UCI ML Drug Review dataset

EL CONJUNT DE DADES (Drug Review dataset)

El conjunt de dades escollit es el "UCI ML Drug Review dataset" publicat a Kaggle.

Les dades proporcionen les ressenyes de pacients sobre el efecte de fàrmacs per una determinada afecció. També inclou una valoració de 0 a 10 sobre la satisfacció general del pacient i el nombre d'usuaris que van considerar útil la ressenya.

El conjunt de dades està ja dividit en un conjunt d'eternament i en un de proves els qual ajuntarem per fer una anàlisis exploratori inicial

CARACTERÍSTIQUES GENERALS

- ***Tipologia***:

  El conjunt de dades combina tant dades qualitatives (***review***, **condition**) com quantitatives (***usefulcount***, ***rating***,).
  Nombre d'atributs: 6
  Nombre de ressenyes 206461 (161297 d'entrenament i 53766 de prova
  Son dades sense processar que es van obtenir rastrejant dades de ressenyes en línia publicades per webs d'empreses farmacèutiques
- ***Sector:*** Mèdic/farmacèutic .
- ***Tipus de dades:*** El conjunt consisteix en una serie de variables numèrics i de text
- ***Font:*** Surya Kallumadi , Kansas State University , Manhattan, Kansas, USA
- Aquestes dades es van publicar en un estudi sobre l'anàlisi de sentiments sobre l'experiència de fàrmacs en múltiples facetes, com per exemple sentiments apresos sobre aspectes específics com l'eficàcia i els efectes secundaris.
- *Reconeixements:*

  El conjunt de dades es va publicar originalment al repositori d'aprenentatge automàtic de la UCI.

  Citació: -Felix Gräßer, Surya Kallumadi, Hagen Malberg i Sebastian Zaunseder. 2018. Anàlisi de sentiments basat en aspectes de revisions de fàrmacs aplicant aprenentatge entre dominis i dades creuades. A Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, Nova York, NY, EUA, 121-125

  Quan utilitzeu aquest conjunt de dades, accepteu que:

  - Utilitzeu les dades només amb finalitats de recerca
  - No utilitzeu les dades amb finalitats comercials
  - No distribuïu les dades a ningú més

DEFINICIO DE VARIABLES

Els atributs del conjunt de dades son els següents:

- **drugName** (categòrica): Nom del farmac/s
- **condition** (categòrica): Nom de l'afecció
- **review** (text): Ressenya del pacient
- **rating** (numèrica): Avaluació del pacient 1-10
- **date** (data): Data en que es va efectuar la ressenya.
- **usefulCount** (numèrica): Nombre d'usuaris que van considerar útil la ressenya

	uniqueID	drugName	condition	review	rating	date	usefulCount	trainTest
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27	True
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192	True
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17	True
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10	True

OBECTIUS

1 - Anàlisi exploratòria i visualització de dades.

Importació i concatenació de conjunt d'entrenament i el de proves per tal de fer una exploració de conjunt sencer. Farem una exploració bàsica per identificar valors nuls,repetits i preprocessar algunes variables de cara a facilitar la fase de visualització i processament posterior de les dades

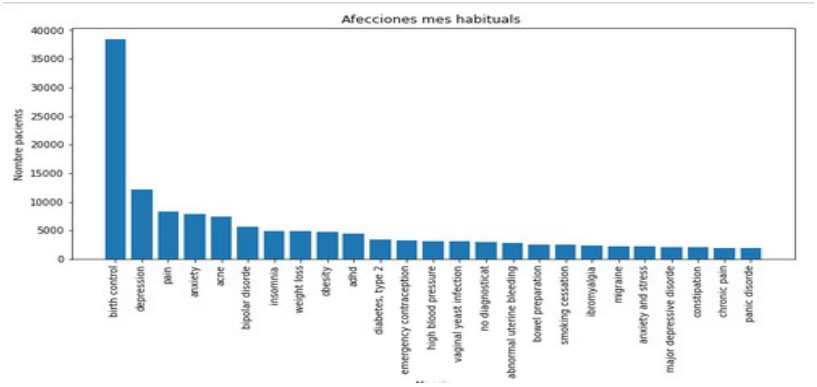
El passos que seguiré son:

