UNIVERSITÀ DI PISA

Progetto di Data Mining 1 a.a. 2021/2022

The Glasgow Norms

A cura di:
Altavilla Federica
Giovino Salvatore
Gorgoglione Ismaele
Petrucci Aura

Università di Pisa

16 febbraio 2022

Indice

1 Introduzione

The Glasgow Norms è un dataset linguistico che raccoglie le cosidette normative ratings di utenti anglofoni riguardo 5553 parole inglesi. Per normative ratings si fa riferimento ad uno strumento valutativo in cui l'intervistato sceglie punteggi da associare ad una serie di elementi, in questo caso alle parole presenti nel corpus. Lo scopo della seguente indagine è quello di analizzare il dataset per osservare eventuali correlazioni tra i termini valutati e i valori ad essi associati. Per fare ciò abbiamo suddiviso il lavoro in differenti fasi: *Data Understanding, Data Preparation*, l'implementazione di tre metodi di *Clustering*, *Classificazione* e *Pattern* e *Association Rules Mining*.

2 Data Understanding

2.1 Data semantics

Il dataset si compone di 4682 record e 13 features delle quali 12 sono di tipo numerico, mentre la tredicesima è di tipo categorico (*word*) avente come valori le parole che compongono il corpus. Nove delle tredici variabili corrispondono a differenti dimensioni psicolinguistiche che gli utenti hanno dovuto valutare per ognuna delle parole presenti nel corpus. Tali parametri sono:

1. **Arousal (AROU)**, che misura l'eccitazione;
2. **Valence (VAL)** che indica un grado di valenza o moralità;
3. **Dominance (DOM)**, che rappresenta il senso di controllo che suscita la parola;
4. **Concreteness (CNC)**, misura di quanto un concetto sia astratto o concreto;
5. **Imageability (IMG)**, che valuta quanto sia facilmente immaginabile una parola;
6. **Familiarity (FAM)**, misura la familiarità che suscita la parola;
7. **Age of acquistion (AOA)**, che indica a che età la parola è stata acquisita dall'utente;
8. **Semantic size (SIZE)**, misura le dimensioni di qualcosa;
9. **Gender association (GEND)**, che misura quanto il significato di una parola possa essere associato a genere maschile o femminile.

Le variabili restanti sono:

1. **Length**, che indica la lunghezza della parola presa in considerazione;
2. **Polysemy**, quando una parola può avere più di un significato, per questo motivo è rappresentato da un valore binario;
3. **Web-corpus-freq** indica quanto frequentemente una specifica parola appare nel *Google Newspapers Corpus*.

Le prime tre variabili sopracitate - *arousal, valence e dominance* - sono state valutate su una scala di 9 punti, mentre gli altri attributi su un range di 7.

Classificazione delle variabili		
Tipo		**Attributo**
Categoriche	Nominali	Word
Numeriche	Binari	Polysemy
	Discreti	Length
	Continui	Arousal, Valence, Dominance, Concreteness, Imageabilty, Familiarity, Aoa, Semsize, Gender, Web_Corpus_Freq

Tabella 1: Attributi categorici e numerici nel dataset

2.2 Distribuzione statistica delle variabili

In questa prima fase sono state analizzate in maniera più approfondita le distribuzioni statistiche di alcune delle *features* più rilevanti, sia individualmente che in correlazione tra loro.

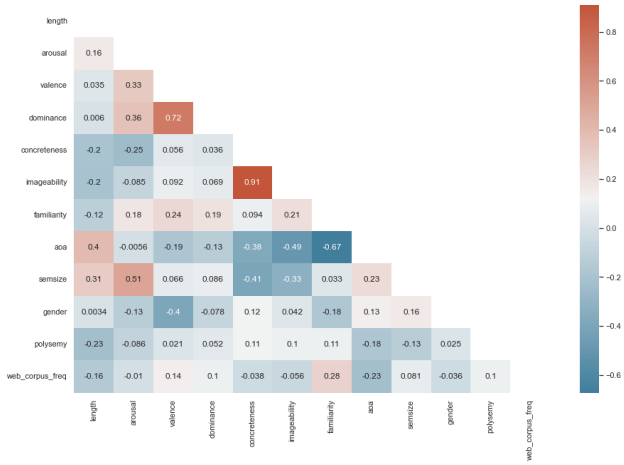


Figura 1: Correlazione delle variabili

Come si può osservare dalla figura 1, la correlazione tra le variabili *concreteness* ed *imageability* è estremamente elevata: oltre al dato analitico, semanticamente le due variabili riconduono ad un significato similare, ossia di quanto una parola possa essere concreta o astratta, facile o difficile da immaginare. Da queste valutazioni si è deciso successivamente di eliminare dal dataset l'attributo *imageability*.

Length e polysemy

Nella figura 2, viene riportata la distribuzione dell'attributo *polysemy* all'interno del dataset. Essa mostra che solo una piccola quantità delle parole presenti sono polisemiche, ovvero veicolano più di un significato: in particolare si tratta di 379 parole. In seguito, abbiamo deciso di osservare la correlazione tra questi termini e la loro lunghezza: come si nota dall'istogramma (figura 3) la maggior parte di esse è composta dai 3 ai 5 caratteri.

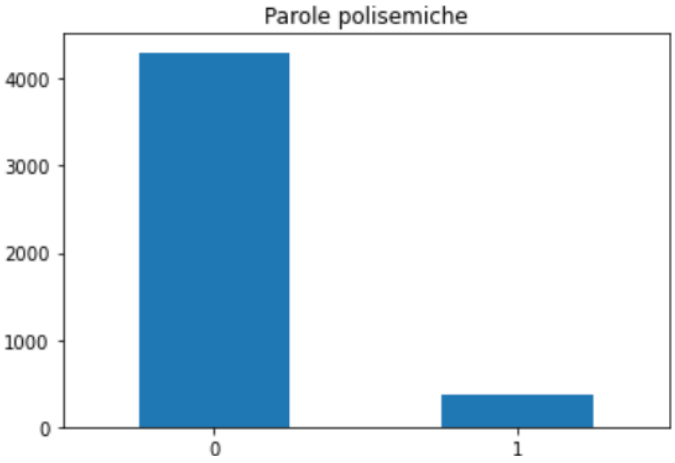


Figura 2: Polysemy

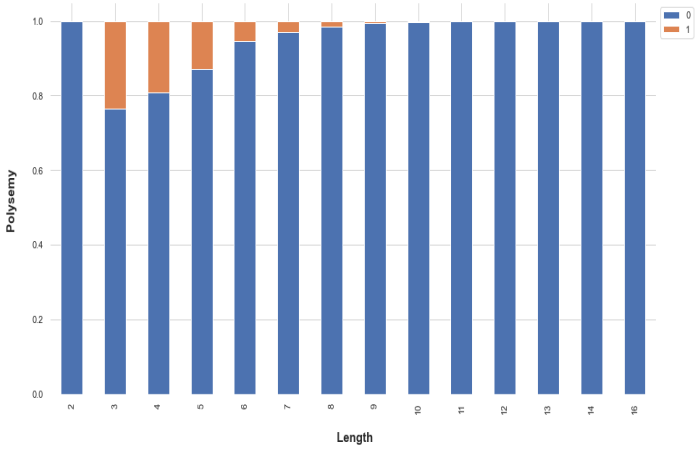
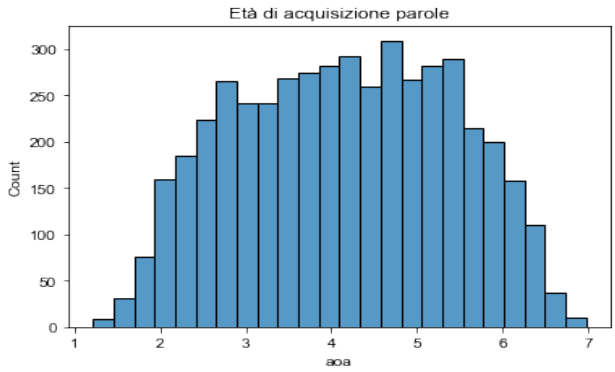


Figura 3: Lunghezza delle parole e polisemia

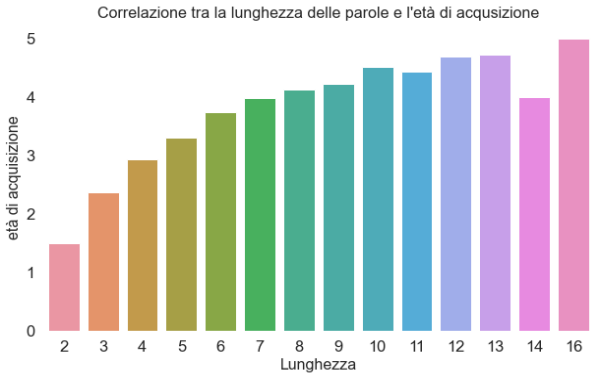
Length e Aoa

Successivamente, abbiamo visualizzato la distribuzione della variabile *aoa* (*acquisition of age*) attraverso la quale si evince che i valori più alti di apprendimento si hanno tra i 3 e i 6 punti (figura 4 (a)). Oltre a ciò è stata analizzata la correlazione tra questa variabile e la lunghezza delle parole:

il grafico (figura 4(b)) mostra il risultato atteso, vale a dire che le parole con maggiori caratteri vengono apprese più facilmente all'aumentare dell'età.



