Clinical Data Science – Final Assignment

One Year Survival Analysis of Sepsis Patients

Gal Shlomo Ben David Ohayone

Abstract

In this work I am trying to reproduce the published study of Zhao et al. Zhao have shown that platelets count in sepsis patients can be a predictor for one year survival after admission at the ICU. The statistical analysis using MIMIC-III data. In this work we took into consideration the population demographic details, vital signs, Laboratory test results, calculated risk scores and some admission outcomes.

Introduction

Background

Sepsis is "life-threatening organ dysfunction caused by a dysregulated host response to infection" (Singer et al, 2016) When body's immune system repones to the infection is exaggerated or not effective enough, this incidence can lead to multi-organ failure, shock, and death.

Sepsis causes many deaths worldwide, with a very high (16.7 billion US dollars) cost of caring to those patients that suffer from this incidence. (Walkey et al, 2015) Those hospitalized partienents conservative estimated death rate reach 30%. (Reinhart et al, 2017)

Objectives

Zhao et al have created a Proportional hazards regression model, our goal is to create a logistic regression model that will outperform Zhao model, and will dill with some of its limitations. In our work we would like to be based on data that can be obtained within first 24 hours of ICU admission. We will validate our model with 10 folds cross-validation on train data (70%) and test our model on test data (30%).

Methods

Source of data

For the research we have used the MIMIC (Medical Information Mart for Intensive Care) Database, the same one that were used in the original research. This database include data of patients that have admitted to Intensive Care Unit of Beth Israel Care Unit in 2001 until 2012.
I have extracted the data using SQL with google Big Query platform and processed with python data science libraries using Google's Collaboratory online notebook. The MIMIC III database (version 1.4) is publicly available in https://mimic.physionet.org/ under filling the access requirements.

Participants

As explained above, The population in this research have admitted to Intensive Care Unit, therefore they can represent the whole population only partially.

Outcome

- One year survival –

I have extracted the date of death (DOD) for each patient (when available) and calculated the distance in time[days] from this admission. When DOD wasn't available, I have assumed that the patients is still alive and fill the data cell with a 'very big number' in the terms of the problem (10,000). After that, I have screened the values to less than one year (0) and one year or more (1).

- Length of stay – in hours
- In hospital mortality (binary)

Predictors

- Partial thromboplastin time (PTT)
- International normalized ratio (INR)
- Prothrombin time (PT)
- White blood cell count (WBC)
- Haemoglobin
- Platelet
- Blood urea nitrogen (BUN)
- Creatinine (Cr)
- Spo2

For all of the above I have made an assumption that their mean value of lab test is the mean between their max value and their mean value of the first 24h of staying at the ICU.

- Sodium (minimal and maximal from first 24 hours of ICU stay)
- Potassium (minimal and maximal from first 24 hours of ICU stay)
- Glucose (minimal and maximal from first 24 hours of ICU stay)

For those 3 I have also calculated "mean"

- Admission type (Emergency, Elective, or Urgent) 2. Age
  Sex
  Race (White, African American, Asian, or Others)
- Heart rate (HR)
  Systolic blood pressure (SBP)
- Diastolic blood pressure (DBP)
- Respiratory rate (RR)
  Temperature

- Glasgow coma scale (GCS) - 'GCS', 'GCSMotor', 'GCSVerbal', 'GCSEyes'
    - (I have used more than one feature of GCS in order to try and improve results)
- Patients' simplified acute physiology score (SAPSII)
- Sequential organ failure assessment (SOFA)

Sample size

There were 5152 admission that were taken into consideration, (I had another try with less restricted excluding rules and were able to create better model in some way)

Missing data

I have eliminated rows with missing data at the lab columns.

For censored data, like 'days untill death'– I have filled by myself as mention above.

Statistical analysis methods

I have used Kaplan-Meier curve in order to compare survival groups, I have created a logistics regression model in order to forecast one year survival.

