[2023] Machine Learning Projects (SC)

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 clean your data, applying pre-processing, feature engineering, regression, and classification methods. Each project will be delivered in milestones.

- The best three teams for each project will be honored.
- Registration starts: Friday 31/3/2023.
- Registration ends: Tuesday 4/4/2023.
- Delivering Milestone 1: 18/4/2023 11:59 PM Online.
- Delivering Milestone 2: Practical exam.
- Minimum number of members is 5 and the maximum is 6 or 7 with teams as 7 having an extra task mandatory.
- You must deliver a detailed report for each milestone contains all your work (feature analysis, algorithms used in each module and the achieved accuracy for each one)

  Note : Each report will be graded

In the first milestone, you will apply the followings :-

Preprocessing: Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Regression: Apply different regression techniques (at least two) to find the model that fit your data with minimum error.

Milestone 1: 50%

- Preprocessing, Regression.

Milestone 1 Report Must Include:

- You must explain in details the **preprocessing techniques** you needed to apply on your dataset and how you implemented them.
- Perform **analysis** on the dataset as studied and explain how the features affect and relate to each other.
- You must explain what **regression techniques** you used (at least two).
- Mention the **differences** between each model and the acquired **results** (accuracy/error and so on).
- You must clearly mention **what features** you used or discarded to create your regression models.
- Explain what the **sizes** of your training, testing and validation sets are, if exist.
- Mention any further techniques that were used to **improve** the results (if exist).
- You should include **screenshots** of the resultant(s) regression line plots.
- 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

Can we predict a shipment's profit based on a number of factors such as order categories, order date and address among other factors?. If a megastore could predict the profit on each order before shipment, they can understand which factors affect their revenue and control them as such.

Dataset Snapshot:

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal Coc	Region
1	CA-2016-1	11/8/2016	11/11/2016	Second Cl	CG-12520	Claire Gut	Consumer	United Sta	Henderso	Kentucky	42420	South
2	CA-2016-1	11/8/2016	11/11/2016	Second Cl	CG-12520	Claire Gut	Consumer	United Sta	Henderso	Kentucky	42420	South
3	CA-2016-1	6/12/2016	6/16/2016	Second Cl	DV-13045	Darrin Van	Corporate	United Sta	Los Angele	California	90036	West
4	US-2015-1	10/11/2015	10/18/2015	Standard (	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South
5	US-2015-1	10/11/2015	10/18/2015	Standard (	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South
6	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
7	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
8	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
9	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
10	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
11	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West
12	CA-2014-1	6/9/2014	6/14/2014	Standard (	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West

~Dataset header Continued:

Product ID	CategoryTree	Product Name	Sales	Quantity	Discount	Profit
FUR-BO-10001798	{'MainCategory': 'Furniture', 'SubCategory': 'Bookcases'}	Bush Somerset	261.96	2	0	41.9136
FUR-CH-10000454	{'MainCategory': 'Furniture', 'SubCategory': 'Chairs'}	Hon Deluxe Fab	731.94	3	0	219.582
OFF-LA-10000240	{'MainCategory': 'Office Supplies', 'SubCategory': 'Labels'	Self-Adhesive A	14.62	2	0	6.8714
FUR-TA-10000577	{'MainCategory': 'Furniture', 'SubCategory': 'Tables'}	Bretford CR450	957.5775	5	0.45	-383.031
OFF-ST-10000760	{'MainCategory': 'Office Supplies', 'SubCategory': 'Storag	Eldon Fold 'N R	22.368	2	0.2	2.5164
FUR-FU-10001487	{'MainCategory': 'Furniture', 'SubCategory': 'Furnishings'	Eldon Expressic	48.86	7	0	14.1694
OFF-AR-10002833	{'MainCategory': 'Office Supplies', 'SubCategory': 'Art'}	Newell 322	7.28	4	0	1.9656
TEC-PH-10002275	{'MainCategory': 'Technology', 'SubCategory': 'Phones'}	Mitel 5320 IP Ph	907.152	6	0.2	90.7152
OFF-BI-10003910	{'MainCategory': 'Office Supplies', 'SubCategory': 'Binder	DXL Angle-View	18.504	3	0.2	5.7825
OFF-AP-10002892	{'MainCategory': 'Office Supplies', 'SubCategory': 'Applia	Belkin F5C206V	114.9	5	0	34.47
FUR-TA-10001539	{'MainCategory': 'Furniture', 'SubCategory': 'Tables'}	Chromcraft Rec	1706.184	9	0.2	85.3092

Dataset Description:

Feature	Description
Row ID	
Order ID	Unique Order ID for each Customer.
Order Date	The rating of the application given by users on the play store, it ranges from 0 - 5
Ship Date	The number of reviews given by users for the application.
Ship Mode	Shipping Mode specified by the Customer.
Customer ID	Unique ID to identify each Customer.
Customer Name	Name of the Customer.
Segment	The segment where the Customer belongs.
Country	Country of residence of the Customer.
City	City of residence of the Customer.

State	State of residence of the Customer.
Postal Code	Postal Code of every Customer.
Region	Region where the Customer belong.
Product ID	Unique ID of the Product.
CategoryTree	List of the main category of the order and the subcategory
Product Name	Name of the Product
Sales	Sales of the Product.
Quantity	Quantity of the Product.
Discount	Discount of the Product.
Profit	Profit/Loss incurred.

