Text Mining

Anna Mattiroli - 826381
Beatrice Barzaghi - 831829
Guglielmo Muoio - 826029

Domande di ricerca

TASK 1

È possibile classificare la gravità degli effetti collaterali di un farmaco basandosi sulle opinioni espresse dagli utenti online?

TASK 2

È possibile dividere in gruppi i farmaci in base agli effetti collaterali che gli utenti riportano nelle loro recensioni?

Workflow

Acquisizione dati

Download del dataset Druglib da UCI ML

Analisi esplorativa

Ricodifica delle variabili di interesse

Data manipulation

Pulizia
Text Representation

Text Classification

SVC
KNN
Random Forest
MLP

Text Clustering

k-means
Gerarchico
DBSCAN

Analisi Esplorativa

Variabili di interesse

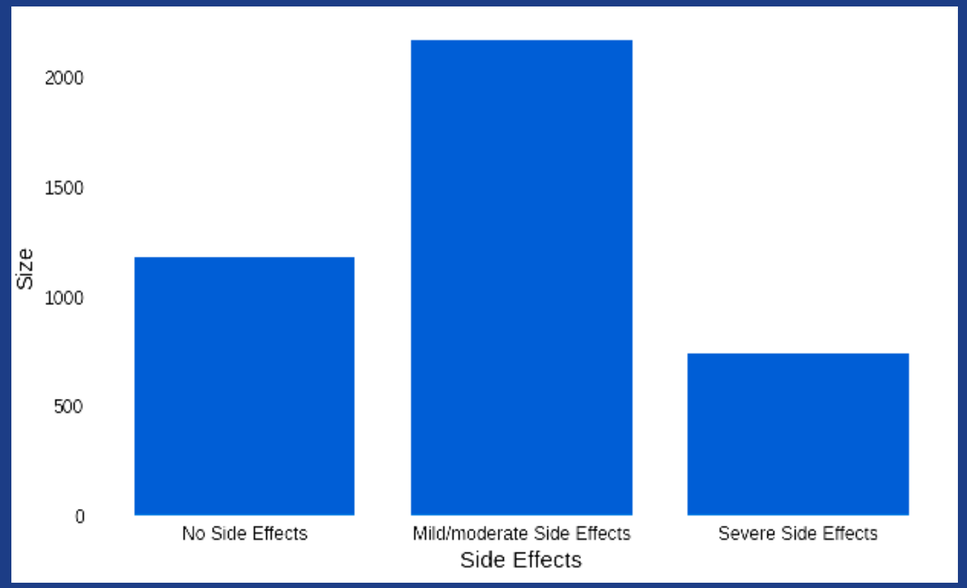
urlDrugName	nome del farmaco (540 valori unici)
condition	condizione clinica per cui il farmaco è stato assunto
benefitsReview	review sui benefici ottenuti dal paziente
sideEffectsReview	review sugli effetti collaterali avuti dal paziente
commentsReview	commento generale del paziente
rating	valutazione da 1 a 10 del paziente
sideEffects	valutazione in una scala di 5 classi degli effetti collaterali
effectiveness	valutazione in una scala di 5 classi dell'efficacia

Ricodifica variabile SideEffects

- No Side Effects → No Side Effects
- Mild Side Effects → Mild/Moderate Side Effects
- Moderate Side Effects → Mild/Moderate Side Effects
- Severe Side Effects → Severe Side Effects
- Extremely Severe Side Effects → Severe Side Effects

NA values: 11

Osservazioni duplicate: 50



Processing dei dati

PREPROCESSING

- Lower Case
- Rimozione di spazi in eccesso, numeri e punteggiatura
- Rimozione parole contratte
- Rimozione URLs
- Tokenization
- Rimozione Stopword
- Stemmatization
- Lemmatization