(a) Età d'acquisizione



(b) Correlazione tra lunghezza parole e età acquisizione

Figura 4: Distribuzione della variabile aoa e della correlazione con la lunghezza

Gender e Dominance

Infine, è parso interessante valutare l'eventuale relazione tra le variabili *gender* e *dominance*. Per fare ciò abbiamo modificato i valori di *gender* inserendoli in un range di 3 assegnando *female* a 0, *unisex* a 1 e *male* a 2. In tal modo è possibile osservare che le parole aventi un valore di *dominance* elevato (3-6) sono state più frequentemente associate a genere maschile.

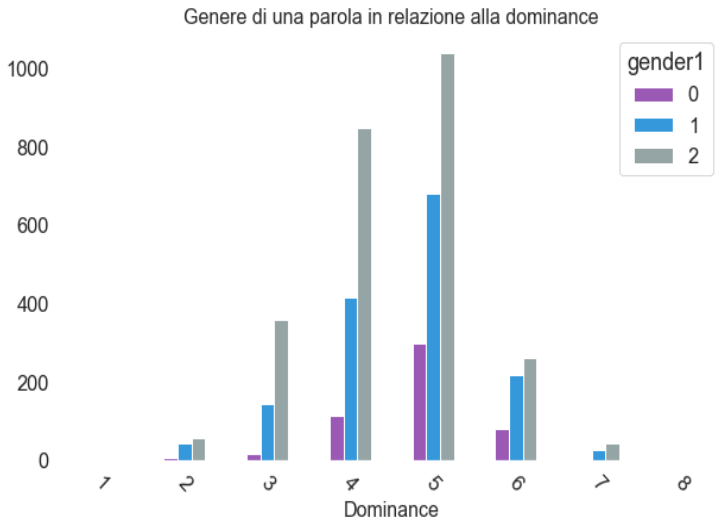


Figura 5: Correlazione tra gender e dominance

3 Data Preparation

La qualità del dataset utilizzato incide profondamente sull'efficienza e la coerenza dello studio effettuato: quanto più pulito è questo, tanto più facile ed affidabile sarà l'indagine. In particolare è necessario individuare:

- Missing values
- Outliers
- Valori duplicati
- Inconsistenza semantica dei dati (i.e. l'incongruenza dei dati inseriti nel database, la cui conseguenza è l'inaffidabilità degli stessi)
- Inconsistenza sintattica dei dati (i.e. eventuali errori di sintassi presenti nel dataset, e.g. "fmale" invece di "female").

3.1 Discretizzazione delle variabili

Per una migliore visualizzazione ed interpretazione dei grafici, alcune variabili sono sta-

te discretizzate. Ad esempio, nell'analisi della distribuzione degli *outliers* di *Dominance* con *Concreteness*, la prima è stata divisa in due intervalli, dal valore minimo al valor medio (*Low_control*) e dal valore medio al valore massimo (*High_control*). Lo stesso procedimento è stato effettuato per la variabile *Gender*. Ugualmente, in seguito sarà utilizzata la medesima procedura nel *clustering* per lo studio della distribuzione delle features.

3.2 Missing values

Analizzando il dataset si riscontra la presenza di 14 *missing values*, tutti appartenenti alla variabile *web_corpus_freq*. Per gestirli si può procedere in diversi modi: in primo luogo bisogna valutare la grandezza del dataset in relazione alla quantità dei dati mancanti.

Missing values	
word	0
length	0
arousal	0
valence	0
dominance	0
concreteness	0
imageability	0
familiarity	0
aoa	0
semsize	0
gender	0
polysemy	0
web_corpus_freq	14

Tabella 2: Missing values

In un caso pratico di lavoro, si potrebbe pensare di non considerare le celle vuote, di eliminarle o ancora di rimpiazzare tali dati con media o mediana del dataset.

Ai fini della nostra indagine, abbiamo pensato che data la natura della variabile *web_corpus_freq*, utilizzare media o mediana sia poco affidabile considerando la distanza di 5 ordini di grandezza dal valore minimo al massimo; stimare tali dati mancanti in questa maniera non darebbe dei risultati attendibili. Per questo motivo abbiamo optato per sostituire i *missing values* tramite interpolazione lineare di tutti i record relativi alla feature *web_corpus_freq*.

3.3 Outliers

Per quanto riguarda la ricerca di eventuali *outliers*, si è provveduto alla visualizzazione grafica dei dati tramite il diagramma boxplot, che permette di individuare qualitativamente la loro presenza nel dataset. Si può notare subito come solo alcune variabili presentino valori anomali, in particolare *arousal, valence, dominance, gender* e *familiarity*; le restanti risultano invece "pulite". Successivamente, si è proseguito individuando le parole corrispondenti agli outliers reperiti.

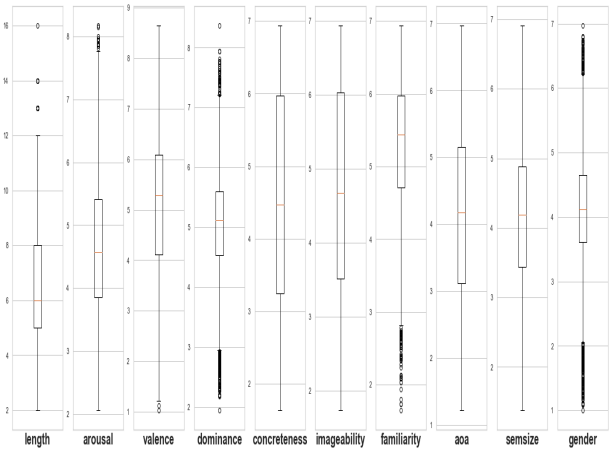


Figura 6: Boxplots per gli outliers delle variabili

Words Glasgow non presenta dati provenienti da misurazioni strumentali per le quali si potrebbe incombere in errori causati da malfunzionamento dei dispositivi utilizzati. Per questo motivo non si può parlare di veri e propri *outliers* come da definizione, ma di valori "eccezionali" nella "potenza semantica" che veicolano. Si prendono in esame la variabile *arousal* per la quale sono stati individuati 11 outliers (*adventure, aroused, enthusiastic, erotic, euphoria, excited, kiss, love, orgasm, passionate, spectacular*), e la variabile *valence*, che ne presenta solamente due (*genocide, rape*). In entrambi i casi, i risultati ottenuti mostrano come tali parole siano l'enfatizzazione del concetto stesso che esprimono, di fatto per la prima feature enfatizzano "positivamente" il concetto, in quanto essi si trovano al di sopra del valore massimo del boxplot (i.e. parole fortemente eccitanti). Per *valence* invece i due outliers reperiti si localizzano al di sotto del valore minimo, rappresentando delle parole con una forte accezione negativa. In riferimento a ciò, quindi, non avrebbe senso parlare di gestione o correzione di questi, ma risulta migliore darne una caratterizzazione teorica osservando la loro distribuzione in relazione ad altre variabili. Innanzitutto, dopo averli individuati non solo qualitativamente ma anche quantitativamente, sono stati salvati in un nuovo dataframe, per ogni variabile. In seguito sono stati trattati singolarmente, in relazione ad alcune features specifiche (e.g analisi dell'andamento della concretezza per gli outliers di *arousal*). In questo modo abbiamo potuto vedere quanto queste parole "inconsuete" nella variabile considerata, fossero più o meno concrete o astratte.

Formula per identificare gli outliers:

```
Outliers = df.iloc[df['arousal'] > Upperfence1]
```

Questa procedura è stata utilizzata per tutte le variabili aventi outliers, facendo una distinzione, se necessaria, per valori minori del limite inferiore o maggiori del limite superiore.