- Convertir atributs al tipus de dades mes adients
- Identificació de valors nuls i duplicats
- Dividiren la data en dia , mes i any
- Eliminació de espais superflus en variables de tipus text
- Anàlisis de variables Afeccions, Fàrmacs i relacions entre elles

Variable afecció(condition)

Veure quantes afeccions s'han tractat i quines son les mes freqüents
Nombre total d'afeccions 836, unes poques afeccions concentres una gran quantitat de ressenyes , la mediana de la serie es 13 , per tant hi ha moltes afeccions amb poques ressenyes

	condition	frequency
0	birth control	38428
1	depression	12163
2	pain	8245
3	anxiety	7808
4	acne	7431
5	bipolar disorde	5601
6	insomnia	4902
7	weight loss	4855
8	obesity	4757
9	adhd	4509

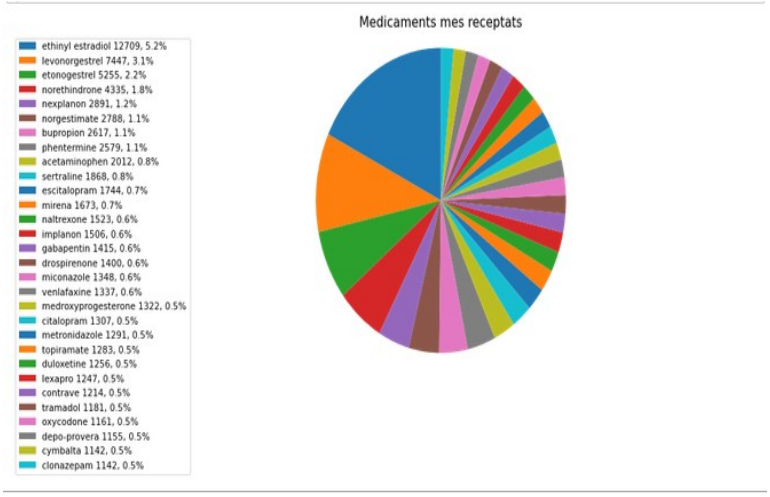


Variable fàrmacs(drugName)

Veure quants fàrmacs o combinacions de fàrmac s'han utilitzat, quins son els mes freqüents en el conjunt de dades. Finalment si hi han varis fàrmacs en una mateixa ressenya separar-los de forma individual (fen servir ***regex***)

```
#Top 15 farmacs
drugRank.head(10)
```

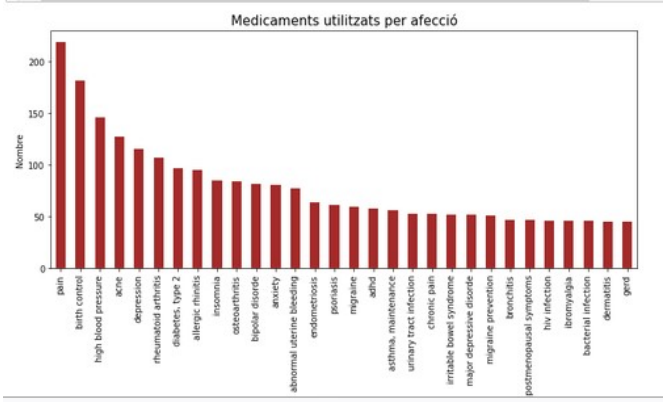
```
drugName
levonorgestrel                      4925
etonogestrel                        4420
ethinyl estradiol / norethindrone   3752
nexplanon                           2891
ethinyl estradiol / norgestimate    2788
ethinyl estradiol / levonorgestrel  2500
phentermine                         2085
sertraline                          1868
escitalopram                        1744
mirena                              1673
Name: review, dtype: int64
```

