Machine Learning Projects (CS)

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.
- Team and Projects' Registration **starts**: Thursday 24/3/2022 11:00PM.
- Registration **ends**: Tuesday 5/4/2022 11:59PM.
- Delivering Milestone 1: 21/4/2022.
- 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:

- 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(1): Player Value Prediction

What factors affect football player values. Given this dataset, we would like to understand and predict a player's value based on the provided data.

Dataset Snapshot:

id	name	full_name	birth_date	age	height_cm	weight_kg	positions	nationality	overall_ra	potential	wage	preferred_	internatio	weak_foo	skill_move	work_rate	body_type
158023	L. Messi	Lionel Anc	6/24/1987	31	170.18	72.1	CF,RW,ST	Argentina	94	94	565000	Left	5	4	4	Medium/	Messi
190460	C. Eriksen	Christian	2/14/1992	27	154.94	76.2	CAM,RM,CM	Denmark	88	89	205000	Right	3	5	4	High/ Mec	Lean
195864	P. Pogba	Paul Pogb	3/15/1993	25	190.5	83.9	CM,CAM	France	88	91	255000	Right	4	4	5	High/ Mec	Normal
198219	L. Insigne	Lorenzo In	6/4/1991	27	162.56	59	LW,ST	Italy	88	88	165000	Right	3	4	4	High/ Mec	Normal
201024	K. Koulibal	Kalidou Ko	6/20/1991	27	187.96	88.9	CB	Senegal	88	91	135000	Right	3	3	2	High/ High	Normal
203376	V. van Dijk	Virgil van I	7/8/1991	27	193.04	92.1	CB	Netherlan	88	90	215000	Right	3	3	2	Medium/	Normal
231747	K. Mbappé	Kylian Mb	12/20/1998	20	152.4	73	RW,ST,RM	France	88	95	100000	Right	3	4	5	High/ Mec	Lean
153079	S. AgÃ¼er	Sergio Leo	6/2/1988	30	172.72	69.9	ST	Argentina	89	89	300000	Right	4	4	4	High/ Mec	Stocky
167495	M. Neuer	Manuel Ne	3/27/1986	32	193.04	92.1	GK	Germany	89	89	130000	Right	5	4	1	Medium/	Normal
179813	E. Cavani	Edinson Re	2/14/1987	32	185.42	77.1	ST	Uruguay	89	89	200000	Right	4	4	3	High/ High	Lean
189511	Sergio Bus	Sergio Bus	7/16/1988	30	187.96	76.2	CDM,CM	Spain	89	89	315000	Right	4	3	3	Medium/	Lean
192119	T. Courtoi	Thibaut Co	5/11/1992	26	198.12	96.2	GK	Belgium	89	90	240000	Left	4	2	1	Medium/	Courtois
192448	M. ter Ste	Marc-And	4/30/1992	26	187.96	84.8	GK	Germany	89	92	240000	Right	3	4	1	Medium/	Normal
194765	A. Griezm	Antoine G	3/21/1991	27	175.26	73	CF,ST	France	89	90	145000	Left	4	3	4	High/ High	Lean
209331	M. Salah	Mohamed	6/15/1992	26	175.26	71.2	RW,ST	Egypt	89	90	265000	Left	3	3	4	High/ Mec	PLAYER_B
211110	P. Dybala	Paulo Brur	11/15/1993	25	152.4	74.8	CAM,RW	Argentina	89	94	205000	Left	3	3	4	Medium/	Normal

Dataset Snapshot ~Cont'd:

