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**: Monday 30/11/2020 11:00PM.
- Registration **ends**: Friday 4/12/2020 11:59PM.
- Delivering Milestone 1: 25/12/2020.
- 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: 50%

- Preprocessing, Regression.

Milestone 1 Report Must Include:

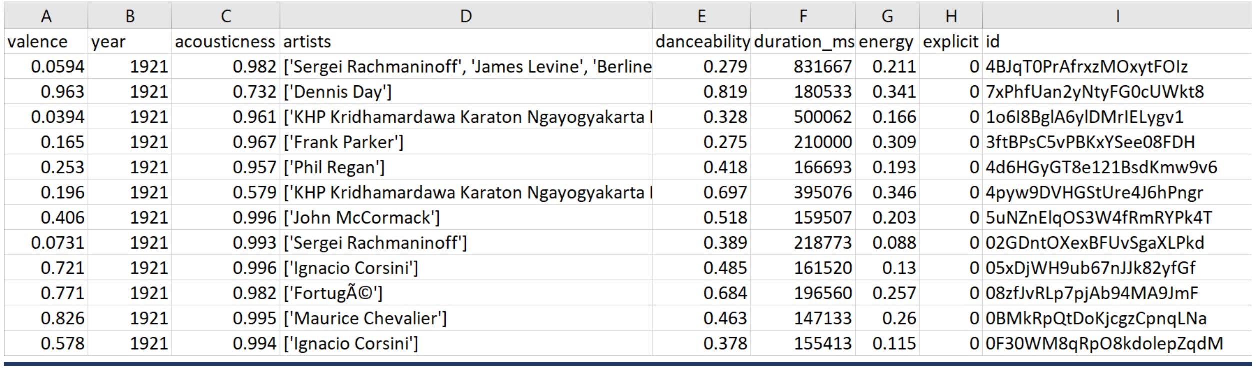
- You must explain in details the **preprocessing techniques** you needed to apply on your dataset and how you implemented them.
- Perform **analysis** on the dataset as studied and explain how the features affect and relate to each other.
- You must explain what **regression techniques** you used (at least two).
- Mention the **differences** between each model and the acquired **results** (accuracy/error and so on) 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): Predicting Song Popularity

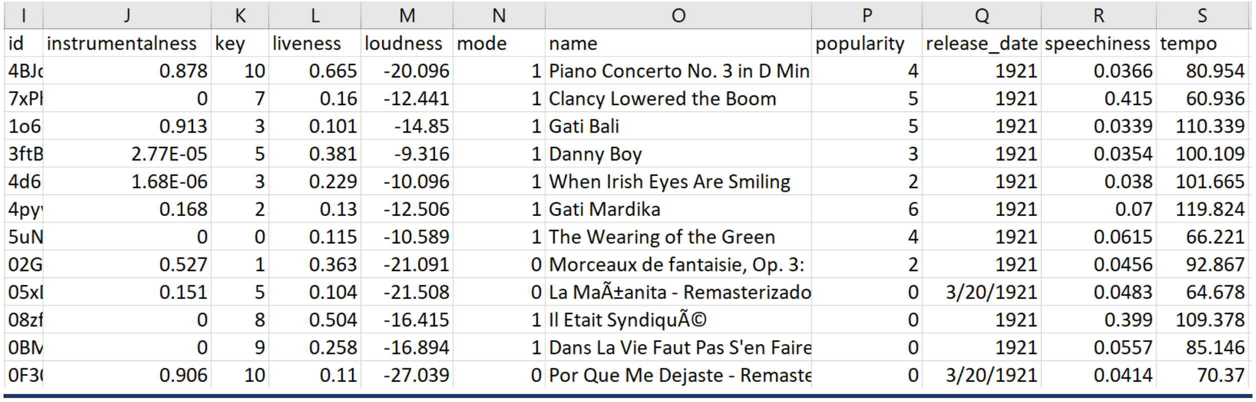
Can you predict a certain song’s popularity before it is even published to an audience? This dataset asks this question. It contains audio features of songs published between 1920 and 2020 along with a popularity score ranging from 0 to 100. Using the given data, try analyzing which features play the most important role in determining the popularity of a song.

Dataset Snapshot:



