Data Mining & Machine Learning
Report di gruppo

Giovanni Cusumano Martina Formichini Giuseppe Minardi
Gilda Pepe Carmela Sgroi

Maggio 2021

INDICE

Introduzione

Il dataset contiene dati ricavati dal social network OkCupid e presenta 2,001 record con 21 attributi (tabella 1). Due di questi (user_id e username), tuttavia, sono stati immediatamente eliminati in quanto non rilevanti ai fini di una indagine sul dataset. L'obbiettivo di questo lavoro è di fare un'iniziale analisi esplorativa sui dati, seguita da clustering, pattern mining ed infine classificazione. In particolare, a seguito di una osservazione approfondita del dataset e di una serie di prove, abbiamo deciso di costruire un classificatore per la predizione del sesso dell'utente.

Data understanding

Assessing data quality

Il dataset presenta soltanto due features con missing values: height e location_preference. Al fine di eliminare i missing values, abbiamo sostituito i valori di height con il valore medio raggruppato per sesso; per la variabile location_preference, invece, abbiamo inserito "anywhere" nei casi in cui questa mancava in quanto abbiamo assunto che l'utente non intendeva esprimere preferenze particolari.

Data semantics

Da un'analisi preliminare del dataset abbiamo visto come tutte le variabili siano categoriche ad eccezione di age e height che sono continue.

Age e height

Per quanto riguarda le due variabili numeriche presenti nel nostro dataset, abbiamo plottato la distribuzione della variabile height facendo uno scatterplot con age allo scopo di accertarci dell'esistenza o meno di una correlazione tra le uniche due variabili continue (figura 2). Quello che emerge è che le due variabili numeriche non sono correlate. Non ci sono differenze sostanziali fra uomini e donne per quanto riguarda l'età, ma, come ci aspetteremmo, la distribuzione dell'altezza è molto diversa per entrambi i sessi.

Location

Dal momento che la variabile location è in realtà composta da due variabili, ovvero lo Stato e la città, abbiamo per prima cosa diviso la variabile in modo tale da poter meglio procedere con il data understanding.

Contando gli individui per Stato notiamo un forte sbilanciamento: mentre la California raccoglie la stragrande maggioranza degli utenti, ciascuno degli altri Stati (Arizona, New York, Ohio, Vietnam) presenta un solo elemento. Andando ad analizzare la location_preference, quello che emerge è che chi vive in Ohio, Arizona, New York e in Vietnam vuole uscire solo con persone o della stessa città o dello stesso Stato. Anche per quanto riguarda chi vive in California, la maggioranza delle persone (987) esprime una preferenza per la propria città e il proprio Stato, 416 utenti indicano invece preferenza soltanto per lo stesso Stato, mentre 594 non manifestano alcuna preferenza.

Education Level

Per quanto riguarda la variabile education_level, i cui valori vanno da 1.0 a 5.0 in ordine di livello educativo crescente, abbiamo raggruppato per sex e dropped out. Dai dati è emerso come siano in particolare gli uomini ad avere un tasso maggiore di abbandono degli studi rispetto alle donne. Tuttavia, va ricordato che i dati in questione non sono stati normalizzati e gli uomini nel nostro dataset sono in maggior numero (tabella 2).

Body Profile

La variabile body_profile è composta dai seguenti valori: "a little extra", "athletic", "average", "curvy", "fit", "full figured", "jacked", "overweight", "rather not say", "skinny", "thin", "used up". Come prima cosa abbiamo raggruppato la variabile body_profile per sesso. Dai risultati ottenuti è stato possibile notare come alcuni valori siano fortemente caratterizzati per genere. Due esempi sono "curvy", utilizzato quasi esclusivamente da donne, e "a little extra", usato invece tipicamente da uomini. È anche emerso che in generale gli uomini sono più atletici o con body_profile nella norma rispetto alle donne, le quali si dichiarano maggiormente in sovrappeso rispetto agli uomini. Inoltre, altri valori quali "jacked", "used up", "overweight" e "rather not say" sono scarsamente popolati. Per tali ragioni, siamo giunti alla conclusione che body_profile potrebbe essere un'ottima variabile per la predizione del sesso dell'utente.

Tabella 1: Features presenti nel dataset

Feature	Tipo	Descrizione
age	ratio	Età dell'utente.
status	categorica	Se impegnato, sposato, in una relazione o libero.
sex	categorica	Sesso dell'utente.
orientation	categorica	Orientamento sessuale dell'utente.
drinks	categorica	Abitudini nel consumo di bevande alcooliche.
drugs	categorica	Abitudini nel consumo di droghe.
height	continua	Altezza dell'utente.
location	categorica	Luogo di residenza dell'utente.
pets	categorica	Preferenza per cani e gatti.
smokes	categorica	Abitudini di fumo dell'utente.
language	categorica	Lingue parlate con annesso livello.
new_languages	categorica	Se l'utente è disponibile a imparare nuove lingue.
body_profile	categorica	Costituzione fisica dell'utente.
education_level	categorica	Scolarizzazione dell'utente.
dropped_out	categorica	Se l'utente ha lasciato la scuola.
bio	testo	Biografia scritta dall'utente.
interests	categorica	Interessi dell'utente.
other_interests	categorica	Altri interessi dell'utente.
location_preference	categorica	Luogo di incontro preferito dall'utente.

Tabella 2: Numero di persone che ha lasciato la scuola, divise per livello di educazione e sesso

	Not dropped out		Dropped Out	
Education Level	Female	Male	Female	Male
1	10	20	9	33
2	167	258	14	38
3	411	589	1	4
4	233	171	3	1
5	14	25	0	0

Orientation, Sex e Status

Per quanto riguarda la variabile "orientation" i valori presenti sono: "bisexual", "gay" e "straight" mentre lo status presenta i valori "available", "married", "seeing someone" e "single". In questo processo di data understanding siamo andati quindi a vedere come sono distribuite le persone in base all'orientamento, allo stato relazionale e al sesso.

In tabella 3, come ci si aspetterebbe da un sito di dating, abbiamo osservato che la stragrande maggioranza degli utenti di entrambi i sessi è "single"; solo un esiguo numero di persone è registrata con le etichette: "available", "married" e "seeing someone". In modo da diminuire la cardinalità delle variabili abbiamo fatto confluire i "married" in "seeing someone" e i "single" in "available".

