[ML Fall'22] Project Outline

The objective of the project is to prepare the student to apply different machine learning algorithms to real-world tasks. This will help the student to increase their knowledge about the workflow of the machine learning tasks. Students will learn how to clean data, apply pre-processing, feature engineering, regression, and classification techniques. The project will consist of **two competitions** that will be delivered in milestones.

- ➢ The best three teams will be honoured.
- ➢ Registration **ends: Friday 4/11/2022 11:59 PM**.
- ➢ Delivering Milestone 1: To Be Announced.
- ➢ Delivering Milestone 2: Practical Exam.
- ➢ Minimum number of members is 3 and the maximum is 5.
- ➢ You must deliver a presentation for each milestone containing all your work for the **top 2 submissions** (feature analysis, algorithms used in each module and the achieved accuracy for each one)

Note: Each presentation 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.

MSE

Milestone 1:

➢ Preprocessing, Regression.

Milestone 1 Presentation Must Include:

- ❖ You must explain in detail the **preprocessing techniques** you needed to apply to 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 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.

Rules:

1) **Don't share code outside of the team (you will get 0 in the milestone)**
2) **Don't use external data**
3) **Each team will have 5 submissions per day, all members should merge in the leaderboard with team name titled by (Team [ID])**
4) **Don't use Advanced Architectures of neural networks (out of scope) .**
5) **Save the implementation of the top 2 submissions you will get that reproduce the same score.**
6) **You should define a seed for each model (fixed) .**
7) **#Hint 1 : Use random_state parameter in train_test_split function**
8) **#Hint 2 : check this link for definition of random seed here**

Project: Car Market Prediction

What factors affect a car's price? Given this dataset, we would like to understand and predict the price of a car based on the provided data.

Dataset Snapshot:

car_id	car-info	condition	mileage(kilometers)	fuel_type	volume(cm3)	color	transmission	drive_unit	segment	price(USD)
0	[(90),(audi),(1986)]	with mileage	319999	PETROL	2200	gray	mechanics	front-wheel drive	D	450
1	[(rapid),(skoda),(2016)]	with mileage	53000	petrol	1600	blue	mechanics	front-wheel drive	C	9600
2	[(primera),(nissan),(1992)]	with mileage	350000	PETROL	2000	blue	mechanics	front-wheel drive	D	1700
3	[(combo),(opel),(1997)]	with mileage	299709	petrol	1400	white	mechanics	front-wheel drive	M	205
4	[(zafira),(opel),(2007)]	with mileage	110000	PETROL	1800	silver	mechanics	front-wheel drive	M	5300
5	[(s-klass),(mercedes-benz),(2011)]	with mileage	182000	PETROL	4700	brown	auto	all-wheel drive	S	20500
6	[(vectra),(opel),(1998)]	with mileage	300000	petrol	1800	burgundy	mechanics	front-wheel drive	D	2800
7	[(5-seriya),(bmw),(1998)]	with mileage	273000	petrol	2500	gray	mechanics	rear drive	E	4100
8	[(605),(peugeot),(1996)]	with mileage	320000	PETROL	2000	burgundy	mechanics	front-wheel drive	E	950
9	[(323),(mazda),(1998)]	with mileage	306500	petrol	1500	green	mechanics	front-wheel drive	C	1550
10	[(partner),(peugeot),(2003)]	with mileage	233248	diesel	1900	white	mechanics	front-wheel drive	M	3800
11	[(juke),(nissan),(2013)]	with mileage	93000	petrol	1600	burgundy	auto	front-wheel drive	J	12400
12	[(a6),(audi),(1996)]	with mileage	282200	PETROL	2600	blue	mechanics	front-wheel drive	E	3680
13	[(a6-allroad),(audi),(2009)]	with mileage	200	diesel	2700	black	auto	front-wheel drive	E	19000
14	[(golf),(volkswagen),(1992)]	with mileage	0	diesel	1900	black	mechanics	front-wheel drive	C	704
15	[(transporter),(volkswagen),(2002)]	with mileage	355000	diesel	2500	blue	mechanics	front-wheel drive	M	7600
16	[(406),(peugeot),(2001)]	with mileage	350000	diesel	2000	silver	mechanics	front-wheel drive	D	4400
17	[(accent),(hyundai),(2013)]	with mileage	94700	PETROL	1600	silver	auto	front-wheel drive	B	8050
18	[(passat),(volkswagen),(1993)]	with mileage	410200	petrol	2000	white	mechanics	front-wheel drive	D	1770
19	[(ascona),(opel),(1988)]	with mileage	9999	petrol	1300	silver	mechanics	front-wheel drive	D	200
20	[(santa-fe),(hyundai),(2011)]	with mileage	135000	PETROL	2400	brown	auto	part-time four-wheel drive	J	13750
21	[(laguna),(renault),(2011)]	with mileage	252000	DIESEL	1500	gray	mechanics	front-wheel drive	D	8100

Milestone 1 tasks:

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 the prediction of the "price" (Deliver at least two regression models with a significant difference).

3. Finish Milestone 1 Presentation.

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

Milestone 1 Presentation Agenda:

- Preprocessing
  - Techniques/Analysis
  - Categorical Encoding
  - Data Imputation
- Feature Selection
- Regression Models
- Results (Top 2 Submissions)

These are the main points you should cover, but you can have more detailed sections as you see fit.