Risk groups

blood platelets level ≥ 100 (×109/L) and ≤300 (×109/L) were considered Normal. (the final risk group were :Thrombocytosis, Normal, and Thrombocytopenia)

Development vs. validation

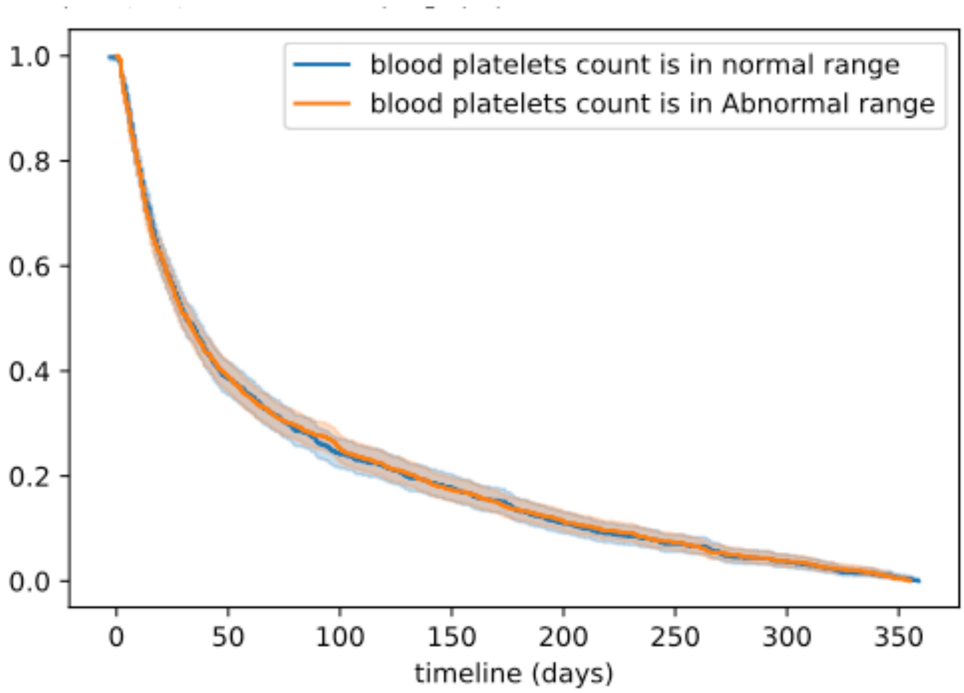
Results

Baseline patient characteristics and outcomes. I have painted the features that I believe that have big enough p-value in order to be correlated to our risk group.

