[MS2] Acquire and Understand the Data [03/14 - 04/2]

Key dates:
Projects assigned: Friday, March 14th
Data acquisition and verification due: April 2nd

Note that **submissions are due at 9:59pm** the day of the deadline (submission window closes at 10pm sharp). There are no late days for any of the project milestones.

Objective:

In this milestone, you will focus on preparing your dataset for subsequent analysis. This involves ensuring its quality and suitability through initial inspection. Key tasks will include identifying issues regarding missingness, class imbalance, scale of the features, and other dataset-specific challenges. Your goal for your group is to outline and execute the steps that can be taken to resolve them to ensure a robust foundation for your analysis and modeling efforts.

Milestone Checklist:

- **Access:** Download, collect, or scrape* the dataset from relevant source(s). In other words, can you get the data on your local machine or in a cloud based repository?
- **Load:** Start a new Jupyter Notebook, import necessary Python libraries (e.g., pandas, numpy, sklearn), and load your dataset.
- **Understand:** Examine the dataset. Ensure you understand what different columns/rows represent or the image/text intricacies.
- **Preprocess:** Propose or perform basic dataset cleaning to make it suitable for analysis, visualization, and modeling which you will pursue in later milestones. Document each step in your Jupyter Notebook to justify the preprocessing decisions made. Reference the next section for details on what comprehensive data cleaning and preprocessing should include.

*Projects relying entirely on data that have yet to be scraped will not be given credit. You must grab all of the data you intend to use for this milestone.

Cleaning and Preprocessing:

Below is a non-exhaustive list of issues you might want to check for and address in your dataset preprocessing:

Missing Data:

Missing data may arise due to a range of factors, such as human error (e.g., intentional non-response to survey questions), malfunctioning electrical sensors, or other causes. When data is missing, a significant amount of valuable information can be lost. Investigate the extent

and pattern of missing data. Determine the nature of missingness (Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR)) , these are CS1090a concepts, and apply the most suitable technique to address it. Options include data deletion, mean/mode imputation, or more advanced methods like multiple imputation or k-NN imputation. Justify your choice based on the dataset's characteristics.

Data Imbalance:

Imbalanced data is a common issue in classification problems when one class has significantly fewer samples than the other. When dealing with imbalanced data, machine learning models may learn to favor the majority class and make predictions that prioritize accuracy for that class. This can result in unsatisfactory performance for the minority class and reduced overall model effectiveness.

Assess the class distribution in your dataset, especially for classification tasks. If a significant imbalance is present, consider resampling techniques (oversampling minorities or undersampling majorities) or applying synthetic data generation methods like SMOTE to achieve a balanced dataset, another CS1090a content piece.

Feature Scaling:

Scaling the data is a crucial step in improving model performance and avoiding bias, as well as enhancing interpretability. When features are not appropriately scaled, those with larger scales can potentially dominate the analysis and result in biased conclusions. Standardize or normalize numerical features to ensure equal weighting in analytical models. Choose the most appropriate scaling method (e.g., Min-Max normalization, Z-score standardization) based on your data distribution and the models you plan to use.

Deliverables

To complete Milestone 2, students must submit a 1-2 page document (PDF or `.ipynb` only) by **Wednesday, April 2nd by 9:59pm** addressing the following points:

- Provide a description of the dataset.
- Discuss any potential data issues outlined above.
- Explain how these issues have been addressed before the next milestone.

The document should clearly demonstrate to the teaching staff that all potential issues outlined above have been investigated, even if the conclusion is that they are not present or do not pose significant problems.

Note: if submitting an `.ipynb` file, the code cell content will not count toward the 1-2 page document length.

Your submission should demonstrate a thorough understanding of the dataset and readiness for further analysis and modeling. This is what your TA/TF/CA will be grading you on. Please submit your document on time.

-