[2024] Machine Learning Projects (SC)

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 starts: Wednesday 3/4/2024 11:59 PM.
- Registration ends: Saturday 6/4/2024.
- Delivering Milestone 1: 23/4/2024 11:59 PM.
- Delivering Milestone 2: Practical exam.
- Minimum number of members is 4 and the maximum is 6, **however teams with 6 members will be assigned extra requirements to be announced later**.
- 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 followings :-

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 fit 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).
- 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 applicable.
- 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): Song Popularity Prediction

Can you predict a certain song's popularity before it is even published to an audience? This dataset asks this question. It contains audio features of songs with a popularity score ranging from 0 to 100. Using the given data, try analyzing which features play the most important role in determining the popularity of a song.

Dataset Snapshot:

Song	Album	Album Release Date	Artist Names	Artist(s) Genres	Hot100 Ra	Hot100 Ra	Song Leng	Spotify Lin	Song Imag	Spotify UR
Prisoner of Love - Rer	Today & Yesterday	10/25/1993	['Perry Como']	['adult standards', 'easy listenir	1946	1	211866	https://op	https://i.sc	spotify:trac
To Each His Own	The Best Of The M	1/1/1996	['Eddy Howard']	['british dance band']	1946	2	184293	https://op	https://i.sc	spotify:trac
The Gypsy	The Anthology	6/16/1998	['The Ink Spots']	['vocal harmony group', 'lounge	1946	3	164533	https://op	https://i.sc	spotify:trac
Five Minutes More	The Columbia Yea	10/8/1993	['Frank Sinatra']	['adult standards', 'lounge', 'eas	1946	4	154773	https://op	https://i.sc	spotify:trac
Rumors Are Flying	1946 Broadcasts	11/30/2006	['Frankie Carle']	['space age pop', 'honky-tonk pi	1946	5	183133	https://op	https://i.sc	spotify:trac
Oh! What It Seemed t	12 Double-Barrele	9/12/1964	['Frankie Carle, His Piano	[]	1946	6	115133	https://op	https://i.sc	spotify:trac
Personality	Accentuate The P	1/1/1957	['Johnny Mercer', 'The Piec	['vocal harmony group', 'swing',	1946	7	169626	https://op	https://i.sc	spotify:trac
South America, Take	Bing-His Legenda	1/1/1993	['Bing Crosby', 'The Andrev	['torch song', 'vocal harmony gr	1946	8	187293	https://op	https://i.sc	spotify:trac
The Gypsy - 78rpm Ve	The Essential Dina	4/18/2014	['Dinah Shore']	['adult standards', 'torch song',	1946	9	177266	https://op	https://i.sc	spotify:trac
Oh, What It Seemed T	The Columbia Yea	1993	['Frank Sinatra']	['adult standards', 'lounge', 'eas	1946	10	178760	https://op	https://i.sc	spotify:trac
Surrender	Perry Como's Gre	9/28/1999	['Perry Como']	['adult standards', 'easy listenir	1946	11	191306	https://op	https://i.sc	spotify:trac
Doctor, Lawyer, India	A Square In The Sc	1/1/1950	['Betty Hutton']	['vintage hollywood']	1946	12	186186	https://op	https://i.sc	spotify:trac
Let It Snow, Let It Sno	Presenting Vaugh	12/3/1949	['Vaughn Monroe']	['swing', 'vaudeville', 'deep adul	1946	13	184760	https://op	https://i.sc	spotify:trac
To Each His Own	Presenting Freddy	5/6/1932	['Freddy Martin & His Orch	['big band', "man's orchestra"]	1946	14	191293	https://op	https://i.sc	spotify:trac
Ole Buttermilk Sky	The Best Of Kay Ky	5/9/2000	['Kay Kyser & His Orchestr	["man's orchestra"]	1946	15	177506	https://op	https://i.sc	spotify:trac

~Dataset header Continued:

