UNIVERSITÀ DI PISA

Data Mining a.a 2021-22

Analisi Dataset Glasgow Norms

A cura di
Alice Isola, Eleonora Rossi, Gianmario Ercoli, Lorenzo Casarosa

Indice

1 Introduzione

Il dataset, oggetto delle analisi riportate in questo report, è il **Glasgow Norms**[1], composto da un insieme di valutazioni normative svolte su più di quattromila parole inglesi, valutate secondo *nove dimensioni psicolinguistiche* a cui si aggiungono altri tre attributi molto importanti: la lunghezza in termini di caratteri, la polisemia e la frequenza. Alcune di queste variabili verranno illustrate nello specifico nella sezione dedicata alla *Data Understanding*.
L'obiettivo di questo progetto è di analizzare queste dimensioni linguistiche, cercando di cogliere tutti quegli aspetti che possono condurci ad osservazioni interessanti sul dataset.

2 Data Understanding

In questa sezione si tratterà della **Data Understanding**, in particolare analizzeremo le *features* che presentano una maggior correlazione con la variabile target *polysemy*, anche mediante l'ausilio di alcuni grafici.

2.1 Il dataset

Il dataset **Glasgow Norms** è composto da 4682 parole per ciascuna delle quali sono state analizzati 12 attributi. Tali attributi sono sintetizzati nella Tabella 1.

Variabile	Tipologia	Descrizione	Valori
length	Numerica	Lunghezza di una parola	Numero di caratteri
arousal	Numerica	Attivazione interna della parola (se suscita nel soggetto eccitazione o calma)	Scala da 0 (calma) a 9 (eccitazione)
valence	Numerica	Valenza positiva o negativa della parola	Scala da 0 (valenza negativa) a 9 (valenza positiva)
dominance	Numerica	Grado di controllo della parola (dominante, controllato)	Scala da 0 (dominante) a 9 (controllato)
concreteness	Numerica	Concretezza di una parola	Scala da 0 (astratto) a 7 (concreto)
imageability	Numerica	Capacità di una parola di generare un'immagine mentale nel soggetto	Scala da 0 (capacità debole di generare immagini mentali) a 7 (capacità forte di generare immagini mentali)
familiarity	Numerica	Conoscenza di una parola	Scala da 0 (non familiare) a 7 (familiare)
aoa (age of acquisition)	Numerica	Età in cui il soggetto ha acquisito quella parola	Scala da 0 (acquisito in età precoce) a 7 (acquisito in età avanzata)
semsize (semantic size)	Numerica	Grandezza percepita di una parola sia concreta che astratta	Scala da 0 (molto piccola) a 7 (molto grande)
gender (association)	Numerica	La parola è collegata alla sfera maschile o femminile	Scala da 0 (sfera femminile) a 7 (sfera maschile)
polysemy	Categorica	Variabile binaria che stabilisce se una parola assume più significati	0 (non polisemico) 1 (polisemico)
web_corpus_freq	Numerica	Frequenza di una parola all'interno del Google Newspaper Corpus	Valore della frequenza

Tabella 1: Variabili presenti all'interno del Glasgow Norms

Per quanto concerne le dimensioni psicolinguistiche, il dataset analizzato ha natura prevalentemente soggettiva, in quanto i valori dei loro attributi sono stati determinati mediante questionari. Le uniche due variabili oggettive sono la polisemia e la lunghezza.

2.2 Distribuzione delle features in relazione alla variabile target

Polysemy: questa variabile indica se una parola può assumere una pluralità di significati o no, assegnandole il valore 1 e 0 nei rispettivi casi. Rappresentando la distribuzione delle parole rispetto alla polisemia è stato

[1]Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). *The Glasgow Norms: Ratings of 5,500words on nine scales.Behavior Research Methods*,51(3), 1258–1270.

possibile notare che c'è una netta minoranza di parole polisemiche (solo 379 su 4682 totali). (Figura 1). Questo risultato era abbastanza prevedibile in quanto, in generale, all'interno di una lingua non sono presenti molte parole polisemiche.

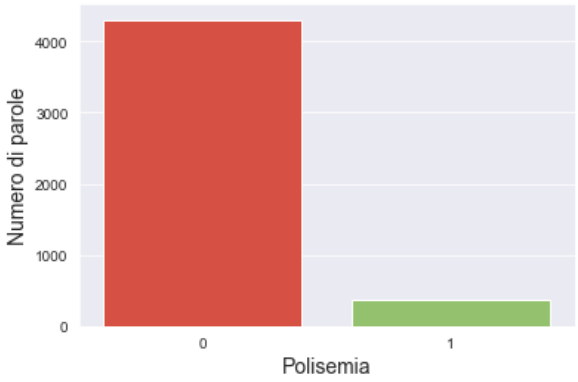
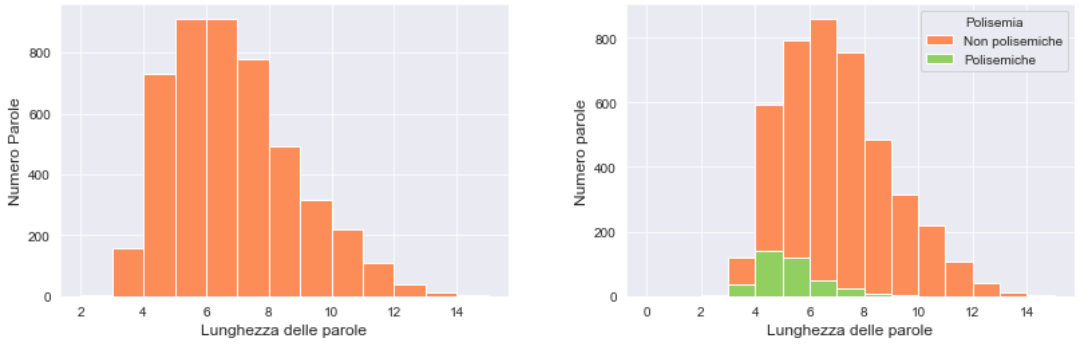


Figura 1: Distribuzione delle parole polisemiche all'interno del dataset

Length: questa variabile rappresenta la lunghezza di ciascuna parola in termini di caratteri.
Come si può osservare dall'istogramma rappresentato nella figura 2a, all'interno del dataset la maggior parte delle parole ha una lunghezza compresa tra i 4 e gli 8 caratteri.

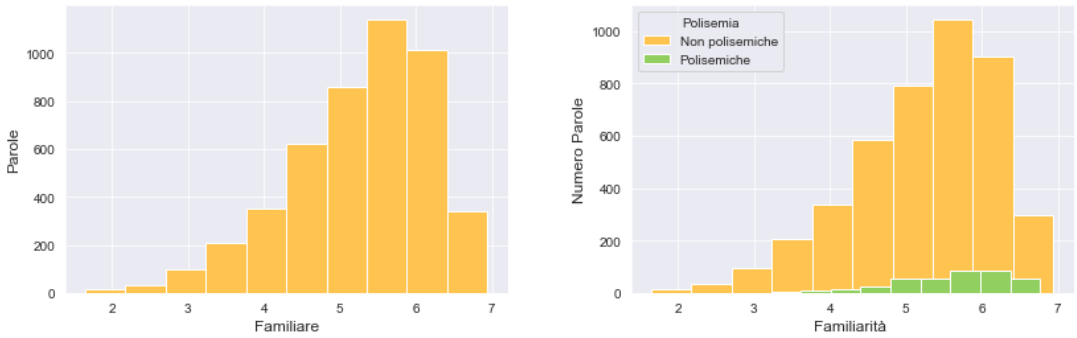


(a) **(b)**

Figura 2: Distribuzione della lunghezza delle parole (2a) e della polisemia rispetto alla lunghezza (2b)

Abbiamo poi analizzato la distribuzione della polisemia in relazione a questo attributo al fine di stabilire quale fosse la lunghezza che le parole polisemiche tendono ad avere. Il grafico (Figura 2b) mostra un risultato abbastanza prevedibile: le parole più corte sono quelle più soggette ad essere polisemiche.

Familiarity: la *familiarity* è l'attributo che, in psicolinguistica, indica la facilità con la quale si percepisce una parola. In questo dataset tale variabile può assumere valori compresi tra 1 e 7, a indicare la non familiarità e la familiarità, rispettivamente.
L'istogramma presente in figura 3a mostra una distribuzione mono-modale piuttosto asimmetrica a rappresentazione del fatto che un maggior numero di parole all'interno del dataset viene percepito come molto familiare. Questo risultato è ulteriormente provato dal fatto che la mediana di tale variabile rispetto alla popolazione si attesta intorno a 5.4 con un grado di dispersione molto basso pari a 0.92 calcolato attraverso la deviazione standard.

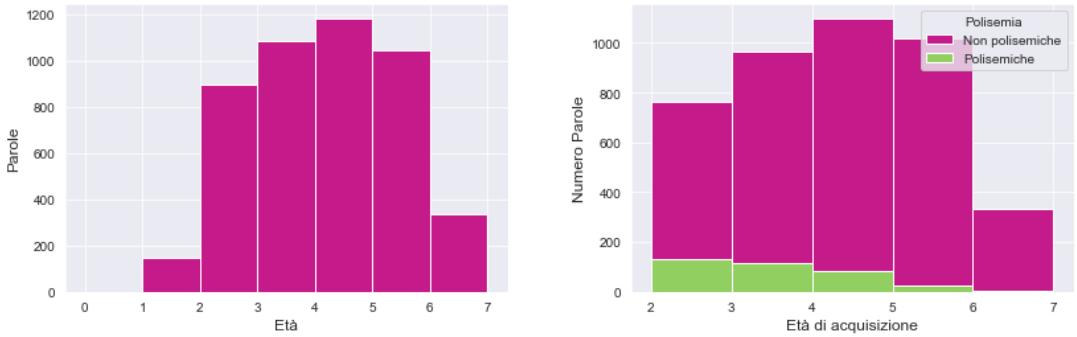


(a) **(b)**

Figura 3: Distribuzione della familiarità delle parole 3a e della polisemia rispetto alla familiarità (3b)

Osservando, tramite grafico (Figura 3b), il comportamento che la variabile target assume in relazione a *familiarity* è stato possibile concludere che, come da aspettative, la maggior parte delle parole polisemiche tende ad essere percepita come più familiare, attestandosi mediamente all'interno del range 5-7. Questo aspetto potrebbe essere spiegato dal fatto che l'utilizzo maggiore delle parole nella propria quotidianità può portare ad uno sviluppo di diversi significati di queste.

aoa - Age Of Acquisition: in psicolinguistica la *Age of Acquisition* rappresenta l'età in cui, solitamente, una parola viene appresa e che può influenzare in maniera importante il processo di apprendimento del soggetto.
Nello studio per la compilazione di questo dataset la *Age of Acquisition* assume valori in un intervallo 1-7 e come mostra l'istogramma nella figura 4a, l'andamento della variabile risulta essere mono-modale, con il numero di parole apprese che tende a crescere all'aumentare dell'età fino ai 5 anni, superati i quali la tendenza si inverte e inizia a decrescere.