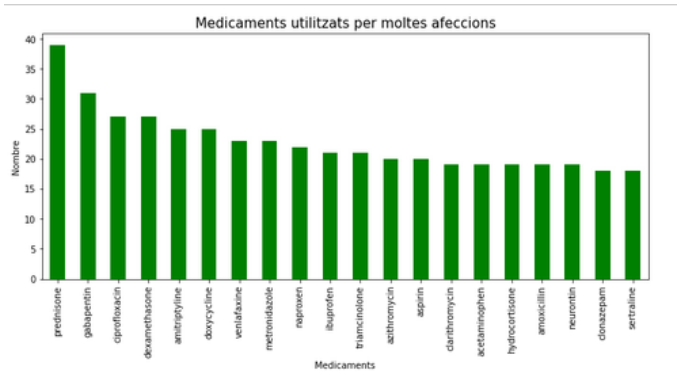


Relacions afeccions/fàrmacs

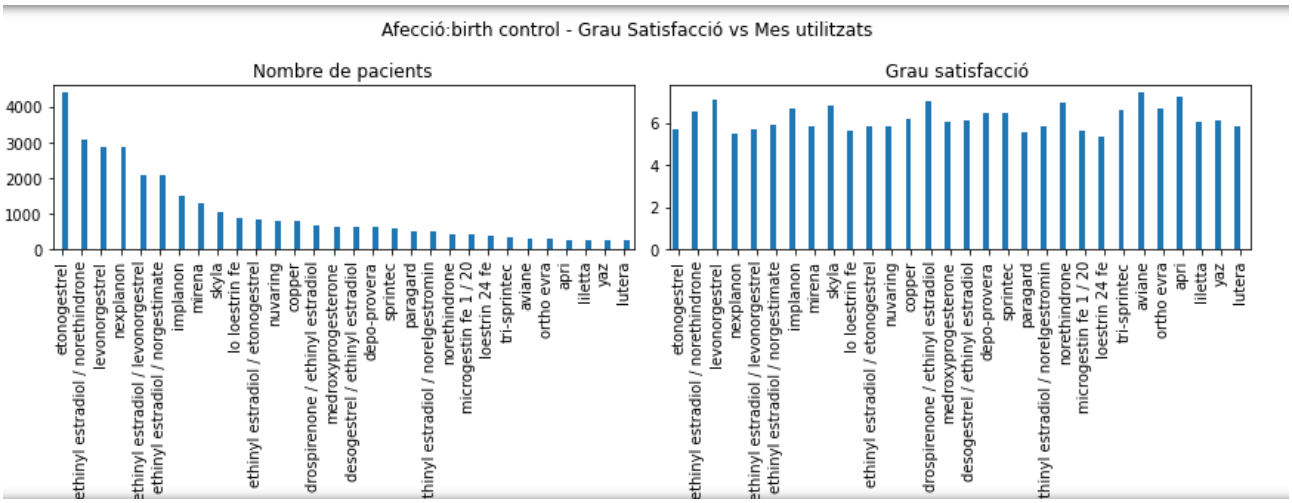
Tractarem de veure les veure les interrelacions entre els fàrmacs i les afeccions , per exemple

- Quants fàrmacs s'han utilitzats per tractar una mateixa cada afecció
- Els fàrmacs que mes s'utilitzen per diferents afeccions
- Veure com es la distribució de ressenyes per fàrmacs o afeccions. Veiem clarament que no es una distribució normal. En els dos casos la mediana es baixa i la mitja es alta. Es a dir poques afeccions tenen moltes ressenyes