REPRESENTATION

- Binary
- Raw Frequency
- Tf-Idf

Per classificazione

DIMENSIONALITY REDUCTION

- Feature selection: test Chi-quadro
- Feature synthetization: PCA

Text Classification

Viene sviluppata una procedura di Cross-Validation con 5 folds

Vengono testati i seguenti modelli:

- Support Vector Classifier
- Random Forest
- K-Nearest Neighbour
- Multi-Layer Perceptron

Si seleziona il modello **Binary lemmed**, per il quale vengono effettuate le operazioni di dimensionality reduction:
7666 features → 1000 features, che sono in grado si spiegare più del 90% della varianza

	SVM	RF	KNN	MLP
Bin stem	0.69	0.73	0.59	0.71
Bin lemm	0.70	0.73	0.59	0.71
RFreq stem	0.68	0.73	0.61	0.70
RFreq lemm	0.69	0.73	0.61	0.70
Tf-Idf stem	0.72	0.73	0.56	0.69
Tf-Idf lemm	0.72	0.73	0.56	0.68

Tabella 1. Valori di accuracy sul cross-validation

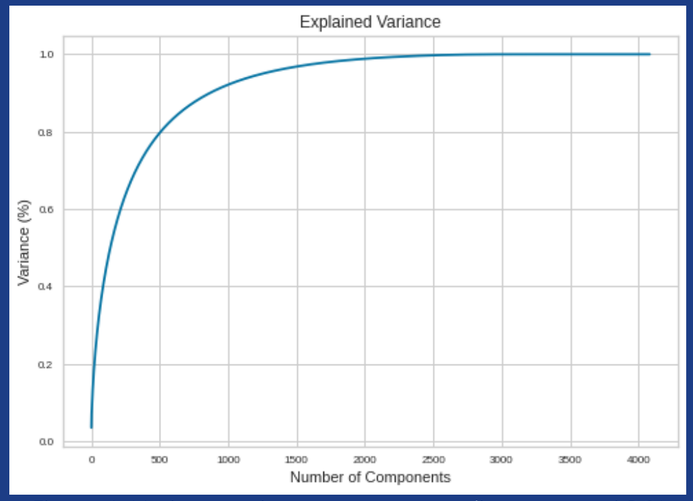


