ECE 443 (Fall 2024) – Term Project Instructions (100 points)

Last updated: October 20, 2024

The term project has three main phases, with three distinct deadlines, as described below. All submission times, unless otherwise stated, are 11:59 PM Eastern Standard Time and there is no possibility of an extension. A student/team failing to deliver on one of the phases will forfeit points for the subsequent phase(s).

1 Phase I (Due: November 15, 2024)

Phase I of the term project involves the following deliverables in a single PDF file.

1.1. (2 points) **Formation of teams:** Each team must have no more than three and no less than two members, with a designated point-of-contact (POC); the POC is the only one who would submit the files required for grading of the different phases of this term project.

Specific deliverable: Provide a chosen name for the team and a list of names of the team members, with the POC of the team clearly marked on the list, as part of this Phase I submission.

1.2. (5 points) **Declaration of project datasets:** The project needs to revolve around four different datasets for a three-person team and three different datasets for a two-person team, with each dataset being different from the other in terms of nature/modality, as per the following requirements:

- Each one of the datasets can be a tabular dataset, a time-series dataset, an imaging dataset, a video dataset, a text dataset, a sound/speech dataset, a multimodal dataset, etc.
- One of the declared datasets must be a tabular dataset, with the number of samples at least 200; i.e., $n \geq 200$, and the number of raw features (attributes) at least 20; i.e., $p \geq 20$.
- It is important to note that:
  - **Tabular dataset:** While many datasets are provided in a row-and-column format, tabular datasets specifically refer to structured data where each row represents an independent sample (or instance), and each column represents a distinct feature (or attribute). Examples include demographic data, financial transactions, or any dataset where relationships between samples are not dependent on sequence or time. For instance, an image is not considered tabular data, even though pixel values can be represented in a matrix, because the pixel arrangement has spatial relationships and is not treated as independent features.
  - **Time-series dataset:** A time-series dataset consists of data points collected at successive intervals, where the temporal order is important. The data points are connected over time, and their sequence matters for analysis. Simply having a column with timestamps does not make a dataset time-series; the relevance of time between data points is crucial. Examples include stock prices over time, weather data, or sensor readings where changes and patterns over time are significant.

Specific deliverable: Provide a brief summary of each dataset, as well as the source URL for each dataset, as part of this Phase I submission.

1.3. (3 points) **Declaration of project tasks:** Specify the machine learning tasks that will be performed in relation to each dataset (one, and exactly one, task per dataset). Collectively, when looking at the declared datasets together, you must tackle two out of three tasks of classification, regression, and clustering. That is, all classification (or regression or clustering) tasks are not acceptable. Note that you do not need to finalize (or declare) the methods you will use to solve the tasks.

Specific deliverable: Briefly discuss what motivated you to select the declared datasets and corresponding tasks.

 waheed.bajwa@rutgers.edu

Some online resources for datasets:

- `https://www.kaggle.com/datasets`
- `https://archive.ics.uci.edu/datasets`
- `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
- `https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research`

2 Phase II (Due: December 18, 2024)

Phase II of the term project involves the following set of deliverables.

2.1. **Creation and submission of separate notebooks, one for each of the datasets/tasks:** The declared machine learning task for each dataset must be carried out in a well-commented Jupyter notebook, with careful discussions/explanations in markdown cells of the notebook. Additional guidelines for this part of Phase II of the term project include:

- While you are allowed to use packages such as `sklearn` for this phase of the project, you *must only* use those modules/functions that you fully understand **and** you must explain the usage of those modules/functions in markdown cells.
- You must name each of the notebooks as `<TeamName>_Dataset<n>.ipynb`, where you would replace `<TeamName>` with your team's name, and `<n>` with 1, 2, 3, 4 for each dataset. As an example, for a two-person team named Anaconda, the notebooks should be named as `Anaconda_Dataset1.ipynb`, `Anaconda_Dataset2.ipynb`, and `Anaconda_Dataset3.ipynb`.
- You must ensure that your submitted notebooks are fully executed, so that they are not required to be rerun during grading. *Please double- and triple-check this to ensure compliance with this requirement.*
- The following breakdown of points for this part of Phase II of the term project should guide your code development for each notebook.

(a) (3 points) **Brief exploration of each dataset:** Carry out a brief exploration of the dataset, such as the number of samples, the number of raw features, the fraction of missing values (if any), the number of categorical variables (if any), histograms of different variables, etc. This exploration should be accompanied with detailed commentary in markdown cells.