Table 1

	Grouped by group						
		Missing	Overall	Normal	Thrombocytopenia	Thrombocytosis	P-Value
n			5152	3063	962	1127	
GCS, mean (SD)		28	13.5 (2.8)	13.4 (2.9)	13.6 (2.6)	13.6 (2.6)	0.009
GCSMotor, mean (SD)		86	5.0 (1.6)	4.9 (1.7)	5.2 (1.5)	5.1 (1.6)	<0.001
GCSVerbal, mean (SD)		71	2.9 (2.2)	2.8 (2.2)	3.3 (2.1)	2.8 (2.2)	<0.001
GCSEyes, mean (SD)		44	3.0 (1.1)	2.9 (1.2)	3.1 (1.1)	3.0 (1.1)	<0.001
sapsii, mean (SD)		0	43.6 (15.4)	43.2 (14.9)	46.4 (16.8)	42.1 (15.0)	<0.001
sapsii_prob, mean (SD)		0	0.3 (0.3)	0.3 (0.3)	0.4 (0.3)	0.3 (0.3)	<0.001
HEART_RATE_HR, mean (SD)		910	85.6 (18.9)	84.6 (18.8)	87.0 (19.5)	86.9 (18.3)	0.001
SYSTOLIC_BLOOD_PRESSURE_SBP, mean (SD)		442	118.3 (33.7)	118.8 (34.0)	116.5 (42.0)	118.5 (23.2)	0.224
DIASTOLIC_BLOOD_PRESSURE_DBP, mean (SD)		442	64.8 (34.7)	64.3 (32.6)	65.7 (37.3)	65.3 (37.7)	0.495
SOFA, mean (SD)		0	6.8 (3.5)	6.5 (3.2)	9.0 (3.9)	5.7 (3.0)	<0.001
sepsis, mean (SD)		0	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	nan
organ_failure, mean (SD)		0	0.9 (0.3)	0.9 (0.3)	0.9 (0.3)	0.9 (0.3)	<0.001
respiratory, mean (SD)		0	0.4 (0.5)	0.4 (0.5)	0.4 (0.5)	0.4 (0.5)	0.161
cardiovascular, mean (SD)		0	0.4 (0.5)	0.4 (0.5)	0.5 (0.5)	0.4 (0.5)	0.19
renal, mean (SD)		0	0.7 (0.5)	0.7 (0.5)	0.7 (0.5)	0.6 (0.5)	0.005
hepatic, mean (SD)		0	0.1 (0.3)	0.1 (0.3)	0.2 (0.4)	0.1 (0.2)	<0.001
hematologic, mean (SD)		0	0.2 (0.4)	0.2 (0.4)	0.6 (0.5)	0.1 (0.3)	<0.001
metabolic, mean (SD)		0	0.2 (0.4)	0.2 (0.4)	0.3 (0.4)	0.2 (0.4)	<0.001
neurologic, mean (SD)		0	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)	0.256
SEX, n (%)	F	0	2140 (41.5)	1256 (41.0)	382 (39.7)	502 (44.5)	0.053
	M		3012 (58.5)	1807 (59.0)	580 (60.3)	625 (55.5)	
AGE, mean (SD)		0	64.5 (15.2)	65.6 (15.1)	59.5 (14.9)	65.6 (14.5)	<0.001
RACE, n (%)	ASIAN	0	119 (2.3)	86 (2.8)	17 (1.8)	16 (1.4)	0.001
	BLACK		574 (11.1)	365 (11.9)	84 (8.7)	125 (11.1)	
	OTHER		751 (14.6)	431 (14.1)	178 (18.5)	142 (12.6)	
	WHITE		3704 (71.9)	2178 (71.1)	682 (70.9)	844 (74.9)	
ADMISSION_TYPE, n (%)	ELECTIVE	0	199 (3.9)	130 (4.2)	29 (3.0)	40 (3.5)	0.201
	EMERGENCY		4816 (93.5)	2856 (93.2)	900 (93.6)	1060 (94.1)	
	URGENT		137 (2.7)	77 (2.5)	33 (3.4)	27 (2.4)	
icustay_id, mean (SD)		0	250229.4 (28903.8)	250058.8 (28866.5)	252291.3 (28894.9)	248932.8 (28945.2)	0.026
CREATININE, mean (SD)		12	2.1 (2.0)	2.2 (2.1)	2.0 (1.7)	2.0 (2.0)	0.007
GLUCOSE, mean (SD)		10	147.9 (62.0)	148.6 (62.3)	144.1 (61.2)	149.1 (62.1)	0.107
HEMOGLOBIN, mean (SD)		13	10.3 (1.8)	10.5 (1.8)	9.8 (1.7)	10.0 (1.7)	<0.001
PLATELET, mean (SD)		14	216.5 (134.1)	192.1 (54.6)	62.8 (23.6)	413.8 (113.0)	<0.001
PTT, mean (SD)		378	41.5 (19.9)	41.3 (20.0)	44.4 (20.1)	39.5 (18.9)	<0.001
INR, mean (SD)		351	1.8 (1.2)	1.8 (1.2)	1.8 (0.9)	1.8 (1.4)	0.264
PT, mean (SD)		350	17.9 (8.4)	17.8 (8.8)	18.3 (6.3)	17.8 (8.9)	0.245
BUN, mean (SD)		12	36.7 (25.5)	36.4 (24.9)	39.7 (27.0)	35.0 (25.7)	<0.001
WBC, mean (SD)		17	13.9 (14.0)	13.6 (8.4)	10.8 (26.3)	17.4 (9.5)	<0.001
POTASSIUM_min, mean (SD)		8	3.8 (0.6)	3.8 (0.6)	3.7 (0.6)	3.8 (0.6)	<0.001
POTASSIUM_max, mean (SD)		8	4.7 (1.0)	4.7 (1.0)	4.5 (0.9)	4.7 (0.9)	<0.001
SODIUM_min, mean (SD)		9	136.1 (5.5)	136.3 (5.4)	135.7 (6.1)	136.0 (5.5)	0.006
SODIUM_max, mean (SD)		9	140.0 (5.4)	140.2 (5.2)	139.6 (6.0)	139.6 (5.2)	<0.001
tempc_mean, mean (SD)		58	36.9 (0.8)	36.9 (0.8)	36.8 (0.8)	36.9 (0.7)	<0.001
tempc_max, mean (SD)		58	37.7 (1.0)	37.8 (1.0)	37.6 (1.0)	37.7 (0.9)	0.001
tempc_min, mean (SD)		58	36.1 (0.9)	36.1 (0.9)	35.9 (1.0)	36.1 (0.8)	<0.001
resprate_mean, mean (SD)		20	20.6 (4.7)	20.4 (4.6)	21.0 (5.3)	20.5 (4.7)	0.005
resprate_max, mean (SD)		20	29.2 (7.5)	29.1 (7.3)	29.6 (8.2)	29.2 (7.3)	0.181
resprate_min, mean (SD)		20	13.3 (4.2)	13.2 (4.1)	13.6 (4.4)	13.2 (4.3)	0.025
RRT, mean (SD)		0	0.2 (0.4)	0.2 (0.4)	0.3 (0.4)	0.2 (0.4)	<0.001

