Project 2: Classification

Total Points: 100
Given Out: Week 5: Friday, Sept 20th, 2024
Due Date for Dataset Selection: Week 6: Wednesday, Sept 25th, 2024
Due Date: Week 7: Tuesday, Oct 4th, 2024
Project is to be done in pairs.

Background & Introduction: In the realm of machine learning and data mining, classification stands as one of the most fundamental and widely pursued tasks. At its core, classification involves identifying the category or class to which a new observation belongs, based on a training set of data containing observations whose categories are already known. It's akin to sorting items into predefined buckets or groups based on certain features or criteria.

In this project, we're setting our sights beyond just understanding the basic concept. We aim to deeply immerse ourselves in the task of classification, unraveling its intricacies from start to finish. By navigating through the entire process, from data selection and preprocessing to model training and evaluation, we'll gain invaluable hands-on experience. This is our opportunity to not only grasp the theoretical foundations of classification but also to familiarize ourselves with practical tools and techniques, like building and tuning classification models. Through this project, we aim to transform theoretical knowledge into actionable skills and to gain a better understanding of the classification task.

Project group: The project is to be done in pairs. Before we move on to the details of the project, please remember to join the project group on Canvas. (Canvas → People → Group → Choose a group and join in)

Keypoints: Just like in Project 1, we would like to emphasize the importance of creativity in the project. Most of the questions are open-ended. The more innovative your explorations, the higher the scores you stand to achieve. Dive deep, and let your analytical creativity lead the way!

Datasets: For this project, you have two exciting options when it comes to choosing your dataset:

- **Combine Datasets:** Collaborate with your partner and think creatively about how you can merge your individual datasets. This is a fantastic opportunity to discover new insights and uncover patterns you might not see in a single dataset alone!
- **Select a New Dataset:** Alternatively, discuss with your partner and choose a new dataset that meets the requirements outlined below. This gives you the chance to explore fresh data and dive into areas that intrigue both of you.

There are no restrictions on data type. Whether you're interested in text from articles or reviews, images of different subjects, or tabular data with structured information, it's all up to you. You're encouraged to find datasets that are **interesting and challenging to you**! Also, consider using this opportunity to work with textual/image data if this is the domain you are interested in. Make sure the data you choose is suited for **classification**, try to think of a classification task you are interested in solving over this dataset.

Dataset requirements: If you choose tabular data, it should contain at least 20 different features of different types and at least 200,000 records (rows). No particular requirements for textual/image/other data.

!!! Dataset approval: You do *NOT* need approval for your dataset if it contains at least 20 different features of different types and at least 200,000 records(rows).

However, if your dataset doesn't meet these requirements but you're passionate about working with it, you must ask Professor or TA for the dataset approval. In this case, submit your dataset on Canvas or ask during office hours by Wednesday, Sept 25th.

Submission and Deliverables: You will submit the project via Canvas. The submission includes two separate files (DO NOT compress your files to .zip): (1) A report that provides the tasks detailed below accompanied by explanations and (2) execution source codes that can reproduce the reported results.

1. Your report doc can be a .pdf file or .docx file. The file needs to contain the results (e.g., figures, tables) of each task and your analysis. If your answer contains figures/images, please take a screenshot or download them, and then insert them into your results doc.
2. Your code can be .py file or .ipynb file.
3. **Rename the files using your names** (student1's first name_student1's last name_student2's first name_student2's last name)**.** Eg., roee_shraga_yao_su.pdf roee_shraga_yao_su.ipynb

Using Generative-AI (e.g., ChatGPT): I encourage responsible and sensible use in assignment. Please report if, how, and what you used to complete the assignment. Explain how you validated the trustworthiness of the solution, which prompt/s you used, and you used the output of the model (basic code/documentation/etc.)

Demonstration: Once completed, several students may be asked to provide a brief demonstration of their results to your classmates. The instructor/TA may also request a one-on-one demonstration to support grading.

Task 1 – BYOD (Bring Your Own Data) [10 points]

Craft a compelling introduction for your project. Describe the dataset that you are going to use, introduce the background, explain its significance, and share your motivation for choosing it. You don't have to finalize the specific problem that you are going to work on just yet but you should have a sense of what you are targeting.

Highlight the potential impact of a successful solution and anticipate any challenges ahead. This introduction should encapsulate the essence of your project, drawing readers in with clarity and conviction, while setting the stage for the details to follow.

Task 2 – EDA [10 points]
Perform exploratory analysis on the data. Research online for ideas, and then show analysis on **at least five** different aspects of the dataset. Analyze your findings.
Note 1: This part is open-ended, any valid analysis is fine as long as it useful for you to understand the data better! Use visualizations when necessary.
Note 2: If you choose to combine your Project 1 datasets, make sure the EDA highlights the integrated nature of the dataset (**do not reuse the two independent EDAs from Project 1)**.

Task 3 – Problem Definition (Revisiting Task 1) [5 points]
Based on Task 2, revisit Task 1 to write a compelling introduction for your project that also highlights the problem that you are going to solve and some observations to support the choice, motivation and significance based on your EDA.

Task 4 – Preprocessing [15 points]
Perform data processing on the data. Dive deep into data preparation, ensuring that the dataset is primed and ready for model training. Document all preprocessing steps and include them in the report, specify what happened to the data after the step, why you performed it and if you are going to use it moving forward.
For example:

1. **Data Cleaning:** Address any inconsistencies within the data. This includes dealing with missing values, anomalies, and duplicated entries.
2. **Data Transformation & Normalization:** Depending on the nature and distribution of your data, apply necessary transformations. Ensure features are on a similar scale, making them more amenable to analysis and model training.
3. **Specific Data Type Processing:**
   a. Text Data: Implement tokenization to break down text into smaller chunks, remove stop words to filter out common but uninformative words, and utilize feature extraction techniques, like TF-IDF or word embeddings, to represent text in a form suitable for machine learning.
   b. Image Data: Normalize image pixel values, ensuring they fall within a consistent range (typically 0 to 1). Consider image augmentation techniques to artificially enhance your dataset, creating varied representations of the same image to improve model robustness.

Task 5 – Model Selection, Training, and Optimization [30 points]
In this task, you'll identify and train **at least five** distinct machine learning models apt for your classification task. Following that, try to boost the performance of the models using suitable techniques. For example, feature engineering, cross-validation, grid search, tuning model hyperparameters, etc. Try **at least three** techniques and compare the performance with the vanilla ones.

Task 6 – Model Evaluation [30 points]

1. Use **at least four** metrics to evaluate the performance of your selected models.
2. Analyze the experimental results and report your conclusion. The analysis should compare results (using the metrics you chose above) the different pre-processing steps (Task 4), the different models and optimization steps (Task 5) in a table/graph. Report your observations discuss their implications (e.g., Model X with Optimization Y after preprocessing step Z was the most/least effective because ABC…).