[2023] 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 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.

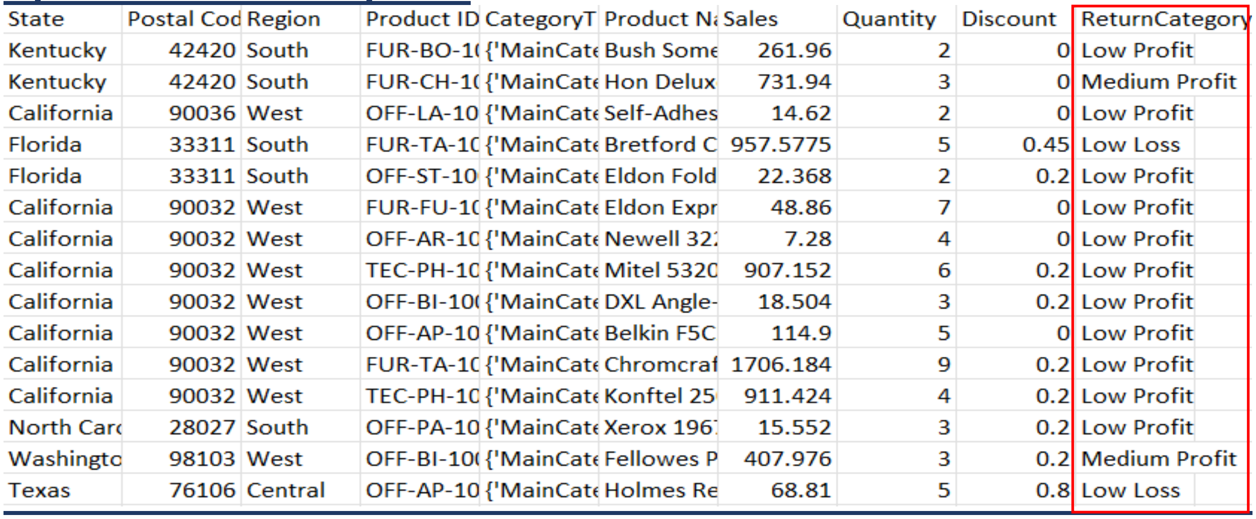
- Explain in details how **hyperparameter tuning** affected your models' performance.

- ❖ 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): Megastore Profit Prediction

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

Updated Dataset Snapshot:



State	Postal Cod	Region	Product ID	CategoryT	Product Na	Sales	Quantity	Discount	ReturnCategory
Kentucky	42420	South	FUR-BO-10	{'MainCate	Bush Some	261.96	2	0	Low Profit
Kentucky	42420	South	FUR-CH-10	{'MainCate	Hon Delux	731.94	3	0	Medium Profit
California	90036	West	OFF-LA-10	{'MainCate	Self-Adhes	14.62	2	0	Low Profit
Florida	33311	South	FUR-TA-10	{'MainCate	Bretford C	957.5775	5	0.45	Low Loss
Florida	33311	South	OFF-ST-10	{'MainCate	Eldon Fold	22.368	2	0.2	Low Profit
California	90032	West	FUR-FU-10	{'MainCate	Eldon Expr	48.86	7	0	Low Profit
California	90032	West	OFF-AR-10	{'MainCate	Newell 32	7.28	4	0	Low Profit
California	90032	West	TEC-PH-10	{'MainCate	Mitel 5320	907.152	6	0.2	Low Profit
California	90032	West	OFF-BI-10	{'MainCate	DXL Angle-	18.504	3	0.2	Low Profit
California	90032	West	OFF-AP-10	{'MainCate	Belkin F5C	114.9	5	0	Low Profit
California	90032	West	FUR-TA-10	{'MainCate	Chromcraf	1706.184	9	0.2	Low Profit
California	90032	West	TEC-PH-10	{'MainCate	Konftel 25	911.424	4	0.2	Low Profit
North Car	28027	South	OFF-PA-10	{'MainCate	Xerox 196	15.552	3	0.2	Low Profit
Washingto	98103	West	OFF-BI-10	{'MainCate	Fellowes P	407.976	3	0.2	Medium Profit
Texas	76106	Central	OFF-AP-10	{'MainCate	Holmes Re	68.81	5	0.8	Low Loss

Updated Dataset Description:

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

- A New **"ReturnCategory"** column has been added instead. Each player can have a level that is either {High Loss, Low Loss, Low Profit, Medium Profit, High Profit}.

