Documentation

This project is based on prediction of drug name with ENR and sentiment of reviews. We have collected the dataset from Kaggle 'DrugReview Dataset' that consist on 100K+ samples. The short description of the work in points are below.

1) Dataset From Kaggle
2) NER (Named Entity Recognizer) is used for medical entities extraction from drug review.
3) We have considered only first medical entity from drug review text.
4) We have selected top 30 drug name as response variable and its corresponding samples for explanatory variables.
5) We have performed sentiment analysis on drug review text. But before sentiment analysis we have clean the text with preprocessing functions (remove punctuation, stop words, lower case conversion, special character removing).
6) TextBlob is used for sentiment scores (1: Positive, -1 for negative, 0 for neutral).
7) At this point we have 3 features a s input (condition, sentiment, NER Entity). Ans response variable drug names.

	drugName	condition	sentiment	ENRS
0	Levonorgestrel	Emergency Contraception	1	Normal
1	Ethinyl estradiol / levonorgestrel	Birth Control	1	acne
2	Nexplanon	Birth Control	1	acne
3	Etonogestrel	Birth Control	1	depressed
4	Sertraline	Depression	1	zoloft anxiety mood

8) we have 30 classes and this problem is called multi-class problem.
9) Label Encoding is used to convert condition, ENR, and Drug Names into integer representation.

	drugName	condition	sentiment	ENRS
0	16	55	1	0
1	10	36	1	51
2	22	36	1	51
3	13	36	1	1546
4	25	51	1	4857

10) Then we have split the dataset in 20, 80 ratios for test and training respectively.
11) We have applied different machine learning models with grid search cv to find best parameters. And present the models here which give us effective prediction accuracy.
12) Random Forest accuracy is **42**%.

	precision	recall	f1-score	support
0	0.66	0.61	0.63	200
1	0.34	0.37	0.35	189
2	0.45	0.68	0.54	203
3	0.44	0.51	0.47	164
4	0.30	0.24	0.27	177
5	0.21	0.09	0.13	169
6	0.38	0.27	0.32	182
7	0.56	0.21	0.30	179
8	0.25	0.26	0.26	162
9	0.26	0.48	0.33	233
10	0.45	0.18	0.26	388
11	0.26	0.25	0.26	548
12	0.37	0.31	0.34	428
13	0.24	0.63	0.34	639
14	0.73	0.61	0.66	216
15	0.07	0.01	0.02	231
16	0.52	0.65	0.58	753
17	0.16	0.06	0.09	189
18	0.36	0.39	0.37	207
19	0.96	0.98	0.97	188
20	1.00	0.99	1.00	188
21	0.35	0.19	0.25	257
22	0.12	0.03	0.05	446
23	0.84	0.87	0.86	332
24	0.67	0.77	0.71	183
25	0.41	0.41	0.41	263
26	0.35	0.15	0.21	151
27	0.63	0.82	0.71	152
28	0.49	0.37	0.42	226
29	0.22	0.08	0.12	166
accuracy			0.42	8009
macro avg	0.43	0.42	0.41	8009
weighted avg	0.42	0.42	0.40	8009