TW2

Submit Assignment

Due Friday by 11:59pm **Points** 10 **Submitting** a text entry box or a file upload
File Types txt, pdf, doc, docx, png, jpg, jpeg, and zip

Learning objectives:

- Be able to understand data preprocessing process.
- Be able to identify issues exist in datasets.
- Be able to apply Python package functions for preprocessing data.

Problem-solving problems

This work should be done by your assigned team.

- Starter code: **TW2-preprocessing.zip**
  - **tw2_data_clearning.ipynb**: this is the file you will work on today.
  - data_preprocessing,ipynb: examples that show different methods for data transformation, normalization and discretization.

Your team can decide how to collaborate on solving problems.

Part 0: (http://localhost:8889/notebooks/A-CPSC4310/TW/week1/TW1/tw1_data_analysis.ipynb#Part-1:)

- Run two given examples in the notebook and understand the process of data cleaning.

Part 1:

- Apply methods discussed in Part 0 on a new datasets.
- Instructions can be found in the notebook.

Notes: Students should push an updated notebook file to his/her/their Git repo.

Part 2:

- (http://localhost:8889/notebooks/A-CPSC4310/TW/week1/TW1/tw1_data_analysis.ipynb#Part-2) Write a summary of what your team has learned from this process.

Resources:

- Lecture notes: **03_dataPreprocessing.pdf**
- Python libraries for data preprocessing

- Working with mssing data, in Pandas:

https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html (https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html)

- How to inteporate the data, in Pandas:

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html)

- Imputation of missing values, in Scikit-learn:

https://scikit-learn.org/stable/modules/impute.html#impute (https://scikit-learn.org/stable/modules/impute.html#impute)

- Preprocessing, in Sciktit-Learn:

https://scikit-learn.org/stable/modules/preprocessing.html (https://scikit-learn.org/stable/modules/preprocessing.html)

Submission(s)

- Students should push an updated notebook file to his/her/their Git repo.
  - **You do not need to submit any notebook files to Canvas**.
  - I will visit your Github to check the file.
- Part 2: Submit a summary of your learning to Canvas. Your document should include:
  - Full names of your team members who work on the assignment.
  - A summary of what you learn from the process.
  - One submission of your learning per team.