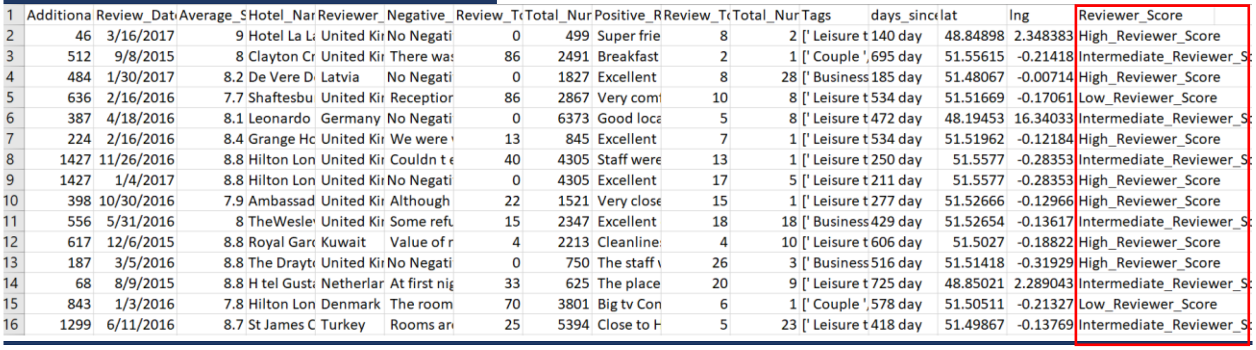
Milestone 2 Task:

Classify a return category into one of five categories: {High Loss, Low Loss, Low Profit, Medium Profit, High Profit} based on the provided features in **the updated dataset.**

Project(2): Hotel Rating Prediction

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

Updated Dataset Snapshots:



1	Additiona	Review_Dat	Average_S	Hotel_Nar	Reviewer_	Negative_	Review_Tc	Total_Nur	Positive_R	Review_Tc	Total_Nur	Tags	days_since	lat	lng	Reviewer_Score
2	46	3/16/2017	9	Hotel La La	United Kir	No Negati	0	499	Super frie	8	2	[' Leisure t	140 day	48.84898	2.348383	High_Reviewer_Score
3	512	9/8/2015	8	Clayton Cr	United Kir	There was	86	2491	Breakfast	2	1	[' Couple ',	695 day	51.55615	-0.21418	Intermediate_Reviewer_S
4	484	1/30/2017	8.2	De Vere D	Latvia	No Negati	0	1827	Excellent	8	28	[' Business	185 day	51.48067	-0.00714	High_Reviewer_Score
5	636	2/16/2016	7.7	Shaftesbu	United Kir	Reception	86	2867	Very comf	10	8	[' Leisure t	534 day	51.51669	-0.17061	Low_Reviewer_Score
6	387	4/18/2016	8.1	Leonardo	Germany	No Negati	0	6373	Good loca	5	8	[' Leisure t	472 day	48.19453	16.34033	Intermediate_Reviewer_S
7	224	2/16/2016	8.4	Grange Hc	United Kir	We were	13	845	Excellent	7	1	[' Leisure t	534 day	51.51962	-0.12184	High_Reviewer_Score
8	1427	11/26/2016	8.8	Hilton Lon	United Kir	Couldn t e	40	4305	Staff were	13	1	[' Leisure t	250 day	51.5577	-0.28353	Intermediate_Reviewer_S
9	1427	1/4/2017	8.8	Hilton Lon	United Kir	No Negati	0	4305	Excellent	17	5	[' Leisure t	211 day	51.5577	-0.28353	High_Reviewer_Score
10	398	10/30/2016	7.9	Ambassad	United Kir	Although	22	1521	Very close	15	1	[' Leisure t	277 day	51.52666	-0.12966	High_Reviewer_Score
11	556	5/31/2016	8	TheWesley	United Kir	Some refu	15	2347	Excellent	18	18	[' Business	429 day	51.52654	-0.13617	Intermediate_Reviewer_S
12	617	12/6/2015	8.8	Royal Gard	Kuwait	Value of r	4	2213	Cleanline:	4	10	[' Leisure t	606 day	51.5027	-0.18822	High_Reviewer_Score
13	187	3/5/2016	8.8	The Drayto	United Kir	No Negati	0	750	The staff v	26	3	[' Business	516 day	51.51418	-0.31929	High_Reviewer_Score
14	68	8/9/2015	8.8	H tel Gusta	Netherlar	At first nig	33	625	The place	20	9	[' Leisure t	725 day	48.85021	2.289043	Intermediate_Reviewer_S
15	843	1/3/2016	7.8	Hilton Lon	Denmark	The room	70	3801	Big tv Con	6	1	[' Couple ',	578 day	51.50511	-0.21327	Low_Reviewer_Score
16	1299	6/11/2016	8.7	St James C	Turkey	Rooms are	25	5394	Close to H	5	23	[' Leisure t	418 day	51.49867	-0.13769	Intermediate_Reviewer_S

Updated Dataset Description:

- The **"Reviewer_Score"** column used in the previous milestone as the actual output has been removed.
- A New **"Reviewer_Score"** column has been added instead. Each review can result in {High_Reviewer_Score, Intermediate_Reviewer_Score or Low_Reviewer_Score}.

Milestone 2 Classification task:

Classify each review(row) into one of three categories: High_Reviewer_Score, Intermediate_Reviewer_Score or Low_Reviewer_Score based on the provided features **in the updated dataset**