GK_kicking	GK_positic	GK_reflex	tags	traits	LS	ST	RS	LW	LF
15	14	8	#Dribbler,#Distance S	Finesse Shot,Long Shot Taker (CPU AI Only),Speed Dribbler (CPU A	89+2	89+2	89+2	93+2	93+2
7	7	6	#PlaymakerÂ ,#Cross	Flair,Long Shot Taker (CPU AI Only),Playmaker (CPU AI Only),Tech	79+3	79+3	79+3	85+3	84+3
2	4	3	#Dribbler,#Playmakei	Flair,Long Passer (CPU AI Only),Long Shot Taker (CPU AI Only),Pla	81+3	81+3	81+3	82+3	83+3
14	9	10	#Speedster,#Dribbler	Finesse Shot,Long Shot Taker (CPU AI Only),Speed Dribbler (CPU A	78+3	78+3	78+3	86+3	85+3
7	13	5	#TacklingÂ ,#Tacticiar	Power Header	53+3	53+3	53+3	53+3	54+3
13	11	11	#TacticianÂ ,#Strengt	Injury Free,Leadership,Power Header	68+3	68+3	68+3	66+3	67+3
7	11	6	#Speedster,#Dribbler	Finesse Shot,Flair,Speed Dribbler (CPU AI Only),Technical Dribbler	85+3	85+3	85+3	87+3	87+3
6	11	14	#Dribbler,#Acrobat,#I	Beat Offside Trap,Leadership,Chip Shot (CPU AI Only),Technical D	86+3	86+3	86+3	86+3	87+3
91	87	87		GK Long Throw,1-on-1 Rush,Rushes Out Of Goal,Comes For Crosses					
13	13	10	#Engine	Beat Offside Trap,Power Header	85+3	85+3	85+3	81+3	83+3
13	9	13	#PlaymakerÂ ,#TacticianÂ		71+3	71+3	71+3	74+3	76+3
72	86	88		GK Long Throw,Comes For Crosses					

Milestone 1 tasks:

1. 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 prediction of the "value" (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

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

Project(2): Movie Revenue Prediction

What factors affect the success or failure of a movie. Given this dataset we would like to predict a movie's revenue based on the provided features.

Dataset Snapshots – File 1 (movies_revenue.csv):

movie_title	release_date	genre	MPAA_rat	revenue
Snow White and the Seven Dwarfs	21-Dec-37	Musical	G	$5,228,953,251
Pinocchio	9-Feb-40	Adventure	G	$2,188,229,052
Fantasia	13-Nov-40	Musical	G	$2,187,090,808
Song of the South	12-Nov-46	Adventure	G	$1,078,510,579
Cinderella	15-Feb-50	Drama	G	$920,608,730
20,000 Leagues Under the Sea	23-Dec-54	Adventure		$528,279,994
Lady and the Tramp	22-Jun-55	Drama	G	$1,236,035,515
Sleeping Beauty	29-Jan-59	Drama		$21,505,832
101 Dalmatians	25-Jan-61	Comedy	G	$1,362,870,985
The Absent Minded Professor	16-Mar-61	Comedy		$310,094,574
Babes in Toyland	14-Dec-61	Musical	G	$124,841,160
Bon Voyage!	17-May-62	Comedy	Not Rated	$109,581,646

Dataset Snapshots – File 2 (movies_director.csv):

name	director
Snow White and the Seven Dwarf	David Hand
Pinocchio	Ben Sharpsteen
Fantasia	full credits
Dumbo	Ben Sharpsteen
Bambi	David Hand
Saludos Amigos	Jack Kinney
The Three Caballeros	Norman Ferguson
Make Mine Music	Jack Kinney
Fun and Fancy Free	Jack Kinney
Melody Time	Clyde Geronimi
The Adventures of Ichabod and M	Jack Kinney
Cinderella	Wilfred Jackson

Dataset Snapshots – File 3 (movies_voice_actors.csv):

character	voice-actor	movie
Abby Mallard	Joan Cusack	Chicken Little
Abigail Gabble	Monica Evans	The Aristocats
Abis Mal	Jason Alexander	The Return of Jafar
Abu	Frank Welker	Aladdin
Achilles	None	The Hunchback of Notre Dame
Adella	Sherry Lynn	The Little Mermaid
Adorabeezle Winte	None	Wreck-It Ralph
The Agent	Greg Germann	Bolt
Agent Wendy Pleak	Kevin McDonald	Lilo & Stitch
Ajax the Gorilla	None	Donald Duck and the Gorilla
Akela	John Abbott	The Jungle Book
Al the Alligator	Thurl Ravenscroft	Lady and the Tramp

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess all the features provided in all data files and add the information in the second and third files to the information in the first file)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the “revenue” (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

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