Pets

La variabile Pets è composta da una combinazione di valori che descrivono se l'utente ama i gatti o/e i cani, se possiede entrambi, solo uno o nessuno dei due e se odia o ama l'altra categoria (ad esempio: "has dogs and dislikes cats").

Al fine di semplificare i valori, li abbiamo rimappati creando una stringa dove il primo elemento è 0 per i dislikes e 2 per i likes e 1 per l'indifferenza, mentre il secondo elemento è 1 per chi possiede e 0 se non possiede il cane o il gatto. Abbiamo creato quindi una matrice che ci ha permesso di plottare i risultati per genere. Dal grafico in figura 3 è possibile vedere come la maggioranza dichiari di amare sia i cani che i gatti con valori più netti per quanto riguarda i cani. Essendo la variabile ben distribuita tra i due sessi, abbiamo deciso di non utilizzarla nella classificazione.

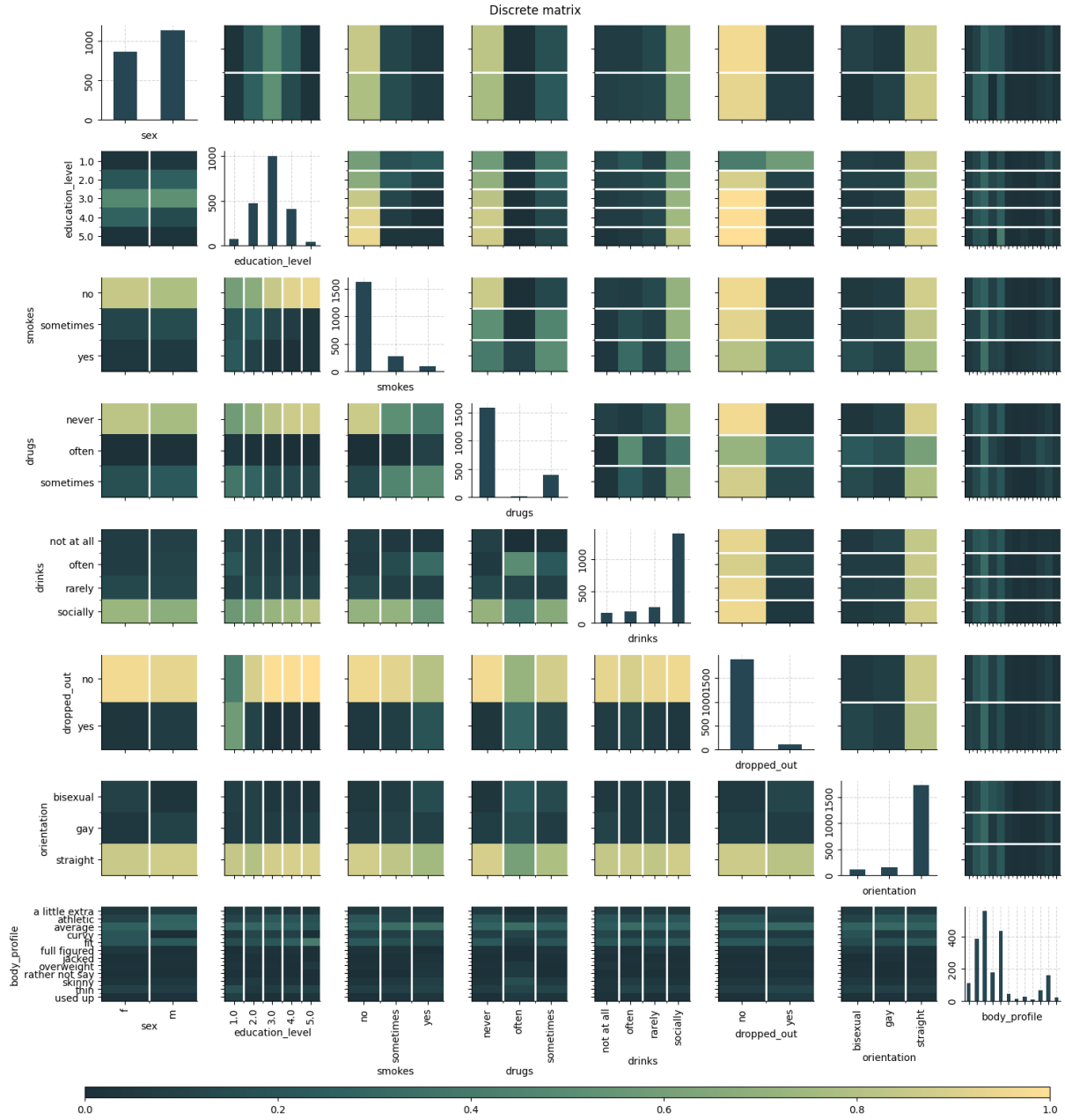


Figura 1: Matrice della distribuzione delle variabili categoriche. Sulla diagonale è presente il *countplot* delle singole variabili, mentre nella parte inferiore e superiore della matrice più grande sono presenti delle matrici più piccole che rappresentano un **joint countplot** di due variabili categoriche normalizzate (la normalizzazione è lungo le colonne per la parte inferiore e lungo le righe per la parte superiore della matrice)

Drinks, Drugs e Smokes

La variabile Drinks presenta i seguenti valori: “not at all”; “rarely”; “socially”; “often”; “very often”; “desperately”. Osservando però la distribuzione, abbiamo notato che pochissimi utenti dichiaravano di bere molto spesso o “disperatamente”. Per questo motivo - e assumendo che il “desperately” abbia una valenza scherzosa - abbiamo ridotto la dimensionalità della variabile facendo confluire “very often” in “often” e “desperately” in “socially”. A questo punto possiamo osservare la distribuzione tra uomini e donne (figura 1). Si nota che sono molti di più gli uomini che dichiarano di non bere o di bere spesso, mentre la differenza si attenua per i valori “rarely” e “socially”. Va ricordato, tuttavia, che il numero delle donne nel dataset è inferiore a quello degli uomi-

Tabella 3: Distribuzione in base ad orientamento, sesso e status

	females			males		
	bisexual	gay	straight	bisexual	gay	straight
available	20	2	7	7	6	24
married	2	0	5	1	0	6
seeing someone	10	2	12	2	4	24
single	52	36	714	18	103	944

