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

➢ **In the practical exam**:
   ▪ We will you give you an unseen test set.

   ▪ 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 :-
1- Classification:
   ▪ Split your dataset into 80% training and 20% testing.

   ▪ Train at least 4 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.

2- Dimensionality Reduction:
- Apply PCA on the training set to reduce the dimensionality. You need to study at least **three** different values for the reduced dimensionality. Explain your choice.

- Train and evaluate your classifiers (for the reduced features) on your validation set.

Milestone 2:

➢ Classification, Hyperparameter tuning, Dimensionality reduction effect on classification results.

Milestone 2 Report <u>Must</u> 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.

❖ Explain your choices for 3 different values for the **reduced dimensionality** and their effect on the data.

❖ After applying PCA and training the 4 classification models on the new reduced features, Summarize the **classification accuracy**, **total training time**, and **total test time** using three bar graphs.

❖ 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 Movie Success

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

Updated Dataset Snapshot:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	budget	genres	homepag	id	keywords	original_l	original_t	overview	popularity	productio	productio	release_d	revenue	runtime	spoken_la	status	tagline	title	vote_cour	rate
2	2.37E+08	[{"id": 28,	http://ww	19995	[{"id": 146	en	Avatar	In the 22n	150.4376	[{"name":	[{"iso_316	########	2.79E+09		162	[{"iso_639	Released	Enter the	Avatar	11800	High
3	3E+08	[{"id": 12,	http://dis	285	[{"id": 270	en	Pirates of	Captain Ba	139.0826	[{"name":	[{"iso_316	########	9.61E+08		169	[{"iso_639	Released	At the enc	Pirates of	4500	High
4	2.45E+08	[{"id": 28,	http://ww	206647	[{"id": 470	en	Spectre	A cryptic r	107.3768	[{"name":	[{"iso_316	########	8.81E+08		148	[{"iso_639	Released	A Plan No	Spectre	4466	Intermediate
5	2.5E+08	[{"id": 28,	http://ww	49026	[{"id": 849	en	The Dark I	Following	112.313	[{"name":	[{"iso_316	########	1.08E+09		165	[{"iso_639	Released	The Leger	The Dark I	9106	High
6	2.6E+08	[{"id": 28,	http://mo	49529	[{"id": 818	en	John Carte	John Carte	43.927	[{"name":	[{"iso_316	3/7/2012	2.84E+08		132	[{"iso_639	Released	Lost in ou	John Carte	2124	Intermediate
7	2.58E+08	[{"id": 14,	http://ww	559	[{"id": 851	en	Spider-Ma	The seem	115.6998	[{"name":	[{"iso_316	5/1/2007	8.91E+08		139	[{"iso_639	Released	The battle	Spider-Ma	3576	Intermediate
8	2.6E+08	[{"id": 16,	http://dis	38757	[{"id": 156	en	Tangled	When the	48.68197	[{"name":	[{"iso_316	########	5.92E+08		100	[{"iso_639	Released	They're ta	Tangled	3330	High
9	2.8E+08	[{"id": 28,	http://ma	99861	[{"id": 882	en	Avengers:	When Ton	134.2792	[{"name":	[{"iso_316	########	1.41E+09		141	[{"iso_639	Released	A New Ag	Avengers:	6767	High
10	2.5E+08	[{"id": 12,	http://har	767	[{"id": 616	en	Harry Pott	As Harry b	98.88564	[{"name":	[{"iso_316	7/7/2009	9.34E+08		153	[{"iso_639	Released	Dark Secre	Harry Pott	5293	High
11	2.5E+08	[{"id": 28,	http://ww	209112	[{"id": 849	en	Batman v	Fearing th	155.7905	[{"name":	[{"iso_316	########	8.73E+08		151	[{"iso_639	Released	Justice or	Batman v	7004	Intermediate
12	2.7E+08	[{"id": 12,	http://ww	1452	[{"id": 83,	en	Superman	Superman	57.92562	[{"name":	[{"iso_316	########	3.91E+08		154	[{"iso_639	Released		Superman	1400	Intermediate
13	2E+08	[{"id": 12,	http://ww	10764	[{"id": 627	en	Quantum	Quantum	107.9288	[{"name":	[{"iso_316	########	5.86E+08		106	[{"iso_639	Released	For love, f	Quantum	2965	Intermediate
14	2E+08	[{"id": 12,	http://dis	58	[{"id": 616	en	Pirates of	Captain Ja	145.8474	[{"name":	[{"iso_316	########	1.07E+09		151	[{"iso_639	Released	Jack is bac	Pirates of	5246	High
15	2.55E+08	[{"id": 28,	http://dis	57201	[{"id": 155	en	The Lone I	The Texas	49.04696	[{"name":	[{"iso_316	7/3/2013	89289910		149	[{"iso_639	Released	Never Tak	The Lone I	2311	Intermediate
16	2.25E+08	[{"id": 28,	http://ww	49521	[{"id": 83,	en	Man of Ste	A young b	99.39801	[{"name":	[{"iso_316	########	6.63E+08		143	[{"iso_639	Released	You will b	Man of Ste	6359	Intermediate
17	2.25E+08	[{"id": 12,	"name": "/	2454	[{"id": 818	en	The Chron	One year a	53.9786	[{"name":	[{"iso_316	########	4.2E+08		150	[{"iso_639	Released	Hope has	The Chron	1630	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.**

Project(2): Predicting Mobile Game Success

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 Pu	Descriptic	Develope	Age Ratin	Language:	Size	Primary G	Genres	Original Rele:	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 c	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, Pt	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		AWARDEL	Mikko Kar	4+	EN	3643392	Games	Games, Pt	28/08/2008	5/5/2019	High
15	https://ap	2.89E+08	Tetravex Lite		https://is:	2358	0		Play the c	Futrell So	4+	EN	731525	Games	Games, Pt	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**