[2024] 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.

- **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' 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 different 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.
- [**Extra Requirement Mandatory for Teams of 6 Only**]: Apply (heteregenous) ensemble learning using different machine learning models to get the output. You should try both voting and stacking approaches.

(Note: Ensemble methods based on the same base model e.g. random forest will not be counted as doing the extra task)

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): Song Popularity Prediction

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

Updated Dataset Snapshots:

Song	Album	Album Rele	Artist Nam	Artist(s) Ge	Hot100 Ra	Hot100 Ra	Song Lengt	Spotify Link	Song Image	Spotify URI	PopularityLevel	Acousticne	Danceabili	Energy	Instrument	Liveness	Loudness	Speechine	Tempo	Valence
I Need You	I Need You	########	['LeAnn Rir	['country d	2000	44	229826	https://ope	https://i.sc	spotify:trac	Average	0.02	0.478	0.736	9.58e-05	0.118	-7.124	0.0367	144.705	0.564
Sweet Lad	Tyrese	########	['Tyrese']	['hip pop', '	1999	43	290600	https://ope	https://i.sc	spotify:trac	Average	0.233	0.588	0.522	0	0.24	-6.254	0.0383	66.024	0.584
You Take M	Sooner or L	1979	['Rex Smith	['bubblegu	1979	86	197453	https://ope	https://i.sc	spotify:trac	Average	0.476	0.313	0.6	2.56e-05	0.523	-7.913	0.0357	141.912	0.272
If I Give My	Golden Gir	1945	['Doris Day	['adult star	1954	20	169066	https://ope	https://i.sc	spotify:trac	Average	0.973	0.503	0.059	0	0.103	-16.131	0.0497	76.642	0.331
Don't Forg	Drew's Far	########	['The Karac	['karaoke']	2006	50	251013	https://ope	https://i.sc	spotify:trac	Not Popular	0.0149	0.843	0.348	0.00128	0.247	-10.669	0.0545	143.595	0.385
Always	Cross Roa	########	['Bon Jovi']	['glam met	1995	17	353026	https://ope	https://i.sc	spotify:trac	Popular	0.117	0.383	0.659	0	0.0778	-5.558	0.0312	140.795	0.327
I Saw Red	Cherry Pie	5/1/1990	['Warrant']	['glam met	1991	96	226920	https://ope	https://i.sc	spotify:trac	Average	0.269	0.495	0.677	0	0.381	-6.139	0.0271	84.879	0.437
Hands Cle	Under Rug	########	['Alanis Mo	['canadian	2002	95	269400	https://ope	https://i.sc	spotify:trac	Average	0.00192	0.513	0.82	2.83e-06	0.504	-5.428	0.0299	99.952	0.52
Here (In Yo	Zombies! A	8/8/2006	['Hellogoo	['pop punk'	2007	81	240546	https://ope	https://i.sc	spotify:trac	Average	0.197	0.7	0.607	0.00173	0.272	-6.804	0.0359	126.045	0.774
One More T	Faith	########	['George M	['new wave	1988	11	350666	https://ope	https://i.sc	spotify:trac	Average	0.434	0.551	0.291	3.78e-05	0.11	-12.544	0.0283	119.005	0.0823
Back To De	Speak Now	########	['Taylor Sw	['pop']	2011	74	293026	https://ope	https://i.sc	spotify:trac	Popular	0.117	0.529	0.67	0	0.334	-4.663	0.0303	141.893	0.286
Fast Car	Tracy Chap	4/5/1988	['Tracy Cha	['lilith', "wo	1988	76	296800	https://ope	https://i.sc	spotify:trac	Popular	0.313	0.711	0.292	0	0.131	-15.523	0.037	103.951	0.194
Throwing It	Invisible To	6/9/1986	['Genesis']	['art rock', '	1986	84	229560	https://ope	https://i.sc	spotify:trac	Average	0.144	0.377	0.606	0.00877	0.0879	-7.409	0.0262	83.972	0.291
Hurt	The Best O	1/1/1992	['Timi Yuro'	['northern s	1961	48	148466	https://ope	https://i.sc	spotify:trac	Average	0.707	0.493	0.244	0.000462	0.121	-11.487	0.0277	87.044	0.222

Updated Dataset Description:

- The **"**popularity**"** column used in the previous milestone as the actual output has been removed.

- A New **"PopularityLevel"** column has been added instead. Each apartment can have a category of {Not Popular, Average or Popular}.

Milestone 2 Classification task:

Classify each song into one of three categories: (Not Popular, Average or Popular) based on the provided features **in the updated dataset.**

Project(2): Online Articles Popularity Prediction

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

Updated Dataset Snapshot:

LDA_03	LDA_04	global_sul	global_ser	global_rat	global_rat	rate_posit	rate_nega	avg_positi	min_positi	max_posit	avg_negat	min_negat	max_nega	Article Popularity
0.02505	0.025101	0.460128	0.08733	0.051434	0.025717	0.666667	0.333333	0.334834	0.033333	1	-0.29583	-1	-0.05	Above Average
0.020103	0.020001	0.425505	0.162121	0.068323	0.018634	0.785714	0.214286	0.291667	0.033333	0.5	-0.275	-0.4	-0.125	Average
0.050082	0.470683	0.470696	0.190751	0.033816	0.009662	0.777778	0.222222	0.411395	0.214286	0.5	-0.3	-0.4	-0.2	Average
0.022304	0.325403	0.301515	0.201407	0.035714	0	1	0	0.3133	0.033333	0.5	0	0	0	Average
0.04	0.040002	0.323407	-0.00708	0.028078	0.036717	0.433333	0.566667	0.265297	0.0625	0.5	-0.22794	-0.8	-0.1	Not Popular
0.033334	0.366563	0.440358	-0.00753	0.028791	0.042226	0.405405	0.594595	0.32496	0.136364	0.7	-0.22027	-0.66667	-0.08333	Average
0.171315	0.742847	0.432119	0.124721	0.042254	0.023474	0.642857	0.357143	0.40404	0.1	0.6	-0.25333	-0.5	-0.1	Average
0.033482	0.033334	0.477963	0.036481	0.055556	0.061728	0.473684	0.526316	0.46713	0.033333	0.7	-0.30333	-0.8	-0.1	Above Average
0.910955	0.022226	0.666667	0.18869	0.034384	0.014327	0.705882	0.294118	0.541667	0.1	1	-0.4475	-1	-0.1875	Above Average
0.04	0.040032	0.54008	0.08932	0.034707	0.021692	0.615385	0.384615	0.380919	0.136364	0.6	-0.24792	-0.6	-0.1	Average
0.022224	0.022224	0.366468	0.057837	0.020737	0.009217	0.692308	0.307692	0.311111	0.1	0.5	-0.29514	-0.6	-0.125	Average
0.025622	0.139304	0.355629	0.095582	0.027027	0.009653	0.736842	0.263158	0.291342	0.033333	0.5	-0.23873	-0.6	-0.07143	Not Popular

Updated Dataset Description:

- The **"shares"** column used in the previous milestone as the actual output has been removed.

- A New column is added **"Article Popularity"**. An article can have a rating category of {Not Popular, Average, Above Average or Very Popular}.

Milestone 2 Classification task:

Classify a device into one of four categories: {Not Popular, Average, Above Average or Very Popular} based on the provided features in **the updated dataset.**