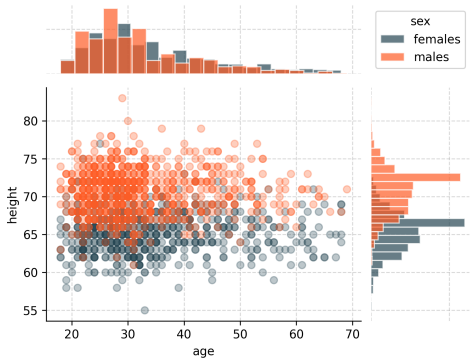


Figura 2: Jointplot delle variabili age e height

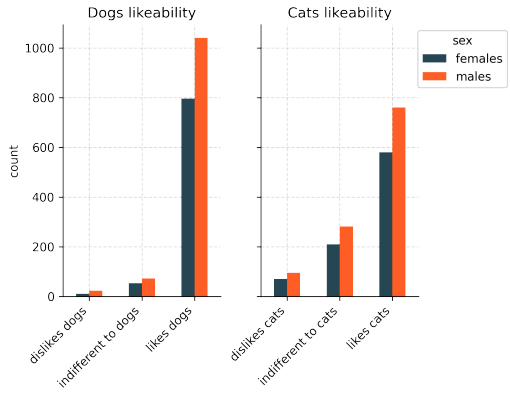


Figura 3: Gradimento per gli animali domestici

ni. In aggiunta, abbiamo messo in relazione la variabile Drinks con Education level. Sempre in figura 1 si può notare che all'aumentare del grado di istruzione aumentano gli utenti che bevono solo in compagnia ("socially"), mentre diminuiscono le persone che bevono spesso e quelle che non bevono.

Abbiamo condotto la stessa analisi con la variabile Drugs che contiene i valori: "never"; "often"; "sometimes". È emerso come la stragrande maggioranza sia degli uomini che delle donne dichiara di non aver mai fatto uso di droghe mentre una minima parte dichiara di averne fatto uso qualche volta e quasi nessuno mai. Anche in questo caso abbiamo messo in relazione la nostra variabile Drugs con l'education level ed è emerso che chi consuma droghe ogni tanto o spesso diminuisce al crescere del livello di istruzione.

Per quanto riguarda la variabile Smokes i valori sono: "no"; "sometimes"; "yes"; "when drinking"; "trying to quit". Anche in questo caso abbiamo rimappato alcuni valori, considerando "when drinking" e "trying to quit" molto simili a "sometimes". I valori risultanti sono quindi tre: "no"; "sometimes"; "yes". Anche qui vedendo le distribuzioni normalizzate, è possibile rilevare le stesse considerazioni assunte in precedenza, con gli uomini che fumano di più e le percentuali che diminuiscono in base al livello di istruzione.

In tutti e tre i casi, Drinks, Drugs e Smokes, non abbiamo rilevato una grande differenza per genere, seppur i maschi dichiarino di bere, assumere droga e fumare più spesso delle donne, mentre al contrario abbiamo una significativa differenza per livello di istruzione, dove chi ha un livello più basso di istruzione dichiara maggiormente di bere, assumere droghe e fumare più spesso rispetto a chi ha un livello di istruzione più elevato.

Interests

Per quanto riguarda gli interessi nel nostro dataset abbiamo due variabili categoriche, Interests e other_interests, contenenti gli interessi di ogni utente. Vista l'impossibilità di determinare la presenza o meno di una forte preferenza del primo interesse dichiarato sul secondo, abbiamo deciso di riunire i due attributi inserendo i valori in una matrice. Questo ci ha permesso di plottare i risultati in modo tale da osservare la frequenza degli interessi e le differenze tra i due sessi. Osservando il grafico in figura 4 è possibile vedere come la musica sia l'interesse più frequente mentre, dal grafico con le proporzioni sulla destra, emerge

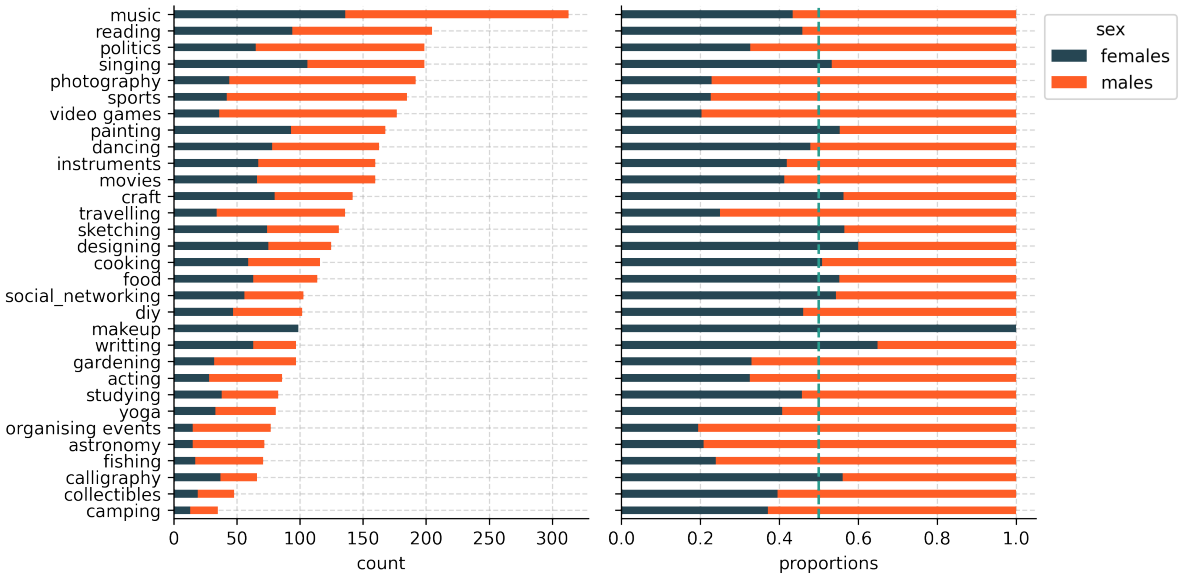


Figura 4: Interessi più comuni

con chiarezza come molti interessi siano diversi per genere, come ad esempio "makeup" che è esclusivamente femminile o "photography", "sports", "videogames" a prevalenza maschile. Proprio per queste differenze, Interests costituisce un ottimo attributo per il nostro classificatore.

