Milestone 1 – (Max 20 points)

In this part of the project, you would build your team and establish your goals.

1. **Team**: The team must comprise 3 members of your choice. Select team members who are responsible and responsive and who will pull their weight in the final project. You must also provide a detailed breakdown of each team member's roles.

2. **Context**: Next, establish the context of the problem you would work on for your final project. When writing the context, provide detailed information on why the problem is important, and what aspects can benefit from applying machine learning to it. Please include all information sources which helped the team formulate the problem. An example of how to write detailed context can be found in Assignments I and II. The context must provide accurate and descriptive responses to the following questions:
    a. What is the problem you have identified?
    b. Why is it a challenging problem?
    c. How can it be solved through machine learning?
    d. What aspects of the problem are you going to solve?
    e. Why is it relevant to the world? How does it help

3. **Dataset**: Once the context is established, please indicate from where and how the team will gather the data. If the dataset for your project already exists, please include the link to the dataset. Here you must also include metadata like how this dataset was gathered, and it's size (number of features, number of samples). For instance, if you download your dataset from Kaggle, you must include the Kaggle link and the original data source link from where the dataset was downloaded and posted on Kaggle. Do not work on the datasets for which no source information is available. Please indicate whether you will use a small subset of the data or features for your project or the entire dataset, and why.

4. **Proposed Solution**: Provide the specifics of the ML challenge and your proposed solution. This part should answer the following questions.
    a. Is the problem predictive or inferential?
    b. Which variables are involved as features and target variables?
    c. Which ML technique do you plan to use for the project and why? You may use a technique taught in class (e.g. SVM's, Logistic Regression, Neural Networks, CNNs, RNNs) or select a new technique from the book. Please note that using simplistic models like Linear Regression and Polynomial Regression is NOT permitted for the final project.
    d. How will you pre-process the data, and what will be your evaluation strategy? Why is this evaluation strategy optimal for this problem?
    e. Are there any existing solutions to the problem? Unless you have come up with an entirely novel problem and solution, you must indicate sources from

where your solution (and problem) was inspired. Research papers, books, or problems taken from ongoing ML challenges are all acceptable sources. The solutions must be well thought out and should not be copied or plagiarized from open-source notebooks.

f. Libraries you intend to use for the project.

Please discuss with TAs and me on what comprises of an acceptable project solution.

Milestone 1 deliverables:

A **pdf** file containing all the above details in Times New Roman, Font Size 12, Maximum 2 pages. Any files larger than 2 pages or with different font sizes and styles WILL NOT be evaluated. Please name the file like <member1_id>_<member2_id>_< member3_id>.pdf. Keep your report direct, and brief, and include accurate and to-the-point responses.

Please make sure your project is non-trivial, and has some **visible** effort put into either 1) data collection techniques, 2) modeling approach, or 3) evaluation strategy. For instance, replicating example projects from sklearn library clearly shows that you have put no effort into your project. Below are some examples of acceptable final projects that require effort and creative thinking.

Example 1: You may select a dataset from a public source (e.g. Canada Open Data) and build a predictive model to support some of the ongoing issues of public interest. Read up on the data page, to find the list of issues that data might address.

Example 2: You may select an open-source dataset from the UCI machine learning repository and perform a comparative analysis of (at least 3) different regression/classification techniques on the same dataset. This project will focus on how different techniques compare against each other when addressing the same problem.

Example 3: You may use any open-source "toy" datasets to implement proposed new modeling or visualization techniques in Machine Learning from research papers. See papers published at ICML, NeurIPS, AAAI, and IEEE VIZ conferences and journals.

Example 4: You may design novel creative and engaging tasks (please refer to lectures on crowdsourcing) to gather data on human intelligence. This data may be used to build predictive or inferential models you have learned in class that replicate human intelligence.

Example 5: You may compare how reducing the number of samples using different active learning strategies can impact model performance for neural networks (see active learning lectures for this). Here you focus on quantifying uncertainty for neural networks and may use uncertainty sampling to reduce the number of samples necessary to achieve the same performance measures as with a full/large dataset.

