brainstorm, ideate, experiment, and code classification techniques to get the best results. The goal of this challenge is to create awareness about the applications of data analysis and machine learning in the chemical engineering domain, especially air quality analysis for air pollution control applications among the student community and provide them a platform to showcase their ideas and innovations.

Data Set Information

The training dataset contains 7485 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.

Attribute Information

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
4. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
6. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
7. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. True hourly averaged NOx concentration in ppb (reference analyzer)
9. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

11. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
12. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13. Temperature in Â°C
14. Relative Humidity (%)
15. AH Absolute Humidity

Target Information

CO level is given as five broad categories Very High, High, Moderate, Low and Very Low. The target is to predict this class of CO level based on all the attributes listed

Timeline & Submission Details

- Training dataset is made available here
- Test data will be provided on 31st March 2020
- Report + Source code should be submitted to technicalaffairs@iitp.ac.in with subject "[Chem Learning Challenge] Entry from <house name>". This email should be sent by House Leaders by 23:59 hrs of 10th April 2020

Submission Guidelines

- Participating teams are expected to use any of the following programming languages for implementation:
    - Python
    - Java
    - Matlab
    - C
    - C++
    - R
- Participating Team may use standard machine learning frameworks such as TensorFlow, Caffe, Theano, Keras, PyTorch, etc
- The source code should be appropriately commented and must be accompanied by a 'Read-Me' file containing instructions to run the code. The "Read-Me' files must also specify any additional packages/resources if used. Please provide the link to download the same
- Participating Teams are expected to submit a single .zip file containing the following:
    - Source Code Files
    - Read Me File
    - Classified Output (.csv)
    - Report (.pdf)
- Naming Convention: The .zip file must have the same name as your <house>.zip.
- The report is expected to follow the given format:
    - Team Details
        - Names and Contact details of the participants
    - Introduction
    - Describe the problem statement and the need for air quality analysis

    - Classification Approach
    - Motivation
    - Methodology
    - Implementation
    - Results
        - F1-score
        - Confusion Matrix
        - Kappa Coefficient
        - Overall Accuracy
    - Conclusion

Judging Criteria

- 60% weightage is assigned to accuracy which would be the primary evaluation criterion for validating the classification technique used by the teams. The F1-score, Kappa Coefficient, Confusion Matrix and the Overall Accuracy would be the primary parameters for evaluation.
- 30% weightage will be assigned to the quality of research, novelty, and innovation in the classification technique used would be other parameters for evaluation.
- 10% weightage is assigned to the presentation, demo and the knowledge of the teams in the subject addition to the above-mentioned parameters.

Rules and Regulations

- Max Team size: 4
- The submissions will be scrutinized for forgery. Any sort of ethical misconduct will not be tolerated and will result in the disqualification of the team
- In case of any dispute, the decision of the judges or the expert panel will be final and binding on all.
- The team must adhere to the spirit of healthy competition.