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.
- Team and Projects' Registration **starts**: Tuesday 16/11/2021 11:00PM.
- Registration **ends**: Saturday 20/11/2021 11:59PM.
- Delivering Milestone 1: 11/12/2021.
- 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): Loan Risk Prediction

The aim of this project is to predict loan credit risk and determine the probability of non-payment of bank financial services (e.g., whether a person will pay back a loan or not). By using scoring models that are AI-based and use deep learning, banks and financial institutions can access more realistic predictions on credit risk, using customers' credit history and the power of big data. This way credit can be approved to the right people and better pricing options offered to people who deserve it.

Dataset Snapshot:

ListingNun	CreditGr	Term	LoanStatu	BorrowerAPR	BorrowerRate	ListingCategory	BorrowerState	EmploymentStatus	EmploymentStatusDuration	IsBorrowerHom	CreditScoreRange	CreditScoreRangeL
193129	C	36	Completed	0.16516	0.158	0	CO	Self-employed	2	TRUE	640	659
1209647		36	Current	0.12016	0.092	2	CO	Employed	44	FALSE	680	699
81716	HR	36	Completed	0.28269	0.275	0	GA	Not available		FALSE	480	499
658116		36	Current	0.12528	0.0974	16	GA	Employed	113	TRUE	800	819
909464		36	Current	0.24614	0.2085	2	MN	Employed	44	TRUE	680	699
1074836		60	Current	0.15425	0.1314	1	NM	Employed	82	TRUE	740	759
750899		36	Current	0.31032	0.2712	1	KS	Employed	172	FALSE	680	699
768193		36	Current	0.23939	0.2019	2	CA	Employed	103	FALSE	700	719
1023355		36	Current	0.0762	0.0629	7	IL	Employed	269	TRUE	820	839
1023355		36	Current	0.0762	0.0629	7	IL	Employed	269	TRUE	820	839
587746		60	Current	0.27462	0.2489	1	MD	Employed	300	FALSE	640	659
213551	C	36	Completed	0.15033	0.1325	0		Full-time	19	FALSE	640	659
1081604		36	Past Due (	0.17969	0.1435	1	AL	Employed	1	FALSE	680	699
840820		36	Current	0.13138	0.1034	1	AZ	Employed	98	TRUE	740	759
757359		60	Current	0.11695	0.0949	1	VA	Employed	35	FALSE	740	759
577164		36	Defaulted	0.35797	0.3177	13	FL	Other	121	TRUE	700	719

~Dataset header Continued:

RevolvingCreditBalance	BankcardUtilization	AvailableBankcardCredit	TotalTrades	DebtToIncomeRatio	IncomeRange	StatedMonthlyIncome	TotalProsp	LoanNumI	LoanOriginalAı	LoanRiskScore
0	0	1500	11	0.17	$25,000-49,999	3083.333333		19141	9425	
3989	0.21	10266	29	0.18	$50,000-74,999	6125		134815	10000	7.761652039
				0.06	Not displayed	2083.333333		6466	3001	
1444	0.04	30754	26	0.15	$25,000-49,999	2875		77296	10000	9.731208558
6193	0.81	695	39	0.26	$100,000+	9583.333333	11	102670	15000	4.539452696
62999	0.39	86509	47	0.36	$100,000+	8333.333333		123257	15000	10.6356456
5812	0.72	1929	16	0.27	$25,000-49,999	2083.333333		88353	3000	2.363887913
1260	0.13	2181	10	0.24	$25,000-49,999	3355.75		90051	10000	4.550680139
9906	0.11	77696	29	0.25	$25,000-49,999	3333.333333		121268	10000	9.764699768
9906	0.11	77696	29	0.25	$25,000-49,999	3333.333333		121268	10000	11.81191958
387	0.51	363	47	0.12	$75,000-99,999	7500		65946	13500	7.418640271
1220	0.32	2580	7	0.27	$1-24,999	1666.666667		20907	1000	
8624	0.7	3626	20	0.18	$25,000-49,999	2416.666667		125045	4000	4.633045336
9171	0.32	19129	18	0.09	$50,000-74,999	5833.333333		96202	8500	8.708976155
32898	0.43	42204	48	0.2	$100,000+	10833.33333		90060	19330	8.763824373
9103	0.97	178	17	0.49	$50,000-74,999	5500		63982	4000	5.244586899

Dataset Description:

Note: This dataset contains 24 variables that are explained in detail in the LoanRiskDataDescription.csv file found with the dataset
Dependent variable: LoanRiskScore

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must use an encoding technique on at least one feature)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the loan risk score (Deliver at least two different regression techniques).
3. Finish Milestone 1 Report.

Project(2): Video Likes Prediction

To determine the popularity of a video on a video sharing website such as YouTube, several factors are taken into consideration, not just the number of views but also how much the viewers like the content of the video. This dataset contains statistics regarding videos that trended at a certain date and the amount of views, comments and likes they obtained. The aim is to analyze which factors affect the popularity of a video and predict the number of likes obtained by a video.