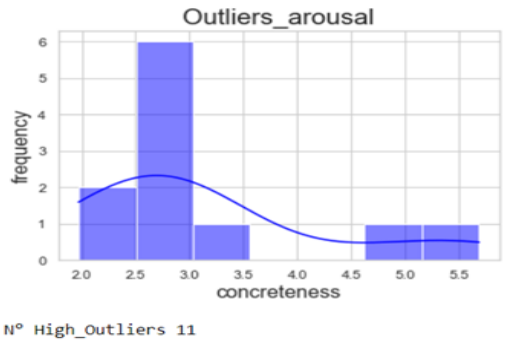


Figura 7: Frequenza outliers di arousal in concreteness

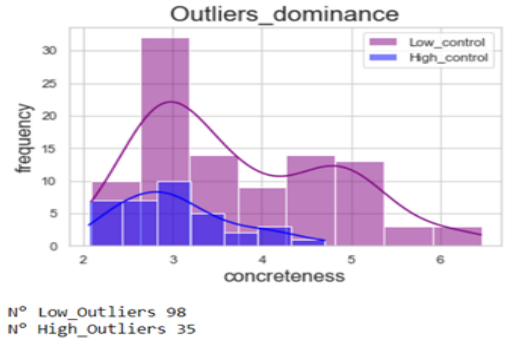


Figura 8: Frequenza outliers di dominance in concreteness

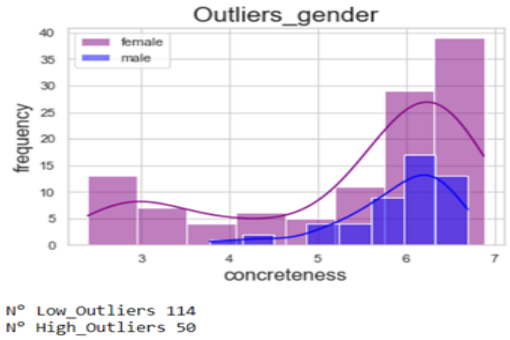


Figura 9: Frequenza outliers di gender in concreteness

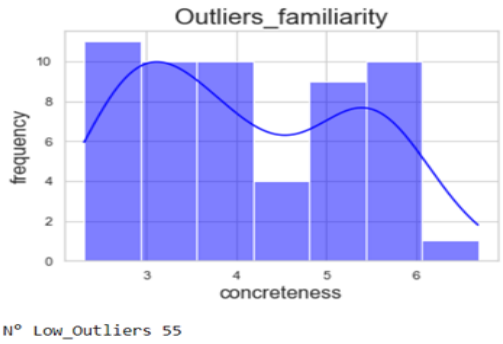


Figura 10: Frequenza outliers di familiarity in concreteness

Dai grafici sopra riportati, si può notare come le parole "eccitanti" abbiano un livello d'astrazione rilevante rispetto alla concretezza, (fig. 7). In merito alla variabile *dominance* (fig. 8) osserviamo lo stesso comportamento, sia per un alto che per un basso grado di controllo (*high_control* e *low_control*): è maggiormente rilevante l'astrazione di queste parole piuttosto che la concretezza. Nel terzo grafico invece notiamo la situazione opposta: sono stati suddivisi gli outliers di *gender* in "*male*" e "*female*", rispettivamente valori elevati e valori più bassi. Indipendentemente dal genere della parola, si evince una distribuzione statistica della concretezza superiore rispetto ai risultati precedenti. Una situazione ibrida invece, riguarda l'ultimo istogramma in cui le parole scarsamente familiari hanno una distribuzione più eterogenea: non vi è alcun tipo di correlazione semantica tra le due variabili.

3.4 Inconsistenze semantiche

Con riferimento all'analisi delle inconsistenze semantiche, nel caso del nostro dataset, si intende una parola che assume un valore inaspettato rispetto alla variabile considerata. Un esempio è la parola "*power*" (it. potenza, forza) che assume una valutazione molto bassa se si considera la variabile *dominance*; l'incongruenza a livello semantico è lampante dal momento in cui il significato della parola stessa evoca l'idea di "dominanza".
Condurre un'esame del genere risulta però molto complesso, in quanto la natura del dataset lascia ampio spazio all'interpretazione personale del valore da assegnare ad ogni attributo. L'analisi è stata dunque limitata alla verifica della coerenza tra la lunghezza della parola nel dataset, riportata nella feature *length* e quella reale: l'output vuoto ottenuto dall'esecuzione dei comandi conferma l'assenza di inconsistenze semantiche.

4 Clustering

Abbiamo proseguito la nostra analisi attraverso il clustering, ovvero processi di raggruppamento di oggetti simili, denominati cluster; le metodologie utilizzate sono: **K-means**, **Hierarchical clustering**, e **DB-SCAN**.
I risultati dei grafici ottenuti dopo la clusterizzazione hanno subito lo stesso approccio di discretizzazione delle variabili come descritto nel cap 3.1. Da notare bene che l'effettiva esecuzione del clustering è stata effettuata con le variabili numeriche e non discretizzate.

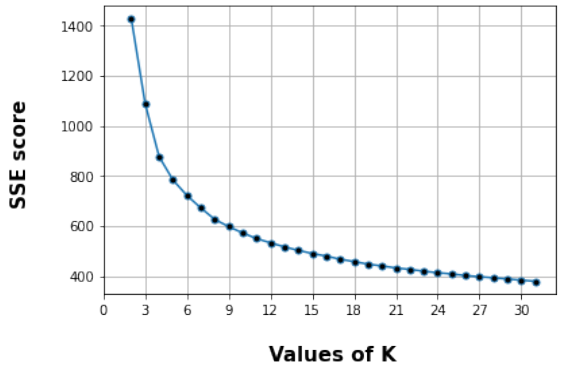
4.1 K-means

Lo studio è stato condotto partendo dal metodo **K-means**, limitandolo soltanto alle variabili semantiche e quindi non considerando *word, length, polisemy* e *web_corpus_freq.* In seguito, i dati sono stati normalizzati con una min-max normalization per poi essere elaborati dall'algoritmo, il quale richiede alcuni

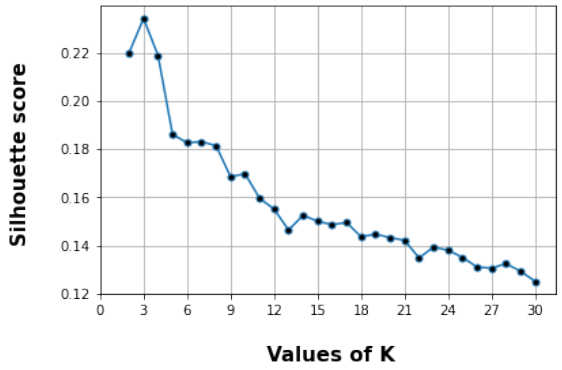
paramentri, tra cui K, ovvero il numero di cluster.

4.1.1 Scelta di K

Per valutare il numero ottimale per K, è stato calcolato l'*SSE* (Sum of Squarred Error) e il *Silhouette Score*:



(a) SSE Elbow method



(b) Silhouette score

Figura 11: SSE e Silhouette

Eseguendo l'algoritmo per un range di valori di K compreso tra 0 e 30, come si può osservare dalla figura 11(a), i risultati migliori si attestano tra 3 e 6, e grazie alla figura 11(b) si nota come il valore ottimale del coefficiente si ha per K=3 a cui corrisponde il valore massimo di *Silhouette*. Nella seguente tabella si riporta una sintesi dei risultati:

N° di cluster	SSE	Silhouette
K = 3	877.30	0.23
K = 4	784.16	0.22
K = 5	722.50	0.19

Tabella 3: Risultati dell'SSE e della Silhouette Score

In base ai risutati ottenuti si è deciso di scegliere K=3 che rappresenta il miglior compromesso tra il numero di cluster e il valore dei due coefficienti. Di seguito si elenca la dimensione per ogni gruppo e si mostrano i valori dei corrispondenti centroidi all'interno del grafico delle coordinate parallele:

1. **Cluster 1**: 1084;
2. **Cluster 2**: 1860;
3. **Cluster 3**: 1738.

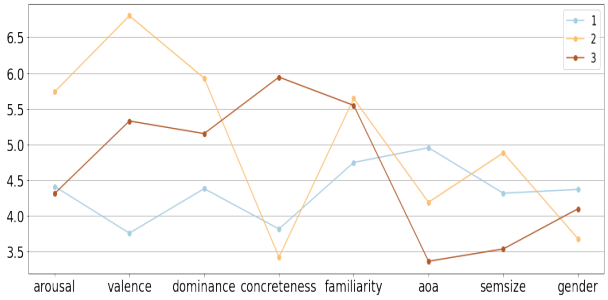


Figura 12: Coordinate parallele

Dalla figura sopra riportata si evince che:

- per il cluster 1 vi è un andamento omogeneo nella distribuzione dei valori associati ad ogni variabile (3.8-5.0);
- il secondo cluster, invece, presenta una tendenza discontinua: in particolare si osservano dei valori molto alti legati alla variabile *valence*, in opposizione alla variabile *concreteness* che presenta dei valori minimi;
- anche nel caso del cluster numero 3 si hanno dei valori discontinui: vi è una

somiglianza per le prime 5 variabili, in seguito si raggiunge il minimo in corrispondenza di *aoa*, continuando con valori ugualmente bassi per *semsize* e *gender*.

In generale, si nota una buona ripartizione nella distribuzione dei cluster per le diverse variabili, ad eccezione di *arousal* per C1 e C3 e di *familiarity* per C2 e C3. La migliore differenziazione dei tre gruppi la si osserva per la variabile *valence*.
Nella figura 13, grazie allo scatterplot, abbiamo una visualizzazione ideale della distribuzione spaziale dei tre cluster prendendo in considerazione le features *valence* e *concreteness*.