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess all the features even if you won't use them later after feature selection)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the "Profit" (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

Bonus Task: To be announced later.

Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)

Project(2): Hotel Rating Prediction

Can you make your trip more cozy by using data science?. Can you predict what score a reviewer will give a hotel using features about the hotel in combination with the reviewer history and each review's language?.

Dataset Snapshots:

	A	B	C	D	E	F	G	H
1	Hotel_Address	Additional_Number_of_Scori	Review_Date	Average_S	Hotel_Name	Reviewer_Nation	Negative_Review	Review_Total_Negative_Word_Counts
2	s Gravesandestraat 5	194	8/3/2017	7.7	Hotel Arena	Russia	I am so angry that i ma	397
3	s Gravesandestraat 5	194	8/3/2017	7.7	Hotel Arena	Ireland	No Negative	0
4	s Gravesandestraat 5	194	7/31/2017	7.7	Hotel Arena	Australia	Rooms are nice but fo	42
5	s Gravesandestraat 5	194	7/31/2017	7.7	Hotel Arena	United Kingdom	My room was dirty an	210
6	s Gravesandestraat 5	194	7/24/2017	7.7	Hotel Arena	New Zealand	You When I booked w	140
7	s Gravesandestraat 5	194	7/24/2017	7.7	Hotel Arena	Poland	Backyard of the hotel	17
8	s Gravesandestraat 5	194	7/17/2017	7.7	Hotel Arena	United Kingdom	Cleaner did not chang	33
9	s Gravesandestraat 5	194	7/17/2017	7.7	Hotel Arena	United Kingdom	Apart from the price fa	11
10	s Gravesandestraat 5	194	7/9/2017	7.7	Hotel Arena	Belgium	Even though the pictu	34
11	s Gravesandestraat 5	194	7/8/2017	7.7	Hotel Arena	Norway	The aircondition make	15
12	s Gravesandestraat 5	194	7/7/2017	7.7	Hotel Arena	United Kingdom	Nothing all great	5
13	s Gravesandestraat 5	194	7/6/2017	7.7	Hotel Arena	France	6 30 AM started big nc	75
14	s Gravesandestraat 5	194	7/6/2017	7.7	Hotel Arena	United Kingdom	The floor in my room v	28
15	s Gravesandestraat 5	194	7/4/2017	7.7	Hotel Arena	Italy	No Negative	0

~Dataset header Continued:

	I	J	K	L	M	N	O	P	Q
1	Total_Number_of_R	Positive_Review	Review_Total_Positive_Word_	Total_Number_of_Revie	Reviewer_Score	Tags	days_since	lat	lng
2	1403	Only the park outside	11	7	2.9	[' Leisure trip ', '	0 days	52.36058	4.915968
3	1403	No real complaints th	105	7	7.5	[' Leisure trip ', '	0 days	52.36058	4.915968
4	1403	Location was good an	21	9	7.1	[' Leisure trip ', '	3 days	52.36058	4.915968
5	1403	Great location in nice	26	1	3.8	[' Leisure trip ', '	3 days	52.36058	4.915968
6	1403	Amazing location and	8	3	6.7	[' Leisure trip ', '	10 days	52.36058	4.915968
7	1403	Good restaurant with	20	1	6.7	[' Leisure trip ', '	10 days	52.36058	4.915968
8	1403	The room is spacious	18	6	4.6	[' Leisure trip ', '	17 days	52.36058	4.915968
9	1403	Good location Set in a	19	1	10	[' Leisure trip ', '	17 days	52.36058	4.915968
10	1403	No Positive	0	3	6.5	[' Leisure trip ', '	25 days	52.36058	4.915968
11	1403	The room was big enc	50	1	7.9	[' Leisure trip ', '	26 days	52.36058	4.915968
12	1403	Rooms were stunning	101	2	10	[' Leisure trip ', '	27 days	52.36058	4.915968
13	1403	Style location rooms	4	12	5.8	[' Business trip ',	28 days	52.36058	4.915968
14	1403	Comfy bed good locat	6	7	4.6	[' Leisure trip ', '	28 days	52.36058	4.915968
15	1403	This hotel is being ren	59	6	9.2	[' Business trip ',	30 days	52.36058	4.915968

Dataset Description:

Feature	Description
Hotel Address	
Review Date	
Average Score	Average Score of the hotel, calculated based on the latest comment in the last year.
Hotel Name	
Reviewer Nationality	
Negative Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
Review Total Negative Word Counts	Total number of words in the negative review.
Positive Review	Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
Review Total Positive Word Counts	Total number of words in the positive review.
Reviewer Score	Score the reviewer has given to the hotel, based on his/her experience.
Total Number of Reviews Reviewer Has Given	Number of Reviews the reviewers has given in the past.
Total Number of Reviews	Total number of valid reviews the hotel has.
Tags	Tags reviewer gave the hotel.
Days since review	Duration between the review date and scrape date.
Additional Number of Scoring	There are also some guests who made a scoring on the service rather than a review. This number indicates how many valid scores without review are in there.

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess all the features even if you won't use them later after feature selection)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the "Reviewer Score" (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)

Bonus Task: Apply sentiment analysis using the review column.