Language

La variabile language si presenta sotto forma di lista della lingue parlate dall'utente con il relativo grado di fluidità. A causa dell'ingente dimensionalità di questa variabile categorica, abbiamo ritenuto opportuno trasformarla in una matrice composta da utenti (record) e attributi (lingue). Ogni valore di questa matrice è un numero che rappresenta il grado di fluidità di quell'utente per quella lingua secondo lo schema seguente: "poorly":1, "okay":2, "fluent":3, "fluently":3 Trattandosi di utenti principalmente residenti in California, risulta un forte sbilanciamento fra il numero di parlanti della lingua inglese e i parlanti di altre lingue (le due lingue più parlate immediatamente dopo l'inglese sono lo spagnolo e il francese, figura 5); abbiamo inoltre eliminato le lingue meno parlate e quelle difficilmente classificabili come "lingue moderne" o "lingue parlabili". Tra queste figurano: "ancientgreek", "c++", "esperanto", "latin" e altre.

La matrice che ci siamo ricavati ha una cardinalità molto alta, pertanto abbiamo deciso di applicare una *non–negative matrix factorization*. In figura 6 è possibile vedere una clustermap delle features create. Abbiamo deciso di estrarre 9 features guardando l'andamento *Frobenius–Norm* all'aumentare delle features. È possibile notare come facendo un clustering gerarchico sulla matrice risultante dalla NMF si creano dei cluster di persone che parlano le stesse lingue con la stessa fluenza.

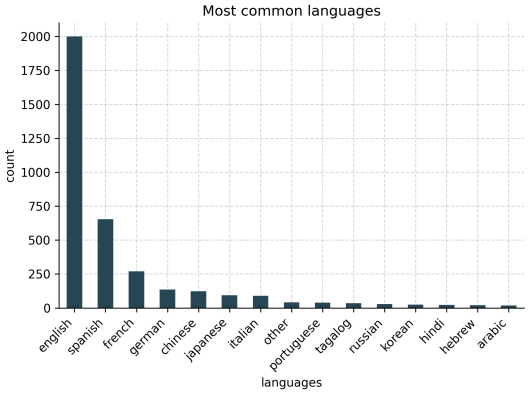


Figura 5: Lingue più parlate

Bio

La variabile Bio è una breve descrizione di se stessi, dei propri interessi e di cosa l'utente cerca in un partner. Abbiamo incluso nelle nostre analisi la lunghezza in caratteri dei testi; notiamo una lunghezza media globale intorno ai 691 caratteri che non varia di molto tra i due sessi (figura 7).

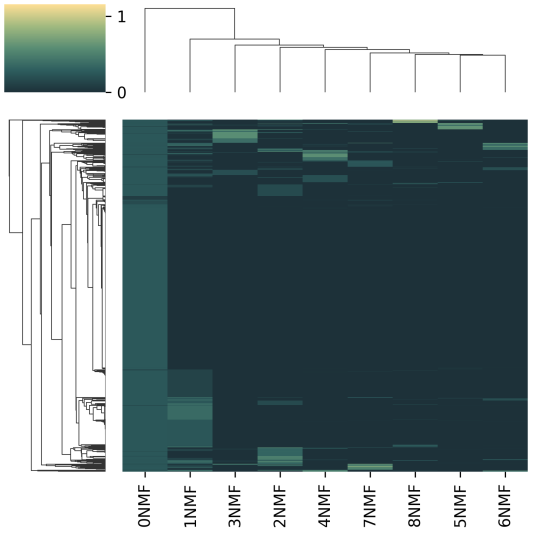


Figura 6: Lingue più parlate

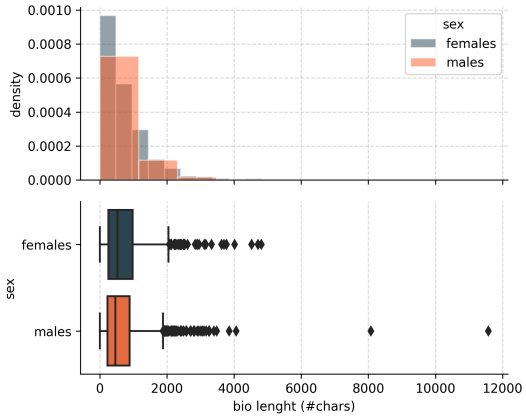


Figura 7: Distribuzione della lunghezza delle bio

Abbiamo effettuato anche una n-gram analysis per valori di n compresi fra 1 e 3, visibile nella figura 8. Notiamo come in cima alle parole più frequenti figurano "love" e "like", un risultato poco sorprendente data la natura dei testi. Fra i binomi più comuni abbiamo invece toponimi come "bay area" e "san francisco", considerando che la predominanza dei californiani anche questo trend ci sorprende poco. Un altro risultato in linea con i precedenti compare nei trigrammi, tra i quali "meeting new people".

In aggiunta ai token più frequenti all'interno dei testi abbiamo ricavato i count di alcuni token associati a determinate emozioni di base secondo lo spettro dello NRC Word-Emotion Association Lexicon e misure di polarity e subjectivity. L'analisi rivela che il sentimento "positività" è quello predominante, seguito da "gioia" e "fiducia"; questa tendenza è pressoché invariata nei due sessi. Per quanto riguarda polarity e subjectivity osserviamo una leggera predominanza di quest'ultima.

Clustering

Una volta terminato il processo di data preparation e data understanding, siamo passati al processo di clustering utilizzando a tal scopo tre algoritmi diversi: K-means, DBSCAN e clustering gerarchico. Al fine di preparare il dataset a questo scopo, come prima cosa, abbiamo creato una matrice con la term frequency di ogni parola nella biografia (dopo aver applicato uno stemming). Una volta creata la matrice abbiamo eliminato le parole meno comuni, abbiamo calcolato la varianza della term frequency per ogni parola e abbiamo eliminato tutte le parole con una varianza inferiore al 10% della distribuzione. Abbiamo unito la term frequency alle altre variabili numeriche come l'altezza, l'età e la matrice ricavata dalla sentiment analysis. Il dataset è così passato dalle 22 features iniziali a 8,814 features totali. Infine abbiamo utilizzato lo *StandardScaler* per la standardizzazione. Il dataset così standardizzato è stato utilizzato in tutte le fasi a seguire.