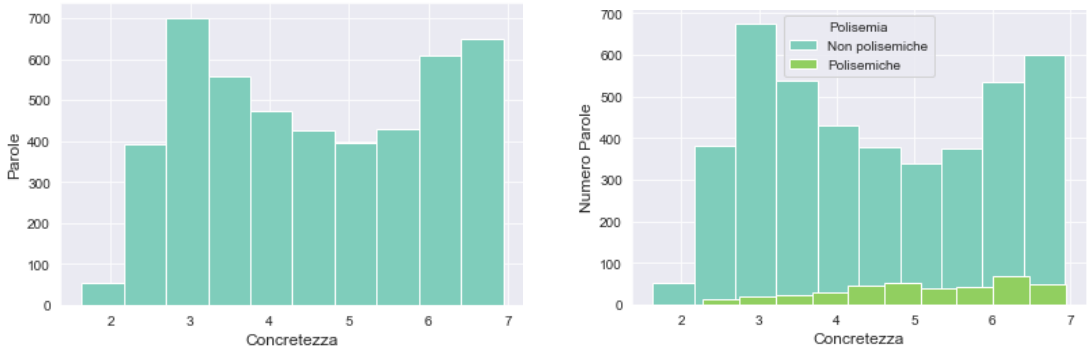


(a) **(b)**

Figura 4: Distribuzione dell'età di acquisizione (4a) e della polisemia rispetto all'età di acquisizione (4b)

Il passo successivo è stato quello di confrontare la polisemia con l'età di acquisizione e, come viene mostrato nell'istogramma nella figura 4b, il maggior numero di parole polisemiche viene appreso non appena si passa dall'età neonatale a quella infantile, cioè quando si impara a parlare e a capire il significato delle parole.

Concreteness: La *Concreteness* indica il grado di concretezza della parola analizzata.
All'interno del Glasgow Norms il grado di concretezza di una parola viene misurata in un range di valori che va da 1 a 7, il quale indica rispettivamente che una parola è fortemente astratta o assolutamente concreta. Come rappresentato in figura 5a, la distribuzione di questa variabile è bi-modale: ciò indica un comportamento della variabile abbastanza prevedibile in quanto abbiamo un numero di parole pressoché equivalente tra quelle abbastanza concrete e quelle più astratte. Un aspetto importante da segnalare è il bassissimo numero di parole totalmente astratte (numero minore di 100).

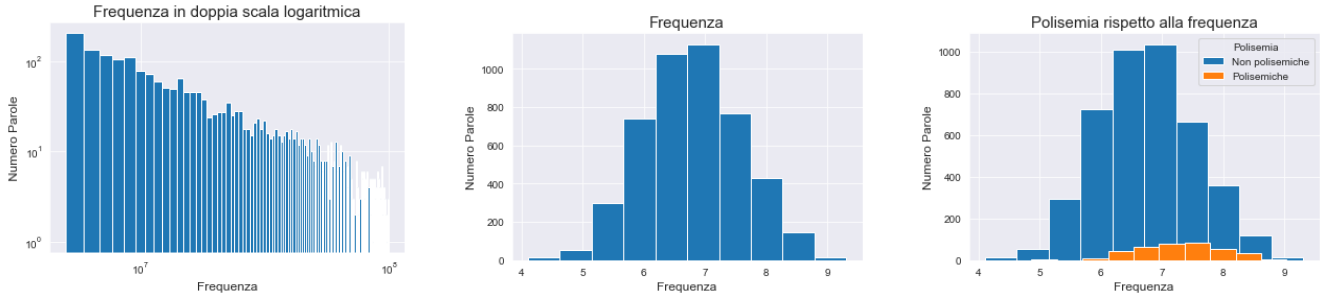


(a) (b)

Figura 5: Distribuzione della concretezza (5a) e della polisemia rispetto alla concretezza (5b)

Possiamo infatti dedurre, contro le nostre sensazioni e previsioni, che le parole che presentano un maggior grado di concretezza hanno, di tendenza, una maggior propensione ad essere polisemiche.

Web_corpus_freq: L'attributo *web_corpus_freq* indica la frequenza delle parole all'interno del Corpus Google Newspapers. L'utilizzo della doppia scala logaritmica ha permesso una migliore visualizzazione della variabile, in quanto i suoi valori oscillano in un intervallo molto grande. Ciò è evidente anche a partire dalla deviazione standard pari a 84901444; la variabile infatti presenta una distribuzione *long tail* estremamente asimmetrica (*skewness* di 9.5). A partire dalla distribuzione in figura 6a è possibile desumere che all'interno del dataset c'è una maggiore concentrazione di parole con frequenza relativamente media, individuabile all'interno della testa, e una minoranza di parole distribuite in un range più ampio a indicare una frequenza maggiore. Abbiamo poi deciso di operare una trasformazione logaritmica di tale variabile ottenendo così una distribuzione normale (Figura 6b) con una media pari a 6.8 e una deviazione standard pari a 0.8.



(a) (b) (c)

Figura 6: Distribuzione della frequenza in scala logaritmica (Figure 6a), della trasformazione in distribuzione normale (Figura 6b) e della polisemia rispetto alla frequenza delle parole (Figura 6c)

Abbiamo poi confrontato l'attributo *polysemy* con la variabile in oggetto. É possibile apprezzare tale confronto grazie alla figura 6c in cui è possibile notare come le parole polisemiche, a differenza di quelle non polisemiche, tendano ad avere una maggiore frequenza attestandosi all'interno del range 6-8.5. Questo potrebbe essere spiegato dal fatto che in generale le parole polisemiche tendono ad occorrere in più contesti, spesso anche diversi, contribuendo così a favorire una maggior frequenza d'uso, ma allo stesso tempo la maggior frequenza di alcuni tipi di parole può essere il fattore d'innesco per lo sviluppo della polisemia stessa.

2.3 Analisi statistiche

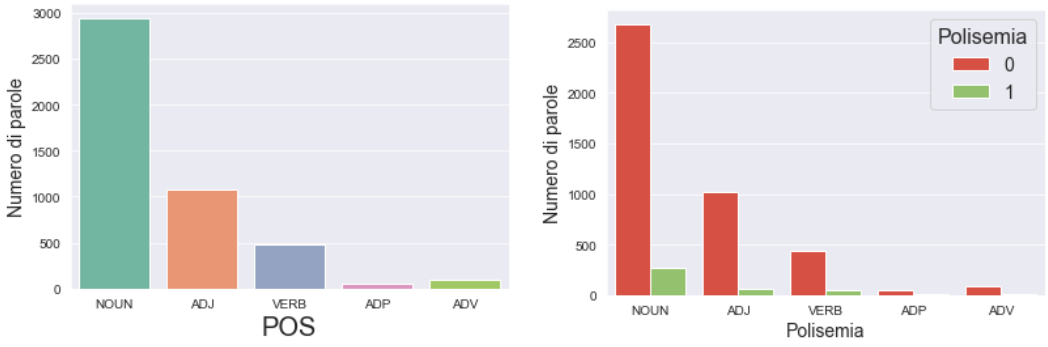
Per approfondire ulteriormente la comprensione dei dati, sono state condotte alcune **Analisi Statistiche** calcolando la moda, media e mediana, varianza e deviazione standard delle variabili e i quantili e percentili delle variabili che compongono il dataset (Tabella 2).

	media	std	min	25%	50%	75%	max
length	6.34	2	2	5	6	8	16
arousal	4.67	1.09	2.05	3.84	4.57	5.41	8.17
valence	5.08	1.59	1.03	4.11	5.29	6.08	8.64
dominance	5.04	0.93	1.94	4.52	5.12	5.60	8.37
concreteness	4.56	1.43	1.63	3.24	4.47	5.97	6.93
imageability	4.72	1.36	1.73	3.51	4.67	6.03	6.94
familiarity	5.27	0.92	1.64	4.70	5.43	5.96	6.93
aoa	4.14	1.25	1.21	3.11	4.17	5.15	6.97
semsize	4.13	1.02	1.37	3.43	4.18	4.88	6.91
gender	4.09	0.91	1	3.60	4.12	4.65	6.97
polisemy	0.08	0.027	0	0	0	0	1
frequency_log	6.78	0.80	4.10	6.22	6.75	7.35	9.30

Tabella 2: Statistica descrittiva: caratteristiche delle variabili

2.4 Analisi delle Parts Of Speech

Essendo il Glasgow Norms un dataset che analizza valutazioni di natura linguistica, abbiamo scelto, infine, di analizzate anche le differenti Parti del Discorso o **Parts Of Speech (POS)** che lo compongono. Per ogni parola sono state quindi estratte le rispettive POS utilizzando il metodo *pos_tag* della libreria NLTK al fine di inserirle successivamente in una nuova colonna del dataframe denominata 'pos'.
Dall'estrazione delle POS, è emerso che quelle più frequenti in assoluto all'interno del dataset sono i *nomi* (NOUN, 2945) seguiti dagli *aggettivi* (ADJ, 1082) , dai *verbi* (VERB, 476) e dagli *avverbi* (ADV,98) come mostrato nella figura 7a.
Questo risultato è abbastanza prevedibile in quanto, in genere, le parole lessicali sono presenti in numero maggiore rispetto alle parole funzionali.



(a) **(b)**

Figura 7: Distribuzione delle POS (7a) e della polisemia rispetto alle POS (7b)

Successivamente, si è scelto di analizzare le POS più frequenti in relazione alla *polisemia* per vedere quante di esse fossero polisemiche. Dalla figura 7b emerge come la maggior parte delle parole polisemiche siano nomi. Tale aspetto potrebbe essere deducibile dalla dominante presenza di nomi nel dataset.