(b) (3 points) **Pre-processing of each dataset:** Carry out preprocessing of each dataset, which should be guided by your exploration of the dataset as well as your forthcoming plans for the datasets. This preprocessing could involve, e.g., replacement of invalid entries with plausible values, centering of the data, standardization of the data, encoding of categorical variables, etc. All of the preprocessing steps should be fully motivated and justified in markdown cells.

(c) (6 points) **Feature extraction / feature learning from each dataset:** Depending on the dataset, engage in either feature engineering or feature learning for that dataset. In the case of text dataset, e.g., this would involve transforming the raw text into numerical features. In the case of large images or correlated numerical variables, e.g., this could involve using something like *principal component analysis* (PCA) to reduce the dimensionality of images or to decorrelate different variables. All of the steps involved in this feature extraction / feature learning component should be fully motivated and justified in markdown cells.

(d) (32 points) **Processing of each dataset using *two different* machine learning methods:** Carry out the declared task on each dataset using two different machine learning methods, with the parameters for each method (where applicable) carefully tuned using cross-validation, the results averaged over multiple validation folds, and the final results presented in an aesthetically pleasing manner. In addition, use markdown cells to justify different steps in your implementations and explain different aspects of the two methods as much as possible.

(e) (8 points) **Comparative analysis of the two methods on each dataset:** Provide a comparison between the two machine learning methods for each dataset across dimensions such as computational complexity,

performance, etc., and a final recommendation on the method that should go into production for each dataset. This comparison should include both coding cells (e.g., overlayed plots, side-by-side confusion matrices, etc.) and markdown cells for discussion.

(f) (5 points) **Discussion on ethical issues for each dataset/task:** Provide a discussion on the ethical aspects of the machine learning tasks that you carried out on the declared datasets. This discussion should be carried out in a markdown cell and should be carefully formatted for readability purposes.

(g) (3 points) **Bibliography for each notebook:** Provide bibliographic references that helped you during the preparation of the notebook. These references, which should be provided in a markdown cell at the end of the notebook, should be referenced within the body markdown cells of each notebook as much as possible.

2.2. (15 points) **Video Presentation:** Prepare a 10-minute (or shorter) video presentation that summarizes your efforts as part of the term project. The presentation should be based entirely on slides, with code snippets integrated into the slides to highlight important points. The purpose of the video is to demonstrate your understanding of the various aspects of the submitted notebooks and to convince the teaching staff of your team's thorough grasp of the project. The presentation should be uploaded on Canvas on its respective assignment. A few reminders:

- This should be a team presentation, so avoid simply recording separate sections and stitching them together. The presentation must flow smoothly as a unified effort from all team members.
- Videos exceeding 10 minutes will lose points for each additional minute, rounded to the nearest minute. For example, 10.4 minutes will count as 10 minutes, but 10.6 minutes will count as 11 minutes.
- Avoid scrolling through your notebooks during the presentation. Instead, use slides with code snippets to keep the presentation professional and clear.

3 Phase III (December 21, December 22, and December 23, 2024)

Phase III of the project will consist of a 10- to 15-minute in-person Question-and-Answer session involving the teaching staff and each team that submitted a video presentation. In other words, teams that do not submit a video presentation will lose points for this phase of the project, and this may also affect their grade for Phase II. Teams will have the opportunity to sign up for the Q&A session on a first-come, first-served basis, with sessions scheduled from December 21 to December 23, 2024.

3.1. (15 points) Come prepared to the Q&A session with a thorough understanding of your project. Based on your video presentation and submitted notebooks, the teaching staff will ask you five to ten technical questions. *The answers to these questions will also be used to assess parts of the submitted notebooks*.

4 Grading Rubric Guidelines

Each component of the project will receive a letter grade, which will then be converted to a numerical grade according to the scale in Table 1. The grading will follow the rubric outlined below:

- **Baseline grade:** Each component of the project starts with a baseline grade of **B (80%)**. If all the required elements are adequately addressed, the grade will remain at B.
- **Earning above B:** To achieve a grade above B, teams must demonstrate extra effort, depth, and clarity in their work. This includes detailed analysis, proper documentation, thoughtful preprocessing, rigorous evaluation of models, and similar contributions that go beyond basic expectations.
- **Falling below B:** Grades drop below B when critical components are missing, insufficiently explained, or executed incorrectly.
- **Grade for each dataset:** Every dataset is evaluated separately, meaning that teams receive feedback and grades for each dataset.