Spotify URI	Popularity	Acousticn	Danceabil	Energy	Instrumen	Liveness	Loudness	Speechine	Tempo	Valence	Key	Mode	Time Signature
spotify:trac	19	0.767	0.247	0.182	0.00209	0.167	-11.121	0.0328	80.064	0.185	2	1	4
spotify:trac	25	0.947	0.344	0.0596	5.95e-05	0.181	-16.766	0.0394	81.037	0.15	3	1	4
spotify:trac	33	0.812	0.531	0.125	0	0.103	-15.463	0.0552	76.056	0.222	7	1	4
spotify:trac	31	0.794	0.67	0.0625	0	0.0762	-20.393	0.0611	142.894	0.569	9	1	4
spotify:trac	27	0.951	0.332	0.204	0.00252	0.638	-15.557	0.0437	72.355	0.377	11	1	4
spotify:trac	1	0.567	0.644	0.463	0.0392	0.412	-14.6	0.0304	105.297	0.811	7	1	3
spotify:trac	51	0.894	0.741	0.147	0	0.126	-16.563	0.0613	134.465	0.827	5	1	4
spotify:trac	22	0.742	0.748	0.4	0	0.107	-11.47	0.117	78.382	0.842	1	1	4
spotify:trac	10	0.957	0.365	0.143	0.0011	0.126	-14.636	0.0407	80.105	0.231	9	1	4
spotify:trac	19	0.94	0.331	0.0901	0	0.173	-18.921	0.0393	122.16	0.251	1	1	4
spotify:trac	18	0.987	0.273	0.106	0.141	0.265	-17.009	0.0386	80.062	0.112	7	1	1
spotify:trac	30	0.822	0.7	0.291	0	0.0639	-9.72	0.231	176.253	0.943	8	1	4
spotify:trac	15	0.985	0.591	0.148	0.00463	0.516	-10.665	0.0298	117.689	0.501	2	1	4
spotify:trac	8	0.986	0.218	0.293	0.0354	0.373	-8.174	0.032	92.151	0.267	9	0	3
spotify:trac	20	0.975	0.572	0.184	7.29e-05	0.159	-14.434	0.0438	168.596	0.829	0	1	4
spotify:trac	29	0.975	0.489	0.0729	4.61e-05	0.104	-16.442	0.0826	70.464	0.202	3	1	4

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 “Popularity” (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): Online Articles Popularity Prediction

In today's digital age, the popularity of online articles plays a crucial role in the success of digital publishers and content creators. Predicting the popularity of an article before it's published can greatly assist publishers in optimizing their content strategy and resource allocation. This project aims to develop a predictive model that can forecast the popularity of online articles based on various features.

Dataset Snapshots:

url	title	timedelta	n_tokens_(	n_unique_	n_non_sto	n_non_sto	num_hrefs	num_self_	num_imgs	num_videc	average_t	num_keyw	channel type
http://mashable.com/2013/01/(	amazon-instant-video-br	731	219	0.663594	1	0.815385	4	2	1	0	4.680365	5	data_channel_is_entertainm
http://mashable.com/2013/01/(	ap-samsung-sponsored-	731	255	0.604743	1	0.791946	3	1	1	0	4.913725	4	data_channel_is_bus
http://mashable.com/2013/01/(	apple-40-billion-app-dov	731	211	0.57513	1	0.663866	3	1	1	0	4.393365	6	data_channel_is_bus
http://mashable.com/2013/01/(	astronaut-notre-dame-b	731	531	0.503788	1	0.665635	9	0	1	0	4.404896	7	data_channel_is_entertainm
http://mashable.com/2013/01/(	att-u-verse-apps	731	1072	0.415646	1	0.54089	19	19	20	0	4.682836	7	data_channel_is_tech
http://mashable.com/2013/01/(	beewi-smart-toys	731	370	0.559889	1	0.698198	2	2	0	0	4.359459	9	data_channel_is_tech
http://mashable.com/2013/01/(	bodymedia-armbandgets	731	960	0.418163	1	0.549834	21	20	20	0	4.654167	10	data_channel_is_lifestyle
http://mashable.com/2013/01/(	canon-poweshot-n	731	989	0.433574	1	0.572108	20	20	20	0	4.617796	9	data_channel_is_tech
http://mashable.com/2013/01/(	car-of-the-future-infogra	731	97	0.670103	1	0.836735	2	0	0	0	4.85567	7	data_channel_is_tech
http://mashable.com/2013/01/(	chuck-hagel-website	731	231	0.636364	1	0.797101	4	1	1	1	5.090909	5	data_channel_is_world
http://mashable.com/2013/01/(	cosmic-events-doomsda	731	1248	0.49005	1	0.731638	11	0	1	0	4.617788	8	data_channel_is_world
http://mashable.com/2013/01/(	crayon-creatures	731	187	0.666667	1	0.8	7	0	1	0	4.657754	7	data_channel_is_lifestyle
http://mashable.com/2013/01/(	creature-cups	731	274	0.609195	1	0.707602	18	2	11	0	4.233577	8	[]
http://mashable.com/2013/01/(	dad-jokes	731	285	0.744186	1	0.84153	4	2	0	21	4.34386	6	[]
http://mashable.com/2013/01/(	downton-abbey-tumblrs	731	259	0.562753	1	0.644444	19	3	9	0	5.023166	7	[]
http://mashable.com/2013/01/(	earth-size-planets-milky	731	682	0.459542	1	0.634961	10	0	1	0	4.620235	6	data_channel_is_world

