CS-414 Intro to Machine Learning Final Project

Here are the steps to follow:

1. **Form a group**: Each group should have up to 4 students. Make sure to select team members who have complementary skills and with whom you can work effectively (there will be a Peer-evaluation at the end)
2. **Select a problem:** You will need to pick a problem that you are interested in and that you can solve with machine learning. It can be any problem that can be addressed with supervised, unsupervised, and/or reinforcement learning. You could consider problems in areas like healthcare, finance, sports, e-commerce, or any other area that interests you.
3. **Gather data:** Once you have identified a problem, you will need to find relevant data that can be used to train and test your machine-learning model. You could use datasets that are already available (recommended) or collect your data if necessary (be advised that data collection is very time-consuming, so it is only recommended for testing).
4. **Implement an ML model**: You will need to implement a machine learning model to solve the problem you have selected. You can use pre-trained models or train your models from scratch. You should choose the model that best suits your problem and dataset.
5. **Create a self-sustained notebook:** You will need to create a self-sustained notebook that runs/implements your model. This notebook should include all the code necessary to reproduce your experiments and results. It should also have text explaining all your steps and rationale. Be advised that your code can pull from a GitHub repository your or an existing model weights (no need to repeat the training every time someone runs the notebook).
6. **Record a video:** You will need to record a video explaining your notebook and hosted on the cloud(like on YouTube or GitHub). This video should provide an overview of your problem, data, model, and results. You should also highlight any challenges you faced and how you overcame them, as well as future next steps.
7. **Write an executive report:** You will need to write an executive report that explains the rationale for your decisions and how they address the problem you selected (the key here is you need to show you address the problem not that you really solved it 100%). This report should be clear, concise, and well-organized (max 5 pages total).
8. **Create a GitHub Repo:** You will need to create a GitHub repository that contains all the material you have produced for this project, including the self-sustained notebook, video, executive report, and any other material you need for your project (like models weights, etc.).
9. **Create a GitHub static website:** Finally, you will need to create a GitHub static website to showcase your project. This website should include a brief overview of your problem, data, model, and results, as well as links to your notebook, video, and executive report (If you don't want to use your own GitHub account to host the static page, just let me know).

Overall, the final project should demonstrate your understanding of machine learning concepts and your ability to apply them to real-world problems. You should also showcase your ability to work collaboratively with others and to communicate your ideas effectively. I would highly encourage you to share with me ideas of problems and/or datasets as soon as possible to ensure you are on the right track and/or it would be something feasible that would meet the scope of this final project. Be advised this is 25%!!! Of your final grade, which should tell you know much time and effort I am expecting you to put into this. While you are encouraged to use GenAI to help you with all this, remember it is ultimately your responsibility to ensure your project meets the requirements. If you need ideas we can also discuss this on a 1-group basis.

Where can you find datasets for your final project?

Here are some websites that you can use to find datasets for your final project:

1. Kaggle: https://www.kaggle.com/datasets Kaggle is a popular platform for machine learning competitions and also hosts a large collection of datasets on various topics.
2. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for empirical analysis of machine learning algorithms.
3. Google Dataset Search: https://datasetsearch.research.google.com/ Google Dataset Search is a search engine for finding datasets hosted by a wide variety of organizations, including universities, governments, and research institutions.
4. Data.gov: https://www.data.gov/ Data.gov is the home of the U.S. government's open data, providing access to thousands of datasets on topics ranging from health to transportation to energy.
5. OpenML: https://www.openml.org/ OpenML is an online platform for discovering and sharing machine learning datasets and experiments. It provides a range of datasets, benchmarks, and competitions that are updated regularly.
6. Data.world: https://data.world/ Data. World is a platform for discovering, sharing, and collaborating on datasets. It contains a large collection of datasets on various topics, including business, sports, and politics.
7. Reddit Datasets: https://www.reddit.com/r/datasets/ The Reddit Datasets community is a forum for sharing and discovering datasets on a wide range of topics. Members of the community post links to datasets and share information on how to use them.

Where can you find pre-trained models for your final project?