 waheed.bajwa@rutgers.edu

Letter Grade	Numerical Grade (%)
A+	100%
A	95%
A-	90%
B+	85%
B	80%
B-	75%
C	65%
C-	60%
D+	55%
D	50%
D-	45%
E+	40%
E	35%
E-	30%
F+	25%
F	20%
F-	15%
G+	10%
G	5%
G-	0%

Table 1: Letter Grade to Numerical Grade Conversion

- **Team grades:** In general, team members will receive the same grade. However, if one or more team members fail to contribute meaningfully (e.g., lack of participation, missed deadlines), they may be excluded from the team's final grade and assessed separately. *Communication and documentation are key to verifying each member's contributions.*

Here is some more information on what might earn you points in different categories:

- **Dataset Exploration (3 points – B baseline, 80%):**
  - **Going Above B:** Students can earn higher grades by providing thorough insights into the data, such as identifying trends, anomalies, or correlations and discussing their potential impact on future modeling (e.g., receiving an A- or A).
  - **Dropping Below B:** If the dataset exploration lacks depth or insight and simply includes basic visualizations without meaningful commentary, the grade will drop (e.g., C or lower).
- **Preprocessing (3 points – B baseline, 80%):**
  - **Going Above B:** Well-structured preprocessing that demonstrates an understanding of data issues (e.g., handling missing data, avoiding data leakage, scaling features) will earn higher grades. Proper explanation and reasoning for preprocessing steps are essential for an A- or A.
  - **Dropping Below B:** Minimal preprocessing, or steps that are incorrectly applied (e.g., applying standardization before splitting data), results in lower grades (C or lower).
- **Feature Extraction / Learning (6 points – B baseline, 80%):**
  - **Going Above B:** A thorough approach to feature engineering (e.g., dimensionality reduction, transforming text data) with solid explanations and justified decisions will earn a higher grade. Innovative or well-applied methods will push this section to A-, A, or A+.
  - **Dropping Below B:** A lack of meaningful feature extraction, or not performing feature engineering when it is needed, will lower the score (C or lower).

 waheed.bajwa@rutgers.edu

- **Two Machine Learning Methods (32 points – B baseline, 80%):**
  - **Going Above B:** Effective implementation of machine learning methods, including parameter tuning and performance evaluation, will raise the grade. Detailed markdown cells explaining model choices, tuning, and cross-validation will lead to A- or higher.
  - **Dropping Below B:** Poorly implemented models, lack of parameter tuning, or missing evaluation metrics will drop this section to a C or lower. Incomplete or incorrect code, especially without justification, may result in an F.
- **Comparative Analysis of Methods (8 points – B baseline, 80%):**
  - **Going Above B:** A detailed comparison of the models based on metrics like accuracy, runtime, or confusion matrices will earn a higher grade. Thorough, quantitative analysis with clear visualizations can lead to an A.
  - **Dropping Below B:** If the comparison is lacking or based on incomplete results, the grade will drop (e.g., C or lower). A superficial comparison without actual results could earn a failing grade (F).
- **Discussion of Ethical Issues (5 points – B baseline, 80%):**
  - **Going Above B:** A thoughtful, detailed analysis of ethical concerns specific to the project, such as bias in data or fairness in predictions, will earn a higher grade (A- or A).
  - **Dropping Below B:** A vague or irrelevant discussion of ethics will lower the grade (C or lower). Surface-level commentary that does not engage with the ethical challenges specific to the data or models used will receive a low grade.
- **Bibliography (3 points – B baseline, 80%):**
  - **Going Above B:** A well-organized, properly formatted bibliography with inline citations throughout the project will earn a higher grade (A- or higher). Demonstrating thorough research through quality references is key.
  - **Dropping Below B:** A minimal or poorly formatted bibliography, or failure to include inline citations, will result in a lower grade (C or lower).
- **Video Presentation (15 points – B baseline, 80%):**
  - **Going Above B:** A well-structured presentation that is engaging, well-paced, and clearly presents the key points using slides and code snippets will earn a higher grade (A- or A). Clear explanations of results and explicit discussion of ethical issues or challenges faced will also raise the grade.
  - **Dropping Below B:** Disjointed presentations, overly long presentations, or failure to use slides and code snippets effectively (e.g., scrolling through notebooks) will result in lower grades (C or lower).

 waheed.bajwa@rutgers.edu