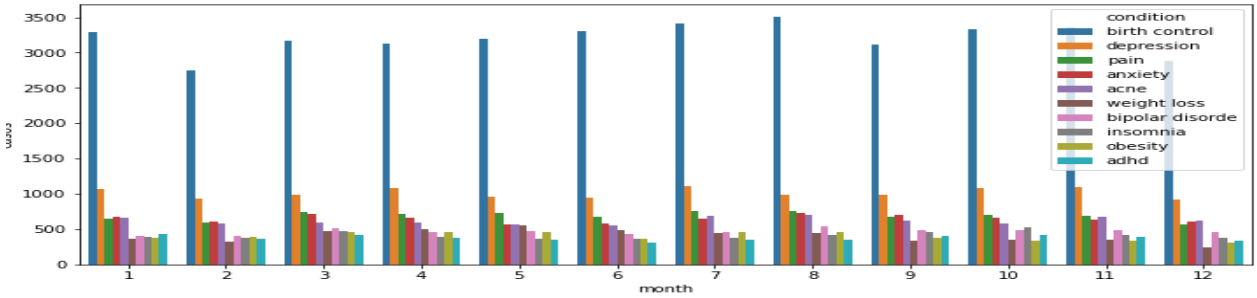


- Visualitzar en dues gràfiques de costat els fàrmacs mes utilitzats per afecció i dels mateixos fàrmacs i afeccions la ràting mitja



Freqència afeccions/mesos

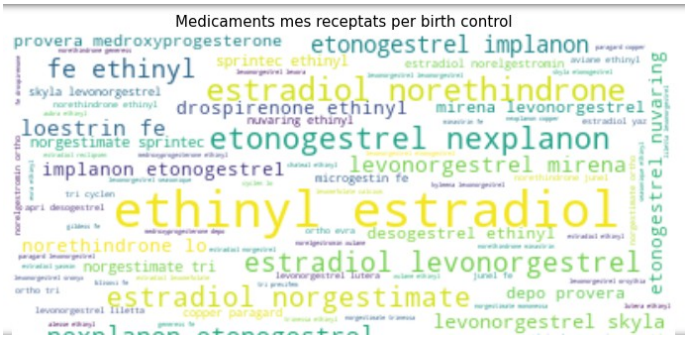
- Veure la evolució de caos mes a mes per les diferents afeccions, per exemple prenem les 10 afeccions amb mes caos per observar estacionalitats

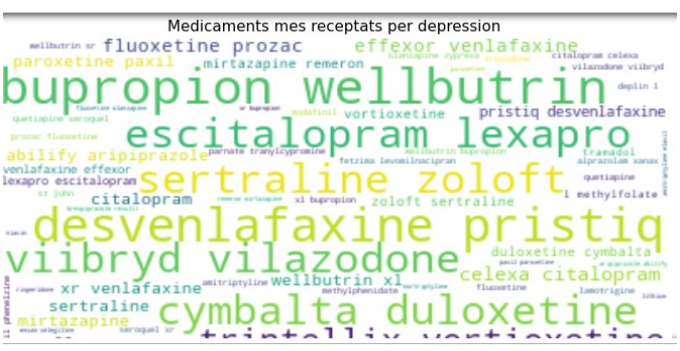


2- Analisi de Sentiments

Word Cloud Fàrmacs per afecció

Farem un funció que ens generi un ***wordCloud*** donada una afecció per poder visualitzar per cada afecció els fàrmacs mes habituals