3 Data Preparation

Al fine di valutare e migliorare la qualità dei dati presenti nel dataset, abbiamo effettuato diverse analisi preliminari come la sostituzione di valori mancanti, la ricerca di eventuali errori ed infine l'analisi e gestione degli outliers.

3.1 Valori mancanti

All'interno del dataset sono stati rilevati *14 valori mancanti* nell'attributo *web_corpus_freq*. Vista l'alta correlazione tra frequenza e familiarità si è deciso di utilizzare quest'ultima variabile come mezzo per sostituire

i valori mancanti. È stato necessario *in primis* discretizzare tale variabile, individuando quattro categorie: *very low* per valori all'interno del range 0-1.75, *low* per valori da 1.76-3.5, *medium* per valori da 3.51-5.25, *high* per valori da 5.26-7.
Attraverso il metodo *groupby* abbiamo sostituito ogni valore mancante della *web_corpus_freq* con il valore della mediana dell'intervallo di valori (della variabile discretizzata *familiarity*) a cui quel valore mancante appartiene. Più specificatamente se un valore mancante della *web_corpus_freq* ha un valore della variabile discretizzata *familiarity* che appartiene al primo range di valori (0-1.75) tale valore verrà sostituito con la rispettiva mediana. É stato poi possibile verificare la correttezza di questa operazione confrontando la distribuzione della variabile *web_corpus_freq* prima che venisse attuata la modifica e post modifica.
Abbiamo ulteriormente approfondito la questione andando ad analizzare i dati statistici (media, mediana e deviazione standard) delle due distribuzioni create da cui non è emersa alcuna differenza sostanziale. Ciò a ulteriore dimostrazione che la sostituzione dei valori mancanti ha avuto esito positivo.

3.2 Errori

Abbiamo successivamente ricercato eventuali errori tra i valori degli attributi. Data la forte soggettività delle *features*, relative alle nove dimensioni psicolinguistiche, l'analisi si è limitata alla verifica della correttezza dell'intervallo dei valori di tali variabili, in quanto questi devono necessariamente cadere entro una precisa scala (0-9 per *arousal*, *valence*, *dominance* e 0-7 per tutti gli altri).
Tale verifica ha poi riguardato il confronto tra i valori indicati nella variabile *length* e la lunghezza effettiva delle parole. Non sono emersi errori, per questo motivo non è stata attuata nessuna modifica.

3.3 Gestione degli Outliers

Dopo aver analizzato l'eventuale presenza di valori mancanti e di errori, siamo passati all'*analisi degli outliers*, ossia tutti quei valori distanti rispetto alle altre osservazioni disponibili. Attraverso la creazione di un boxplot per ciascuna variabile è stato possibile individuare chiaramente tutti gli outliers delle *features*. Dai boxplots riportati in figura 8 notiamo chiaramente che le variabili con una maggiore quantità di outliers sono: *length*, *arousal*, *dominance*, *gender*, *web_corpus_freq*.

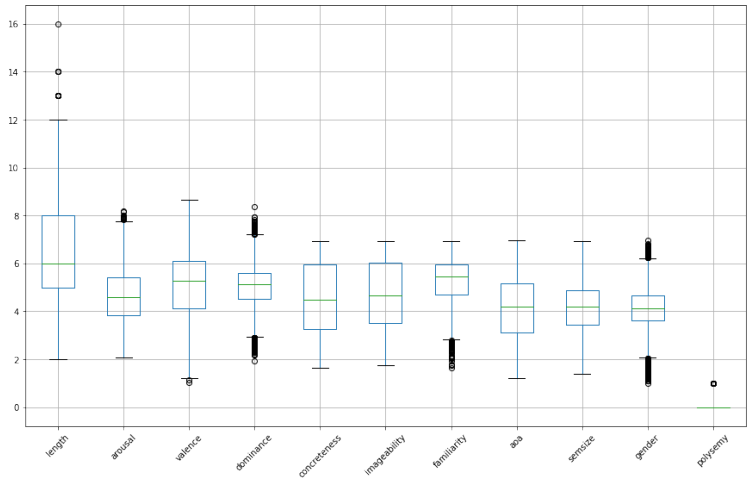
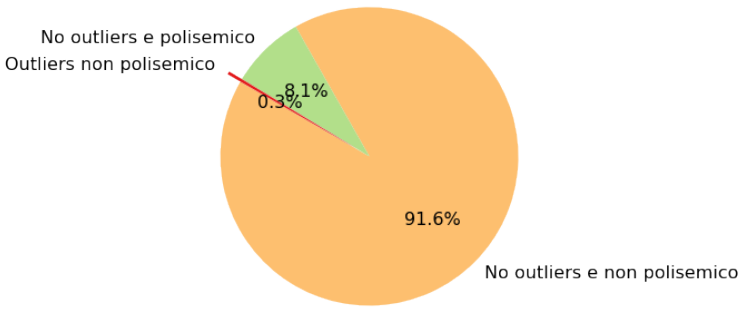


Figura 8: Boxplots contenenti gli outliers delle variabili del dataset

Tuttavia, in questo report ci limiteremo alla trattazione degli outliers della variabile **lunghezza** in quanto gli outliers delle altre variabili assumono un comportamento analogo. Data l'elevata soggettività che caratterizza i valori delle variabili presenti nel data set che stiamo analizzando non è stato possibile rimuovere gli outliers che sono emersi nella nostra analisi (perché non riconducibili ad errori di misurazione, raccolta dati, campione inadeguato



etc.).
Abbiamo quindi deciso di rappresentare all'interno di un grafico a torta (Figura 9) gli outliers relativi all'attributo Lunghezza poiché risulta essere una delle variabili maggiormente correlata con la nostra variabile target (polisemia). Tuttavia, l'analisi non ha mostrato niente di inaspettato, anzi: come si evince dal grafico a fianco (Figura 9) i pochi outliers presenti sono tutti non polisemici. Ciò conferma ulteriormente la correlazione negativa tra lunghezza e polisemia: infatti tutti gli outliers relativi a lunghezza (sedici) hanno un numero di caratteri elevato.

3.4 Analisi delle correlazioni

Per costruire la matrice di correlazione in figura 10 è stato utilizzato l'indice di correlazione di Spearman in quanto non è presente una correlazione lineare tra le variabili. Da essa è possibile notare che non esiste una significativa correlazione tra le variabili del dataset tranne per quanto riguarda *concreteness* e *imageability* (0.9). Considerando come valore soglia 0.9, è stata eliminata la variabile *imageability* poiché rispetto a *concreteness* presenta una minore correlazione con la variabile target *polysemy*; essa costituirebbe quindi una variabile ridondante.

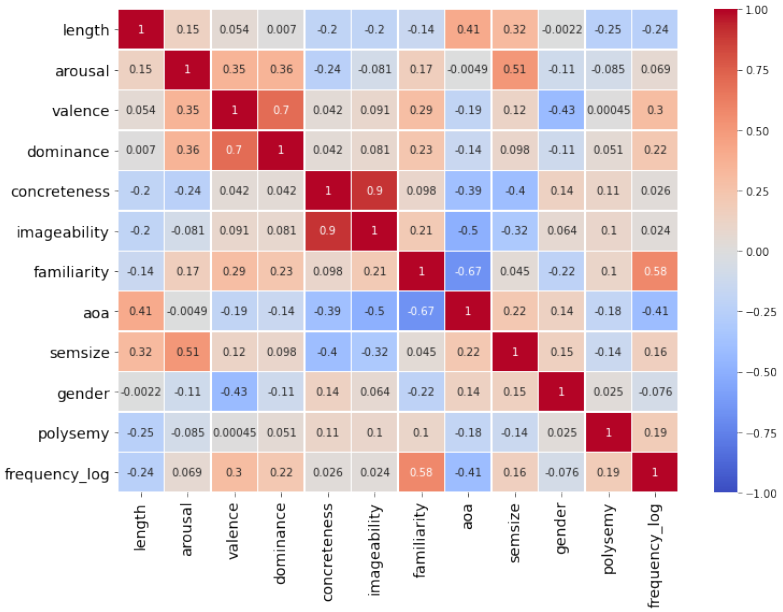


Figura 10: Matrice di correlazione di tutte le features del dataset

4 Clustering

Successivamente alla fase di Data Cleaning, sono stati implementati tre algoritmi di Clustering: **K-means**, **Clustering Gerarchico** e **DBSCAN**. Per essi verranno utilizzati i medesimi attributi al fine di favorire una migliore comparazione tra i risultati raggiunti. Per l'applicazione di ciascun algoritmo, è stato necessario effettuare la normalizzazione di tutte le variabili presenti nel dataset attraverso il *min_max_scaler*.

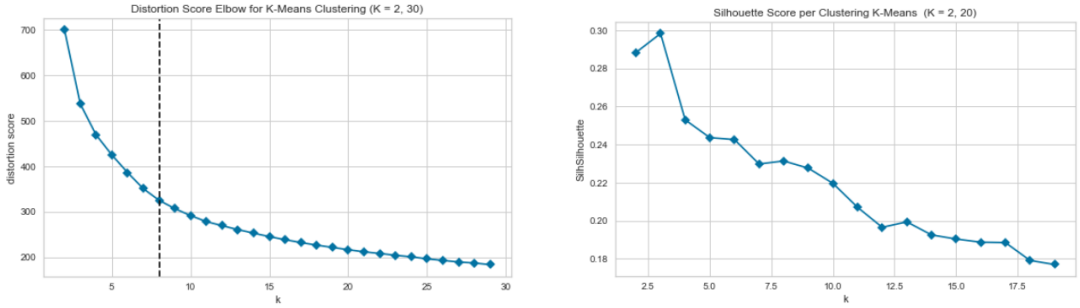
4.1 K-means

Il primo algoritmo di clustering implementato è il **K-means** per il quale si è reso necessario scegliere il numero di clusters e le features che garantissero il miglior funzionamento possibile di tale algoritmo. Studiando varie iterazioni dell'algoritmo K-means sull'intero dataset normalizzato (esclusa la variabile *imageability*) e avvalendoci di *Elbow method* e *Silhouette method* è stato possibile individuare tre come valore ottimale di k (numero di clusters).

