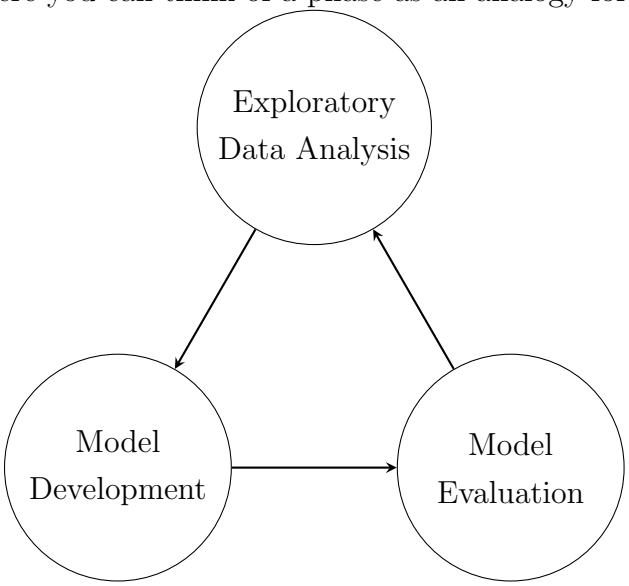
Practice Competition

1 Practical Machine Learning Development

Students are expected to follow a development process for practical Machine Learning project which is structured with three key steps: Exploratory Data Analysis, Model Development, and Model Evaluation. The following figure illustrates this development process as an iterative phase where you can think of a phase as an analogy for Agile's sprint in casual Software Development.



1.1 Exploratory Data Analysis (EDA)

This step concerns the analysis of obtained data to understand its characteristics and to identify potential issues that may affect modeling. There are several tasks related to this step, including but not limited to:

- **Data collection:** Gather data either to have a preliminary dataset for the first time or to supplement the data from previous phase(s).
- **Data pre-processing:** Clean the data by handling missing values, normalizing or standardizing features, and encoding categorical variables as needed.
- **Feature engineering:** Extract or construct new features as you need to augment the newly obtained data or to mitigate problems from prior phase(s).

1.2 Model Development

This is when you develops models to capture the patterns revealed during the analysis. This step involves selecting appropriate modeling approach and design a detailed algorithm.

- **Model selection:** Research to find an appropriate modeling algorithm that may help you capture the patterns within this data as well as to define relevant information about the model, including its hyper-parameters and training procedure.
- **Training:** Train the selected algorithm with chosen approach on the prepared dataset to obtain a final model.

1.3 Model Evaluation

You will then evaluate that model's performance using metrics relevant to the problem as well as desired results of the project, and perform analysis on the trained model's limitations to identify rooms for improvement.

- **Evaluation:** Assess the model using metrics such as mean squared error, mean percentage error, or root-mean-squared logarithmic error for regression problems and accuracy, precision, recall, or AUC-ROC for classification problems..
- **Limitation Analysis:** Analyze the erroneous cases that are incorrectly predicted by the model, which are reflected by the above metrics, and identify their properties as well as reasons for their inaccuracy.

2 Competition Description

2.1 Platform

We will use Kaggle to host a competition where students can participate either as **individual** entrants or in teams of **two**. The naming of your Kaggle team, be it 1 or 2 members, should follow the **TeamName** convention specified under the final section of this project description. Students and teams are expected to use Kaggle-based competition's **discussion channel** for discussing relevant information and raising project-related questions. The competition details and submission portal are available at [https://www.kaggle.com/t/1ad984cd42bc40b7bc76980accffef5d].

2.2 Project Schedule

The project is scheduled to occur over a total of 3 weeks, structured as 3 distinct phases. Each phase spans one week and encompasses the full cycle of the three main steps: Exploratory Data Analysis (EDA), Model Development, and Evaluation and Error Analysis. Each phase requires students and/or teams to explore, experiment, and evaluate a specific framework of Machine Learning as specified under the section of Assessment.

This progressive approach is designed to encourage steady progress—students or teams should actively work on the project each week rather than postponing efforts until the final week. A minimum of one submission per week is required, with a maximum limit of one submission per day. Prior to the end of the Kaggle deadline, each student or team must select one submission as the final entry for scoring, which will then be used to calculate the bonus point.

3 Submission

Your report and source code must be submitted in a compressed ZIP file (`.zip` extension) named according to the format **TeamName.zip**. If the submission file is large, you may upload it to Google Drive and provide a text file with the shared link (ensure the "last updated" field is before the deadline).

3.1 Source Code (Jupyter Notebooks)

The source code and results should be reported in Jupyter Notebooks with the following requirements:

- Include student information (Student ID, full name, etc.).
- Provide detailed explanations for each step with illustrative images, diagrams, and equations.
- Fully comment every processing step, and print intermediate results for observation.
- Ensure the notebooks are well-formatted.
- Before submitting, re-run the notebook (`Kernel` → `Restart \& Run All`).

3.2 Report

Prepare a detailed report that includes:

- Student information (Student ID, full name, etc.).
- Self-evaluation of the assignment requirements.
- **Phases' documentation:** Organize your analysis into separate sections where each of them corresponds to a phase, then provide detailed results for each step performed within that phase:
  - **EDA:** Describe the data characteristics and any potential issues identified.
  - **Model Development:** Present the model selection, architecture design, and training setup along with the reasoning behind your choices.
  - **Model Evaluation:** Report on model performance using appropriate metrics, also including observations on running time, computational resource usage. Provide detailed examination on any kind of errors that student and/or team has identified as the model's weakness.
- Discussion of the Pros and Cons for different modeling approaches and frameworks that the student and/or team has experimented with.
- Ensure the report is well-formatted and exported to PDF.
- Include any references in a properly formatted bibliography section.

4 Assessment

No.	Details	Score
1	Phase 1: Scikit-learn	3 points (1 point for each step)
2	Phase 2: Choose one or more framework between PyTorch and TensorFlow	3 points (1 point for each step)
3	Phase 3: Choose at least two frameworks among XGBoost, LightGBM, and CatBoost	3 points (1 point for each step)
4	Final Project Report	1 point
5	Bonus (Competition Ranking: Gold/Silver/Bronze)	0.6 / 0.4 / 0.2 points
	Total	**10 points + Bonus (max 0.6)**

5 Notices

- This is an **INDIVIDUAL** or **TEAM** project.
  For **INDIVIDUAL** submissions, **TeamName** is your StudentID, .i.e **StudentID** and **StudentID.zip** respectively for your Kaggle team and the final submission.
  For **TEAM** submissions, the **TeamName** is **StudentID1_StudentID2**. Aside from being graded with stricter standards, the bonus earned from rankings in the competitions will be evenly *divided* among the team's members. Specifically, each member of the team ranked at Gold/Silver/Bronze will receive a bonus of 0.3 / 0.2 / 0.1 points.
- Duration: Approximately 3 weeks.
- Any plagiarism, use of tricks, or dishonesty will result in a zero score for the course grade.

The end.