STINTSY
Machine project specifications

If there are any questions, or the specifications seem vague, approach your teacher immediately for clarifications.

GROUPING
Each group can have at most 3 members.

DATASETS
Each group may choose a Kaggle dataset option, or find an existing dataset.

Option 1: **Kaggle dataset**

You may choose one of the following datasets in Kaggle. These datasets have different domains that require different kinds of preprocessing. It is imperative that you learn the domain the dataset belongs to perform the preprocessing and make analysis on the data.

Pros
You don't need to collect data.

Cons
You need to compare your work with the top ranking results in Kaggle's leaderboard. You also have to at least use the evaluation metric the original challenge requires.

Option 2: **Your own found dataset**

If you found another dataset that you want to use for your project, you are very welcome to do so. Just make sure that the dataset will be enough for exploration.

Pros
You love the dataset.

Cons
You may need to come up with your own idea of a "goal" (e.g. predicting the price of the house, detecting whether a dog is peeing or not).

Some datasets you may like

Image datasets
[Classification] https://www.kaggle.com/c/dog-breed-identification/rules
https://www.kaggle.com/c/invasive-species-monitoring/data
[Fine-grained classification] https://www.kaggle.com/competitions/sorghum-id-fgvc-9/overview/description
[Fine-grained classification] https://www.kaggle.com/competitions/herbarium-2022-fgvc9/overview/description
[Fine-grained classification] https://www.kaggle.com/competitions/inaturalist-2019-fgvc6/overview/evaluation
[Classification] https://www.kaggle.com/competitions/cropharvest-crop-detection/overview/description
[Classification] https://www.kaggle.com/competitions/sportify-physdl/overview/evaluation
[Object detection] https://cocodataset.org/#explore

Text datasets
[Classification] https://www.kaggle.com/c/spooky-author-identification
[Classification] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[Sentiment analysis]
https://www.kaggle.com/code/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews/data

Other datasets
[Transportation, Regression] https://www.kaggle.com/c/nyc-taxi-trip-duration#evaluation
[Transportation, Regression (accident severity)] https://www.kaggle.com/competitions/datascienceandgis2018

[Education, Classification] https://www.kaggle.com/competitions/mooc-dropout-prediction-17
[Education, Classification] https://www.kaggle.com/competitions/delana

[Personality test, Clustering] https://www.kaggle.com/datasets/tunguz/big-five-personality-test

[Stock prices, Regression] https://www.kaggle.com/competitions/the-winton-stock-market-challenge/overview
[PH stock prices (with crypto), no task] https://github.com/enzoampil/fastquant

[Movie dataset, Regression] https://www.kaggle.com/competitions/tmdb-box-office-prediction/overview
[Movie dataset, no task] https://www.the-numbers.com/daily-box-office-chart (daily box office numbers per movie are available but automated scraping is banned)

[Music, no task] https://developer.spotify.com/discover/

[Video game sales, no task] https://www.kaggle.com/datasets/gregorut/videogamesales

[Sports, Regression] https://www.kaggle.com/competitions/sports-trading-will-there-be-more-goals

[PH Covid, no task] https://doh.gov.ph/covid19tracker (check the data dump)

[Commodity, no task] https://www.kaggle.com/datasets/unitednations/global-commodity-trade-statistics

DELIVERABLES

Groups must submit a **Jupyter notebook** containing the whole pipeline of their methodology. This includes and is not limited to data preprocessing, model building, evaluation (validation and testing), and fine tuning.

The notebooks are expected to be **verbose**. It should walk the reader on the steps the group made to make their model work. The notebooks must also show the authors' efforts to make their model work.

The notebooks are expected to come with **their latest checkpoints** (check hidden files of your notebook's directory). Groups must also make sure their notebook will run properly from start to bottom in a single kernel run.

Offshoots (a new task different from the original task because the new task seems interesting) are also encouraged, but make sure that the original task has already been completed.

* A technical paper is not required for this project

MILESTONES

The group is expected to demonstrate the progress in the following dates:

Week	Date (Estimated)	Milestone	Expected Output
8th	May 19, 2022 (H)	Proposal	Proposal PPT
14th	July 4, 2022 (M)	Final	Technical Report (notebook)
14th	July 5 - 6, 2022 (TW)	Final	Oral Report

STINTSY
Machine project specifications

Technical report

Deliverables

- Jupyter Notebook (ipynb)
- Checkpoints
- Images/other files, if necessary

Suggested outline for the technical report

1. Title + authors (your names)
2. Introduction to the problem/task and dataset
   Link to the Kaggle dataset or other sources
3. Description of the dataset
   Describe your data and show what kind of initial features you are dealing with
   Describe what each instance of your dataset represents
4. List of requirements (Python libraries, datasets/files)
5. Simple exploratory data analysis
   Point out anomalies/outliers in the data
   Show graphs to quickly digest data distributions and possible patterns
6. Data preprocessing/cleaning
   Explain why the data was preprocessed that way
   If you removed data, explain why removing the data was necessary
7. Feature engineering
   Explain (even if briefly) what these features are, and why they may help
8. Model training
   Explain why you chose the algorithms you are training with
9. Feature selection + Hyperparameter tuning
   You may use grid/random search for hyperparameter tuning
10. Model selection
    Present a summary of your best model configuration
    You can use this website to make tables easier to manage
11. Insights and conclusions
    Explain what insights you have learned from the models (why they failed/succeeded)
    Summarize your conclusions on which model performed the best and why
12. Ask for the user's input and show prediction
    Please refer to the Naive Bayes notebook, where you can test if a phrase is ham/spam
13. References
    You are encouraged to look at existing solutions online and learn from them (please cite)

