ML Projects (SC) – Milestone 2

The objective of the projects is to prepare you to apply different machine learning algorithms to real-world tasks. This will help you to increase your knowledge about the workflow of the machine learning tasks. You will learn how to apply pre-processing, feature engineering, regression, and classification methods.

Delivering Milestone 2: Practical exam.

- ➤ You must deliver a detailed report for milestone 2 contains all your work in this phase. Combine both reports and deliver a complete report for the project (Hardcopy).
- Each team should work on their project's updated dataset for milestone 2. The **updated dataset for each project** can be found <u>here</u>

➤ In the practical exam:

- We will give you two unseen test sets, one for regression and one for classification.
- Make sure you save your trained model and create a test script that takes the new csv file, loads the saved models, and outputs predictions. This is to allow us to test your model without retraining.

Hint 1: You can use libraries such as 'pickle' to save and load your models.

Hint 2: Any model that you need to 'fit' during training means you need to save it and reload it for the test to work correctly.

- You should be able to handle missing values for features in a test sample. (You can't drop an entire test sample row).
- You must Show the MSE and R2 score of the regression models and the classification accuracy of each classifier on the test set.

- Each team member will be graded individually according to their response to the oral questions related to their project.
- > In the second milestone, you will apply the following: -

Classification:

- Split your dataset into 80% training and 20% testing.
- Train at least 3 models to classify each sample into distinct classes.
- 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.

Milestone 2:

Classification and Hyperparameter tuning.

Milestone 2 Report Must Include:

- Summarize the classification accuracy, total training time, and total test time using three bar graphs.
- ❖ Note that your **Feature Selection** process may differ in this phase (classification) than the previous (regression), If so, explain your feature selection process and how it was proved or disproved.
- * Explain in details how **hyperparameter tuning** affected your models' performance.
- ❖ Finally, write a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

Project(1): Loan Risk Prediction

An **updated dataset** will be provided for each project in the second milestone.

Updated Dataset Snapshot:

Revolving(Bankcardl	. Available B	TotalTrade	DebtToInd	IncomeRa	StatedMo	Total	LoanNum	LoanOriginalA	ProsperRating	(Alpha)
0	0	1500	11	0.17	\$25,000-49	3083.33		19141	9425		
3989	0.21	10266	29	0.18	\$50,000-74	6125		134815	10000	Α	
				0.06	Not displa	2083.33		6466	3001		
1444	0.04	30754	26	0.15	\$25,000-49	2875		77296	10000	Α	
6193	0.81	695	39	0.26	\$100,000+	9583.33	11	102670	15000	D	
62999	0.39	86509	47	0.36	\$100,000+	8333.33		123257	15000	В	
5812	0.72	1929	16	0.27	\$25,000-49	2083.33		88353	3000	E	
1260	0.13	2181	10	0.24	\$25,000-49	3355.75		90051	10000	С	
9906	0.11	77696	29	0.25	\$25,000-49	3333.33		121268	10000	AA	
9906	0.11	77696	29	0.25	\$25,000-49	3333.33		121268	10000	AA	
387	0.51	363	47	0.12	\$75,000-99	7500		65946	13500	С	
1220	0.32	2580	7	0.27	\$1-24,999	1666.67		20907	1000		
8624	0.7	3626	20	0.18	\$25,000-49	2416.67		125045	4000	В	
9171	0.32	19129	18	0.09	\$50,000-74	5833.33		96202	8500	Α	
32898	0.43	42204	48	0.2	\$100,000+	10833.3		90060	19330	Α	
9103	0.97	178	17	0.49	\$50,000-74	5500		63982	4000	HR	

Updated Dataset Description:

- The "LoanRiskScore" column used in the previous milestone as the actual output has been removed.
- A New "ProsperRating (Alpha)" column has been added instead. Each loan can have a risk rate that is either {A, AA, B, C, D, E or HR}.

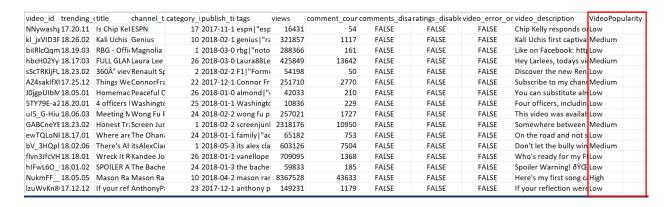
Milestone 2 Task:

Classify a loan into one of seven categories: {A, AA, B, C, D, E or HR} based on the provided features in **the updated dataset.**

Project(2): Video Likes Prediction

An **updated dataset** will be provided for each project in the second milestone.

Updated Dataset Snapshots:



Updated Dataset Description:

- The "likes" column used in the previous milestone as the actual output has been removed.
- A New "VideoPopularity" column has been added instead. Each video can have a popularity level that is either {Low, Medium, or High}.

Milestone 2 Classification task:

Classify each sample (row) into one of three categories {Low, Medium, or High} based on the provided features in the updated dataset.