ML Projects (CS) – 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 (not the original dataset). The **updated dataset for each project** will be included in Lab 8 materials.

- **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 re-training.

    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 can 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): Predicting Song Popularity

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

Updated Dataset Snapshot:

valence	year	acousticne	artists	danceabili	duration_	energy	explicit	id	instrumen	key	liveness	loudness	mode	name	tempo	release_date	speechine	popularity_level
0.895	1997	0.454	["Adolesce	0.737	251351	0.614	0	0f6VT4JZx	0	8	0.285	-11.871	1	Arrepentic	110.929	1997	0.0542	Intermediate
0.846	1996	0.258	["Adolesce	0.451	280189	0.785	0	3AK05RIuu	0	2	0.12	-4.701	0	Me Tengo	96.831	1/10/1996	0.0362	Intermediate
0.84	1997	0.437	["Adolesce	0.464	283273	0.511	0	0u2LbmKA	0	2	0.149	-11.815	0	Persona Ic	96.856	1997	0.0329	Intermediate
0.877	2008	0.516	["Adolesce	0.709	242965	0.747	0	6vYGqxYC	0	0	0.352	-4.62	0	Se Acabo l	153.151	2008	0.0484	Intermediate
0.803	1987	0.321	["Alexand	0.75	309400	0.59	0	7DVTNkZR	0	5	0.115	-11.799	1	Never Kne	104.101	7/29/1987	0.0346	Intermediate
0.762	1988	5.70E-05	["Alexand	0.725	356800	0.91	0	1R7yjdyoC	0.135	1	0.322	-9.357	1	Sleigh Ride	111.605	1/1/1988	0.031	Low
0.273	1987	0.496	["Alexand	0.649	312733	0.245	0	7uGRDTQj	0	11	0.174	-17.585	1	Crying Ove	112.709	7/29/1987	0.0286	Low
0.735	1987	0.06	["Alexand	0.701	241107	0.715	0	77PvXsp1z	0	11	0.354	-14.999	0	Hearsay	101.515	7/29/1987	0.0291	Low
0.953	1987	0.0103	["Alexand	0.719	277773	0.65	0	3zkhVuJ6X	4.36E-05	7	0.133	-12.947	0	The Lovers	105.141	7/29/1987	0.0265	Low
0.85	1991	0.108	["Alexand	0.761	304427	0.702	0	6nhCjIwKC	3.13E-05	7	0.0935	-8.523	1	All True M	104.773	1/1/1991	0.0389	Low
0.902	1987	0.276	["Alexand	0.661	264933	0.828	0	0pRMtIwr	1.40E-06	11	0.287	-12.673	0	(What Car	115.078	7/29/1987	0.039	Low
0.755	1987	0.0011	["Alexand	0.735	235493	0.694	0	3wN45r0L	0.0521	6	0.218	-13.515	0	Fake	112.846	7/29/1987	0.0348	Intermediate

Updated Dataset Description:

- The **"popularity"** column used in the previous milestone as the actual output has been removed.
- A New column is added **"popularity_level"**. A song can have popularity that is {Low, Intermediate or High}.

Milestone 2 Classification task:

Classify a song into one of three categories: Low, Intermediate or High based on the provided features in **the updated dataset.**

Project(2): Predict Mobile App Success

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

Updated Dataset Snapshots:

id	track_name	size_bytes	currency	price	rating_cou	rating_cou	vpp_lic	ver	cont_ratin	prime_genre	sup_devic	ipadSc_ur	lang.num	rate
281656475	PAC-MAN Pre	100788224	USD	3.99	21292	26	1	6.3.5	4+	Games	38	5	10	Intermediate
281796108	Evernote - sta	158578688	USD	0	161065	26	1	8.2.2	4+	Productivity	37	5	23	Intermediate
281940292		100524032	USD	0	188583	2822	1	5.0.0	4+	Weather	37	5	3	Intermediate
282614216	eBay: Best Ap	128512000	USD	0	262241	649	1	5.10.0	12+	Shopping	37	5	9	Intermediate
282935706	Bible	92774400	USD	0	985920	5320	1	7.5.1	4+	Reference	37	5	45	High
283619399	Shanghai Mał	10485713	USD	0.99	8253	5516	1	1.8	4+	Games	47	5	1	Intermediate
283646709	PayPal - Send	227795968	USD	0	119487	879	1	6.12.0	4+	Finance	37	0	19	Intermediate
284035177	Pandora - Mu	130242560	USD	0	1126879	3594	1	8.4.1	12+	Music	37	4	1	Intermediate
284666222	PCalc - The Be	49250304		9.99	1117	4	1	3.6.6	4+	Utilities	37	5	1	High
284736660	Ms. PAC-MAN	70023168	USD	3.99	7885	40	1	4.0.4	4+	Games	38	0	10	Intermediate
284791396	Solitaire by M	49618944	USD	4.99	76720	4017	1	4.10.1	4+	Games	38	4	11	High
284815117	SCRABBLE Pre	227547136		7.99	105776	166	1	5.19.0	4+	Games	37	0	6	Intermediate

Updated Dataset Description:

- The **"user_rating"** column used in the previous milestone as the actual output has been removed.
- A New **"rate"** column has been added instead. Each app can have a rate that is either {High, Intermediate or Low}.

Milestone 2 Classification task:

Classify each app(row) into one of three categories: High, Intermediate or Low based on the provided features **in the updated dataset.**