Figura 1. Varianza spiegata per numero di componenti PCA

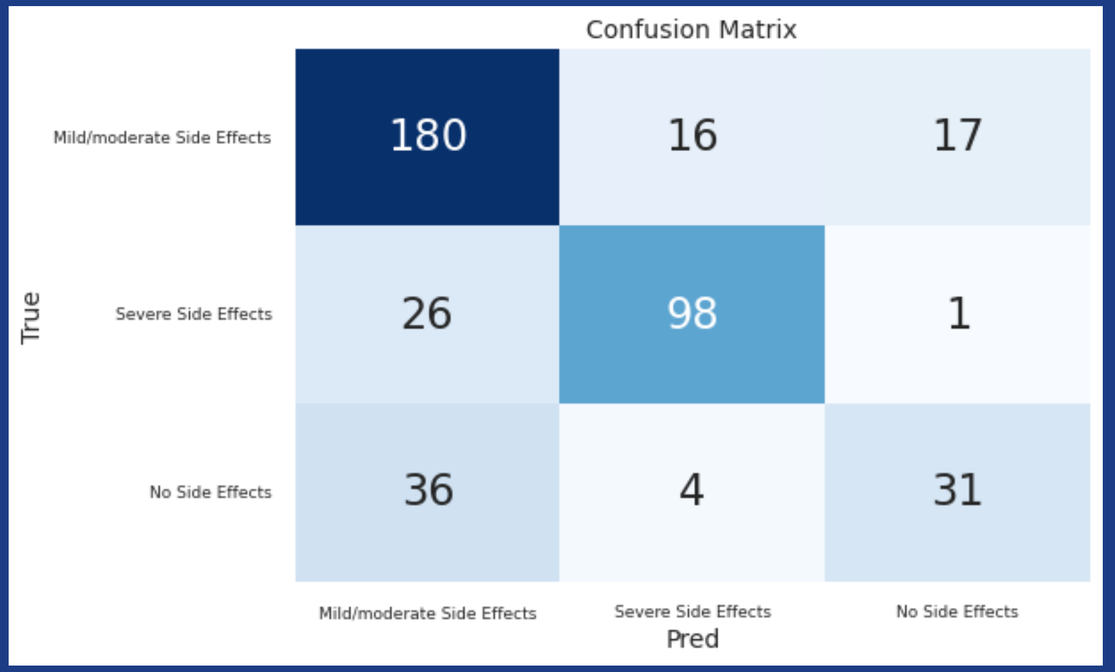
Text Classification

Binary lemmed con 1000 features

	Accuracy CV	Accuracy Test	Execution Time
SVC	0.71	0.76	36 sec
RF	0.68	0.67	44 sec
MLP	0.71	0.74	84 sec

Tabella 2. Metriche per il confronto dei modelli

Support Vector Classifier su Binary lemmed



Il modello performa meglio nella classicazione delle recensioni appartenenti alle classi mild/moderate side effects e severe side effects, rispetto a quelle che non riportano alcun effetto collaterale.

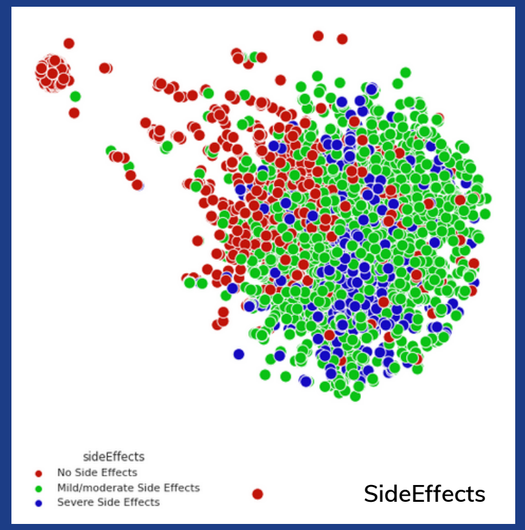
Text Clustering

Le tecniche di clustering prese in considerazione sono:

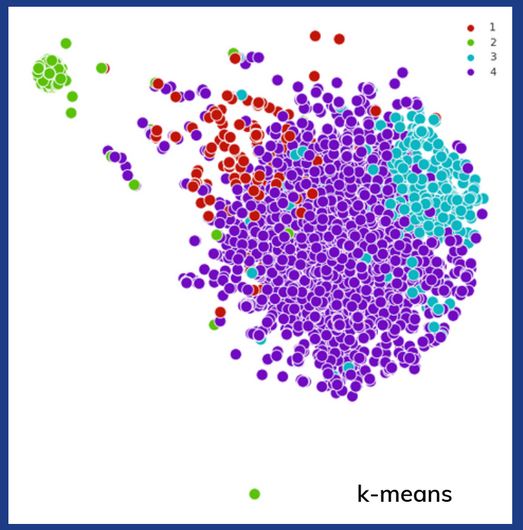
- K-means
- Gerarchico agglomerativo
- DBSCAN

Per lo sviluppo di questo task si è deciso di sfruttare la rappresentazione testuale Tf-Idf stemmed.

