COMP 6981 – Data Preparation Techniques

Project Description

Winter 2022

This deliverable is due on the April 11th, 2022 at 10:00 PM Newfoundland time. No submissions done outside D2L will be marked (e.g., email). Please organize yourself with your team to submit the document on time. Grade deductions for late submissions will be applied, check the syllabus for the detailed grade decrease.

The main goal of the final project of COMP 6981 is to make sure that you can use and reason around all content seen in this course. You must find a dataset with some interesting topic you want to work with. Although the data set is open, your project must follow several guidelines that I will list throughout this document. First, it is recommended that you find a dataset with some real data. Having a dataset with real data is encouraged because you might want to use it in your personal portfolio if you develop an exciting tool. Some public dataset libraries include, but are not limited to, the ones listed in google public datasets (https://cloud.google.com/public-datasets), community list catalogs such as this one https://github.com/awesomedata/awesome-public-datasets#datachallenges or if you want to develop something with Canadian public datasets you may want to try this one https://open.canada.ca/en/open-data. The dataset chosen must have at least 6 attributes and 1,000 instances (i.e., examples).

Your deliverable is going to be a **Jupyter Notebook** with several cells that must described and discussed with your code commented **(10 marks)**.

Part 1 – Dataset Presentation (15 marks)

The first part of your notebook must describe your dataset. Start by describing your variables and contextualizing the dataset. The minimal requirements to have full marks in this part are:

- You must generate attribute's descriptors for all variables and discuss your results.
- You must generate plots that can give hints about the data distribution and a discussion about the results must be done.
- The decision of any descriptor statistics or plots must be justified. You will lose marks if plots are made without any justification.
- Does your data contain missing data? Are outliers present in your data? **The answer must be yes**, you should point it out the attribute(s) and how you will handle it.
    - If you are using your own dataset, please make sure you create a procedure to remove data and 'simulate' that data is missing or to replace some values with outliers.
    - At least two attributes must contain missing data (between 10%-20%) and outliers (3%-5%). Make sure you use a proper technique to 'decide' which rows will be considered outliers.
- Make sure you remove the missing data rows and the outliers from your DataFrame.

Part 2 – Data scaling (25 marks)

The second part of your project will consist of exploring data scaling techniques. The following tasks must be performed in this part:

- Choose at least 3 data scaling techniques and discuss why you are testing them. At least one data scaling technique must change the attributes' distribution.
- Plot the result, side by side, of the original and the data scaling techniques. Discuss the figure.

Part 3 – Handling missing data and outliers (25 marks)

The third part of your project must advance your pre-processing strategy to handle the missing data and outliers selected in Part 1.

- Choose one technique provided by pandas for filling missing data (e.g., bfill, ffill, etc) and fill the missing data. Plot the data and discuss potential problems with it.

- Create models using the two classifiers or regressors discussed in class (e.g., linear or KNN). Validate them and decide which one must be chosen. You must decide the best combination Classifier, Regressor + Data scaling technique. Discuss why you chose the combination. Don’t forget to justify which evaluation metric you will use to make such a decision.
- Fill the missing data and the outliers using the decided Classifier, Regressor + Data scaling technique. Make sure the rows are properly scaled before you forecast the values

Part 4 – Feature selection or Encodings (25 marks)

In the fourth part of your project, you can choose between two paths, (1) solve a feature selection problem or (2) perform data encoding to try to enhance the performance of a classifier.

Feature selection

- Define a target value to forecast (either a classification or a regression task).
- Select two feature selection techniques and justify why you chose them.
- Define as baseline model using all features and store its performance. You must consider using a repeated cross-validation to generate several estimates. You may want to choose as your selected model the one with best performance in Part 3.
- Apply the feature selection techniques and compare their performance with the baseline. Are there improvements? Yes, no, why? Try to discuss your outcomes.

Data encoding

- Find features that are suitable for encoding in your data set. You must use at least two features.
- Define a baseline model using all the attributes and store its performance. You must consider using a repeated cross-validation to generate several estimates. You may want to choose as your selected model the one with best performance in Part 3.
- Apply the feature data encoding technique(s) and compare their performance with the baseline. Are there improvements? Yes, no, why? Try to discuss your outcomes.