4.1.1 Identificazione del miglior valore di K

Rappresentando graficamente (Figura 11a) i valori di *SSE* ottenuti iterando 30 volte l'algoritmo è stato individuato 8 come miglior valore di k. Approfondendo l'analisi, abbiamo deciso di ricorrere ad un'ulteriore

metrica di valutazione per misurare la bontà del clustering: *il coefficiente di Silhouette*. Poiché valori bassi della Silhouette denotano un'alta dissimilarità tra gli elementi che compongono un singolo cluster è sempre preferibile optare per una configurazione dell'algoritmo che garantisca un alto coefficiente di Silhouette. Rappresentando graficamente il valore assunto della Silhouette ad ogni iterazione (Figura 11b), si è individuato 3 come il miglior valore di k.



(a) **(b)**

Figura 11: Elbow (11a) e Silhouette (11b) per l'identificazione del miglior K

4.1.2 Features selection

Di seguito il *parallel coordinates* dei 3 clusters identificati. Dal grafico in figura 12 emerge che i clusters rappresentati presentano per alcune variabili centroidi tra loro molto aderenti o addirittura sovrapposti.

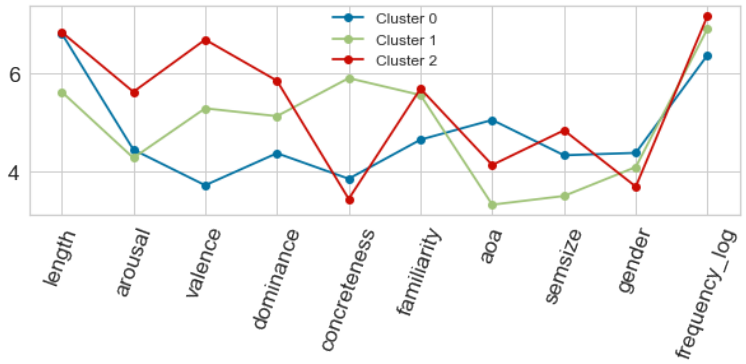


Figura 12: Coordinates Parallel dei centroidi di tutte le variabili del dataset

Abbiamo quindi eliminato gradualmente le seguenti variabili: *lenght*, *arousal*, *familiarity*, *frequency_log*, *gender*. Con questa operazione, sono stati individuati di volta in volta i corrispondenti miglior valori di K e della Silhouette. Ciò ha permesso di ottenere un buon bilanciamento tra un più alto valore di Silhouette e un numero adeguato di *features* (Tabella 3.) In quest'ultima configurazione ottimale, il valore ideale che k assume è ancora una volta 3.

Attributi eliminati	**k ideale**	**SSE**	**Silhouette**
{}	3	1063	0.21
{length}	3	974	0.22
{length, frequency_log}	3	877	0.23
{length, frequency_log, gender}	3	776	0.25
{length, frequecy_log, gender, arousal}	3	659	0.27
{length, frequency_log, gender, arousal, familiarity}	3	537	0.3
{length, frequency_log, gender, arousal, familiarity, semsize}	3	409	0.34
{length, frequency_log, gender, arousal, familiarity, semsize, aoa}	3	219	0.44

Tabella 3: Analisi del coefficiente di silhouette e SSE

4.1.3 Analisi qualitativa dei clusters

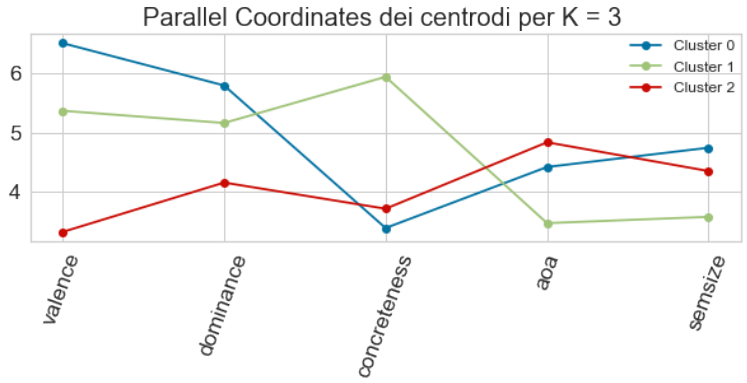


Figura 13: Coordinates Parallel dei centroidi delle variabili utilizzate per il clustering

Dal grafico riportato in figura 13 emerge che questa configurazione risulta essere quella ottimale in quanto i centroidi di tutti gli attributi sono ben separati tra di loro. Inoltre, questo grafico conferma quanto appreso dalla matrice di correlazione 10: ad esempio il cluster 0 raggruppa parole tendenzialmente meno concrete che richiederanno quindi una maggiore età per essere apprese (correlazione negativa tra *concreteness* e *aoa*). Analogo ragionamento per gli altri due clusters.

Dalla figura 14 si evince che tutti i clusters hanno un coefficiente di Silhouette superiore alla media e presentano una popolazione numericamente simile, fatta eccezione per il cluster 1 che ne contiene un numero maggiore (perché presenta osservazioni più concentrate intorno al rispettivo centroide).

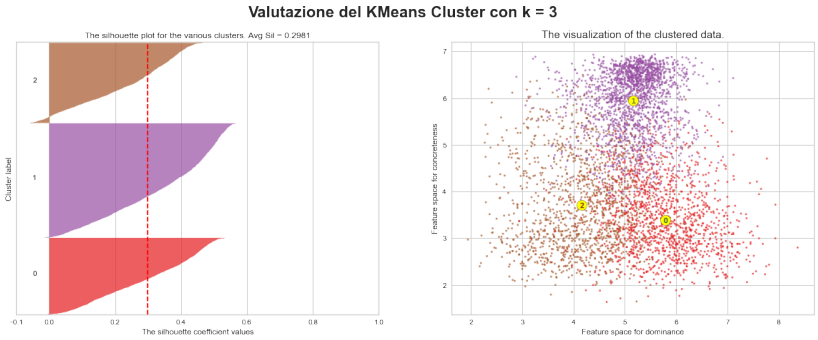
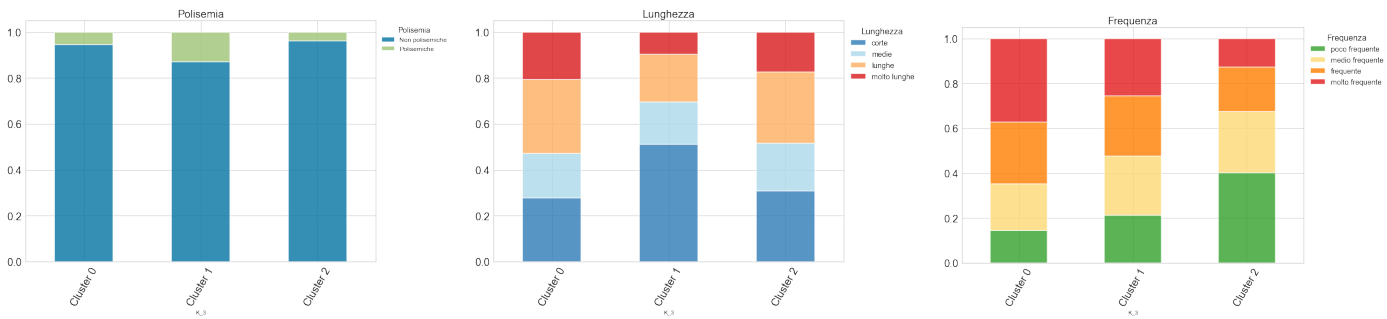


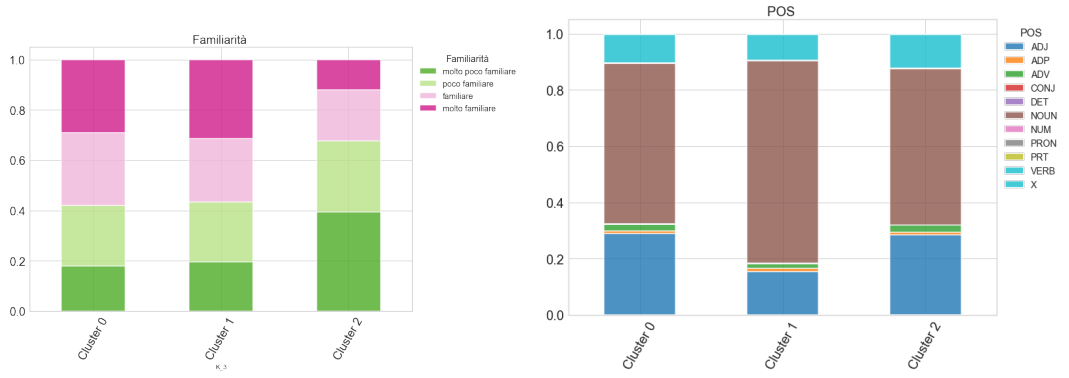
Figura 14: Valutazione della qualità del cluster con k = 3

Il passaggio successivo ha riguardato l'analisi della distribuzione delle variabili utilizzate nella fase di Data Understanding, che sono state escluse nel processo di *Clustering*, al fine di osservare comportamenti significativi che esse potrebbero assumere all'interno dei clusters individuati. Il primo *bar chart* (Figura 15a) mostra come la maggior concentrazione di parole polisemiche risieda nel cluster 1. Mediante la comparazione con gli altri tre altri grafici è stato possibile confermare, in maniera più o meno marcata, quanto studiato e stabilito nelle fasi di *Data Understanding* e nella matrice di correlazione (Figura 10).



(a) **(b)** **(c)**

In relazione alla lunghezza (Figura 15b), ad esempio, è molto evidente come il cluster 1 sia costituito in maniera preponderante da parole corte e di media lunghezza, ad ulteriore conferma della relazione negativa e forte che intercorre tra tale attributo e la variabile target.