Here are some websites where you can find pre-trained machine-learning models:

1. TensorFlow Hub: https://tfhub.dev/ TensorFlow Hub is a repository of pre-trained machine learning models that can be easily integrated into your own projects.
2. PyTorch Hub: https://pytorch.org/hub/ PyTorch Hub is a library of pre-trained models for PyTorch, a popular deep-learning framework.
3. Hugging Face Models: https://huggingface.co/models Hugging Face is a popular open-source library for natural language processing (NLP), and it also provides a large collection of pre-trained models for various NLP tasks.
4. Model Zoo: https://modelzoo.co/ Model Zoo is a repository of pre-trained machine learning models in various frameworks, including TensorFlow, PyTorch, and Keras.
5. NVIDIA NGC Model Repository: https://ngc.nvidia.com/catalog/models The NVIDIA NGC Model Repository is a collection of pre-trained models optimized for NVIDIA GPUs, including models for computer vision, natural language processing, and speech recognition.

How would this project be graded?

Since there would be a large variety of projects (e.g., using complex vs simple pre-trained models vs own models vs very large datasets vs very challenging problems, no need for data preprocessing, etc) some of the grading schemes might vary a bit, but the relative importance of the steps would not change.

Here is an example of how I could distribute grades based on the importance of each step:

1. **Selecting a problem to solve and finding a dataset** - *5 points* This step is crucial as it sets the foundation for the rest of the project. The quality of the problem and dataset selection would determine how well the team can apply machine-learning techniques to solve a real-world problem.
2. **Preprocessing and exploring the data** – *5 points* Preprocessing and exploring the data are crucial for understanding the data and preparing it for use in the machine learning model. The quality of data preparation and exploratory analysis would affect the quality of the final model. For any preprocessing done, the team needs to explain why they did it.
3. **Defining the evaluation metric(s)** - *5 points* Defining the evaluation metric is important as it determines how the performance of the machine learning model would be measured. The team must choose an appropriate evaluation metric based on the problem they are trying to solve and provide an explanation as to why they are using that metric(s).
4. **Training the model/Implementing the appropriate pre-trained model-** *13 points* Training the machine learning model involves using the selected dataset to optimize the model's parameters. The team must ensure that the training process is appropriate and that the model is not overfitting and/or underfitting the data. If using a pre-trained model they need to make efforts to

tune the model to their unique case. If this is not possible (given the magnitude of the model, limited dataset, etc) they need to explain it and provide an explanation of how they would do it given enough resources. Most importantly they need to explain and provide supporting evidence why they decided to use this model.

5. **Evaluating or Fine-tuning the model** - *5 points* Evaluating the machine learning model involves measuring its performance on a validation set or through cross-validation. Whenever possible, your team should also fine-tune your model by performing hyperparameter tuning. The team must ensure that the model is performing well based on the defined evaluation metric.
6. **Testing the model** - *7 points* Testing the machine learning model involves measuring its performance on a test set. The team must ensure that the model is not overfitting to the training data and is performing well on unseen data. If using a pre-trained model, this and the previous step would be one, and the team would need to focus more on showing the model's ability to tackle the problem at hand.
7. **Creating a self-sustained notebook** & **GitHub Repo with Website**- *25 points* Creating a self-sustained notebook involves ensuring that the code is well-documented and can be run independently. The team must ensure that the notebook is organized, and clear, and includes appropriate explanations for all the previous steps/points. All deliverables of the project should be on the GitHub repo.
8. **Creating a report and video** - *30 points* Creating an executive report and video involves summarizing the project's objectives, methodology, and results clearly and concisely. The team must ensure that they can communicate their work effectively to a non-technical audience. You should use references to support your claims for your decision (e.g., why use pre-trained model X instead of Y, etc)
9. **CATME Peer evaluation-** *5 points*. Since this is a group-based project, at the end students will be required to complete a CATME Peer evaluation. Based on the results of the Peer evaluation, your grade on the section above might change. That is, if on average, your peers indicated that you did not contribute significantly to the project, your grade might be reduced compared to the overall group grade.

On Moodle, your team would only need to submit a single link for the repo, which should have everything listed above.

This is a very complex project, so I would advise you to start thinking and working on it as soon as possible. By the end of the semester, we would have time devoted in class to work on this project. If you have any questions or need additional clarification pls post on Piazza for everyone to see.