K-means

Per la fase di clustering abbiamo deciso come prima cosa di utilizzare il K-means con 6 cluster. La scelta del valore k da fornire all'algoritmo è stata effettuata sfruttando il metodo del ginocchio, cioè attraverso il plotting dell'SSE e della silhouette optando per il numero di cluster che minimizzava il primo e massimizzava il secondo (k=6), come è possibile vedere dalla figura 10.

Abbiamo inoltre impostato un n_init=10 (come di default), che rappresenta il numero di volte che l'algoritmo del k-means viene lanciato con un diverso seed per i centroidi, e un max_iter=20 (numero massimo di iterazioni). Gli iperparametri sono relativamente bassi in quanto la computazione è estremamente lenta e abbiamo deciso di avere un algoritmo più veloce che ci aiutasse a ripetere velocemente le analisi. Facendo il plot con t-SNE quello che notiamo è che il clustering è poco significativo in quanto tutti i valori sono andati a popolare solo uno dei 6 cluster creati e la proporzione fra maschi e femmine è simile.

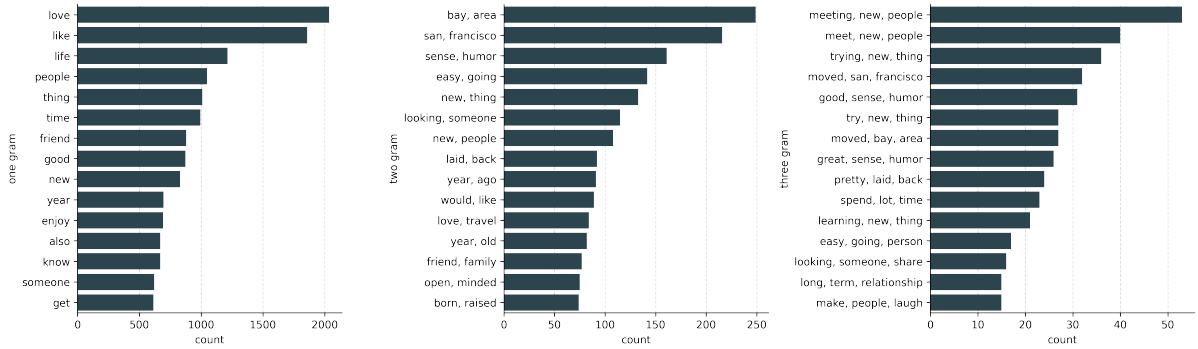


Figura 8: Analisi degli n grammi

Density-based clustering

A questo punto, abbiamo utilizzato un algoritmo di clustering density-based (DBSCAN). Al fine di determinare il valore di eps (la massima distanza stabilita tra due punti che consente di considerarli vicini l'uno all'altro) dell'algoritmo abbiamo plottato la distanza ordinata di ogni punto al suo k nearest neighbor; così facendo abbiamo deciso di adottare il valore 175 (figura11).

Al parametro min_samples, che indica il numero minimo di campioni da considerare nelle vicinanze di un punto (compreso il core-point stesso), abbiamo invece assegnato un valore dato dalla radice quadrata del numero di campioni. Facendo il plot con t-SNE, quello che emerge anche in questo caso, è la presenza di cluster poco significativi in quanto, nell'unico cluster creato la proporzione tra maschi e femmine è perlopiù bilanciata, come anche nel cluster degli outlier (figura 9c).

Hierarchical clustering

Come ultimo tentativo di clustering abbiamo provato ad utilizzare il gerarchico adottando come metrica la distanza euclidea e come metodo quello del Complete Linkage (la prossimità di due cluster è misurata sulla base dei punti più distanti).

Al fine di determinare la threshold t (la massima distanza consentita tra i cluster), abbiamo osservato il dendrogramma in figura 12. Infine, abbiamo plottato i cluster con t-SNE (figura 9b) e anche in questo caso il risultato del clustering è stato poco significativo; nello specifico, abbiamo ottenuto 11 cluster di cui solo uno altamente popolato e con una distribuzione tra maschi e femmine bilanciata.

Valutazione finale

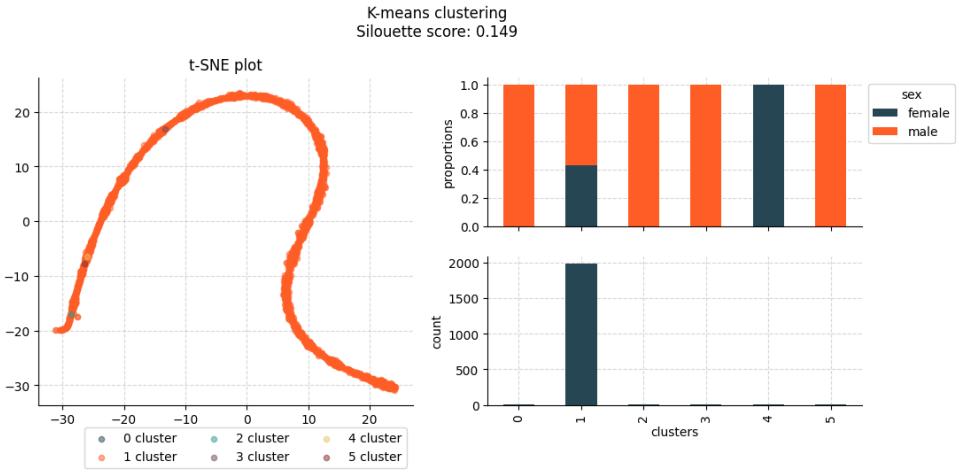
A giudicare dai risultati ottenuti, nessuna tipologia di clustering sembra fornire conclusioni interessanti. Con tutti e tre gli algoritmi, infatti, solo uno dei cluster creati viene popolato con una proporzione tra uomini e donne bilanciata, lasciando gli altri cluster quasi del tutto vuoti. Questo risultato è dovuto alla grande quantità di features presenti nel dataset.