Milestone 2 – (Max 40 points)

For this milestone, submit the preliminary implementation of your solution and share some preliminary results. The deliverables must include:

1. The dataset and its associated preprocessing operations. If you are gathering data by designing a crowdsourcing task, include your fully functional task interface design ready to be deployed for data collection.
2. Feature engineering operations. This may include correlation analysis to select your features, feature augmentation strategies used (image rotation, segmentation), and any dimensionality reduction techniques used.
3. Preliminary model implementation. Identifying the loss/cost function, implementing the gradient descent, and performance measures.
4. Preliminary evaluation strategy. This may include any cross-validation, bias-variance evaluation, or any other systematic evaluation strategy (e.g. evaluating on benchmark dataset). Please remember, that evaluating the model is the MOST important part of any ML system. Submitting your models, without any suitable evaluation will lead to a significant reduction in final grade.
5. You may communicate your preliminary results by printing out performance numbers on the console. Graphs are NOT necessary at this stage.
6. You must also provide a detailed breakdown of what each member of your team worked on. Please note that it will NOT be acceptable if 1 team member works on the report, while others work on code. That is NOT how engineering teams in ML operate in the real world.

Milestone 2 deliverables:

1. This milestone has 2 deliverables
    a. the code and data in a zip file,
    b. a report in a **pdf** format containing all the above details in Times New Roman, Font Size 12, Maximum 2 pages.
2. Any files with greater than 2 pages or with different font sizes, WILL NOT be evaluated.
3. **IMPORTANT**: Please merge these 2 pages with the previous 2 pages of your milestone 1. So you will submit 4 pages of the report in this milestone (max 2 pages from the previous milestone + max 2 pages from this milestone)
4. Please name the files as <member1_id>_<member2_id>_< member3_id>.pdf and <member1_id>_<member2_id>_< member3_id>.zip.
5. The report must provide details about all the techniques used for steps 1-5 and 6.
6. You may provide your data as a CSV file (put it inside the zip folder). If you are pinging a live server, or doing web scrapping, you may provide the code used for data gathering.
7. It is the team's responsibility to ensure that the data and the training model can be reproduced without any glitches by the person who is grading.

Milestone 3 – (Max 40 points)

After you execute some preliminary modeling experiments on your dataset, you may realize that the ML system performance is not what you expected, or your proposed solution is not working the way you had planned. In this milestone, you will identify all the weaknesses of the system you built in the previous milestone and improve them systematically. For instance, you may realize that your target variable is not accurately capturing the problem at hand, and you may alter your data-gathering task to improve the model performance. Deliverables in this milestone may include:

1. Improving model parameters.
2. Changing pre-processing techniques or data sampling strategies.
3. Improving evaluation strategy.
4. Selecting different sets of features.
5. Improving strategies for analyzing features.
6. All necessary graphs and charts to support performance claims.

Please note that at this stage, you will NOT change the problem definition, the dataset, or the project entirely. You may, however, change the modeling technique in case it does not perform well. In this deliverable, you will submit the polished version of your code and pdf file. The code should be well-commented, structured, readable, and free from any runtime errors. The report should be clear, include your charts, and describe all improvements you made from milestone 2, why, and how.

Milestone 3 deliverables:

1. This milestone has 2 deliverables
    a. the code and data in a zip file,
    b. a report in a **pdf** format containing all the above details in Times New Roman, Font Size 12, Maximum 2 pages.
2. Any files with greater than 2 pages or with different font sizes, WILL NOT be evaluated.
3. **IMPORTANT**: Please merge these 2 pages with the previous 4 pages of your milestones 1 and 2. Therefore, you will submit 6 pages of the report here (max 4 pages from previous two milestones + max 2 pages from this milestone)
4. The code should be your final polished version, well commented with clear indications of what you have changed from milestone 2.
5. Please name the files as <member1_id>_<member2_id>_< member3_id>.pdf and <member1_id>_<member2_id>_< member3_id>.zip.
6. The report must provide details about all the techniques used for steps 1-5. It must include the weaknesses identified and strategies used to improve them.
7. In your report, also add a **references** section, that includes references to all sources that helped you build your project. This should include references to lectures, class demo code, book chapters, research articles, and even blog posts.