(d) **(e)**

Figura 15: Distribuzione della *polisemia* (15a), *lunghezza*, (15b) *frequenza* (15c), *familiarità* (15d) e delle *Part Of Speech* (15e) all'interno dei clusters

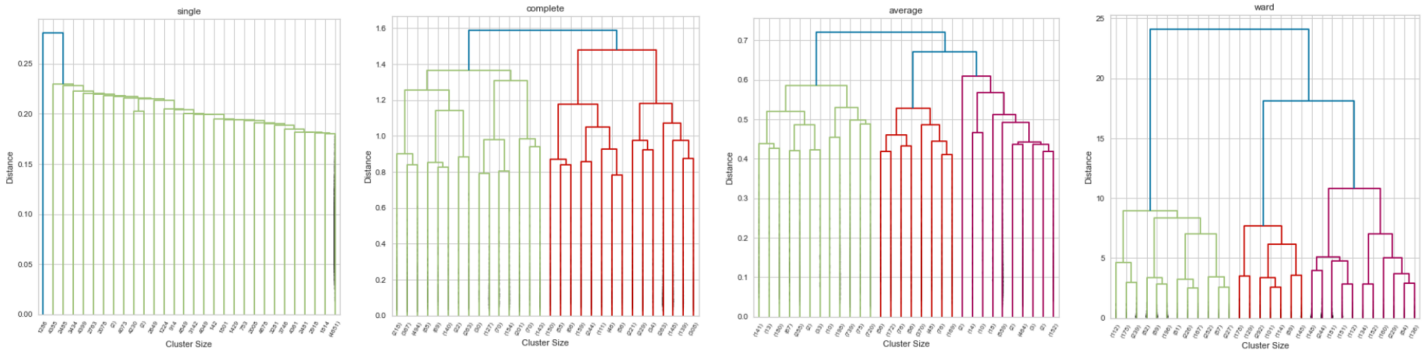
Per quanto concerne le altre due variabili, *frequency-log* (Figura 15c) e *familiarity* (Figura 15d), la minore evidenza è motivata dal fatto che tali attributi presentano una correlazione nettamente inferiore con *polisemy*, rispetto alla lunghezza.
Nell'ultimo grafico (Figura 15e), con lo scopo di approfondire ulteriormente l'analisi, viene mostrata la distribuzione che le *Parts Of Speech*, estratte mediante la libreria NLTK, presentano nei vari clusters. Come da aspettative il cluster 1 presenta una maggior concentrazione di nomi (Figura 7b).

4.1.4 Conclusioni del k-means

Osservando la distribuzione dei centroidi, la popolosità dei clusters e il valore del coefficiente di Silhouette possiamo concludere che l'algoritmo del K-means conduce a una buona configurazione di clusters per il nostro dataset.

4.2 Clustering Gerarchico

Successivamente è stato implementato l'algoritmo del **Clustering gerarchico** utilizzando le stesse *features* del K-means, al fine di agevolare un diretto confronto tra i diversi algoritmi di clustering. Di seguito vengono mostrati i *Dendrogrammi* ottenuti utilizzando i 4 criteri *linkage* (*Single*, *Complete*, *Avarage*, *Ward*).



(a) **(b)** **(c)** **(d)**

Figura 16: Cluster gerarchici ottenuti utilizzando i criteri linkage: *Single* (16a), *Complete* (16b), *Average* (16c) e *Ward* (16d)

I risultati del *Clustering gerarchico* (Figura 16) mostrano che il Single Link presenta un solo cluster, quindi non è adeguato all'analisi condotta. Per quanto concerne *Complete*, *Avarage* e *Ward* riescono ad ovviare a tale problema producendo clusters ben bilanciati. Tuttavia, abbiamo ritenuto opportuno considerare i criteri

Avarage e *Ward* sulla base del numero di clusters che erano stati individuati dall'algoritmo del *K-means* (k = 3), escludendo quindi *Complete* che presenta un numero di clusters pari a due. Al fine di individuare quale fosse il miglior criterio tra questi due ci siamo avvalsi del *coefficiente di Silhouette*.

Linkage	**Silhouette**
Single	0.25
Complete	0.19
Average	0.25
Ward	0.27

Tabella 4: Coefficiente di Silhouette per ciascun criterio di linkage

Nella tabella 4 si mostra come il metodo *Ward*, presentando il più alto coefficiente di Silhouette, è da considerarsi il migliore tra i vari criteri. Nonostante sia il più alto (0.27) risulta comunque essere inferiore a quello individuato dall'algoritmo del K-means (0.3).

Osservando lo scatter-plot delle *features dominance* e *concreteness* (usate anche per il K-means, Figura 17), possiamo comunque concludere che i clusters generati siano molto simili a quelli osservati nel grafico 14 prodotto dall'analisi svolta con l'algoritmo K-means.

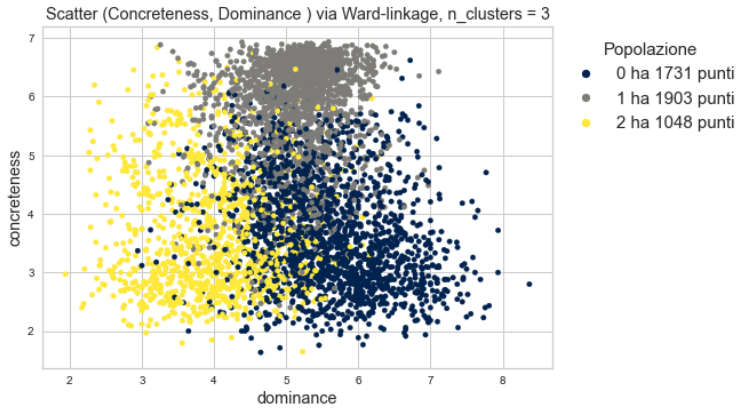


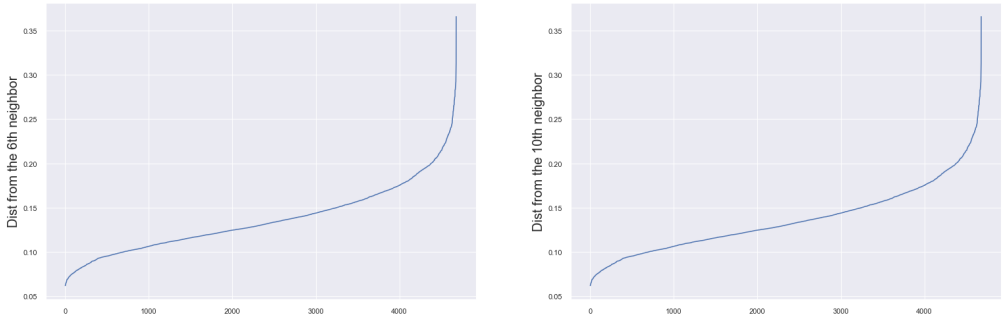
Figura 17: Scatterplot delle variabili *Concreteness* e *Dominance* con il metodo *Ward*

L'analisi qualitativa del clustering gerarchico non presenta differenze sostanziali con quelle emerse per il K-means, all'interno della sezione 4.1.3, a ulteriore dimostrazione che i clusters prodotti dai due algoritmi sono molto simili tra di loro.
In conclusione, l'algoritmo di *Clustering gerarchico* si dimostra valido per il nostro dataset ma basandoci sul coefficiente di Silhouette è possibile concludere che il K-means sia migliore.

4.3 DBSCAN

L'ultimo algoritmo di clustering implementato è il **DBSCAN** per il quale è stato necessario stabilire il *minPts*, ovvero il numero minimo di punti da cui un punto deve essere circondato entro un certo raggio *Eps* (Epsilon) per essere definito cluster. É stato necessario anche stabilire il raggio Epsilon.

Le regole più diffuse per stabilire con precisione il valore di *minPts* sono due: la prima consiste nel sommare 1 alla dimensionalità del dataset che viene passato all'algoritmo; la seconda (quella da noi utilizzata), invece, si basa sul moltiplicare per due la dimensionalità del dataset.
Dato le nostre analisi fin qui condotte (Figura 13), la dimensionalità del dataset è 5, conseguentemente abbiamo considerato come *minPts* 6 e 10.



(a) **(b)**

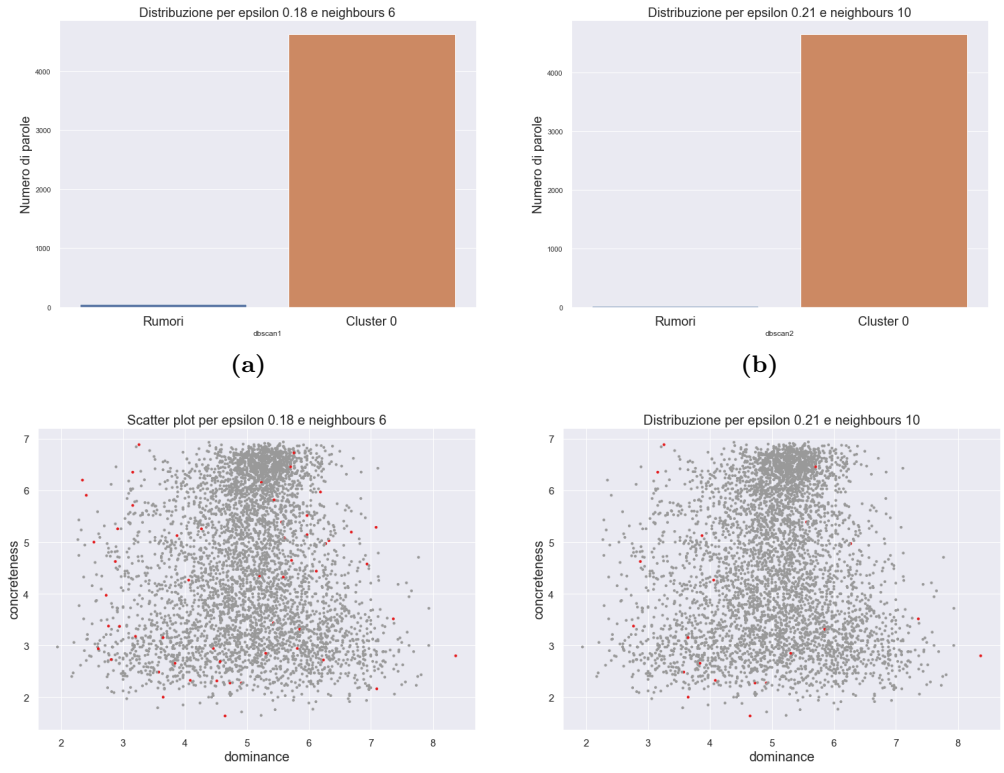
Figura 18: Elbow method

Dopo aver definito 0.18 e 0.21 come valori ottimali di *eps* dati i nostri *minPts*, abbiamo optato per la seconda via, fissando minpts a 10 ed epsilon a 0.21, prediligendo quindi la configurazione che garantisse il più alto coefficiente di Silhouette (tabella 5).