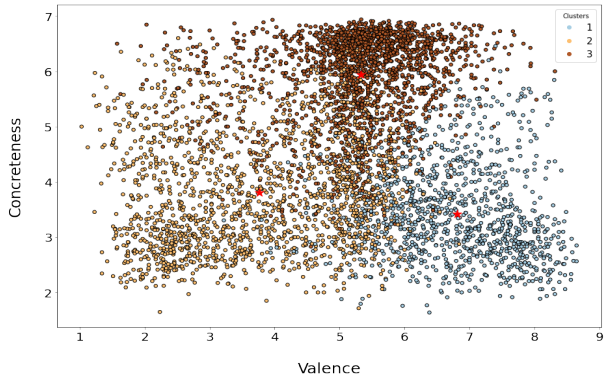
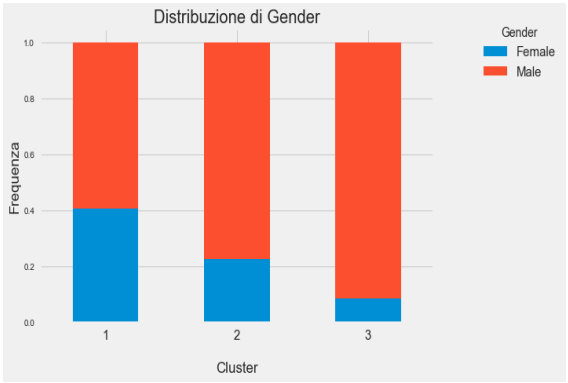


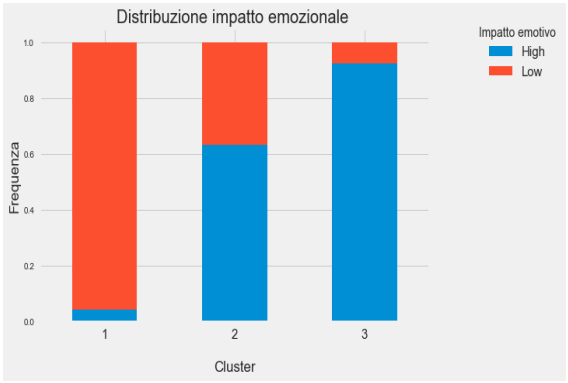
Figura 13: Scatterplot del K-mean cluster

4.1.2 Sintesi dei risultati

Riassumendo i risultati ottenuti, nella distribuzione della variabile *gender* (figura 14 (a)) all'interno dei cluster si può osservare la predominanza di parole di genere maschile. Più nello specifico, se nel cluster 1 si osserva circa il 40% di parole di genere femminile, nel cluster 3 tale percentuale cambia abbassandosi notevolmente, evidenziando la netta prevalenza di *male words*. Nella figura 14 (b), invece, viene mostrata la distribuzione che fa riferimento a parole con un alto o basso impatto emozionale. Per fare ciò è stata creata una nuova feature nel dataset, in cui ogni dato è ricavato dalla media dei valori delle variabili *arousal*, *valence* e *dominance*, che nel loro insieme definiscono l'impatto emotivo di una determinata parola; successivamente la feature è stata discretizzata. Nel caso del primo cluster, vi è la quasi totalità di parole "*low emotional impact*", circa il 90%. D'altra parte invece, il terzo cluster presenta una situazione opposta, in cui vi sono principalmente termini con un "*high emotional impact*". Per il secondo cluster si rimarca una situazione intermedia.



(a) Distribuzione di gender



(b) Distribuzione impatto emozionale

Figura 14: Distribuzione delle variabili nei cluster

4.2 Hierarchical algorithm

Il secondo metodo di clustering utilizzato è quello gerarchico. Le verifiche sono state condotte applicando la tipologia *Agglomerative* sulla base della definizione di distanza Euclidea e di clustering proximity (**Single, Average, Complete** e **Ward**).

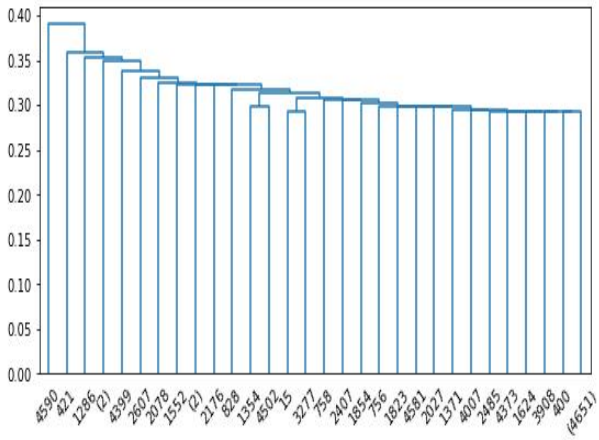


Figura 15: Dendrogramma Single link

Inizialmente è stato utilizzato il **Single** linkage il quale si è rivelato il peggiore per il nostro studio in quanto raggruppa più di 4500 punti in un unico cluster. Si osserva infatti, quanto le distanze tra ognuno di essi siano molto simili, tanto che la radice dell'albero dista dal precedente raggruppamento soltanto di 0.4, il che vuol dire che il distacco dei sottostanti non sarà mai maggiore di tale valore. Per questa ragione non si è ritenuto opportuno svolgere ulteriori analisi.
Il motivo di ciò è la rilevante densità dei punti del dataset che essendo così vicini non permettono una buona clusterizzazione con il **Single** method. Si è quindi proceduto utilizzando l'**Average** method, che risulta migliore, ma comunque non fornisce una differenziazione dei cluster chiara e distinta. Una visualizzazione simile, si è ottenuta con il **Complete** method, da cui si possono trarre conclusioni analoghe a quelle precedenti.
Infine, per progressiva miglioria di clusterizzazione, utilizzando l'approccio gerarchico, i risultati più efficaci sono stati riconosciuti tramite il **Ward** method (figura 16) che si basa sull'agglomerazione dei gruppi, in termini incrementali di SSE (analogo al *K-Means*):

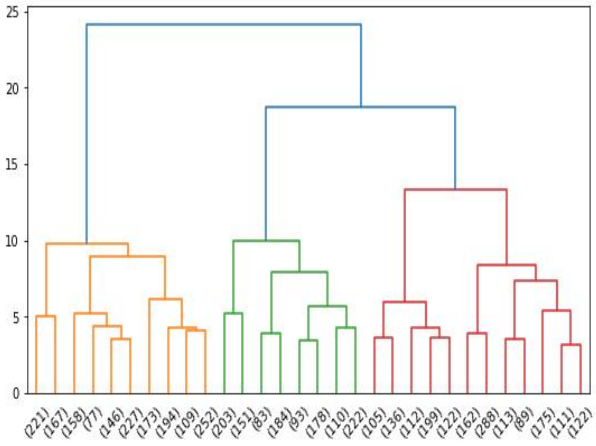


Figura 16: Dendrogramma Ward's method

Come si può osservare dal grafico soprastante, il dataset viene suddiviso in 3 cluster ben distinti. In merito alla Silhouette score si nota come i valori più alti, circa 0.18, si attestano per un numero di cluster uguale a 2 o 3, confermando anche la scelta di K del *K-means* oltre a quella ottenuta tramite il suddetto metodo.

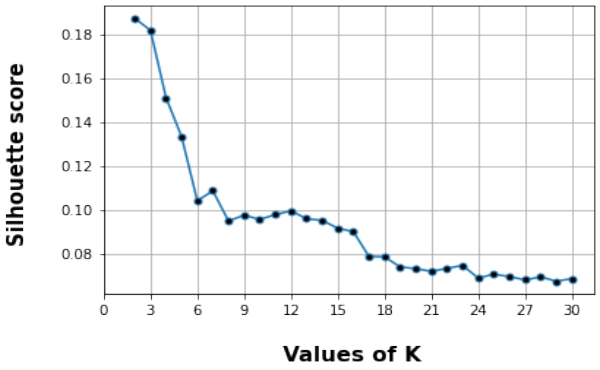
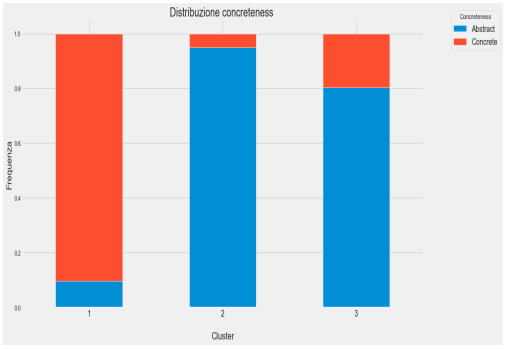


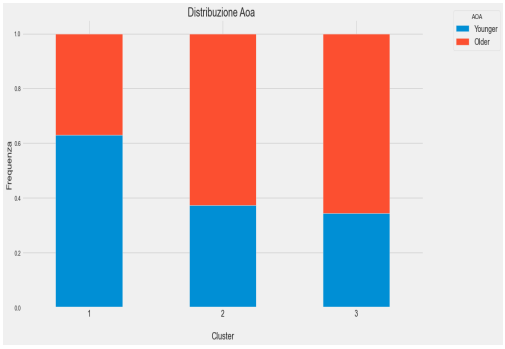
Figura 17: Silhouette Score per il Ward's method

4.2.1 Sintesi dei risultati

Con il secondo metodo di clustering sono state analizzate le distribuzioni delle variabili *concreteness* e *aoa*. Per quanto riguarda la prima, figura 18(a), appare evidente una consistente percentuale di parole considerate concrete nel primo cluster, in contrapposizione a quanto è possibile notare negli altri due, soprattutto nel secondo dove si raggiunge la totalità di astrattezza.