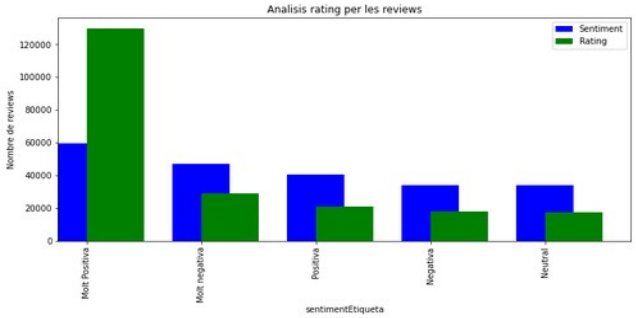


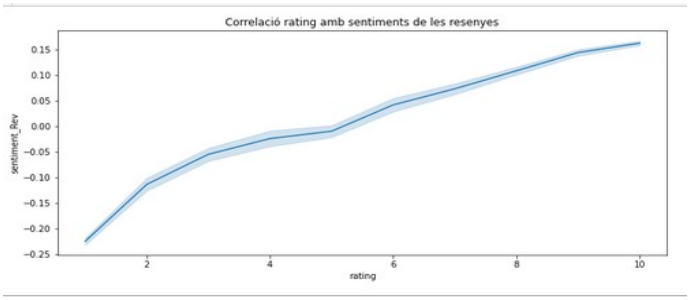
2- Analisi de Sentiments (ressenyes/review)

Una de les components principals d'aquest conjunt de dades es la ressenya (***review***)que expressen la retroacció dels usuaris sobre els fàrmacs utilitzats. Per tant farem una anàlisi de sentiment en aquest atribut.

Faré servir el analitzador ***SentimentIntensityAnalyzer*** i per el ***stemming*** el ***snowballstemmer.stemmer.*** , per la preparació i les anàlisis de sentiment de la variable ressenya. Una vegada processada cada ressenya la avaluarem amb el el ***'polarity_scores'*** i es quedarem la variable 'compund' per tal de tenir el resultat anàlisis de sentiment de cada ressenya. Finalment afegirem al conjunt de dades tant la ressenya neta com el ***compound*** en una nova columna anomenada del dataframe

Veurem la correlació del ràting amb el sentiment calculat i també com es comporten les dues variables en un diagrama amb una agrupació per trams. Veiem que la ràting tendeix a ser mes Molt Positiu mentre que el sentiment està mes balancejat però segueixen la mateixa tendència





3- Establir un model de classificació per predir afeccions(condition) a partir de les ressenyes(reviews)

L'objectiu es identificar el millor model de classificació per predir una afecció a partir de les ressenyes. Les passes a seguir son:

- Mantindrem el conjunt d'entrenament i de test que originalment estava en kaggle.
- Seleccionarem les afeccions que tinguin 90 o mes ressenyes al conjunt d'entrenament, es quedarem només amb les 50 afeccions més habituals (necessitem reduir el conjunt de entrenament per limitacions de processament la màquina)
- Vectoritzarem ell conjunt d'entrenament , ja que es tracta de una variable tipus text.La variable ***review*** ja la tenim preprocessada al dataframe (la hem inclòs en el punt anterior).
- Seleccionen els models ***MultinomialNB(), RandomForestClassifier , KneighborsClassifier,SGDClassifier*** i els evaluem amb un mètode de validació creuada (***cross_val_score)***
- Seleccionem el model amb una ***accuracy*** millor *(**scores.mean())***. En aquest cas he seleccionat el **SGDClassifier.** Una vegada seleccionat l'entrenarem amb tot el conjunt de entrenament i finalment aplicarem el model al conjunt de proves
- Observo que ***l'accuracy*** del entrenament amb ***cross validation*** es 0.68 bastant inferior a la obtinguda en el conjunt de test 0.84 i a mes a mes el F1 Macro F1 Micro son molt elevats la qual cosa fa pensar que el model està sobreestimat . Podria ser degut a la reducció del conjunts o be un cert biaix en les dades que s'hauria d'estudiar mes en profunditat.

```
Exactitud de les dades de prova:  0.8455249466541993
Resultat de Macro F-1 a les dades de prova:  0.8266113591995884
Resultat de Micro F-1 a les dades de prova:  0.8455249466541993
```

Nota:

No he testejat el **RandomForestClassifier j**a que malgrat he reduït els conjunts de entrenament i prova la màquina no ha pogut processar el random Forest