Epsilon	MinPts	Silhouette
0.18	6	0.18
0.21	10	0.20

Tabella 5: Analisi del coefficiente di Silhouette per le due diverse configurazioni

Tuttavia, l'algoritmo DBSCAN non si dimostra adeguato per lavorare con questo dataset; ciò è reso ancora più evidente dai seguenti grafici (Figura 19) i quali mostrano una quasi assoluta concentrazione dei punti in un unico cluster e una classificazione degli altri punti (21 nella configurazione migliore e 54 nell'altra) come rumori.



(a) **(b)**

(c) **(d)**

Figura 19: Popolazione dei clusters ottenuti nelle due configurazioni (figure 19a e 19b) e rappresentazione (tramite Scatterplot) dei clusters ottenuti nelle due configurazioni (figure 19c e 19d)

4.4 Conclusioni sul clustering

Dal confronto tra i tre algoritmi di clustering emerge che il **K-means** sarebbe nettamente il migliore tra i tre se ci soffermassimo esclusivamente sulla valutazione del coefficiente di Silhouette.

Algoritmo	**Silhouette**
K-means	0.3
Gerarchico (Ward)	0.27
DBSCAN	0.2

Tabella 6: Confronto dei tre algoritmi di Clustering sulla base del loro *coefficiente di Sihlouette*

Tuttavia, dall'analisi delle distribuzioni delle variabili all'interno dei vari clusters e soffermandoci su come i clusters sono stati individuati, questa differenza tra K-means e Gerarchico (con il criterio Ward) si assottiglia. Dunque, il cluster gerarchico potrebbe a sua volta essere considerato un buon algoritmo di clustering per il nostro dataset. Il DBSCAN rimane invece inefficace.

5 Classificazione

In questa sezione si è classificato il dataset con lo scopo di predire se una parola risulti essere polisemica o no (nello specifico cerchiamo di predire i valori della nostra variabile target).
A questo scopo abbiamo utilizzato il metodo *Decision Tree* inizialmente utilizzando il dataset originale, ma visti i risultati non ottimali, a causa del dataset stesso troppo sbilanciato, abbiamo applicato la tecnica di *Oversampling*. Successivamente è stato utilizzato un ulteriore algoritmo di classificazione, il *K-Nearest Neighbors (KNN)* per poter confrontare i risultati con quelli ottenuti con il *Decision Tree*.

5.1 Preparazione del dataset

5.1.1 Scelta degli attributi

Riferendoci alla matrice di correlazione analizzata durante la fase di Data Preparation (Figura 10) al fine di determinare gli attributi da utilizzare, sono stati rimossi tutti quelli che presentavano ridondanza con altri (*concreteness*) e irrilevanza (correlazione praticamente nulla) rispetto alla variabile target. Per queste ultime features è stata stabilita una soglia di correlazione minima di 0,05 (Eliminati: *arousal, valence, dominance* e *gender*).
Tra *concreteness* e *imageability* è stata rimossa *concreteness* poiché nel primo tentativo di costruzione del Decision Tree il modello ha prediletto come *Attribute Test Condition* la seconda.

5.1.2 *Hold-out* del Dataset

Per la classificazione abbiamo utilizzato il dataset già preparato come quanto descritto nella sezione di *Data preparation* (Paragrafo 3). Dopo aver provveduto alla sua divisione in *training* (l'80% del dataset) e *test set* (il restante 20%), abbiamo effettuato la ricerca dei parametri ottimali attraverso la tecnica di *Random Grid Search* applicata al training set, la quale opera il metodo di *cross-validation*. Tramite quest'ultimo metodo è possibile estrarre il *validation set* il quale ci permette di valutare la bontà del modello appena addestrato.
I parametri utilizzati nella tecnica del *Random Grid Search* sono i seguenti:

- *Min Sample Split* (numero minimo di elementi per procedere ad una ramificazione)
- *Min Sample Leaf* (numero minimo di elementi per foglia)
- *Criterion* (indice di Gini per misurare l'impurità del nodo)
- *Max Depth* (massima profondità dell'albero decisionale)

Con riferimento a quest'ultimo parametro, sono stati confrontati gli scores ottenuti con una *Max Depth* pari a 12 e con una pari a 6 (i parametri e i risultati dalle due configurazioni sono riassunti nella tabella 7).
La comparazione ha mostrato che entrambe le configurazioni, seppur di differente complessità, presentano

un simile grado di dispersione degli elementi (valutato in base alla deviazione standard); abbiamo perciò optato per la configurazione più semplice ed efficiente (di profondità 6).

Parametri	**Confiigurazione 1**	**Configurazione 2**
Criterio	gini	gini
max_depth	**6**	**12**
min_samples_leaf	48	7
min_samples_split	29	19
Punteggi	**Configurazione 1**	**Configurazione 2**
Mean validation score	0.032	0.161
Dev. standard (validation)	0.043	0.033
Mean training score	0.073	0.462
Dev. standard (training)	0.0089	0.044

Tabella 7: Confronto tra le due configurazioni utilizzate inizialmente all'interno della *Random Grid Search* per la ricerca dei parametri ottimali

5.2 *Decision Tree*

Il modello ottenuto ha generato un albero decisionale che si dimostra efficace nel classificare il training set, ma è completamente inadatto per il test set (in quanto non riesce a predire una percentuale accettabile di parole polisemiche). Ciò può essere giustificato considerando il forte sbilanciamento che caratterizza il nostro dataset. Le metriche ottenute sono riportate nella tabella 8.
Per risolvere lo sbilanciamento è stato applicato il metodo di **Oversampling** al training set; tale metodo si basa sulla creazione di elementi (fittizi) appartenenti alla classe meno numerosa. Il modello è stato poi addestrato su questo nuovo training set ottenendo un albero più efficiente nell'individuazione di parole polisemiche, comprovato anche dalle metriche ottenute applicando il modello al test set.

5.2.1 Interpretazione dell'albero

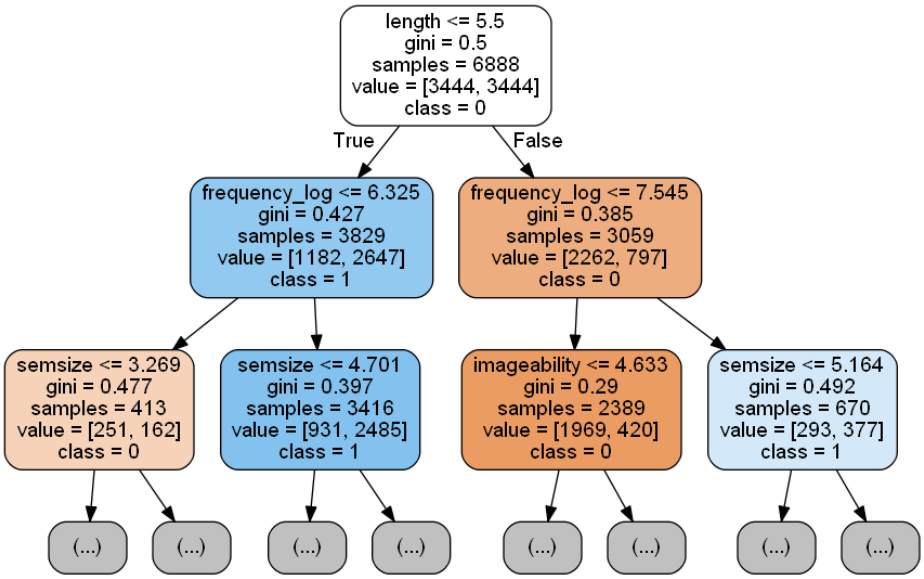


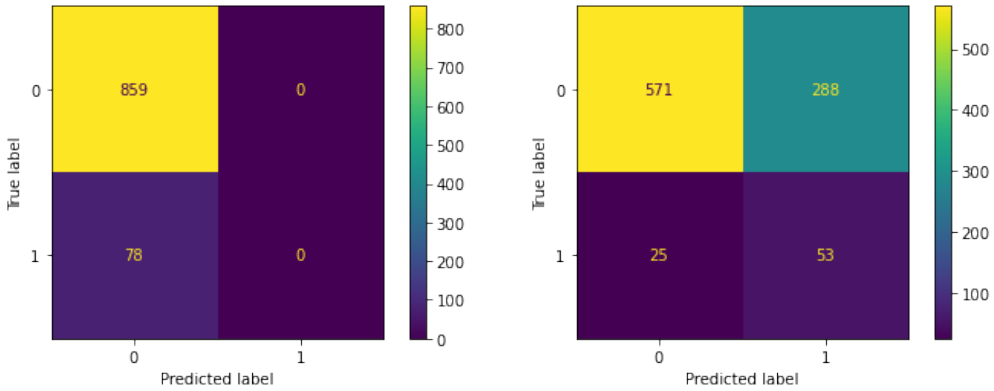
Figura 20: Albero decisionale ottenuto applicando la tecnica di Oversampling

Dall'analisi dell'albero ottenuto (Figura 20) possiamo osservare che gli attributi ritenuti significativi nella *Data Understanding* risultano esserlo anche per la costruzione del *Decision Tree*: *Lenght* e *frequency_log* presentavano infatti una buona correlazione con la variabile target *polisemy*. Tuttavia, dato che per la scelta dell'*Attribute Test Condition* è stato utilizzato il criterio Gain, dal terzo livello di profondità in poi troviamo anche attributi fino ad ora considerati poco funzionali per le analisi sin qui condotte. Questo è ulteriormente dimostrato dai valori di *importance* attribuiti dall'algoritmo alle diverse features:

- *length*: 0.4281066296611338
- *frequency_log*: 0.21669950629873214
- *semsize*: 0.14309158758874202
- *imageability*: 0.09160069322263248
- *aoa*: 0.06088080333534106
- *familiarity*: 0.05962077989341841

Andando più nello specifico possiamo vedere che l'albero ha come radice l'attributo *lenght* che si ramifica nei due nodi figli, aventi come attributo *frequency_log* , sulla base del valore che la lunghezza assume. Nello specifico, se la lunghezza della parola è minore di 5.5 si prosegue sulla parte sinistra dell'albero, giungendo poi sulla base del valore della *frequency_log* al nodo con attributo *sem_size* avente come classe predominante 0 o 1 (terzo livello di profondità). In particolare, una parola che presenti una *sem_size* minore o uguale di 4.7 sarà classificata come polisemica (classe 1). Analoghe considerazioni per la parte destra dell'albero.