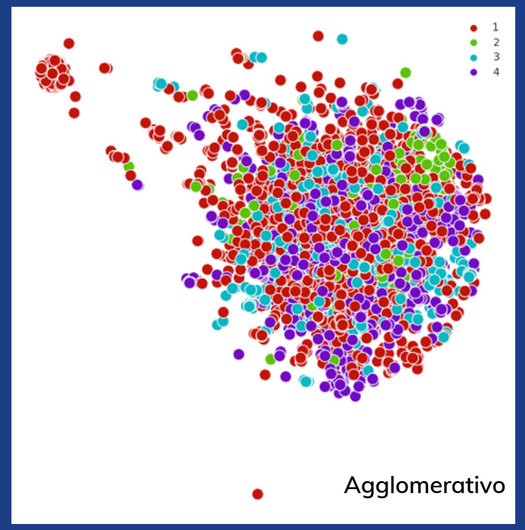
Una volta selezionato il numero di clusters ideale per ciascun algoritmo, si è fatta una prima visualizzazione grafica per mezzo di t-SNE, che è stata messa a confronto con le classi della variabile sideEffects.



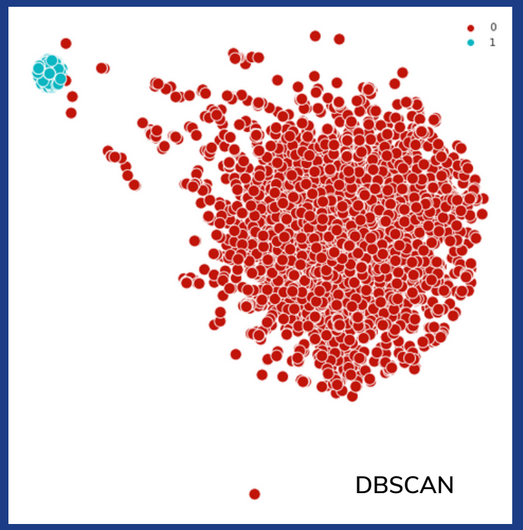
SideEffects



k-means



Agglomerativo



DBSCAN

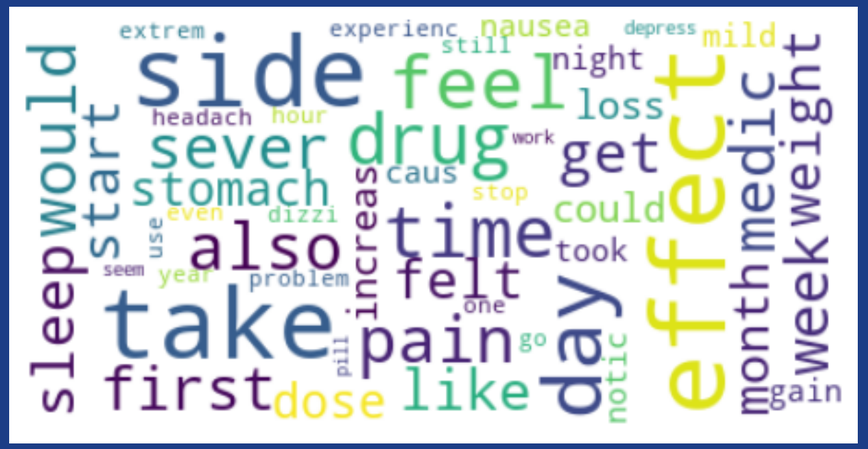
Text Clustering

Successivamente si è deciso di visualizzare le parole più utilizzate per ciascun cluster e la tecnica che ha dato i risultati più apprezzabili è stata k-means