Survival analysis



CoxPHFitter for the two main risk group. It is impossible to infer from this graph if a patients is having more chance to survive due to the fact that he have normal or abnormal blood platelents counts.

One year survival analysis

Table 2 – univariate and multivariate survivals analysis

I have colored the features that have the most statistical significance.

Table2

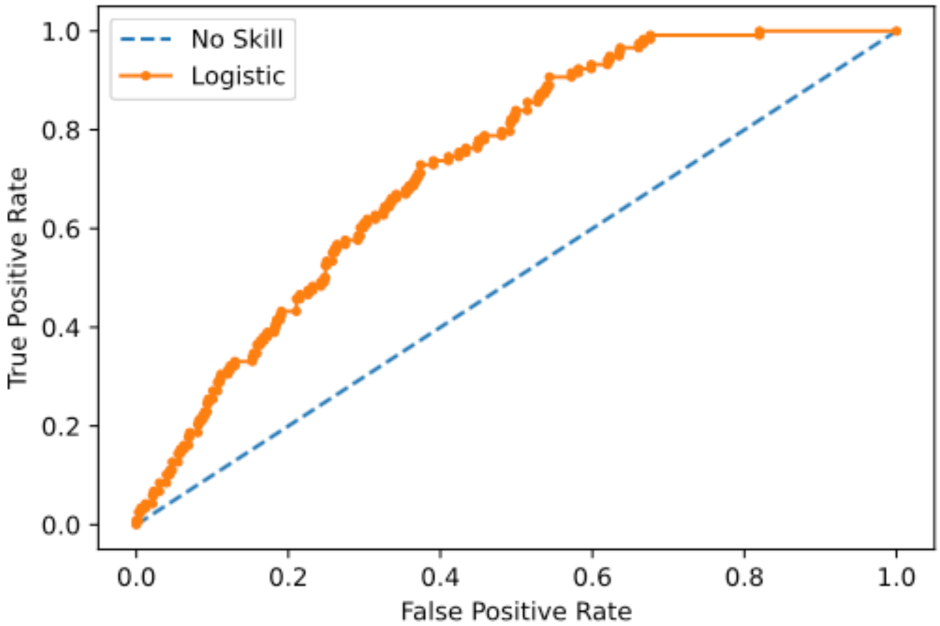
		Univariate analysis				Multivariate analysis			
#	**Feature name**	**Odds Ratio**	**p-value**	**lower bound 95%**	**upper bound 95%**	**Odds Ratio**	**p-value**	**lower bound 95%**	**upper bound 95%**
0	ADMISSION_TYPE_ELECTIVE	0.95	0.50	-0.19	0.09	1.06	0.69	-0.21	0.32
1	ADMISSION_TYPE_EMERGENCY	1.06	0.34	-0.06	0.16	0.99	0.96	-0.26	0.24
2	ADMISSION_TYPE_URGENT	0.94	0.50	-0.23	0.11	0.94	0.67	-0.34	0.22
3	AGE	1.00	1.00	0.00	0.00	1.72	0.00	0.46	0.63
4	BUN	1.00	0.07	0.00	0.00	1.00	0.09	0.00	0.00
5	cardiovascular	1.17	0.00	0.10	0.21	1.03	0.37	-0.03	0.08
6	CREATININE	1.00	0.48	-0.02	0.01	0.97	0.00	-0.05	-0.01
7	DIASTOLIC_BLOOD_PRESSURE_DBP	1.00	0.76	0.00	0.00	1.00	0.20	0.00	0.00
8	GCS	1.00	0.35	-0.01	0.01	1.01	0.04	0.00	0.02
9	GCSEyes	0.99	0.25	-0.04	0.01	0.99	0.55	-0.04	0.02
10	GCSMotor	0.99	0.13	-0.03	0.00	1.01	0.22	-0.01	0.04
11	GCSVerbal	1.00	0.56	-0.02	0.01	0.99	0.20	-0.03	0.01
12	GLUCOSE	1.00	0.43	0.00	0.00	1.00	0.11	0.00	0.00
13	group_Normal	0.92	0.01	-0.13	-0.02	0.98	0.75	-0.14	0.10
14	group_Thrombocytopenia	1.12	0.00	0.04	0.18	1.05	0.48	-0.09	0.18
15	group_Thrombocytosis	1.01	0.73	-0.05	0.08	0.98	0.82	-0.15	0.12
16	HEART_RATE_HR	1.00	0.00	0.00	0.00	1.00	0.03	0.00	0.00
17	hematologic	1.12	0.00	0.05	0.18	1.00	0.91	-0.07	0.06
18	HEMOGLOBIN	0.99	0.08	-0.03	0.00	0.97	0.00	-0.04	-0.01
19	hepatic	1.21	0.00	0.10	0.28	1.03	0.60	-0.07	0.12
20	INR	1.02	0.13	-0.01	0.04	1.08	0.00	0.04	0.11