5.2.2 Analisi metriche



(a) **(b)**

Figura 21: Confusion matrixes senza Oversampling (21a) e con Oversampling (21b)

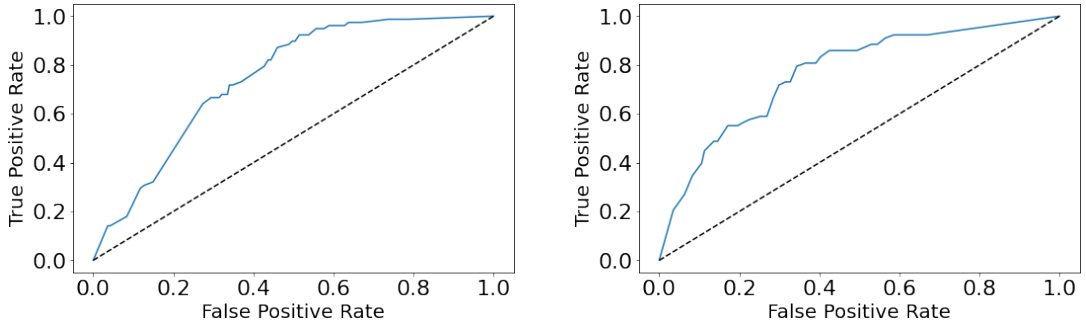
Dalla figura 21a possiamo vedere un aumento dei *False Positive* il cui valore passa 0 a 288, ma soprattutto si osserva un aumento dei *True Positive*, che il modello senza *Oversampling* non riusciva a predire.
Tali matrici saranno di ausilio anche per commentare le seguenti metriche (Tabella 8).

Senza Oversampling				
	Precision	**Recall**	**F1**	**Accuracy**
Parole non polisemiche	0.92	1	0.96	
Parole polisemiche	0	0	0	0.92
Con Oversampling				
	Precision	**Recall**	**F1**	**Accuracy**
Parole non polisemiche	0.96	0.66	0.78	
Parole polisemiche	0.16	0.68	0.25	0.67

Tabella 8: Confronto delle metriche di valutazione con e senza *Oversampling*

Confrontando le metriche ottenute in fase di test prima e dopo l'*Oversampling*, è possibile notare come il metodo di *Oversampling* porti ad un marcato aumento delle parole polisemiche correttamente predette (*True Positive, TP*); osserviamo però anche un maggior numero di *false positive*. A questo aspetto si lega il basso valore della *precision*, che comunque per le parole polisemiche aumenta da 0 a 0.16.
Vediamo, inoltre, un forte aumento della *recall* per le parole polisemiche, scontando però una significativa

diminuzione per le parole non polisemiche (da 1 a 0.66). Di conseguenza si osserva anche l'aumento del valore di F1 per le parole polisemiche e una riduzione per le parole mono-semantiche. Il valore dell'*Accuracy*, nei due casi, ha mostrato una diminuzione sul set bilanciato, a causa dell'aumento dei falsi positivi. Quest'ultimo indice non è stato tuttavia considerato adeguato a causa dello sbilanciamento del dataset tra parole polisemiche e non polisemiche.



(a) **(b)**

Figura 22: ROC-curve del modello con Oversampling (22a) e senza Oversampling (22b)

Dall'analisi del grafico della **ROC-Curve** (Figura 22) possiamo notare alcune differenze tra i due modelli: in particolare la curva generata con dataset bilanciato mostra un andamento più ripido in corrispondenza di alte soglie (tra il 0.4-1.0), ciò a indicare che la maggioranza delle parole che sono state classificate con un alto grado di probabilità come polisemiche effettivamente lo erano. Questo a differenza della curva generata senza oversampling, la quale risulta essere più tendente verso la linea diagonale, in corrispondenza di soglie più alte. Possiamo quindi dire che il modello addestrato sul dataset bilanciato è in grado di predire correttamente una parola come polisemica. Nonostante le differenze di andamento, per entrambe il valore del ROC-AUC è di circa 0,7.

5.3 Confronto tra Decision Tree e K-Nearest Neighbors

É stato poi utilizzato un ulteriore algoritmo di classificazione per poter confrontare i risultati ottenuti nel *Decision Tree*.
Il metodo scelto è stato il **K-Nearest Neighbors** basato sulla classificazione di un oggetto osservando i suoi k-vicini. Come per l'algoritmo dell'albero decisionale abbiamo utilizzato le stesse variabili ed effettuato l'*Oversampling* sul *training set* al fine di ottenere un *dataset* più bilanciato e sopratutto per agevolare il confronto tra i due algoritmi. Dopo aver individuato i migliori parametri con il *Gridsearch* (*'metric': 'euclidean', 'n_neighbors': 20, 'weights': 'uniform'*), sono state calcolate le metriche applicando il modello appena addestrato sul test set.

Decision Tree			
	Precision (test)	**Recall (test)**	**F1 (test)**
Parole non polisemiche	0.96	0.66	0.78
Parole polisemiche	0.16	0.68	0.25
KNN			
	Precision (test)	**Recall (test)**	**F1 (test)**
Parole non polisemiche	0.97	0.64	0.77
Parole polisemiche	0.16	0.76	0.26

Tabella 9: Confronto tra le metriche di valutazione del Decision Tree e del K-Nearest Neighbors

Come è possibile osservare dalla tabella, le metriche di entrambi i modelli sono molto simili, nonostante il *K-Nearest Neighbors* commetta più errori classificando come polisemiche parole che non lo sono (aumento di *False Positive*). L'unico aspetto degno di nota è la maggior *Recall* per le parole polisemiche: questo denota una riduzione dei *False Negative* se applichiamo il KNN rispetto Decision Tree.
Entrambi gli algoritmi si dimostrano validi per predire la variabile target a condizione che venga applicata

la tecnica dell'Oversampling sul dataset. Il KNN, tuttavia, presenta un leggero miglioramento per quanto riguarda la *Recall*.

6 Pattern Mining

Questa sezione è dedicata al Pattern Mining che consiste nell'estrazione di *itemsets* più frequenti, a partire dai quali tramite l'algoritmo *Apriori* vengono generate le regole di associazione. Abbiamo utilizzato le regole più frequenti per risolvere due questioni: la **sostituzione dei valori mancanti** (Sezione 6.4) e la **predizione della variabile target** (Sezione 6.5).

6.1 Preparazione del dataset

Abbiamo utilizzato lo stesso dataset creato in fase di *Data preparation* con la differenza che in questo caso sono stati eliminati i records contenenti i *missing values*. Inoltre sono stati considerati i medesimi attributi utilizzati per il *Decision Tree*. Successivamente alla fase di *Preprocessing* è stato necessario discretizzare tutti gli attributi quantitativi assegnando il valore di tutte le variabili ai relativi quartili.

6.2 *Frequent Itemsets*

Il passaggio successivo è stato quello di applicare l'algoritmo *Apriori* grazie al quale sono stati ottenuti i **Frequent**, i **Closed** e i **Maximal itemsets**. In totale abbiamo ottenuto 1143 *Frequent itemests*, 1126 *Closed itemsets* e 355 *Maximal itemsets*. Il valore del *support* utilizzato è dello 0,03% e il numero minimo di item uguale a 2. La determinazione di un così basso valore della soglia minima di *support* è dovuta al fatto che, impostando un valore più alto, l'algoritmo non era in grado di estrarre regole per le parole polisemiche.

Frequent - Closed Itemsets	**Support**
{'NOUN', 'Non polisemico'}	0.57
{(1.999, 5.0)_length', 'Non polisemico'}	0.32
{(6.0, 8.0)_length', 'Non polisemico'}	0.26
{'(1.999, 5.0)_length', 'NOUN'}	0.24
{(4.1049, 6.223)_frequency', 'Non polisemico'}	0.24
Maximal Itemsets	**Support**
{ADJ', '(1.999, 5.0)_length', 'Non polisemico'}	0.071
{ADJ', '(1.999, 5.0)_length', 'Non polisemico'}	0.064
{'ADJ', '(6.0, 8.0)_length', 'Non polisemico'}	0.064
{(3.118, 4.177)_aoa', '(1.999, 5.0)_length'}	0.062
{(4.672, 6.031)_imageability', '(1.999, 5.0)_length', 'NOUN', 'Non polisemico'}	0.061

Tabella 10: Frequent, Closed e Maximal Itemsets con support minimo = 0.03%

Come è possibile dedurre dalla tabella 10, le parole non polisemiche tendono ad essere nomi, la cui lunghezza è compresa in un range tra 2 e 8 a conferma di quanto osservato durante la fase di *Data Understanding* (Sezione 2). A differenza della prima fase notiamo, tuttavia, che secondo i *frequent itemsets* estratti la maggioranza della parole non polisemiche ha una frequenza tra 4.10 a 6.2.

