[2023] 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. The link can be found [here]

- **Note that milestone 2 requirements can be added to later.**

- **In the practical exam**:

  - We will give you two unseen test sets, **one for regression and one for classification.**
  - In case of the movies dataset you will receive two csv files for regression and two csv files 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' or 'learn' 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): Game Application Success Prediction

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

Updated Dataset Snapshots:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	URL	ID	Name	Subtitle	Icon URL	User Ratin	Price	In-app Pur	Descriptic	Develope	Age Ratin	Language:	Size	Primary G	Genres	Original Relea	Current Versi	Rate
2	https://ap	2.85E+08	Sudoku		https://is	3553	2.99		Join over	Mighty Mi	4+	DA, NL, EN	15853568	Games	Games, St	11/7/2008	30/05/2017	High
3	https://ap	2.85E+08	Reversi		https://is	284	1.99		The classi	Kiss The N	4+	EN	12328960	Games	Games, St	11/7/2008	17/05/2018	High
4	https://ap	2.85E+08	Morocco		https://is	8376	0		Play the cl	Bayou Gar	4+	EN	674816	Games	Games, Bc	11/7/2008	5/9/2017	Intermediate
5	https://ap	2.86E+08	Sudoku (Free)		https://is	190394	0		Top 100 fr	Mighty Mi	4+	DA, NL, EN	21552128	Games	Games, St	23/07/2008	30/05/2017	High
6	https://ap	2.86E+08	Senet Deluxe		https://is	28	2.99		"Senet De	RoGame S	4+	DA, NL, EN	34689024	Games	Games, St	18/07/2008	22/07/2018	High
7	https://ap	2.86E+08	Sudoku - (	Original b	https://is	47	0	1.99	Sudoku w	OutOfThe	4+	EN	48672768	Games	Games, Er	30/07/2008	29/04/2019	Intermediate
8	https://ap	2.86E+08	Gravitation		https://is	35	0		"Gravitati	Robert Fai	4+		6328320	Games	Games, Er	30/07/2008	14/11/2013	Intermediate
9	https://ap	2.86E+08	Colony		https://is	125	0.99		"50 levels	Chris Hayr	4+	EN	64333824	Games	Games, St	3/8/2008	3/10/2018	Intermediate
10	https://ap	2.87E+08	Carte		https://is	44	0		"Jeu simp	Jean-Fran	4+	FR	2657280	Games	Games, St	3/8/2008	23/11/2017	Intermediate
11	https://ap	2.87E+08	"Barrels O' Fun"		https://is	184	0		Barrels O\	BesqWare	4+	EN	1466515	Games	Games, Ca	1/8/2008	1/8/2008	Intermediate
12	https://ap	2.88E+08	Lumen Lite		https://is	5072	0		"The obje	Bridger M	4+	EN	7086403	Games	Games, Pu	18/08/2008	22/11/2008	High
13	https://ap	2.89E+08	BubblePop		https://is	526	0		Are you re	TMSOFT	4+	EN	845008	Games	Games, St	22/08/2008	25/07/2009	Intermediate
14	https://ap	2.89E+08	Marple		https://is	989	0.99		AWARDEI	Mikko Kar	4+	EN	3643392	Games	Games, Pu	28/08/2008	5/5/2019	High
15	https://ap	2.89E+08	Tetravex Lite		https://is	2358	0		Play the cl	Futrell So	4+	EN	731525	Games	Games, Pu	27/08/2008	21/10/2008	Intermediate
16	https://ap	2.89E+08	Awele/Oware - Man		https://is	112	0	0.99	Awele/Ov	SOLILAB	4+	EN, FR, DE	1.23E+08	Games	Games, St	31/08/2008	6/4/2015	Intermediate
17	https://ap	2.89E+08	Awele/Oware - Man		https://is	112	0	0.99	Awele/Ov	SOLILAB	4+	EN, FR, DE	1.23E+08	Games	Games, St	31/08/2008	6/4/2015	Intermediate
18	https://ap	2.89E+08	Chess Game		https://is	504	0		"How abo	Mementic	4+	EN	444163	Games	Games, Bc	2/9/2008	7/10/2009	Intermediate

Updated Dataset Description:

- The **"Average_User_Rating"** column used in the previous milestone as the actual output has been removed.
- A New **"Rate"** column has been added instead. Each application can have a rate of {High, Intermediate or Low}.