Association Rules Mining

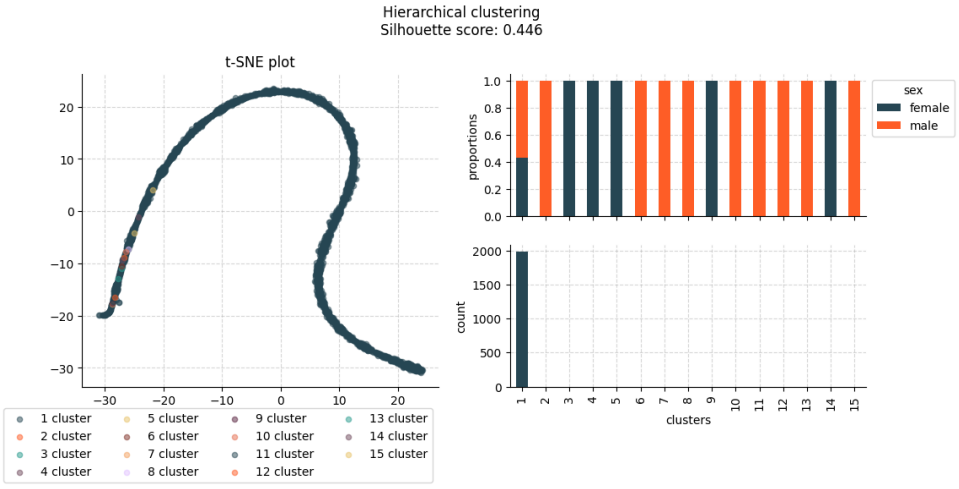
Data la maggiore numerosità di variabili categoriche rispetto alle continue, questo dataset si presta particolarmente bene al pattern mining. Abbiamo tenuto in considerazione tutte le variabili tranne quelle ad elevata dimensionalità, ovvero "bio" (trasformata in lunghezza in caratteri della stessa), "location_state" e "language". È stata necessaria invece un'operazione di binning per l'età e l'altezza degli utenti, e per la lunghezza della bio. L'età è stata raggruppata in bin di 5, con i seguenti range: (17, 28], (28, 38], (38, 49], (49, 59], (59, 69]; l'altezza in 5 quantili: (54.0, 65.0], (65.0, 67.0], (67.0, 70.0], (70.0, 71.0], (71.0, 83.0].

Gruppi frequenti

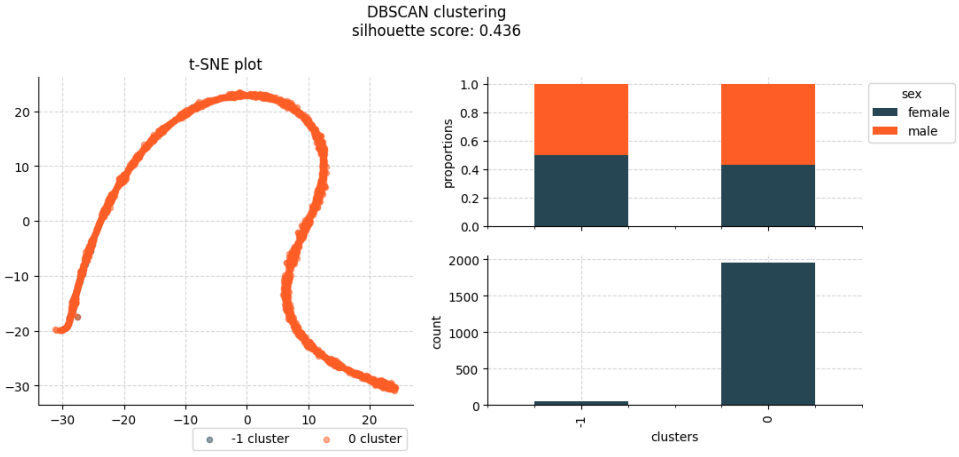
Per quanto riguarda l'estrazione dei pattern frequenti, abbiamo considerato valori del parametro "min_sup" di 20, 40 e 60. Tra i gruppi con supporto più alto spicca il gruppo 1 (tabella 5), il quale riassume bene la popolazione del dataset, composta principalmente da persone single eterosessuali, disponibili e che non fumano. Per quanto riguarda i massimali, ovvero set che non appartengono



(a) K-means plot



(b) Clustering gerarchico



(c) DBSCAN clustering

Figura 9: Cluster plots

a super-set frequenti, notiamo risultati diversi in base al supporto minimo. Uno dei gruppi massimali più frequenti con min_sup di 20 è il gruppo 3, da questa inferiamo che chi parla una sola lingua è generalmente non interessato ad apprenderne un'altra. Un altro massimale con min_sup 60 è il gruppo 4, il quale suggerisce che chi non fa uso di stupefacen-

Tabella 4: gruppi frequenti

	type	min_sup	sup	items
1	groups	20	67%	no smokes, straight orientation, no dropped_out, available status
2	groups	20	66%	never drugs, no smokes, no dropped_out, available status
3	maximal	20	26%	1 n_languages, not interested new_languages, never drugs, no smokes, straight orientation, no dropped_out, available status
4	maximal	60	63%	never drugs, no smokes, no dropped_out, available status
5	closed	60	58%	socially drinks, straight orientation, no dropped_out, available status

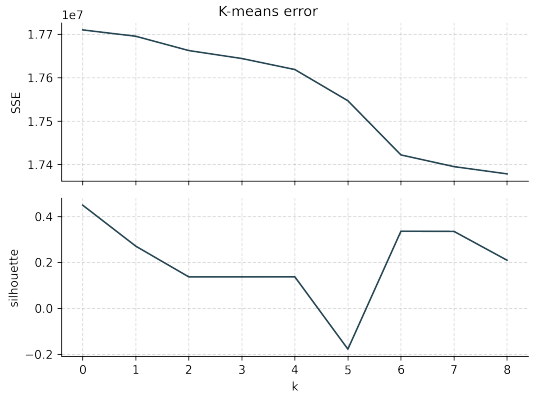


Figura 10: SSE e Silhouette per la decisione del k in K-Means.