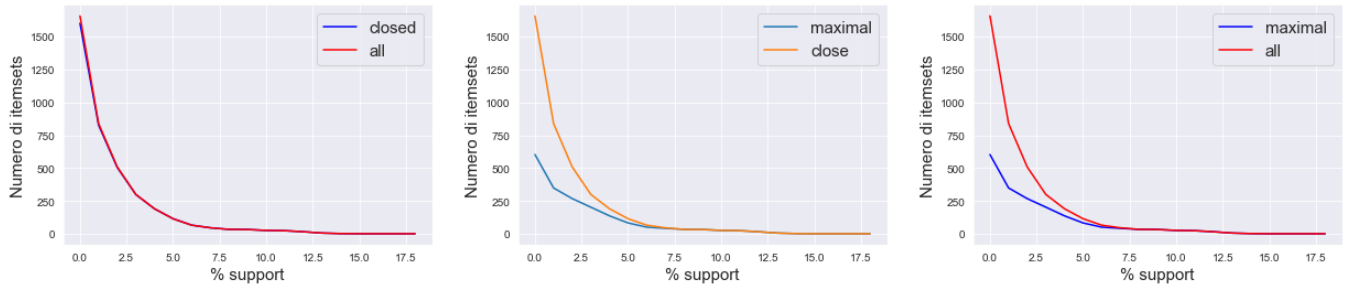
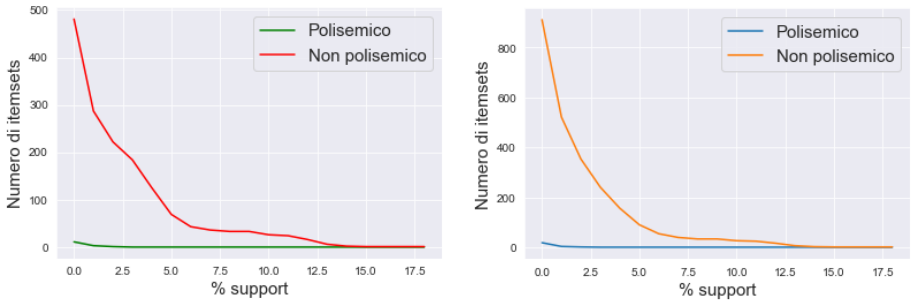


Figura 23: Distribuzione del numero di itemsets in base al valore di support



(a) **(b)**

Figura 24: Distribuzione degli itemset Maximal (24a) e Closed (24b) polisemici in base al valore di support

Nella figura (Fig. 23a) i *frequent* e i *closed itemsets* presentano un andamento quasi similare. É inoltre possibile vedere come all'aumentare del *support* le curve decrescano.
I grafici (Figure 24a e 24b) mostrati evidenziano come il *support* molto basso sia stato necessario per l'individuazione di *frequent itemsets* polisemici.

6.3 *Association Rules*

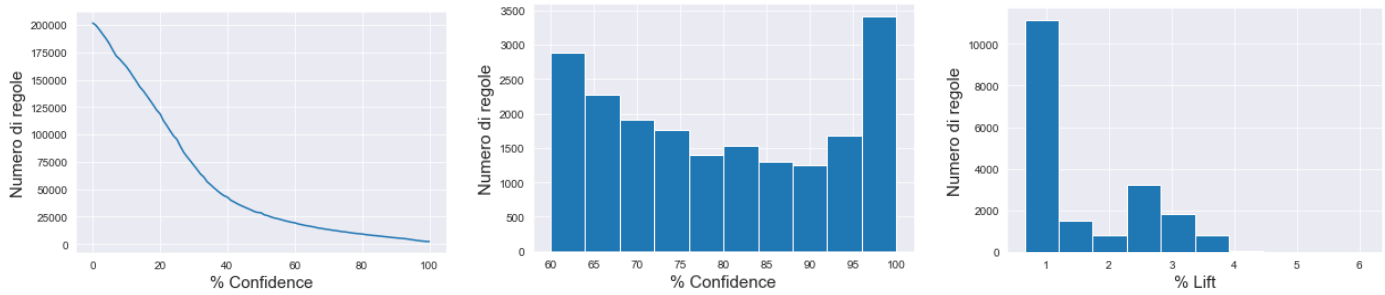
A partire dai frequent itemsets abbiamo estratto le **Association Rules** fissando una soglia per la *confidence* al 60%. Abbiamo così ottenuto 1466 regole, di cui abbiamo riportato solo cinque nella seguente tabella:

Regola	Support	Support (%)	Confidence	Lift
{(8.0, 16.0]_length, (5.152, 6.833]_aoa, (4.1049, 6.223]_frequency_log} ⇒ Non Polisemico	149	3.19	1	1.08
{(8.0, 16.0]_length, (5.152, 6.833]_aoa, NOUN } ⇒ Non Polisemico	196	4.19	1	1.08
{(4.177, 5.152]_aoa, (7.349, 9.306]_frequency_log, (4.672 , 6.031]_imageability, (1.999, 5.0]_length, NOUN} ⇒ Polisemico	8	0.17	0.61	7.57
{(8.0, 16.0]_length, (5.438, 5.969]_familiarity, (4.1049, 6.223]_frequency_log, (3.516, 4.672]_imageability, Non Polisemico} ⇒ VERB	10	0.21	0.62	6.0
{(4.882, 6.912]_semsize, (5.152, 6.833]_aoa, (1.7360, 3.516]_ imageability, (4.1049, 6.223]_frequency_log, (4.706, 5.438]_familiarity, Non polisemico} ⇒ (8.0, 16.0]_length	10	0.21	0.76	5.13

Tabella 11: Regole di Associazione estratte

É possibile notare dalla prima regola che il 100% delle parole che presentano una lunghezza elevata, una bassa frequenza e che vengono apprese in età avanzata hanno natura non polisemica. La seconda regola conduce allo stesso risultato ma ponendo come condizione 'NOUN' al posto di *frequency_log*. Tuttavia, queste due regole presentano un lift molto vicino ad 1, il che denota una correlazione assente tra antecedente e conseguente.
Abbiamo così deciso di rappresentare altre tre regole con maggior *Lift* che però presentano una *confidence* minore. Il *Lift* misura l'importanza di una Regola di Associazione e quindi la sua affidabilità.
Nelle terza regola è possibile constatare che un nome di lunghezza ridotta, con alta frequenza, con un'età di acquisizione medio alta e con un alto grado di *Imageability* ha il 60% di probabilità di essere polisemico. L'aspetto degno di nota è l'alto valore di Lift (7.57), il quale indica una forte correlazione positiva tra antecedente e conseguente. Le ultime due regole non sono funzionali a predire la variabile target, perciò, sono state riportate solo per completezza.

Dopo aver descritto l'andamento decrescente del numero delle regole all'aumentare della *confidence* (Figura 25a), è stato deciso di analizzare nel dettaglio solamente quel range di valori della confidence ritenuti accettabili (60 - 100%). Da quest'ultimo possiamo vedere che la maggior parte delle regole pur avendo una confidence pari al 100% (Figura 25b) presentano un valore di lift intorno a 1 ad indicare un'assenza di correlazione tra l' antecedente e il conseguente di ciascuna regola.



(a) **(b)** **(c)**

Figura 25: Distribuzione delle regole in base ai valori di *Confidence* (25a e 25b) e *Lift* (25c)

6.4 Sostituzione dei *missing values*

Le Association Rules sono un buon metodo per sostituire i **Missing Values** di una feature.
Come osservato in fase di *Data Preparation* (Sezione 3), il nostro dataset contiene 14 *missing values* nella variabile *web_corpus_freq*. Al fine di sostituirli, sono stati fatti alcuni tentativi utilizzando due regole precedentemente estratte.
La prima ha come conseguente la variabile *frequency_log* ed è la seguente:

> (8.0, 16.0)_length, (6.031, 6.941)_imageability, (1.374, 3.433)_semsize, (4.706, 5.438)_familiarity
> Non polisemico ⇒ (4.1049, 6.223)_frequency_log

Tuttavia, a causa dell'antecedente troppo vincolante, la regola non è stata in grado di sostituire alcun *missing values*. Per questo motivo è stata scelta una regola avente *frequency_log* come antecedente e 98% di **confidence**, ed è la seguente:

> (4.1049, 6.223)_frequency_log ⇒ Non polisemico

Dato che tutti gli elementi aventi *missing values* sono di natura non polisemica, questa regola ci ha permesso di sostituire tutti i 14 valori mancanti col valore dell'antecedente (4.1049, 6.223).

6.5 Predizione della *target variable*

Abbiamo, infine, utilizzato le regole estratte per costruire un modello predittivo per la nostra variabile target. La regola utilizzata è la seguente:

> (4.177, 5.152)_aoa, (7.349, 9.306)_frequency_log, (4.672, 6.031)_imageability, (1.999, 5.0)_length,
> NOUN ⇒ Polisemico

Nonostante un valore della *confidence* piuttosto basso (67%), l'alto valore della *lift* (7.6) ci ha permesso di stabilire l'affidabilità della regola. Abbiamo quindi classificato tutti i records che soddisfano la regola come polisemici e gli altri come non polisemici. Successivamente il dataset è stato nuovamente diviso in *training* e *test set* e abbiamo valutato il modello così ottenuto analizzando i valori delle metriche già viste per il Decision Tree.

	Precision	**Recall**	**F1**	**Accuratezza**
Parole non polisemiche	0.92	1	0.96	
Parole Polisemiche	0.25	0.01	0.02	0.92

Tabella 12: Metriche di valutazione delle Regole di Associazione

Il basso valore della Recall e del F1 measure per le parole polisemiche mostrano come questo modello predittivo sia meno efficiente e adatto a predire la variabile target rispetto a quanto fatto dal Decision Tree (Tabella 8)

TP	FN
1	113
FP	**TN**
3	1284

Tabella 13: Elevato numero di falsi positivi e falsi negativi individuati dal modello

Ciò è ulteriormente confermato dall'unico valore polisemico correttamente predetto dal modello e dall'elevato numero di *False Negative* che il modello riconosce.

7 Conclusioni

Tutte le analisi svolte finora hanno fatto emergere molte peculiarità del Glasgow Norms, dovute soprattutto alla natura soggettiva dei valori assunti dalle variabili, alla poca correlazione tra le features e la variabile target e della forte predominanza delle parole non polisemiche rispetto a quelle polisemiche.

Infatti, a causa di questa forte soggettività, non è stato possibile rimuovere gli outliers in fase di Data Preparation, ma è stato necessario studiarli e metterli in relazione con la variabile target. (Figura 9).
Inoltre, il forte sbilanciamento dei valori assunti dalla variabile target polisemia ha causato alcune difficoltà durante la classificazione, per cui è stato necessario applicare la tecnica di oversampling per cercare di bilanciare il dataset, e durante l'estrazione delle regole di associazione in cui si è osservato una certa difficoltà nell'estrarre regole che avessero come conseguente la polisemia.
Al contrario, invece, l'operazione di Clustering ha portato dei buoni risultati soprattutto per quanto riguarda i primi due algoritmi di clustering implementati (K-means e Gerarchico) e ciò dimostra quindi che il dataset contiene al suo interno un certo raggruppamento tra le features che lo compongono.

Per compensare questo sbilanciamento e in generale migliorare le performance degli algoritmi utilizzati in questo report, potrebbe essere efficace introdurre sia nuovi records di parole polisemiche, ma soprattutto nuove *features* maggiormente correlate con la variabile target.