CSCE 590-1: Trusted AI

Prof. Biplav Srivastava, Fall 2021

Quiz 21 / Sep 28, 2021/ Instructions

- Return answer to quiz as report (.pdf) and GitHub link by 5:00 pm on Sunday, October 3, 2021. Post the report as sub-folder “Quiz2” in your shared folder (e.g., Google folder mentioned in spreadsheet) and confirm it being done by email to biplav.s@sc.edu.
- Ask any question by email. Or, office hours and class can be used to clarify questions.

Total points = 100, Obtained =

Student Name:

GitHub link with code in a sub-dir called “Quiz2”

Q1: Understanding of Fairness Issues

[Individual effort][30 points]

Prepare a 1-page summary of your project according to 1-slide template. In particular, please demonstrate your understanding of fairness issues in selected project.

- Slide has to be put at: Google drive: https://drive.google.com/drive/folders/11YFcw42ubyJpOiaHiNZdGn1TB9hx7FT9?usp=sharing
- You need to present in class. We will spend 3-5 mins per student including Q&A.

Q2: Water treatment water data and pH value
[10 + 10 + 10 = 30 points]

Background:

pH is a very important determinant of water quality. However, its safety limits depends on water purpose.

pH considerations:

- EPA: https://www.epa.gov/caddis-vol2/caddis-volume-2-sources-stressors-responses-ph
- Standards collated: https://github.com/biplav-s/water-info/blob/master/dataWaterParameters.json
- Common practice for limit is: within 6.5-8.5 is considered safe, <= 6.5 and > 8.5 is considered unsafe
  - Example: https://www.safewater.org/fact-sheets-1/2017/1/23/tds-and-ph

Datasets:

- **Data:** Weka comes with water treatment data.
  - **Description:** https://archive.ics.uci.edu/ml/datasets/water+treatment+plant
  - **Local cache:** https://github.com/biplav-s/course-tai/tree/main/sample-code/common-data/water-weka
  - Consider the following parameters.

Q-E (input flow to plant)
2 ZN-E (input Zinc to plant)

3 PH-E (input pH to plant)
4 DBO-E (input Biological demand of oxygen to plant)
5 DQO-E (input chemical demand of oxygen to plant)
6 SS-E (input suspended solids to plant)
7 SSV-E (input volatile supended solids to plant)
8 SED-E (input sediments to plant)
9 COND-E (input conductivity to plant)

23 PH-S (output pH)
24 DBO-S (output Biological demand of oxygen)
25 DQO-S (output chemical demand of oxygen)
26 SS-S (output suspended solids)
27 SSV-S (output volatile supended solids)
28 SED-S (output sediments)
29 COND-S (output conductivity)

Things to do:

1. Data exploration: Find correlation between input and output parameter values. Example: pH-E and pH-S.

2. **Data preparation:** Add a new column called 'SAFE-PH-S'. It is 'yes' if pH is within 6.5-8.5 and 'no' otherwise, i.e., <= 6.5 and > 8.5

3. **Train**: Train a classifier to predict SAFE-PH-S using any two classification methods. Show its performance measures.

* Use 20% data for testing

* Use any standard validation method (leave one out, 10-fold cross validation)

Q3: Recent water data and pH value

[10 + 10 + 20 = 40 points]

- **Data : Multi-location data**

Datasets: We will again look at water data from Florida for WaterAtlas project.
Website: https://orange.wateratlas.usf.edu/

Data: Local cache of data
https://github.com/biplav-s/course-tai/blob/main/sample-code/common-data/water/WaterAtlas-ManySites.csv

Things to do:

1. Data preparation: Make a subset which only refers to pH data. Add a new column called 'SAFE-PH'. It is 'yes' if pH is within 6.5-8.5 and 'no' otherwise, i.e., <= 6.5 and > 8.5

2. **Train**: Train a classifier to predict SAFE-PH using any two classification methods. Show its performance measures.

* Use 20% data for testing

* Use any standard validation method (leave one out, 10-fold cross validation)

3. **Explain**: Which places have the most unsafe water (by pH) and which least by occurrence? Show them on a map using latitude longitude information available in each row.
Instructions for Google Earth are at: https://www.google.com/earth/outreach/learn/visualize-your-data-on-a-custom-map-using-google-my-maps/