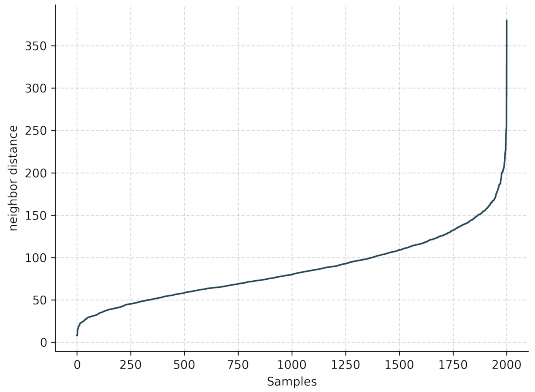


Figura 11: eps DBSCAN

ti tende a non fumare. Infine abbiamo anche analizzato i set "closed", ovvero i set che non hanno super-set con lo stesso supporto. Da quest'ultima analisi non spiccano gruppi che raccontano informazioni aggiuntive rispetto ai gruppi frequenti e i massimali.

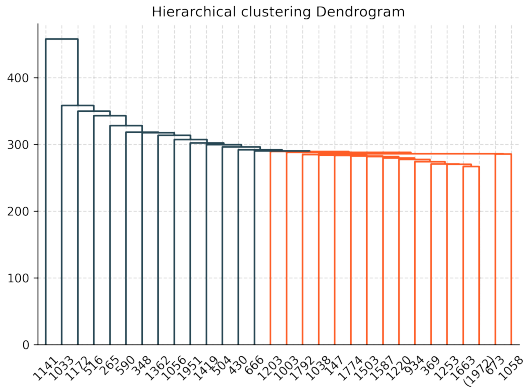


Figura 12: Dendogramma del clustering gerarchico

Association rules

La seguente tabella riassume le regole associative più interessanti ottenute dall'algoritmo Apriori, con valori del parametro min_conf di 90,95 e 99.

Tra le regole con il valore di lift più alto (1.80) spiccano quelle con antecedente "not interested in new_languages" (regole 1 e 2, tabella 6) e tra i loro conseguenti è frequente "1 n_languages", ciò implica che chi non è interessato all'apprendimento di un'altra lingua tende ad essere monolingue. Un'altra regola interessante è la regola 3, la quale dimostra che chi ha completato gli studi tende ad avere un livello di istruzione nella media (3.0). Infine abbiamo anche osservato la distribuzione del lift al variare della min_conf:

Dagli istogrammi si osserva che per i valori di min_conf 80%, 90% e 95%, il lift si distribuisce intorno al valore 1.0 con una coda a 1.8, invece con lift pari al 99%, osserviamo che la coda intorno all'1.8 scompare e i valori si distribuiscono principalmente attorno al valore 1.05. In conclusione, le regole più inte-

Tabella 5: association rules

	min_conf	lift	antecedente	conseguente
1	90	1.82	not interested new_languages	1 n_languages, never drugs, no smokes, no dropped_out, available status
2	95	1.81	not interested new_languages	1 n_languages, never drugs, no smokes
3	99	1.05	no dropped_out	3.0 education_level, never drugs, no smokes

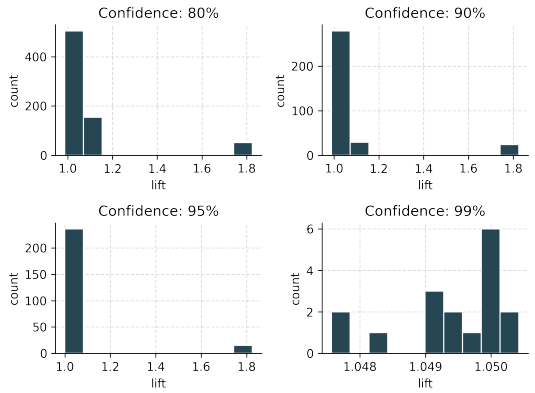


Figura 13: Lift distribution

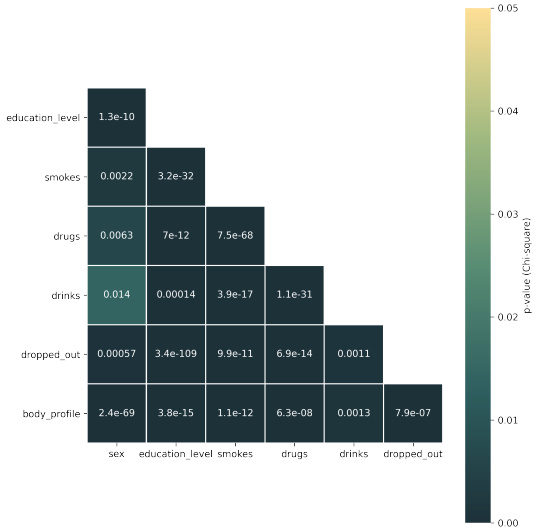


Figura 14: Matrice chi2

ressanti hanno tutte a che fare con il livello di istruzione e di fluenza linguistica; non abbiamo trovato regole utili al riempimento dei valori mancanti o alla predizione della variabile target.

Classification

Prima di procedere alla classificazione abbiamo controllato quali variabili categoriche erano correlate usando un test che Chi–quadro. In figura 14 è possibile osservare la matrice del p–value associato al Chi–quadro per ogni coppia di variabili categoriche. È possibile notare che tutte le coppie di variabili hanno un p–value molto basso, pertanto abbiamo deciso di aggiungere le variabili categoriche a quelle numeriche per il task di classificazione. Abbiamo quindi usato un *one hot encoder* per creare una rappresentazione delle variabili categoriche da usare per la classificazione. A questo punto, abbiamo potuto suddividere il dataframe e le label (nel nostro caso il genere) in training e test set, con una partizione stratificata rispetto alla variabile target del 70% e 30% per train e test rispettivamente.