21	LOS_IN_HOURS	1.00	0.25	0.00	0.00	1.00	0.04	0.00	0.00
22	metabolic	1.14	0.00	0.06	0.20	0.97	0.32	-0.11	0.03
23	neurologic	1.18	0.00	0.09	0.25	0.91	0.02	-0.18	-0.02
24	organ_failure	1.15	0.00	0.06	0.23	1.06	0.24	-0.04	0.16
25	PLATELET	1.00	0.17	0.00	0.00	1.00	0.30	0.00	0.00
26	POTASSIUM_max	1.00	0.77	-0.02	0.03	1.02	0.32	-0.02	0.05
27	POTASSIUM_min	1.02	0.37	-0.02	0.07	1.07	0.01	0.02	0.12
28	PT	1.00	0.01	0.00	0.01	0.99	0.00	-0.01	0.00
29	PTT	1.00	0.59	0.00	0.00	1.00	0.01	0.00	0.00
30	RACE_ASIAN	1.16	0.10	-0.03	0.33	1.03	0.75	-0.18	0.24
31	RACE_BLACK	0.99	0.73	-0.10	0.07	0.93	0.32	-0.22	0.07
32	RACE_CARIBBEAN_ISLAND	1.46	0.60	-1.01	1.76	0.67	0.54	-1.71	0.90
33	RACE_HISPANIC_or_LATINO_HONDURAN	1.46	0.71	-1.58	2.34	0.63	0.62	-2.30	1.38
34	RACE_HISPANIC_or_LATINO_MEXICAN	1.46	0.71	-1.58	2.34	0.75	0.77	-2.14	1.57
35	RACE_OTHER	1.02	0.61	-0.06	0.10	1.02	0.82	-0.12	0.16
36	RACE_WHITE	0.98	0.48	-0.08	0.04	1.02	0.73	-0.11	0.16
37	renal	1.11	0.00	0.04	0.16	1.00	0.96	-0.07	0.07
38	respiratory	1.09	0.00	0.03	0.14	1.03	0.35	-0.03	0.09
39	resprate_max	1.01	0.00	0.00	0.01	1.00	0.31	0.00	0.01
40	resprate_mean	1.01	0.00	0.01	0.02	0.99	0.15	-0.02	0.00
41	resprate_min	1.00	0.81	-0.01	0.01	1.01	0.09	0.00	0.01
42	RRT	1.08	0.03	0.01	0.15	1.19	0.00	0.10	0.26
43	sapsii	1.00	0.00	0.00	0.01	1.01	0.00	0.00	0.01
44	sapsii_prob	1.35	0.00	0.19	0.41	1.36	0.01	0.08	0.54
45	SEX_M	1.02	0.45	-0.03	0.08	1.03	0.34	-0.03	0.08
46	SODIUM_max	1.00	0.07	-0.01	0.00	1.00	0.51	-0.01	0.01
47	SODIUM_min	1.00	0.06	-0.01	0.00	1.01	0.05	0.00	0.01
48	SOFA	1.02	0.00	0.02	0.03	0.99	0.11	-0.02	0.00
49	SYSTOLIC_BLOOD_PRESSURE_SBP	1.00	0.10	0.00	0.00	1.00	0.67	0.00	0.00
50	tempc_max	1.00	1.00	-0.03	0.03	0.96	0.07	-0.09	0.00
51	tempc_mean	0.97	0.15	-0.06	0.01	0.99	0.85	-0.08	0.07
52	tempc_min	0.97	0.09	-0.06	0.00	0.98	0.34	-0.07	0.02
53	WBC	1.00	0.60	0.00	0.00	1.00	0.91	0.00	0.00