(a) Distribuzione Concreteness



(b) Distribuzione Aoa

Figura 18: Distribuzione delle variabili nei cluster

La seconda figura, 18(b), mostra invece una distribuzione più omogenea all'interno dei tre cluster per la variabile *aoa*. Per una visualizzazione ottimale dei risultati abbiamo deciso di realizzare una partizione in classi: "*younger*" per parole apprese prima dei 12 anni e "*older*" per quelle apprese dopo i 13.

Così facendo è stato possibile notare una netta somiglianza delle due classi all'interno dei cluster 2 e 3. Nel primo invece, si osserva una percentuale maggiore di parole acquisite nella prima fascia d'età, da 0-12 anni, ovvero poco più del 60%.

4.3 DBSCAN

Il **DBSCAN** è il terzo ed ultimo algoritmo che è stato implementato ai fini della nostra ricerca; si tratta di un metodo *density-based* che permette di individuare il numero dei cluster di un dataset raggruppando tutti i punti che si trovano in prossimità di un raggio epsilon. In primo luogo, tramite la visualizzazione grafica del **KNN algorithm** (fig. 19) si è evinto che il valore ottimale di EPSILON per il nostro dataset è compreso tra 1.2 e 1.5. In seguito, massimizzando l'SSE si è trovato il numero ottimale di *min_sample* che risulta essere 10 (fig. 20).

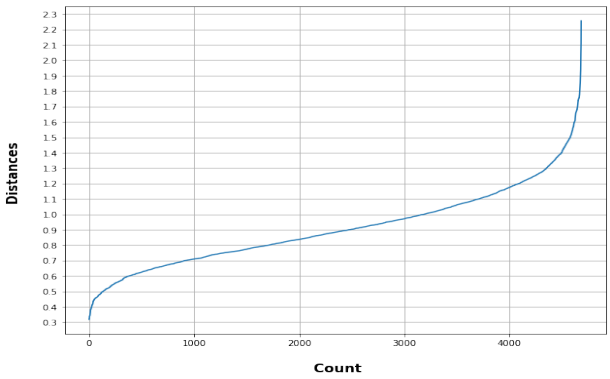


Figura 19: Nearest Neighbors

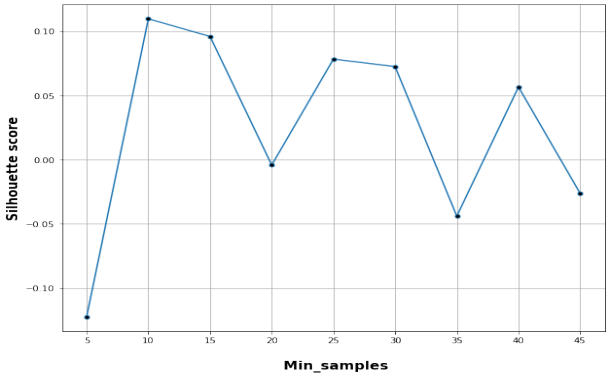


Figura 20: SSE per DBSCAN

Modificando il valore di epsilon all'interno del nostro intervallo utile, possiamo notare come il numero di cluster non muta al variare di epsilon, tranne nel caso in cui il valore sia 1.3 (fig 21) dove troviamo un cluster in più.

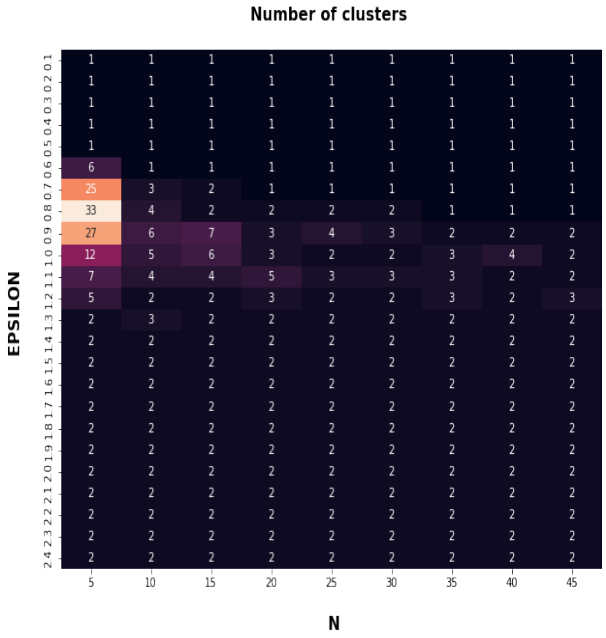
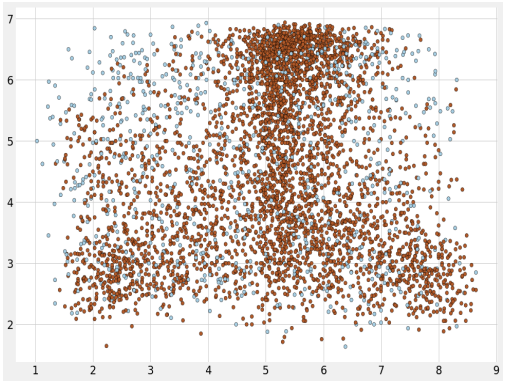


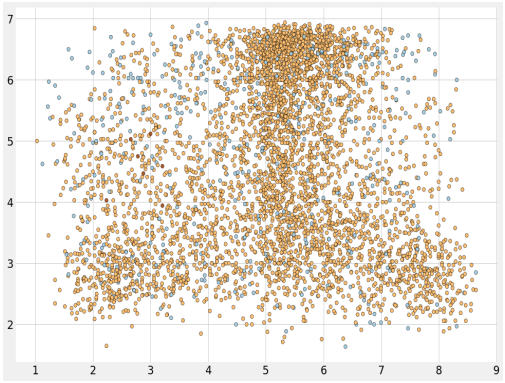
Figura 21: Numero di cluster al variare di N e EPSILON

4.4 Sintesi dei risultati

Il DBSCAN si è rivelato poco adatto al nostro dataset in quanto la densità dei cluster individuati risulta eterogenea, portando dunque ad una distribuzione delle variabili non uniforme all'interno degli stessi. Inoltre, come possiamo osservare nella figura 22 (b), seppur difficilmente, si identifica un terzo cluster: esso risulta appunto poco distinguibile e di bassa densità, in quanto composto da pochi punti sparsi.



(a) Cluster individuati con Eps = 1.2



(b) Cluster individuati con Eps = 1.3

Figura 22: Cluster al variare di Eps

5 Classification

La classificazione è un metodo supervisionato che ha la funzione di assegnare degli oggetti di un dataset a delle categorie predefinite (classi); oltre a ciò, tale procedimento permette la predizione di eventuali *missing values* o valori futuri. In riferimento alla ricerca da noi

condotta, lo scopo di tale procedura è stato quello di predire i valori della variabile *polysemy*, ovvero riuscire a capire, in base agli altri dati del dataset, se una parola fosse polisemica o meno (1 o 0).
Prima di cominciare è stato necessario effettuare una fase di "pre-processing", sostituendo eventuali *missing values* e rimuovendo gli attributi non necessari al task. Oltre a ciò, considerata la natura sbilanciata del nostro dataset rispetto ai valori della variabile *polysemy*, si è dovuto ricorrere a procedure definite *sampling*, ovvero metodi di bilanciamento.

5.1 Preparazione del dataset

Nel nostro caso, si è optato per:

1. Sostituire i *missing values* presenti nella variabile *web_corpus_freq* tramite interpolazione lineare;
2. Eliminare dal dataset le variabili *Word*, *Imageability* e *Web_corpus_freq*.

Per fare ciò il dataset è stato suddiviso in **training set** (70%) tramite cui viene generato e allenato il modello predittivo, e **test set** (30%) per mezzo del quale viene verificata l'efficacia dello stesso, constatando la veridicità dei valori effettivi con quelli previsti dal modello. Per quanto riguarda la procedura di bilanciamento, essa è stata effettuata tramite *Oversampling* e *Synthetic Minority Oversampling Technique* (SMOTE), bilanciando le classi in rapporto 1:3 (30%), in modo da evitare una possibile distorsione dei dati causata da una percentuale di bilanciamento più consistente.

