ML Projects (CS) – 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.

  - 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): House Price Prediction

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

Updated Dataset Snapshot:

PavedDriv	WoodDec	OpenPorc	EnclosedP	3SsnPorch	ScreenPor	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCond	MiscFeature2	PriceRate
Y	0	61	0	0	0	0				0	2	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	298	0	0	0	0	0				0	5	2007	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Feedr'}	moderate
Y	0	42	0	0	0	0				0	9	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	0	35	272	0	0	0				0	2	2006	WD	Abnorml	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap
Y	192	84	0	0	0	0				0	12	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	40	30	0	320	0	0		MnPr	Shed	700	10	2009	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap
Y	255	57	0	0	0	0				0	8	2007	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	235	204	228	0	0	0			Shed	350	11	2009	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'PosN'}	moderate
Y	90	0	205	0	0	0				0	4	2008	WD	Abnorml	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Artery'}	cheap
Y	0	4	0	0	0	0				0	1	2008	WD	Normal	{'f1': 'Artery', 'f2': 'Lvl', 'f3': 'Artery'}	cheap
Y	0	0	0	0	0	0				0	2	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap
Y	147	21	0	0	0	0				0	7	2006	New	Partial	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	expensive
Y	140	0	0	0	176	0				0	9	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap
Y	160	33	0	0	0	0				0	8	2007	New	Partial	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	0	213	176	0	0	0		GdWo		0	5	2008	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	moderate
Y	48	112	0	0	0	0		GdPrv		0	7	2007	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap
Y	0	0	0	0	0	0			Shed	700	3	2010	WD	Normal	{'f1': 'Norm', 'f2': 'Lvl', 'f3': 'Norm'}	cheap

Updated Dataset Description:

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

- Two new columns are added:

  1- **"MiscFeature2"**. Column that contains a variety of features that describe the house in addition to the previous features.
  2- **"PriceRate"**. The actual output. A house can be rated as {cheap, moderate or expensive}.

Milestone 2 Task:

Classify a house into one of three categories: cheap, moderate or expensive based on the provided features in **the updated dataset. (You must also preprocess the new column)**

Project(2): Amazon Product Rating Prediction

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

Updated Dataset Snapshots:

product_n	manufacti	price	number_a	number_c	number_c	amazon_category	sellers	product_information	ProductGrade
XT-XINTE S	XT-XINTE	26.99		1	1	Sports Toys & Outdoor > Kites & Flight Toys		Technical Details Item Weig	A
AFV Club 1	AFV Club		4new	1	1	Hobbies > Model	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
1 X FACE M	ASCRAFTS	2.28	6new	1		Fancy Dress > Accessories > Masks		Technical Details Manufact	D
mobilo sta	Plasticant	10	6new	6	1			Technical Details Item Weig	B
Bristol No	Bristol No	22.46 - 71.01		7	1	Fancy Dress > Costumes > Adults		Technical Details Product D	B
SunsOut C	SunsOut	9.32	42new	1	1	Jigsaws & Puzzles	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
Zoo Anima	Nollmit	23.74	2new	517	1	Puppets & Puppet Theatres > Finger Puppets		Technical Details Additiona	C
Peppa Pig	Peppa Pig	3	4new	4	1	Office Supplies > I	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
Hot Whee	Hot Whee	5.96	28new	4	3	Die-Cast & Toy Ve	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
Bettie Pag	Bettie Pag	16.99		1	5	Hobbies > Collect	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
Matt Gree	Studio Anne Carlton		1new	1	1	Games > Chess	{"seller"=>{"Seller_name_1"=>"I	Technical Details Manufact	A
The Trash	The Trash	25.24	2new	86	1	Figures & Playsets	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	C
Hippie Chi	Wicked	9.61	15new	3	2	Fancy Dress > Cos	{"seller"=>[{"Seller_name_1"=>"	Technical Details Product D	D
Thunderca	Thunderca	29.99		3	3	Hobbies > Collectible Figures & Memorabilia > Col		Technical Details Item Weig	B
Funko POI	FunKo	26.99		1	1	Characters & Brar	{"seller"=>[{"Seller_name_1"=>"	Technical Details Item Weig	A
Flashing P	flashflash	4.46	2new	1	1	Party Supplies > Banners, Stickers & Confetti > Ban		Technical Details Product D	A

Updated Dataset Description:

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

- A New **"ProductGrade"** column has been added instead. Each product can have a grade that is either {A, B, C or D}.

Milestone 2 Classification task:

Classify each product (row) into one of four categories{A, B, C or D} based on the provided features **in the updated dataset.**