Machine Learning Projects

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 **ends**: Monday 20/11/2023 11:59PM.
- Delivering Milestone 1: 4/12/2023 11:59 PM.
- Delivering Milestone 2: Practical exam.
- Minimum number of members is 3 and the maximum is 5
- 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 following:-

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 fits 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) and the **training time** for each model.
- 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 if possible or any data visualization.
- Finally, write a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

Milestone 2 Deliverables will be announced later.

Project: Car Price Prediction

A foreign automobile company aspires to enter another country's market by setting up their manufacturing unit there and producing cars locally to give competition to local manufacturers. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in this new market, since those may be very different from the market where they come from.

Dataset Snapshot:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	carheight	curbweight
1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
3	1	alfa-romero Quadrifog	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823
4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337
5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824
6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507
7	1	audi 100ls	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844
8	1	audi 5000	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954
9	1	audi 4000	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086
10	0	audi 5000s (diesel)	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053
11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395
12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395

~Dataset header Continued:

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
curbweight	enginetype	cylindernumber	enginesize	fuelsystem	boreratio	stroke	compressionratio	horsepowe	peakrpm	citympg	highwaympg	price
2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	13495
2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	16500
2823	ohcv	six	152	mpfi	2.68	3.47	9	154	5000	19	26	16500
2337	ohc	four	109	mpfi	3.19	3.4	10	102	5500	24	30	13950
2824	ohc	five	136	mpfi	3.19	3.4	8	115	5500	18	22	17450
2507	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	15250
2844	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	17710
2954	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	18920
3086	ohc	five	131	mpfi	3.13	3.4	8.3	140	5500	17	20	23875
3053	ohc	five	131	mpfi	3.13	3.4	7	160	5500	16	22	17859.2
2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16430
2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16925

Dataset Description:

DATA DICTONARY		
1	**Car_ID**	Unique id of each observation (Interger)
2	**Symboling**	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	**carCompany**	Name of car company (Categorical)
4	**fueltype**	Car fuel type i.e gas or diesel (Categorical)
5	**aspiration**	Aspiration used in a car (Categorical)
6	**doornumber**	Number of doors in a car (Categorical)
7	**carbody**	body of car (Categorical)
8	**drivewheel**	type of drive wheel (Categorical)
9	**enginelocation**	Location of car engine (Categorical)
10	**wheelbase**	Weelbase of car (Numeric)
11	**carlength**	Length of car (Numeric)
12	**carwidth**	Width of car (Numeric)
13	**carheight**	height of car (Numeric)
14	**curbweight**	The weight of a car without occupants or baggage. (Numeric)
15	**enginetype**	Type of engine. (Categorical)
16	**cylindernumber**	cylinder placed in the car (Categorical)
17	**enginesize**	Size of car (Numeric)
18	**fuelsystem**	Fuel system of car (Categorical)
19	**boreratio**	Boreratio of car (Numeric)
20	**stroke**	Stroke or volume inside the engine (Numeric)
21	**compressionratio**	compression ratio of car (Numeric)
22	**horsepower**	Horsepower (Numeric)
23	**peakrpm**	car peak rpm (Numeric)
24	**citympg**	Mileage in city (Numeric)
25	**highwaympg**	Mileage on highway (Numeric)
26	**price(Dependent variable)**	Price of car (Numeric)

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (Use One-Hot-Encoding for at least one categorical feature)
2. Experiment with regression techniques to reduce the error on prediction of the average price of a car (Deliver at least two techniques).
3. Finish Milestone 1 Report.