5.2 Apprendimento dei diversi algoritmi di classificazione e validazione dei modelli

Gli algoritmi di classificazione che sono stati utilizzati in queste analisi sono il **Decision Tree**, il **Random Forest** e il **K-Nearest Neighbor** (**KNN**). Per ognuno di questi sono stati applicati differenti parametri e criteri per evitare eventuali problemi di *overfitting*. Nella tabella che segue sono riportati i parametri con i relativi valori utilizzati nelle diverse prove effettuate attraverso la funzione **grid_search_estimator**, in modo da trovare la migliore combinazione possibile per massimizzare l'efficacia del modello costruito:

Parametri	**Valori**
Criterio	Gini, Entropy
min_sample_leaf	1 - 300
min_sample_split	2 - 300
max_depth	None, 0 - 30

Tabella 4: Criteri utilizzati per la classificazione

5.2.1 Decision Tree

Per il primo classificatore implementato, il *Decision Tree*, è stata presa in esame la misura della F1-score la quale, essendo la media armonica di **precision** e **recall**, riesce a fornire una buona valutazione complessiva del modello. I principali risultati ottenuti nelle analisi svolte sono riportati nella tabella n°5.

Dai valori presenti si può notare come, in tutti i casi, nel dataset non bilanciato l'F-1 score per la classe *yes* sia sempre il più basso: infatti esso non supera mai il 22%. Ciò

accade perché nel dataset è presente una percentuale di valori della classe *no* pari al 93% e *yes* del 7%, (a conferma della presenza di 379 parole polisemiche nel *Glasgow Norms* Dataset), evidenziando un chiaro sbilanciamento della variabile.

Crit.	Parametri	Class	F-1 score non bil.	F-1 score OS	F-1 score SMOTE
GINI	split 3	no	0,93	0,95	0,93
	leaf 6	yes	0,17	0,11	0,21
	depth 19				
	split 16	**no**	**0,95**	**0,95**	**0,94**
	leaf 15	**yes**	**0,11**	**0,11**	**0,26**
	depth None				
	split 5	no	0,94	0,95	0,93
	leaf 7	yes	0,21	0,11	0,22
	depth 19				
ENTROPY	split 3	no	0,93	0,93	0,92
	leaf 6	yes	0,22	0,11	0,21
	depth 19				
	split 16	**no**	**0,95**	**0,95**	**0,93**
	leaf 15	**yes**	**0,15**	**0,11**	**0,21**
	depth None				
	split 5	no	0,94	0,95	0,93
	leaf 7	yes	0,22	0,11	0,22
	depth 19				

Tabella 5: Risultati ottenuti sulla base dell'F-1 score per il Decision Tree

In ogni caso, sebbene la misura dell'accuratezza sfiori il 90% nel dataset non bilanciato, questo potrebbe portare a pensare che il modello di classificazione funzioni regolarmente, ovvero che 9 volte su 10 viene predetta una parola polisemica. In realtà, è necessario effettuare un'analisi più dettagliata sui risultati ottenuti: se si prendono in considerazione le altre misure ottenute con il dataset non bilanciato, risulta evidente che le parole polisemiche vengano riconosciute correttamente solo nel 14% dei casi. Per questo motivo, si è optato per sovrabilanciare il dataset ed allenare il modello con una percentuale più equilibrata delle due classi, nel nostro caso in rapporto 1:3.

Attraverso la procedura di *sampling* è possibile osservare un miglioramento. Non a caso nella tabella si può osservare che i migliori risultati si ottengono utilizzando il dataset bilanciato attraverso l'*Oversampling* e lo *SMOTE*, sia per misura dell'impurità con *Gini* che con *Entropy*, in entrambi i casi con i parametri:

1. *min_sample_split* = 16;
2. *min_sample_leaf* = 15;
3. *max_depth* = None.

In conclusione, per valutare la configurazione migliore dell'albero decisionale si è confrontato e analizzato il *classification report* con la migliore terna di parametri indicata precedentemente. Secondo le stime di *accuracy*, *recall* e *precision*, **Gini** risulterebbe essere il miglior criterio per il dataset non bilanciato; in particolare i risultati migliori per i 3 score sopracitati si sono ottenuti tramite **Gini** con *SMOTE* nel dataset.

5.2.2 Interpretazione dell'albero di decisione

Come stabilito, il miglior modello per l'interpretazione dell'albero decisionale risulta essere quello ottenuto attraverso lo *SMOTE*, applicando l'indice di **Gini** e con dei valori di *min_sample_split* e *min_sample_leaf* pari rispettivamente a 16 e 15.

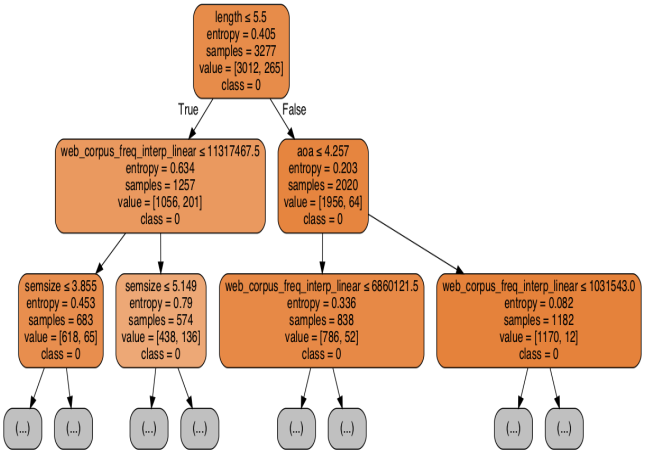


Figura 23: Albero di decisione

La figura 23 mostra parte dell'albero ottenuto, dove si può osservare che i criteri di *split* si basano principalmente sui valori delle variabili *web_corpus_freq*, ma anche sulla lunghezza delle parole e sul loro valore di *gender* e *aoa*. Più nello specifico il primo *split* fa riferimento a parole con una lunghezza minore uguale o maggiore di 5.5, che sono in seguito divisi al successivo livello nel ramo destro in base al valore di *web_corpus_freq* ed in quello sinistro in base a quello di *aoa*.
A questo punto in tutte le ramificazioni, eccetto una, la quale splitta su *gender*, si suddivide nuovamente sull'attributo *web_corpus_freq*.

5.2.3 Random Forest

Ugualmente, per il *Random Forest* sono state utilizzate le stesse modalità di indagine del *Decision Tree*: la tabella n°6 riporta i vari risultati ottenuti.

Come in precedenza, i risultati del dataset non bilanciato non appaiono significativi, in quanto la classe *yes* presenta sempre valori uguali a zero; la situazione cambia notevolmente con l'F-1 score bilanciato presentando di nuovo i risultati ottimali attraverso il criterio **Entropy** con i parametri *min_sample_split* pari a 16, *min_sample_leaf* uguale a 15 e una profondità pari a *None*.

Crit.	**Parametri**	**Classe**	**F-1 score non bil.**	**F-1 score OS**	**F-1 score SMOTE**
Gini	split 3	no	0,96	0,95	0,95
	leaf 6	yes	0	0,13	0,21
	depth 19				
	split 16	**no**	**0,96**	**0,95**	**0,95**
	leaf 15	**yes**	**0**	**0,24**	**0,24**
	depth None				
	split 5	no	0,96	0,95	0,94
	leaf 7	yes	0	0,21	0,21
	depth 19				
Entropy	split 3	no	0,96	0,95	0,95
	leaf 6	yes	0	0,16	0,20
	depth 19				
	split 16	**no**	**0,96**	**0,95**	**0,95**
	leaf 15	**yes**	**0**	**0,27**	**0,25**
	depth None				
	split 5	no	0,96	0,95	0,95
	leaf 7	yes	0	0,20	0,22
	depth 19				

Tabella 6: Risultati ottenuti sulla base del F-1 score per il Random Forest

5.2.4 KNN algorithm

L'ultimo modello di classificazione utilizzato è il **K-Nearest Neightbor** (KNN), implementando il modello al variare del parametro *n-neighbor* (3, 5, 10 e 15). Anche in questo caso è stato fatto riferimento alla misura dell'F-1 score per individuare il risultato migliore:

Parametri	**Classi**	F-1 non bil.	F-1 score Oversampling	F-1 score SMOTE
n_neighbor = 3	no	0,94	0,89	0,90
	yes	0,06	0,14	0,14
n_neighbor = 5	no	0,95	0,88	0,90
	yes	0,03	0,15	0,14
n_neighbor = 10	no	0,96	0,91	0,93
	yes	0	0,16	0,13
n_neighbor = 15	no	0,96	0,91	0,92
	yes	0	0,13	0,14

Tabella 7: Risultati ottenuti con F-1 score per il KNN algorithm

Ancora una volta, anche per il **KNN alogrithm** attraverso un F-1 score non bilanciato, si ottengono risultati non sensibili alla classe *yes*. Questa condizione cambia attraverso l'utilizzo dell'F-1 score in seguito a *Oversampling* e *SMOTE*, che hanno portato a un miglioramento dei valori, seppur non paragonabile a quelli ottenuti precedentemente. Concludendo, i dati ideali si hanno ponendo il parametro *n_neighbor* pari a 10.