Cluster 1

Cluster 2

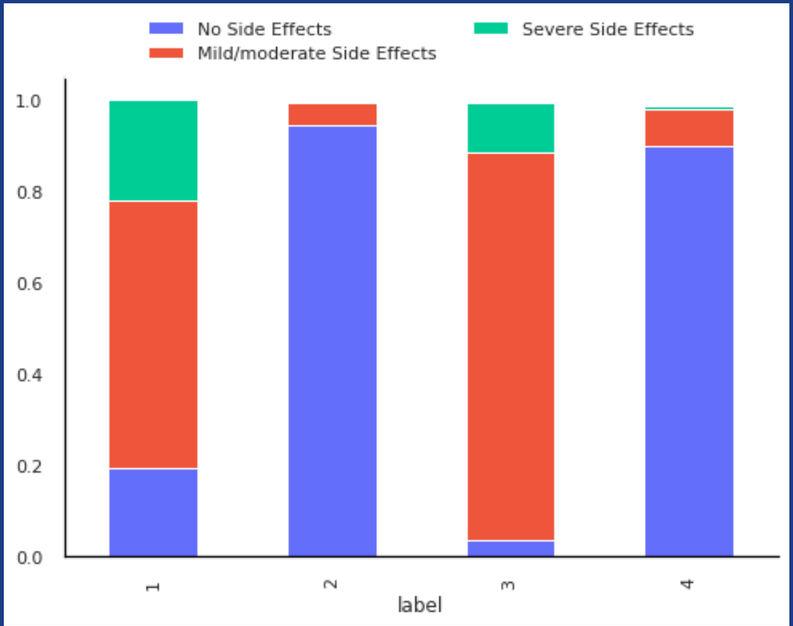
Cluster 3



Cluster 4

Text Clustering

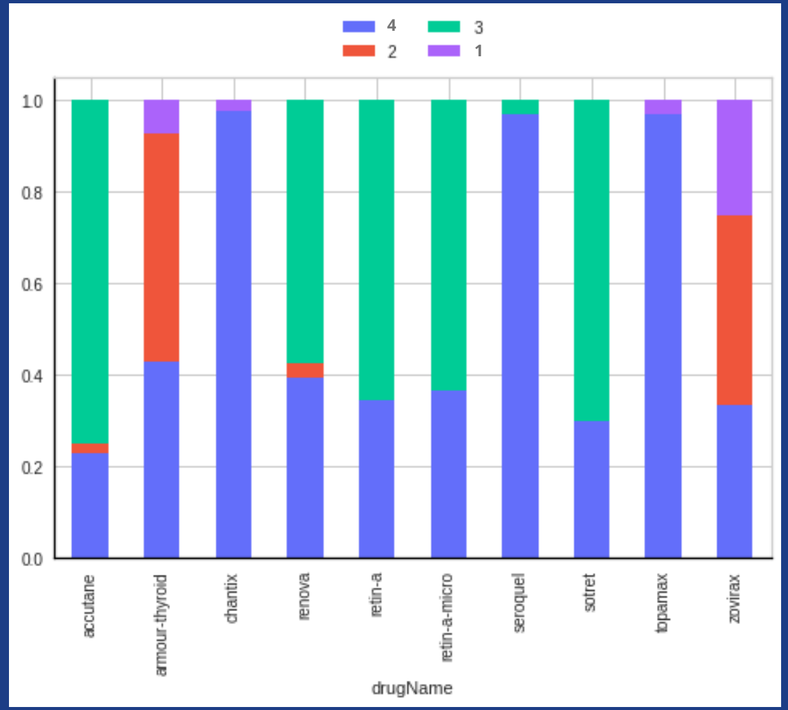
Cluster vs SideEffects



Normalized Mutual Information: 0.2

Ipotesi: correlazione tra clusters e gravità degli effetti collaterali
Risultati: le osservazioni non confermano pienamente l'ipotesi

Cluster vs Top-drugs



Ipotesi: correlazione tra clusters e medicinali più utilizzati
Risultati: relazione tra farmaci per il trattamento di acne e cluster 3

Conclusioni

TASK 1

È possibile classificare la gravità degli effetti collaterali di un farmaco basandosi sulle opinioni espresse dagli utenti online?

I risultati sono soddisfacenti ma è ragionevole pensare che prevedere una variabile multiclasse possa essere complicato in quanto la classe viene assegnata sulla base di una scala soggettiva

TASK 2

È possibile dividere in gruppi i farmaci in base agli effetti collaterali che gli utenti riportano nelle loro recensioni?

Il metodo delle k-means ha permesso di identificare gruppi discretamente significativi in un'ottica di interpretabilità rispetto al tipo di effetti collaterali

Grazie per l'attenzione