Milestone 2 Classification task:

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

Project(2): Movie Popularity Prediction

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

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

Updated Dataset Snapshot:

budget	genres	homepage	id	keywords	original_la	original_tit	overview	viewercou	productior	productior	release_da	revenue	runtin	spoken_la	stati	tagline	title	vote_coun	Rate
2.5E+07	[{"id": 18,	http://ww	33870	[{"id": 432	en	Mao's Last	At the age	1.876811	[{"name":	[{"iso_316	########	20719451	117	[{"iso_639	Rele	æœ€åŽçš„	Mao's L	28	High
3.8E+07	[{"id": 878, "name": "		193	[{"id": 109	en	Star Trek:	Captain Je	14.77904	[{"name":	[{"iso_316	########	1.2E+08	118	[{"iso_639	Rele	Boldly go.	Star Tre	452	Intermediate
2E+07	[{"id": 36,	http://foc	10139	[{"id": 237	en	Milk	The story	30.9097	[{"name":	[{"iso_316	########	54586584	128	[{"iso_639	Rele	Never Bler	Milk	612	High
2.3E+07	[{"id": 18, "name": "D		11632	[{"id": 212	en	Vanity Fair	Beautiful,	6.618149	[{"name":	[{"iso_316	9/1/2004	16123851	141	[{"iso_639	Rele	On Septen	Vanity F	73	Intermediate
5.2E+07	[{"id": 28,	http://ww	26389	[{"id": 90,	en	From Paris	James Ree	27.91628	[{"name":	[{"iso_316	2/5/2010	52826594	92	[{"iso_639	Rele	Two agent	From Pa	675	Intermediate
2.8E+07	[{"id": 18,	http://ww	277216	[{"id": 380	en	Straight Ou	In 1987, fi	61.76233	[{"name":	[{"iso_316	########	2.02E+08	147	[{"iso_639	Rele	The Story	Straight	1355	High
2.6E+07	[{"id": 80, "name": "C		14181	[{"id": 611	en	Boiler Roo	A college c	11.23308	[{"name":	[{"iso_316	########	28780255	118	[{"iso_639	Rele	Welcome	Boiler R	201	Intermediate
0	[{"id": 28, "name": "A		10413	[{"id": 156	en	Nowhere t	Escaped cc	11.68934	[{"name":	[{"iso_316	########	0	94	[{"iso_639	Rele	When the	Nowhe	119	Intermediate
4000000	[{"id": 28, "name": "A		2370	[{"id": 242	en	Topaz	A French ir	5.975604	[{"name":	[{"iso_316	########	6000000	143	[{"iso_639	Rele	Hitchcock	Topaz	77	Intermediate
1.2E+07	[{"id": 99,	http://ww	101267	[{"id": 187	en	Katy Perry	Giving fans	8.410688	[{"name":	[{"iso_316	########	32726956	93	[{"iso_639	Rele	Be yoursel	Katy Pe	85	Intermediate
6E+07	[{"id": 28,	http://ww	35791	[{"id": 445	en	Resident E	In a world	2.143764	[{"name":	[{"iso_316	9/9/2010	3E+08	97	[{"iso_639	Rele	She's back	Residen	1363	Intermediate

Updated Dataset Description:

- The **"vote_average"** column used in the previous milestone as the actual output has been removed.
- A New column is added **"Rate"**. A movie can have a rate of {High, Intermediate or Low}.

Milestone 2 Classification task:

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