[ML Fall’23] Project Outline

Registration details:

- Form Link: https://docs.google.com/forms/d/e/1FAIpQLSe-V94G-pHGIBJiJ0GoSsWNmBBRsDf0I8xzwZIVZaAz-CGBIw/viewform?usp=sf_link

- Minimum number of members is 3 and the maximum is 6.
- Registration ends: **26/4/2023 11:59pm**

Project delivery: Practical Exams week.

Practical Exam: Each team member will be graded individually according to their response to the oral questions related to their project.

Dataset Link:

https://drive.google.com/file/d/1lZW0Cq2KFiEp0aL1R4KDsl9gYPLXqfuD/view?usp=sharing

Project minimum Requirements :

1- Students report their final work via a presentation in a scientific manner.

- e.g., We approached the problem with the assumption that features A, B, C affect the response variable the most … etc.

2- **Preprocessing:** Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Including but not limited to:

- Identifying and handling the missing values
- Categorical encoding

3- **Feature Selection:** You must preprocess all features, but the model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with a valid reason)

4- **Model Selection & training:** You must at least train 4 classification models

5- **Hyper-parameter tuning:** Choose at least two hyperparameters to vary. Study **at least three different choices** for each hyperparameter. When varying one hyperparameter, all the other hyperparameters should be fixed.

6- **Model evaluation:** you should consider at least those evaluation metrics for evaluating the models performance including: accuracy, precision, recall, F1-score and confusion matrix.

7- **Test script:** During the practical exam discussion, you will be given a small test set, and you should classify all its records and print the final evaluation metrics (the same stated in step 6), so you should write a test script that preprocesses the given test data (note you should handle all possible cases to avoid any errors when modeling using this test data) and consider the same features you decided to work with during the feature selection step.

Teams of 6 member: You should apply **one unsupervised learning technique on your data** (e.g., Dimensionality reduction using PCA or any other algorithm, clustering algorithms, etc)

Presentation Must Include:

- ❖ You must explain in detail the **preprocessing techniques** you needed to apply to your dataset and how you implemented them.
- ❖ Perform **analysis** on the dataset as studied and explain how the features affect and relate to each other.
- ❖ You must explain what **classification techniques** you used (at least four).
- ❖ Mention the **differences** between each model and the acquired **results** (accuracy/precision and so on).
- ❖ You must clearly mention **what features** you used or discarded to create your classification models.
- ❖ Explain what the **sizes** of your training and validation sets are (if exist).
- ❖ Mention any further techniques that were used to **improve** the results (if exist).
- ❖ You should include **screenshots** of any data visualization you used for exploratory data analysis (initial investigations on data to discover patterns/spot anomalies and so on).
- ❖ Finally, write a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

.

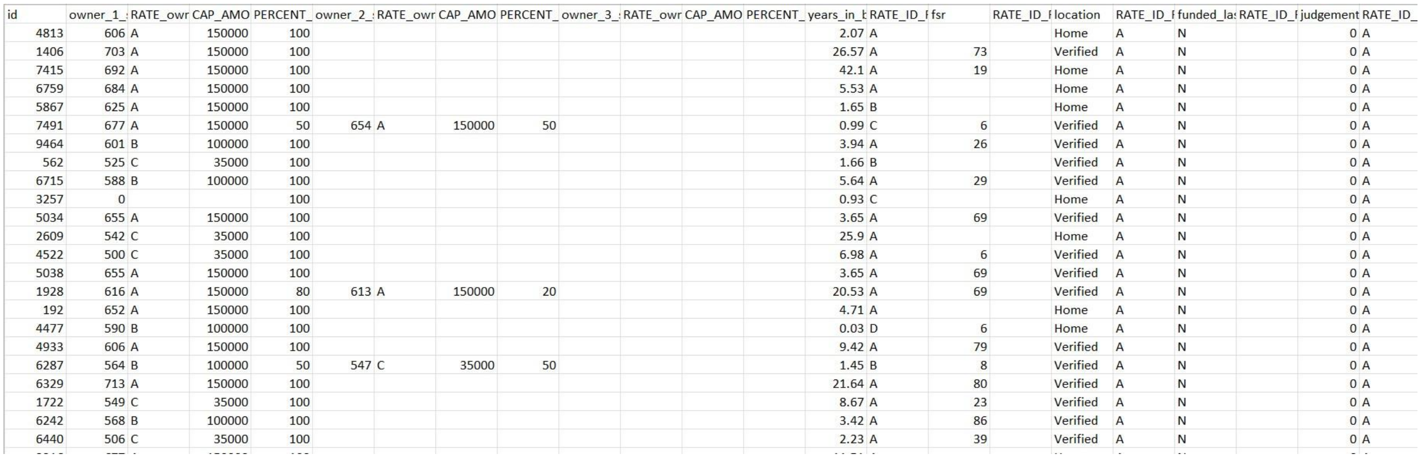
Prepare yourselves to fully present your work in 15 minutes.

Note: All team members must participate in the presentation during the discussion.

Project: Identifying loan completion status

Given this dataset, we would like to understand and predict the completion status regarding a loan based on the provided data.

Dataset Snapshot:



id	owner_1_:	RATE_owr	CAP_AMO	PERCENT_	owner_2_:	RATE_owr	CAP_AMO	PERCENT_	owner_3_:	RATE_owr	CAP_AMO	PERCENT_	years_in_b	RATE_ID_F	fsr	RATE_ID_F	location	RATE_ID_F	funded_las	RATE_ID_F	judgement	RATE_ID_
4813	606	A	150000	100									2.07	A			Home	A	N		0	A
1406	703	A	150000	100									26.57	A	73		Verified	A	N		0	A
7415	692	A	150000	100									42.1	A	19		Home	A	N		0	A
6759	684	A	150000	100									5.53	A			Home	A	N		0	A
5867	625	A	150000	100									1.65	B			Home	A	N		0	A
7491	677	A	150000	50	654	A	150000	50					0.99	C	6		Verified	A	N		0	A
9464	601	B	100000	100									3.94	A	26		Verified	A	N		0	A
562	525	C	35000	100									1.66	B			Verified	A	N		0	A
6715	588	B	100000	100									5.64	A	29		Verified	A	N		0	A
3257	0			100									0.93	C			Home	A	N		0	A
5034	655	A	150000	100									3.65	A	69		Verified	A	N		0	A
2609	542	C	35000	100									25.9	A			Home	A	N		0	A
4522	500	C	35000	100									6.98	A	6		Verified	A	N		0	A
5038	655	A	150000	100									3.65	A	69		Verified	A	N		0	A
1928	616	A	150000	80	613	A	150000	20					20.53	A	69		Verified	A	N		0	A
192	652	A	150000	100									4.71	A			Home	A	N		0	A
4477	590	B	100000	100									0.03	D	6		Home	A	N		0	A
4933	606	A	150000	100									9.42	A	79		Verified	A	N		0	A
6287	564	B	100000	50	547	C	35000	50					1.45	B	8		Verified	A	N		0	A
6329	713	A	150000	100									21.64	A	80		Verified	A	N		0	A
1722	549	C	35000	100									8.67	A	23		Verified	A	N		0	A
6242	568	B	100000	100									3.42	A	86		Verified	A	N		0	A
6440	506	C	35000	100									2.23	A	39		Verified	A	N		0	A

Note: The label (Y) in this dataset is the completion status column

Rules:

1) **Don’t share code outside of the team (any detected plagiarism for any portion of the code will be considered as 0 for the whole project mark)**
2) **Don’t use external data**