~Dataset header Continued:

w_max_n	kw_avg_m	kw_min_a	kw_max_a	kw_avg_av	self_refere	self_refere	self_refere	weekday	isWeekEnc	LDA_00	LDA_01	LDA_02	LDA_03	LDA_04	global_sub	global_ser	global_rat	global_rat
0	0	0	0	0	496	496	496	monday	No	0.500331	0.378279	0.040005	0.041263	0.040123	0.521617	0.092562	0.045662	0.013699
0	0	0	0	0	0	0	0	monday	No	0.799756	0.050047	0.050096	0.050101	0.050001	0.341246	0.148948	0.043137	0.015686
0	0	0	0	0	918	918	918	monday	No	0.217792	0.033334	0.033351	0.033334	0.682188	0.702222	0.323333	0.056872	0.009479
0	0	0	0	0	0	0	0	monday	No	0.028573	0.4193	0.494651	0.028905	0.028572	0.42985	0.100705	0.041431	0.020716
0	0	0	0	0	545	16000	3151.158	monday	No	0.028633	0.028794	0.028575	0.028572	0.885427	0.513502	0.281003	0.074627	0.012127
0	0	0	0	0	8500	8500	8500	monday	No	0.022245	0.306718	0.022231	0.022224	0.626582	0.437409	0.071184	0.02973	0.027027
0	0	0	0	0	545	16000	3151.158	monday	No	0.020082	0.114705	0.020024	0.020015	0.825173	0.51448	0.268303	0.080208	0.016667
0	0	0	0	0	545	16000	3151.158	monday	No	0.022224	0.150733	0.243435	0.022224	0.561384	0.543474	0.298613	0.083923	0.015167
0	0	0	0	0	0	0	0	monday	No	0.45825	0.028979	0.028662	0.029696	0.454412	0.538889	0.161111	0.030928	0.020619
0	0	0	0	0	0	0	0	monday	No	0.04	0.04	0.839997	0.040001	0.040002	0.313889	0.051852	0.038961	0.030303
0	0	0	0	0	0	0	0	monday	No	0.025004	0.287301	0.400829	0.261864	0.025002	0.48206	0.10235	0.038462	0.020833
0	0	0	0	0	0	0	0	monday	No	0.028628	0.028573	0.028596	0.028715	0.885488	0.477165	0.15	0.026738	0.010695
0	0	0	0	0	10700	16200	13450	monday	No	0.150493	0.025934	0.025188	0.304298	0.494088	0.53495	0.100728	0.051095	0.029197

~Dataset header Continued:

LDA_00	LDA_01	LDA_02	LDA_03	LDA_04	global_sub	global_ser	global_rat	global_rat	rate_positi	rate_negat	avg_positi	min_positi	max_positi	avg_negati	min_negat	max_negat	shares
0.500331	0.378279	0.040005	0.041263	0.040123	0.521617	0.092562	0.045662	0.013699	0.769231	0.230769	0.378636	0.1	0.7	-0.35	-0.6	-0.2	593
0.799756	0.050047	0.050096	0.050101	0.050001	0.341246	0.148948	0.043137	0.015686	0.733333	0.266667	0.286915	0.033333	0.7	-0.11875	-0.125	-0.1	711
0.217792	0.033334	0.033351	0.033334	0.682188	0.702222	0.323333	0.056872	0.009479	0.857143	0.142857	0.495833	0.1	1	-0.46667	-0.8	-0.13333	1500
0.028573	0.4193	0.494651	0.028905	0.028572	0.42985	0.100705	0.041431	0.020716	0.666667	0.333333	0.385965	0.136364	0.8	-0.3697	-0.6	-0.16667	1200
0.028633	0.028794	0.028575	0.028572	0.885427	0.513502	0.281003	0.074627	0.012127	0.860215	0.139785	0.411127	0.033333	1	-0.22019	-0.5	-0.05	505
0.022245	0.306718	0.022231	0.022224	0.626582	0.437409	0.071184	0.02973	0.027027	0.52381	0.47619	0.35061	0.136364	0.6	-0.195	-0.4	-0.1	855
0.020082	0.114705	0.020024	0.020015	0.825173	0.51448	0.268303	0.080208	0.016667	0.827957	0.172043	0.402039	0.1	1	-0.22448	-0.5	-0.05	556
0.022224	0.150733	0.243435	0.022224	0.561384	0.543474	0.298613	0.083923	0.015167	0.846939	0.153061	0.42772	0.1	1	-0.24278	-0.5	-0.05	891
0.45825	0.028979	0.028662	0.029696	0.454412	0.538889	0.161111	0.030928	0.020619	0.6	0.4	0.566667	0.4	0.8	-0.125	-0.125	-0.125	3600
0.04	0.04	0.839997	0.040001	0.040002	0.313889	0.051852	0.038961	0.030303	0.5625	0.4375	0.298413	0.1	0.5	-0.2381	-0.5	-0.1	710
0.025004	0.287301	0.400829	0.261864	0.025002	0.48206	0.10235	0.038462	0.020833	0.648649	0.351351	0.40448	0.1	1	-0.41506	-1	-0.1	2200
0.028628	0.028573	0.028596	0.028715	0.885488	0.477165	0.15	0.026738	0.010695	0.714286	0.285714	0.435	0.2	0.7	-0.2625	-0.4	-0.125	1900
0.150493	0.025934	0.025188	0.304298	0.494088	0.53495	0.100728	0.051095	0.029197	0.636364	0.363636	0.37551	0.2	0.7	-0.31042	-0.6	-0.05	823

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 “shares” (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)