[ML Fall'22] Project Outline - 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 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 project delivery, **each team member will present** a part of the whole team's work. The final presentation slides should contain both the work of the first milestone and the second.

➢ Each team should work on their project's updated dataset for milestone 2. Available through the 2nd Kaggle Competition.

➢ **In the practical exam:**

- 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.
- **Teams of 6:** You should apply **at least one dimensionality-reduction technique** to the data, e.g., Principle Component Analysis (PCA).

Milestone 2:

➢ Classification and Hyperparameter tuning.

Milestone 2 Presentation 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) from the previous (regression), If so, explain your feature selection process and how it was proved or disproved.
- ❖ Explain in detail how **hyperparameter tuning** affected your models' performance.
- ❖ Finally, end your presentation with a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

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

An **updated dataset** will be provided in the second milestone competition.

Dataset Snapshot:

car_id	car-info	condition	mileage(kilometers)	fuel_type	volume(cm3)	color	transmission	drive_unit	segment	Price Category
0	[(90),(audi),(1986)]	with mileage	319999	PETROL	2200	gray	mechanics	front-wheel drive	D	cheap
1	[(rapid),(skoda),(2016)]	with mileage	53000	petrol	1600	blue	mechanics	front-wheel drive	C	moderate
2	[(primera),(nissan),(1992)]	with mileage	350000	PETROL	2000	blue	mechanics	front-wheel drive	D	cheap
3	[(combo),(opel),(1997)]	with mileage	299709	petrol	1400	white	mechanics	front-wheel drive	M	cheap
4	[(zafira),(opel),(2007)]	with mileage	110000	PETROL	1800	silver	mechanics	front-wheel drive	M	moderate
5	[(s-klass),(mercedes-benz)]	with mileage	182000	PETROL	4700	brown	auto	all-wheel drive	S	expensive
6	[(vectra),(opel),(1998)]	with mileage	300000	petrol	1800	burgundy	mechanics	front-wheel drive	D	moderate
7	[(5-seriya),(bmw),(1998)]	with mileage	273000	petrol	2500	gray	mechanics	rear drive	E	moderate

Updated Dataset Description:

- The **"price(USD)"** column used in the previous milestone as the actual output has been removed.
- A New **"Price Category"** column has been added instead. Each car can have a category that is either {cheap, moderate, expensive or very expensive}.

Milestone 2 tasks:

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

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)

Final Presentation Agenda:

- Preprocessing *(Your work in Milestone 1 + Milestone 2)*
  - Techniques/Analysis
  - Categorical Encoding
  - Data Imputation
- Feature Selection
  - Regression Task
  - Classification Task
- Regression Models
- Classification Models
- Hyperparameter Tuning
- Results (Top 2 Submissions per milestone)

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

Prepare yourselves to fully present your work in 15 minutes.