5.3 Sintesi dei risultati e miglior modello di previsione

Un'ulteriore valutazione dei modelli osservati finora è avvenuta considerando le **Receiver Operating Characteristic Curve** (**ROC Curve**), infine comparate con i valori migliori ottenuti dai tre classificatori riportati nella tabella n°8.

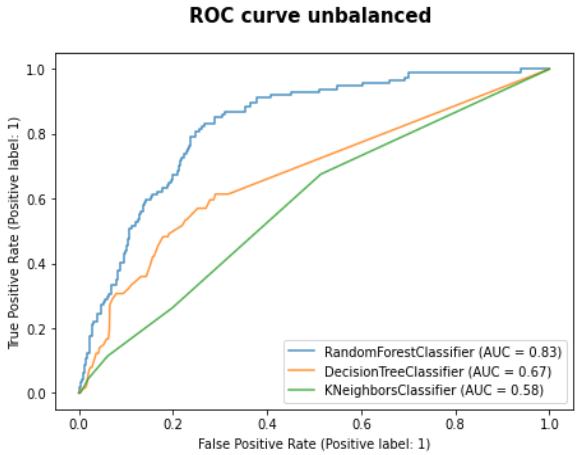


Figura 24: ROC Curve dei diversi algoritmi per il dataset pre bilanciamento

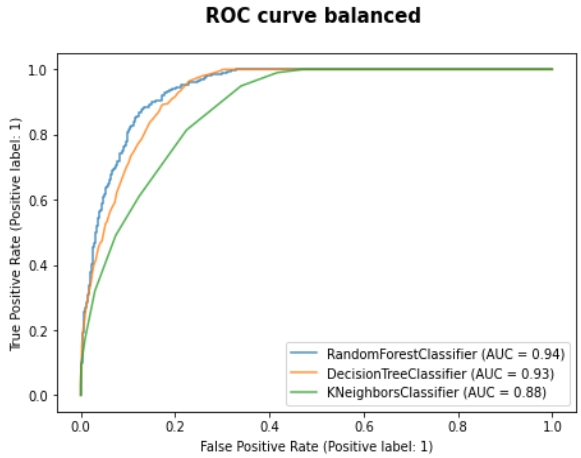


Figura 25: ROC Curve dei diversi algoritmi per il dataset post bilanciamento

Come è possibile osservare, le ROC curve del dataset non bilanciato mostrano dei risultati meno accurati rispetto a quelli ottenuti in seguito al *sampling*. Tuttavia, per entrambi, l'algoritmo meno performante risulta essere il **KNN**, in quanto nel primo caso il valore dell'**Area Under the Curve** (**AUC**) equivale a 0,58 (valore molto vicino alla diagonale), mentre nel secondo raggiunge lo 0,88. Un evidente miglioramento della performance è osservabile nel **Decision Tree**, dove si raggiunge un AUC dello 0,93 rispetto allo 0,67 ottenuto precedentemente. In conclusione, considerando il dataset bilanciato, gli algoritmi che risultano essere più efficaci sono il **Random Forest Classifier** (AUC= 0,94) e il **Decision Tree Classifier** (AUC = 0,93): ciò si nota dalla ripidità iniziale delle rispettive curve le quali evidenziano un elevato tasso di *true positive* rispetto ai *false positive*.

Per terminare quest'analisi, sono stati riportati nella tabella che segue i valori migliori ottenuti dai tre classificatori testati:

Alg.	Acc.	Class	Prec.	Recall	F1score	DB
DT	0,89	No Yes	0,93 0,28	0,95 0,24	0,94 0,26	SMOTE
RF	0,90	No Yes	0,93 0,34	0,97 0,18	0,95 0,27	OS
KNN	0,84	No Yes	0,92 0,12	0,90 0,15	0,91 0,16	OS n_n=10

Tabella 8: Confronto tra i migliori risultati ottenuti

I risultati ricavati confermano l'andamento delle ROC curve, e per tal motivo è stato scelto come miglior modello di classificazione il **Random Forest** in grado di predire al meglio i valori della variabile target definita in precedenza, ovvero *polysemy*. Un'ennesima conferma la si ottiene dalla *Confusion Matrix* del suddetto modello in cui si possono evincere valori quasi ottimali con soli 143 *misclassified*, i quali equivalgono, all'interno del nostro dataset, solo ad un 10% di errore.

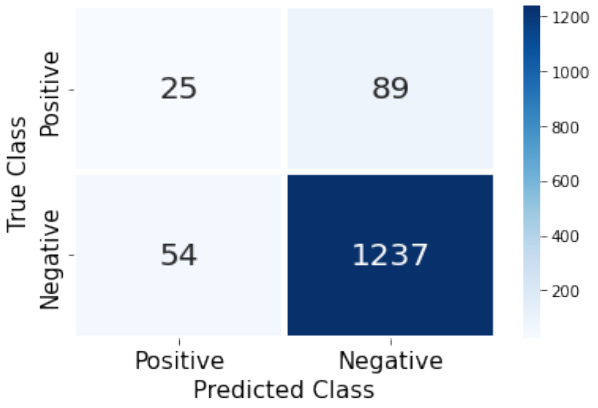


Figura 26: Confusion Matrix del Random Forest

6 Pattern mining e association rules

In questa sezione finale sono individuate ed analizzate le regole di associazione trovate nel dataset, ponendo particolare attenzione all'attributo *polysemy*.

6.1 Preparazione del dataset

Per l'analisi sono stati presi in considerazione i seguenti attributi: *length*, *arousal*, *valence*, *dominance*, *concreteness*, *familiarity*, *aoa*, *gender*, *web_corpus_freq* e *polysemy*. Nello specifico, tutte le variabili, ad eccezione di *polysemy* e *length*, sono state discretizzate: è stata realizzata una discretizzazione su due livelli per *arousal*, *valence*, *dominance*, *concreteness*, *familiarity* e *aoa* suddividendo i valori in due gruppi differenti, *high* e *low*. Per la variabile *gender* si è optato, invece, per una suddivisione su 3 livelli, rappresentando le categorie *male*, *unisex* e *female*. Infine, per quel che riguarda la variabile *web_corpus_freq*, è stata realizzata una ripartizione su 4 livelli indicanti una scarsa, bassa, media o alta frequenza delle parole nel *Google Newspaper Corpus*.

6.1.1 Estrazione dei frequent itemset e valutazione del supporto minimo

Come primo step, è stato individuato il parametro di *supporto minimo* (min_sup) ideale per la rappresentazione del dataset; per fare ciò, si è optato per mettere in relazione il numero di itemset prodotti con la soglia di supporto minimo, facendola variare in un intervallo compreso tra 0 e 100.

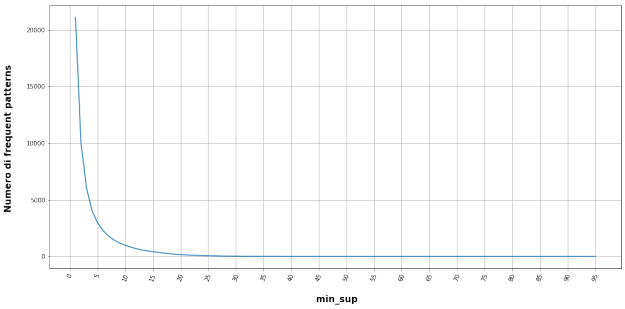


Figura 27: Numero dei Frequent Itemset al variare di min_sup

Si può osservare come la curva ottenuta (fig. 27) presenta il "*gomito*" in valori compresi tra 2 e 10. Di conseguenza, è stato calcolato il numero di itemset in relazione alle differenti soglie di supporto minimo identificate nell'intervallo sopra specificato (tabella n°9):

Min_Support	Numero Itemset
2	9989
4	4044
6	2286
8	1450
10	983

Tabella 9: Frequent Itemset individuati per diversi valori di min_sup

Considerando i risultati ottenuti, si è deciso di utilizzare un *min_support* pari a 2, in modo da poter comprendere un maggior numero di itemset frequenti per l'analisi; questi sono riportati nella tabella 10.

