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 **updated dataset for each project** can be found [here]

- **In the practical exam**:

  - We will give you two unseen test sets, **one for regression and one for classification.**
  - In case of the taxi rides dataset you will receive one main csv file for regression and one main 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' 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): Airline Ticket Price Prediction

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

Updated Dataset Snapshot:

date	airline	ch_code	num_code	dep_time	time_take	stop	arr_time	type	route	TicketCategory
5/3/2022	Vistara	UK	812	9:45	10h 10m	1-stop	19:55	business	{'source':	very expensive
18-03-202	Vistara	UK	975	5:45	06h 30m	1-stop	12:15	business	{'source':	very expensive
9/3/2022	GO FIRST	G8	7537	14:30	08h 10m	1-stop	22:40	economy	{'source':	cheap
15-03-202	GO FIRST	G8	287	10:40	09h 40m	1-stop	20:20	economy	{'source':	moderate
22-03-202	Vistara	UK	826	12:30	07h 25m	1-stop	19:55	economy	{'source':	moderate
13-03-202	Air India	AI	803	6:10	26h 40m	1-stop	8:50	business	{'source':	expensive
14-02-202	Vistara	UK	832	6:55	12h 40m	1-stop	19:35	economy	{'source':	moderate
28-03-202	GO FIRST	G8	392	15:45	08h 05m	1-stop	23:50	economy	{'source':	moderate
########	AirAsia	I5	766	20:55	04h 15m	1-stop	1:10	economy	{'source':	cheap
22-03-202	Indigo	6E	2485	12:45	02h 55m	non-stop	15:40	economy	{'source':	cheap
########	Vistara	UK	641	13:40	19h 55m	1-stop	9:35	business	{'source':	expensive

Updated Dataset Description:

- The **"value"** column used in the previous milestone as the actual output has been removed.
- A New **"TicketCategory"** column has been added instead. Each ticket can have a category that is either {cheap, moderate, expensive or very expensive}.

Milestone 2 Task:

Classify a ticket price into one of four levels: {cheap, moderate, expensive or very expensive} based on the provided features in **the updated dataset.**

Project(2): Taxi Service Price Prediction

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

Updated Dataset Snapshots:

distance	cab_type	time_stam	destinatio	source	surge_mu	id	product_i	name	RideCategory
0.62	Uber	1.54E+12	West End	Haymarke	1	c1b4a572-	8cf7e821-f	Taxi	unknown
2.27	Uber	1.54E+12	Boston Un	Beacon Hi	1	f9e7e7e6-	997acbb5-	UberPool	cheap
2	Lyft	1.54E+12	Back Bay	Haymarke	1	154e8438-	lyft	Lyft	moderate
3.98	Lyft	1.54E+12	Financial D	Northeast	1	6bdc30a6-	lyft_plus	Lyft XL	expensive
1.49	Lyft	1.54E+12	Back Bay	Northeast	1	0cb12fe9-	lyft	Lyft	cheap
1.97	Uber	1.54E+12	Northeast	Beacon Hi	1	8ca92e07-	6d318bcc-	Black SUV	expensive
1.44	Uber	1.54E+12	Boston Un	Back Bay	1	6edfa428-	997acbb5-	UberPool	cheap
1.72	Lyft	1.54E+12	North End	Theatre D	1	e42c821b-	lyft_lux	Lux Black	expensive
1.7	Lyft	1.55E+12	North Stat	South Stat	1	c5177cc5-f	lyft_luxsu	Lux Black	expensive
1.83	Lyft	1.54E+12	West End	South Stat	1	1d6e3ffb-f	lyft	Lyft	moderate
1.5	Uber	1.54E+12	Back Bay	Fenway	1	54f4b84e-	55c66225-	UberX	cheap

Updated Dataset Description:

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

- A New **"RideCategory"** column has been added instead. Each ride can have a category that is either {unknown, cheap, moderate, expensive or very expensive}.

Milestone 2 Classification task:

Classify each row into one of five categories {unknown, cheap, moderate, expensive or very expensive} based on the provided features **in the updated dataset.**