A	B	C	D	E	F	G	H	I
valence	year	acousticness	artists	danceability	duration_ms	energy	explicit	id
0.0594	1921	0.982	['Sergei Rachmaninoff', 'James Levine', 'Berline	0.279	831667	0.211	0	4BJqT0PrAfrxzMOxytFOIz
0.963	1921	0.732	['Dennis Day']	0.819	180533	0.341	0	7xPhfUan2yNtyFG0cUWkt8
0.0394	1921	0.961	['KHP Kridhamardawa Karaton Ngayogyakarta I	0.328	500062	0.166	0	1o6I8BglA6ylDMrIELygv1
0.165	1921	0.967	['Frank Parker']	0.275	210000	0.309	0	3ftBPsC5vPBKxYSee08FDH
0.253	1921	0.957	['Phil Regan']	0.418	166693	0.193	0	4d6HGyGT8e121BsdKmw9v6
0.196	1921	0.579	['KHP Kridhamardawa Karaton Ngayogyakarta I	0.697	395076	0.346	0	4pyw9DVHGStUre4J6hPngr
0.406	1921	0.996	['John McCormack']	0.518	159507	0.203	0	5uNZnElqOS3W4fRmRYPk4T
0.0731	1921	0.993	['Sergei Rachmaninoff']	0.389	218773	0.088	0	02GDntOXexBFUvSgaXLPkd
0.721	1921	0.996	['Ignacio Corsini']	0.485	161520	0.13	0	05xDjWH9ub67nJJk82yfGf
0.771	1921	0.982	['FortugÃ©']	0.684	196560	0.257	0	08zfJvRLp7pjAb94MA9JmF
0.826	1921	0.995	['Maurice Chevalier']	0.463	147133	0.26	0	0BMkRpQtDoKjcgzCpnqLNa
0.578	1921	0.994	['Ignacio Corsini']	0.378	155413	0.115	0	0F30WM8qRpO8kdolepZqdM

~Dataset header Continued:



I	J	K	L	M	N	O	P	Q	R	S
id	instrumentalness	key	liveness	loudness	mode	name	popularity	release_date	speechiness	tempo
4BJ	0.878	10	0.665	-20.096	1	Piano Concerto No. 3 in D Min	4	1921	0.0366	80.954
7xP	0	7	0.16	-12.441	1	Clancy Lowered the Boom	5	1921	0.415	60.936
1o6	0.913	3	0.101	-14.85	1	Gati Bali	5	1921	0.0339	110.339
3ftB	2.77E-05	5	0.381	-9.316	1	Danny Boy	3	1921	0.0354	100.109
4d6	1.68E-06	3	0.229	-10.096	1	When Irish Eyes Are Smiling	2	1921	0.038	101.665
4py	0.168	2	0.13	-12.506	1	Gati Mardika	6	1921	0.07	119.824
5uN	0	0	0.115	-10.589	1	The Wearing of the Green	4	1921	0.0615	66.221
02G	0.527	1	0.363	-21.091	0	Morceaux de fantaisie, Op. 3:	2	1921	0.0456	92.867
05x	0.151	5	0.104	-21.508	0	La MaÃ±anita - Remasterizado	0	3/20/1921	0.0483	64.678
08z	0	8	0.504	-16.415	1	Il Etait SyndiquÃ©	0	1921	0.399	109.378
0B	0	9	0.258	-16.894	1	Dans La Vie Faut Pas S'en Faire	0	1921	0.0557	85.146
0F3	0.906	10	0.11	-27.039	0	Por Que Me Dejaste - Remaste	0	3/20/1921	0.0414	70.37

Dataset Description:

Feature	Description
valence	
year	Ranges from 1921 to 2020
acousticness	
artists	List of artists mentioned (Categorical)
danceability	
duration_ms	Integer typically ranging from 200k to 300k
energy	Ranges from 0 to 1
explicit	0 = No explicit content, 1 = Explicit content
id	Id of track generated by Spotify
instramentalness	
key	All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on… (Categorical)
liveness	
loudness	
mode	0 = Minor, 1 = Major
name	Name of the song
popularity	Ranges from 0 to 100
Release_date	Date of release mostly in yyyy-mm-dd format, however precision of date may vary
speechiness	
tempo	

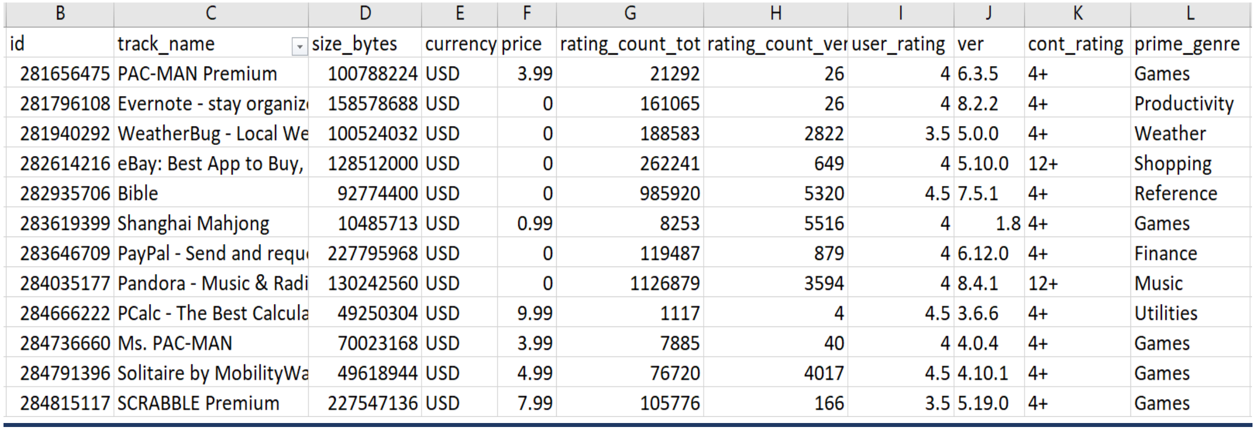
Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (Use One-Hot-Encoding for at least one categorical feature)
2. Experiment with regression techniques to reduce the error on prediction of the average popularity of a song (Deliver at least two techniques).
3. Finish Milestone 1 Report.