Multivariate model – I have created a logistic regression model based on those features.

Multivariate Model discrimination

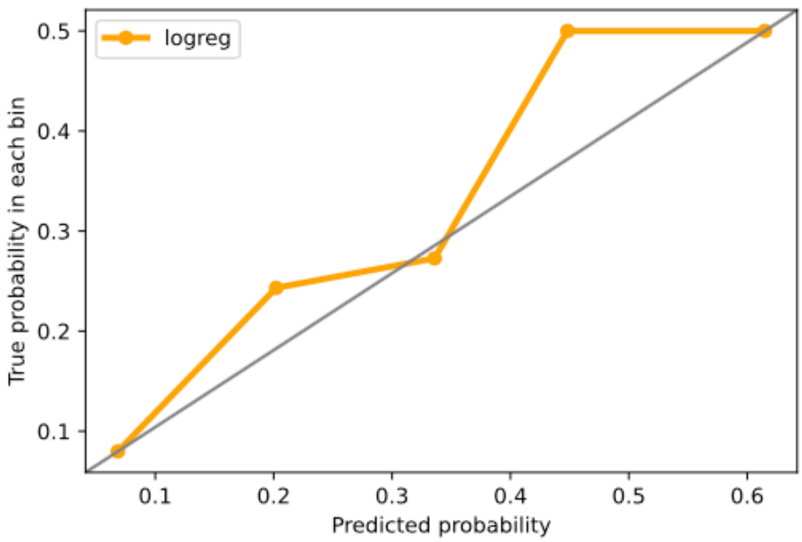
the model preform better than no skilled classifier but doesn't outperform the original article model.

No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.729



Multivariate Model calibration

As it is possible to see, the the model is quite calibrated when being divided into 7 beans, it is possible to see that the calculated risk is very similar to the real risk.



Discussion

Limitations- the population was chosen under very limited criteria as mention above, it might be interesting to make the same research with bigger population. I have followed the instructions and haven't increased the population group. I suggest that that problem of sepsis survival analysis can be defined with less restriction on the population clinical data, and maybe more defined on the demographic data.
It might be interesting to explore other features that appears in this database.

Interpretation- we have gotten some similar results in our bassline table, model performance and on the data analysis odds ratio – to the information that appears in the article.

Implications- I don't think that my model will be more relevant than the one that appears in the article, but exploring the data was an instructive experience.

Bibliography

Reinhart K, Daniels R, Kissoon N, Machado FR, Schachter RD, Finfer SN Engl J Med. 2017 Aug 3; 377(5):414-417.

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC JAMA. 2016 Feb 23; 315(8):801-10.

Walkey AJ, Lagu T, Lindenauer PK Ann Am Thorac Soc. 2015 Feb; 12(2):216-20. Those

I am sorry but I didn't manage to extract my python file as PDF. I have tried to do it in many different ways.

Here is a colab link:
https://colab.research.google.com/drive/1iYTBqcXe8eAJxkwKJ-hNoa9N2UJsXOwj?usp=sharing
another link (git):
https://github.com/galbendavids/project1/blob/master/CDS_ex_final.ipynb