Tips

- Make sure that your notebook can be rerun as a whole without any problems. You can test this out by going to `Kernel` > `Restart and Run All`. If this doesn't give you problems, you should be good.
- Have a train-validation-test set. For Kaggle competition datasets:
    - If you will not join the competition, do not use the hold out test set given because you will need to enter the competition to get the results back. Instead, take their train and validation set, mix them, and then split your train-validation-test set.
    - If you will join the competition, give the results back and your Kaggle ranking.
- Emphasize all your hard work, show how much experimentation you performed
- A lot of focus will be on why you chose the experiments you performed (so there must be a logical explanation on why you did them), and what you did to make it work
- Put enough output to guide the user with what you are doing
- Use images as needed

Presentation

Each group is given 20 minutes: 10 minutes to present, and 10 minutes for Q+A

Presentations will be done via Zoom.

Groups are required to have a PPT ready. You may also open your Jupyter notebook if it helps.

Open up all the necessary ppts/notebooks/files/websites before your allotted presentation time slot, don't wait until the presentation itself to load anything.

You are encouraged to stay in the online meeting to watch the other presentations.

Deadlines

19 May 2022	class time	*Proposal defense*
04 July 2022	2359	*Submission of technical report (AnimoSpace)*
05-06 July 2022	TBA	*Presentation (by appointment)*

Grading matrix

Criteria	Full Marks	Partial Marks	Partial Marks	No Marks
Description of the dataset and the task	**5** An overview or description of the data is provided, including how it was collected, and its implications on the types of conclusions that could be made from the data. A description of the variables, observations, and/or structure of the data is provided. The target task is well introduced and clearly defined.		**3** An overview or description of the data is provided but lacks details. A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset. The task is not clearly defined.	**0** No overview or description of the data is provided. No description of variables, observations, and/or structure is provided. The task is not defined.
Exploratory data analysis	**10** The data is sufficiently explored to get a grasp of the distribution and the content of the data. Appropriate summaries and visualizations are presented. Insights into how the EDA can help the model training is mentioned.	**7** Exploratory data analysis is not sufficiently performed. Summaries and visualizations are presented but have minor issues in terms of methods chosen.	**4** Exploratory data analysis is rudimentary. Inappropriate methods of summarizing and visualizing data are frequently chosen.	**0** No exploratory data analysis is attempted.
Data preprocessing and cleaning	**10** The necessary steps for preprocessing and cleaning are performed, including explanations for every step. If no preprocessing or cleaning is done, there is a justification on why it was not needed.	**7** Preprocessing and cleaning steps are performed but lacks explanation. Or, preprocessing and cleaning done was insufficient for the dataset.	**4** Preprocessing and cleaning steps are incorrectly performed. Or, preprocessing steps do not match the ML model chosen.	**0** No preprocessing and cleaning are done, and no justification was provided as to why it was not done, or the justification is weak or incorrect.
Model training	**15** The appropriate models are used to accomplish the machine learning task. Justification of choosing the models is shown.	**10** A lot of various models are used without proper justification of why they are chosen.	**5** Only one model is generated. Or, all models that are chosen are not appropriate for the task.	**0** No data modelling is done.
Model selection and hyperparameter tuning	**20** Appropriate data-driven error analysis is made, and changes to the model selection and hyperparameters are performed to improve model performance. The study exhausts improvements that can be done to the model	**14** Model selection and hyperparameter tuning is done exhaustively but without proper justification or analysis. Or, improvements to the models are not exhausted.	**6** Model selection and hyperparameter tuning is done, but no efforts to further improve the model are done.	**0** No model selection and hyperparameter tuning are done.
Insights and Conclusions	**5** The study is concluded by effectively summarizing the efforts of the authors. Recommendations on how the model could be further improved are provided.	**3** The study is concluded but misses key insights performed in the study. Or, recommendations on how the model could be further improved are provided without clear justification.		**0** No insights or conclusions are presented.
Presentation Manner	**5** The presenter seldomly looks at notes. The presenter helps the audience visualize through gestures and movements.	**3** The presenter looks at his notes most of the time. The presenter uses very little movement or descriptive gestures. The presenter displays mild tension; has trouble recovering from mistakes.		**0** The presenter reads the entire report from his notes. The presenter does not use movement or descriptive gestures.

	The presenter displays a relaxed, self-confident nature about self, with no mistakes.			The presenter displays tension and nervousness; has trouble recovering from mistakes.
Presentation Organization	**10** Information is presented in a logical and interesting sequence which the audience can follow.	**5** Audience has difficulty following the presentation because the presenter jumps around different topics.		**0** Audience cannot understand the presentation because there is no logical sequence of information.
Presentation Q&A	**20** The presenter demonstrates full knowledge by answering all class questions with explanations and elaboration.	**15** The presenter is at ease with expected answers to all questions, without elaboration.	**5** The presenter is uncomfortable with information and is able to answer only rudimentary questions.	**0** The presenter does not have a grasp of information and cannot answer questions about the subject.

Computing power

https://asti.dost.gov.ph/projects/coare/
https://colab.research.google.com/notebooks/welcome.ipynb
https://cloud.google.com/gcp/