Len	Susp	Itemset
2	43.986	('web_corpus_no_freq', 'gender_unisex')
3	29.047	('familiarity_low', 'aoa_high', 'web_corpus_no_freq')
4	20.034	('familiarity_low', 'aoa_high', 'web_corpus_no_freq', 'gender_unisex')
5	13.519	('familiarity_low', 'aoa_high', 'concreteness_low', 'web_corpus_no_freq', 'gender_unisex')
6	10.038	('valence_low', 'familiarity_low', 'dominance_low', 'aoa_high', 'web_corpus_no_freq', 'gender_unisex')
7	7.326	('valence_low', 'familiarity_low', 'dominance_low', 'aoa_high', 'concreteness_low', 'web_corpus_no_freq', 'gender_unisex')
8	4.912	('valence_low', 'familiarity_low', 'dominance_low', 'aoa_high', 'concreteness_low', 'arousal_low', 'web_corpus_no_freq', 'gender_unisex')

Tabella 10: Lista degli itemsets più frequenti con lunghezze da 2 a 8

Dai risultati presenti in tabella notiamo che l'itemset di lunghezza 5 contiene tutti gli elementi degli itemsets di lunghezza 3 e 4, avente come ulteriore elemento *concreteness_low*. Ciò che si può evincere da questa analisi è che la maggior parte delle parole all'interno del dataset linguisitico *Glasgow Norms*, presenta le seguenti caratteristiche:

- suscita un basso senso di familiarità (*familiarity_low*);
- sono probabilmente termini acquisiti a un'età più "avanzata" (*aoa_high*);
- sono parole che esprimono concetti più astratti (*concreteness_low*);
- appaiono raramente, con una frequenza minore del 25%, nel *Google Newspaper Corpus* (*web_corpus_no_freq*);
- sono principalmente termini neutrali in quanto non appartengono né alla categoria *male* né a quella *female* (*gender_unisex*).

6.1.2 Estrazione delle Association Rules con diversi valori di confidence e discussione delle regole più interessanti

Il passo successivo è stato quello di far variare il valore della *min_conf* partendo da un minimo di 5 fino ad arrivare ad un massimo di 95; la relazione tra le Association Rules e tale valore è riportata nella figura sottostante.

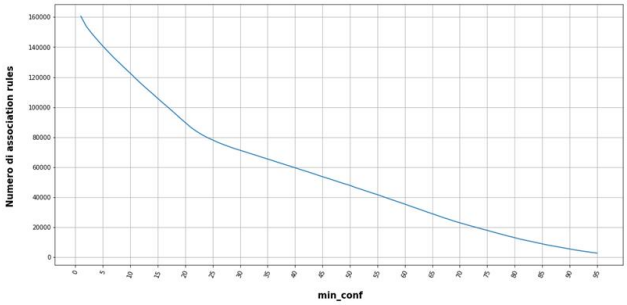


Figura 28: Numero di Association Rules per livelli di confidence

Nella figura 29 sono riportati i valori delle Association Rules per delle misure di confidence che vanno da 80 a 100, così da individuare con maggior dettaglio quale sia l'andamento di queste.

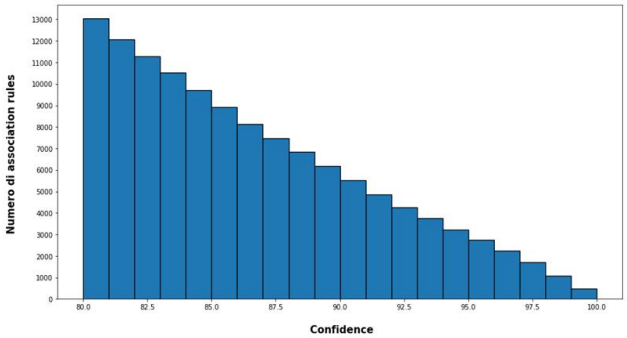


Figura 29: Andamento del valore di confidence

Come è possibile osservare, il maggior numero di regole individuate si ottiene con una *confidence* equivalente a 80, in quanto aumentandone il valore, le AR diminuiscono gradualmente.
Successivamente è stata analizzata la correlazione che intercorre tra i valori con confidence e support, riportata nella fig. 30, la quale mostra che il numero più elevato di Association Rules si ottiene con una confidence abbastanza elevata (circa 80%) e con una misura di support piuttosto bassa (0-5%).

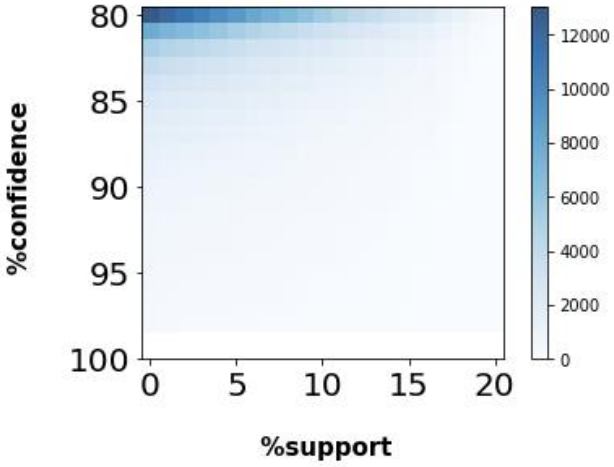


Figura 30: Correlazione tra confidence e support

Attraverso un istogramma è stato valutato l'andamento della *Lift* al variare del numero di Association Rules, ponendo i parametri di supporto pari a 2 e di confidence pari a 80. Si presenta una forte concentrazione di regole intorno a un Lift pari a 1, segno evidente della presenza di regole con antecedenti e conseguenti indipendenti. Molto più interessanti sono invece le regole con una Lift maggiore di 1,5 da cui si nota una maggiore correlazione tra di esse.

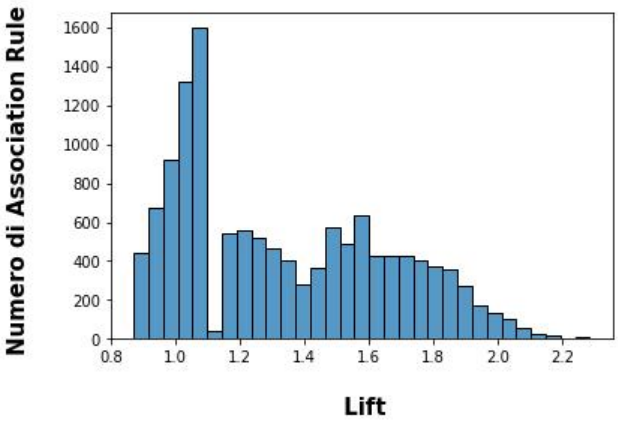


Figura 31: Variazione della lift in base al numero di AR

6.2 Sostituzione dei missing values tramite AR e valutazione della relativa accuratezza

In questa sezione sono stati sostituiti i valori mancanti presenti nel dataset e localizzati nella variabile *web_corpus_freq*, i quali in precedenza, nella fase di Data Preparation (cap. 3.2) sono stati rimpiazzati tramite *interpolazione lineare*. L'obiettivo è stato quello di verificare se l'item più ricorrente, i.e. *web_corpus_no_freq* fosse effettivamente il più "frequente" all'interno delle altre *Association Rules* estratte.

Missing Values	Quantità
web_corpus_no_freq	10
web_corpus_low_freq	2
web_corpus_freq	2
web_corpus_high_freq	0

Tabella 11: Numero dei missing values e la loro ripartizione nella variabile web_corpus_freq

Web corpus	Totale	Frequenza in Web_corpus	Frequenza in AR
no_freq	3794	68,27%	29,1%
low_freq	352	6,33%	2,7%
freq	1411	25,4%	10,8%
high_freq	0	0%	0%

Tabella 12: Frequenza dei valori della variabile *web_corpus_freq*

Osservando i risultati nella tabella n°12, si evince immediatamente che i valori compresi in **web_corpus_no_freq** rispecchiano la classe maggioritaria; in particolare la loro frequenza risulta pari al 68,27% del totale. Comparando questi risultati con la sostituzione dei *missing values* effettuata nella sezione del Data Understanding si nota come essi siano stati sostituiti con un'alta percentuale di correttezza, in quanto su 14 dei totali valori mancanti, 10 risultano appartenere alla classe **web_corpus_no_freq**.

6.3 Previsione della variabile target e valutazione dell'accuratezza

Attraverso l'utilizzo delle regole prodotte nelle precedenti analisi dovrebbe risultare possibile predire le due classi dell'attributo target *Not Polysemic* e *Polysemic*. Tuttavia, a causa dello sbilanciamento del dataset non è stato possibile individuare regole contenenti l'attributo *Polysemic*. D'altra parte invece, per quanto riguarda l'attributo *Not Polysemic* sono state individuate 4961 Association Rules in grado di descrivere parole che non risultano polisemiche. Tale sbilanciamento a favore della classe *Not Polysemic* non permette di predire le sfumature che caratterizzano le parole polisemiche, per tal motivo si è ricorso al bilanciamento del dataset, come effettuato nella fase di Classification (cap.5).

Tale bilanciamento è stato svolto soltanto con *Up-sampling*, ovvero attraverso la previsione di una duplicazine casuale delle osservazioni della classe di minoranza (nel caso in questione della classe *Polysemic*) al fine di rafforzare il segnale. Nonostante il bilanciamento ed una nuova esecuzione dell'algoritmo *Apriori* non si sono ottenuti miglioramenti, quindi non è stato possibile trovare regole aventi come *conseguente Polysemic*. Si è quindi riusciti a classificare le parole solo come non polisemiche.