Dataset Snapshots:

video_id	trending_dat	title	channel_title	category_id	publish_time	tags	views	comment_count	comments_disabled
NNywashg_mw	17.20.11	Is Chip Kelly interested in Gatc	ESPN	17	2017-11-19T14:10:10.000Z	espn|"espn live"|"chip kelly"|"florida	16431	54	FALSE
kl_jxVID3PE	18.26.02	Kali Uchis After The Storm Offi	Genius	10	2018-02-16T18:22:38.000Z	genius|"rap genius"|"verified"|"offic	321857	1117	FALSE
bilRlcQqmOc	18.19.03	RBG - Official Trailer	Magnolia Pictures	1	2018-03-07T20:01:39.000Z	rbg|"notorious rbg"|"ruth bader gins	288366	161	FALSE
hbcH02YynOA	18.17.03	FULL GLAM IN 10 MINUTES M/	Laura Lee	26	2018-03-07T22:00:03.000Z	Laura88Lee|"10 minute makeup"|"gl	425849	13642	FALSE
sScTRKljFUs	18.23.02	360Â° view of Renault R.S.18	Renault Sport	2	2018-02-20T14:55:27.000Z	F1|"Formula 1"|"Formula One"|"Rev	54198	50	FALSE
AZ4sakIfXUE	17.25.12	Things We Should Leave in 20	ConnorFranta	22	2017-12-18T19:17:48.000Z	Connor Franta|"ConnorFranta"|"201	251710	2770	FALSE
J0jgpUIbMEo	18.05.01	Homemade Nut Butter â˜† è‡	Peaceful Cuisine	26	2018-01-04T08:04:03.000Z	almond|"cashew nut"|"butter"|"hor	42033	210	FALSE
5TY79E-a2xU	18.20.01	4 officers hurt in shooting in S	Washington Post	25	2018-01-16T14:59:19.000Z	Washington Post YouTube|"Washing	10836	229	FALSE
uI5_G-Hiu8U	18.06.03	Meeting Mommy	Wong Fu Producti	24	2018-02-28T23:10:41.000Z	wong fu presents|"short film"|"simu	257021	1727	FALSE
GABCneYEE84	18.23.02	Honest Trailers - Justice Leagu	Screen Junkies	1	2018-02-20T18:00:02.000Z	screenjunkies|"screen junkies"|"hon	2318176	10950	FALSE
ewTQLoNPxYg	18.17.01	Where are we?	The Ohana Adven	24	2018-01-16T01:24:41.000Z	family|"adventure"|"the ohana adve	65182	753	FALSE
bV_3HQplzRw	18.02.06	There's Always a Bully	itsAlexClark	1	2018-05-30T22:05:02.000Z	its alex clark|"itsalexclark"|"alex clarl	603126	7504	FALSE
flvn3JfcVHk	18.18.01	Wreck It Ralph: Vanellope Von	Kandee Johnson	26	2018-01-13T20:22:31.000Z	vanellope von schweetz|"wreck it ral	709095	1368	FALSE
hIFwL6O_6aA	18.01.02	SPOILER ALERT! ðŸŒ¹ðŸ Krysta	The Bachelor Insic	24	2018-01-30T01:00:03.000Z	the bachelor insider|"bachelor inside	59833	185	FALSE
NukmFF__G-g	18.05.05	Mason Ramsey - Famous [Lyri	Mason Ramsey	10	2018-04-27T04:00:03.000Z	mason ramsey|"yodel kid"|"yodel bc	8367528	43633	FALSE
lzuWvKn80V8	17.12.12	If your reflection were honest	AnthonyPadilla	23	2017-12-11T16:59:58.000Z	anthony padilla|"padilla"|"anthony p	149231	1179	FALSE

~Dataset header Continued:

ratings_disabled	video_error_or_removed	video_description	likes
FALSE	FALSE	Chip Kelly responds on SportsCenter to sp	151
FALSE	FALSE	Kali Uchis first captivated the industry witl	24682
FALSE	FALSE	Like on Facebook: https://www.facebook.	1609
FALSE	FALSE	Hey Larlees, todays video is a fun collab w	21842
FALSE	FALSE	Discover the new Renault Sport Formula C	355
FALSE	FALSE	Subscribe to my channel here: http://bit.l	38887
FALSE	FALSE	You can substitute almonds with any kind	2420
FALSE	FALSE	Four officers, including three deputies, we	41
FALSE	FALSE	This video was available first to our Patrec	18108
FALSE	FALSE	Somewhere between the awful Suicide Sq	65752
FALSE	FALSE	On the road and not sure where we are? S	3212
FALSE	FALSE	Don't let the bully win. Keep chasing your	37343
FALSE	FALSE	Who's ready for my FIRST VIDEO of 2018!	14595
FALSE	FALSE	Spoiler Warning! ðŸŒ¹ðŸ Sweet Baby Krys	571
FALSE	FALSE	Here's my first song called Famous!Click h	397578
FALSE	FALSE	If your reflection were honest, what woul	16777

Dataset Description:

Feature	Description
video_id	Video ID (could be repeated)
trending_date	The date the video trended on youtube
title	Video Title
channel_title	
category_id	Video category id on youtube
publish_time	The time the video was published
tags	Video tags
views	Number of views the video obtained on this date
comment_count	Number of comments the video obtained on this date
comments_disabled	True or False value
ratings_disabled	True or False value
video_error_or_removed	True or False value
Likes (dependant variable)	Number of likes the video obtained on this date
video_description	(Optional – Read Below)

Additional Optional Data to use:

video_description: Video description on YouTube (Extracting and using information from this column is not mandatory but would be preferrable to experiment with)

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must use an encoding technique on at least one feature)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of number of likes of a video (Deliver at least two different regression techniques).
3. Finish Milestone 1 Report.