Project(2): Predict Mobile App Success

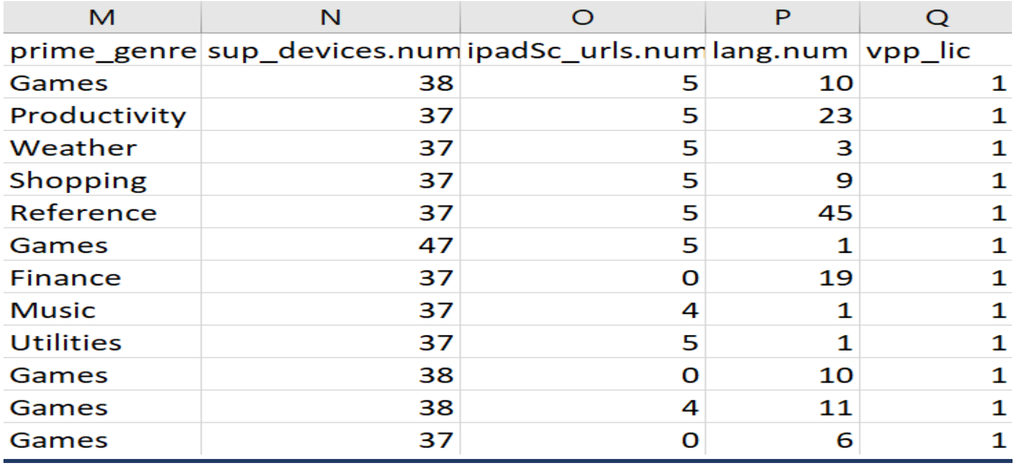
The ever-changing mobile landscape is a challenging space to navigate. The percentage of mobile over desktop is only increasing. Android holds about 53.2% of the smartphone market, while iOS is 43%. To get more people to download your app, you need to make sure they can easily find your app. Mobile app analytics is a great way to understand the existing strategy to drive growth and retention of future users.

Dataset Snapshots:



B	C	D	E	F	G	H	I	J	K	L
id	track_name	size_bytes	currency	price	rating_count_tot	rating_count_ver	user_rating	ver	cont_rating	prime_genre
281656475	PAC-MAN Premium	100788224	USD	3.99	21292	26	4	6.3.5	4+	Games
281796108	Evernote - stay organiz	158578688	USD	0	161065	26	4	8.2.2	4+	Productivity
281940292	WeatherBug - Local We	100524032	USD	0	188583	2822	3.5	5.0.0	4+	Weather
282614216	eBay: Best App to Buy,	128512000	USD	0	262241	649	4	5.10.0	12+	Shopping
282935706	Bible	92774400	USD	0	985920	5320	4.5	7.5.1	4+	Reference
283619399	Shanghai Mahjong	10485713	USD	0.99	8253	5516	4	1.8	4+	Games
283646709	PayPal - Send and requ	227795968	USD	0	119487	879	4	6.12.0	4+	Finance
284035177	Pandora - Music & Radi	130242560	USD	0	1126879	3594	4	8.4.1	12+	Music
284666222	PCalc - The Best Calcula	49250304	USD	9.99	1117	4	4.5	3.6.6	4+	Utilities
284736660	Ms. PAC-MAN	70023168	USD	3.99	7885	40	4	4.0.4	4+	Games
284791396	Solitaire by MobilityWa	49618944	USD	4.99	76720	4017	4.5	4.10.1	4+	Games
284815117	SCRABBLE Premium	227547136	USD	7.99	105776	166	3.5	5.19.0	4+	Games

~Dataset header Continued:



M	N	O	P	Q
prime_genre	sup_devices.num	ipadSc_urls.num	lang.num	vpp_lic
Games	38	5	10	1
Productivity	37	5	23	1
Weather	37	5	3	1
Shopping	37	5	9	1
Reference	37	5	45	1
Games	47	5	1	1
Finance	37	0	19	1
Music	37	4	1	1
Utilities	37	5	1	1
Games	38	0	10	1
Games	38	4	11	1
Games	37	0	6	1

Dataset Description:

Feature	Description
id	App ID
track_name	App Name
size_bytes	Size (in Bytes)
currency	Currency Type
price	Price amount
rating*count*tot	User Rating counts (for all version)
rating*count*ver	User Rating counts (for current version)
user_rating	Average User Rating value (for all version)
ver	Latest version code
cont_rating	Content Rating
prime_genre	Primary Genre
sup_devices.num	Number of supporting devices
ipadSc_urls.num	Number of screenshots showed for display
lang.num	Number of supported languages

Additional Optional Data to use: App Description

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (Use One-Hot-Encoding for at least one categorical feature)
2. Experiment with regression techniques to reduce the error on prediction of user rating of an app (Deliver at least two techniques).
3. Finish Milestone 1 Report.