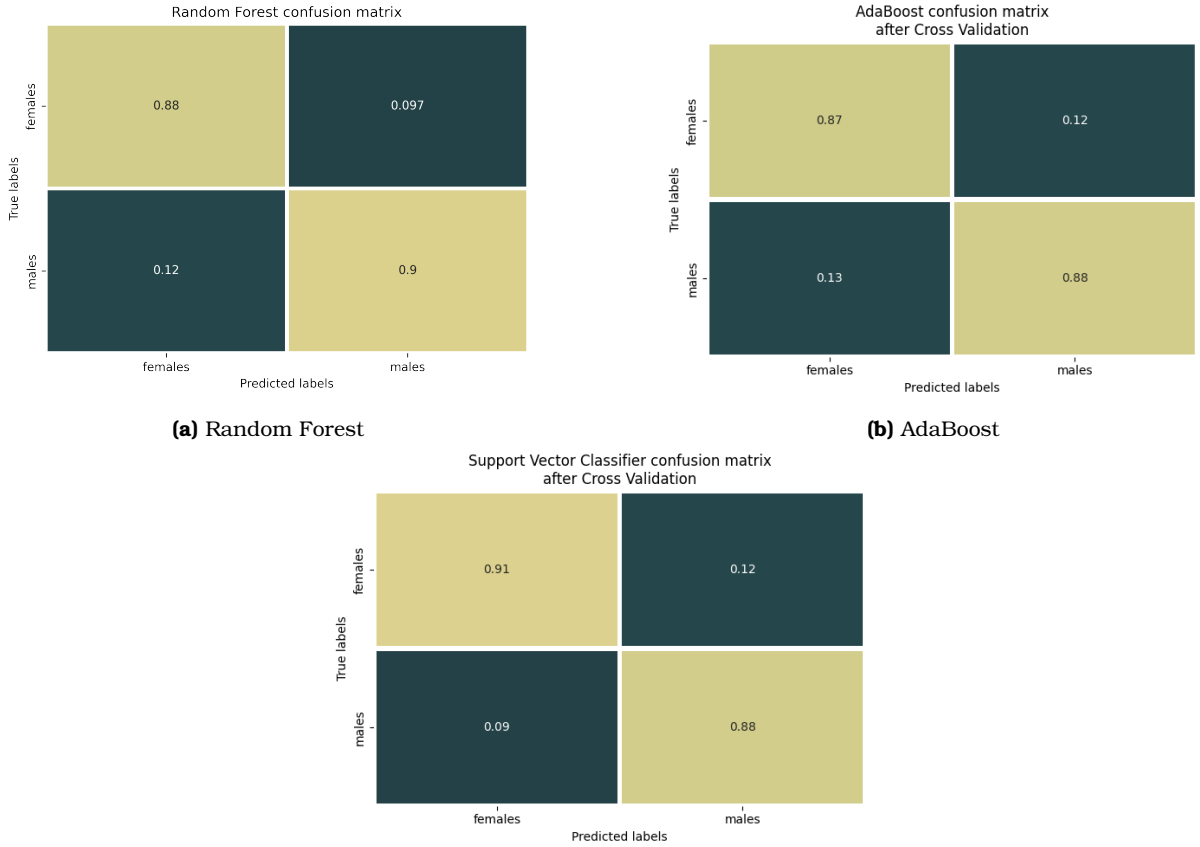
Random Forest Classifier

Il primo algoritmo di classificazione che abbiamo scelto è stato un Random Forest Classifier; utilizzando i parametri di default abbiamo raggiunto un'accuratezza dell'86% e un f1–score dell'85%. Per provare ad aumentare la performance del classificatore abbiamo effettuato una gridsearch cross–validation[1] cercando di massimizzare l'f1–score. Nella grid search abbiamo testato diversi parametri, tra i quali la massima profondità dell'albero (`5, 10, 50, None`), il criterio di splitting degli stump (`gini, entropy`), il numero massimo di feature da considerare ad ogni costruzione degli alberi (`log2, None`) e la presenza o meno di warm start. I migliori parametri si sono dimostrati una massima profondità di 50, l'utilizzo di tutte le feature per costruire gli alberi decisionali, l'utilizzo di worm start e del gini coefficient. In questa maniera abbiamo raggiunto una accuratezza dell'89% e un f1–score

[1] Abbiamo optato per tutte le gridsearch per 3 fold per evitare che la cross validation prendesse troppo tempo.

Tabella 6: Risultati dei classificatori dopo la Cross validation. La Random Forest e AdaBoost usano l'intero dataset, mentre il Support Vector Classifier utilizza le features estratte usando Adaboost.

	Random Forest			AdaBoost			Support Vector Classifier		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
female	0.88	0.87	0.88	0.87	0.84	0.86	0.91	0.84	0.87
male	0.90	0.91	0.91	0.88	0.91	0.89	0.88	0.94	0.91
accuracy			0.89			0.88			0.89
macro avg	0.89	0.89	0.89	0.88	0.87	0.87	0.90	0.89	0.89



(a) Random Forest

(b) AdaBoost

(c) Support Vector Classifier (SVC)

Figura 15: Confusion matrix di Random Forest, AdaBoost e SVC

dell'89%. La confusion matrix si può trovare in figura 15a e le metriche complete si trovano in tabella 6.

Feature Importance

Per provare ad avere una interpretazione di quello che fa il modello abbiamo estratto le feature importance per la Random Forest, visibili in figura 16. Possiamo notare che la feature più importante è l'altezza, come avevamo previsto nella data exploration. Altre feature importanti sono l'essere "curvy" e "makeup" che, come visto nella data exploration, erano a quasi solo appannaggio delle donne.

AdaBoost

Abbiamo provato ad usare AdaBoost che, utilizzando i parametri di default, ha raggiunto le performance della random forest dopo la cross validation (accuratezza e F1 dell'89%). Anche qui, nel tentativo di migliorare il classificatore, abbiamo effettuato una grid–search valutando diversi stimatori (`10, 50,100, 200`) e diversi learning rate (`1, 1e-2, 1e-3`). I migliori parametri si sono dimostrati il learning

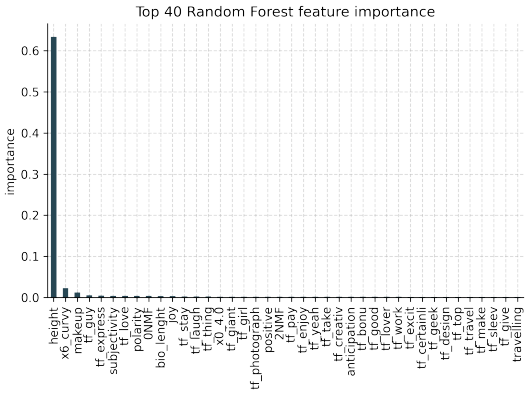


Figura 16: Top 40 feature importance nel Random Forest Classifier

rate a 1 e 200 stimatori. Dopo la cross validation, stranamente le performance dell'algoritmo sembrano peggiorare, in tabella 6 si trovano le metriche complete e in figura 15b la confusion matrix.

Feature Importance

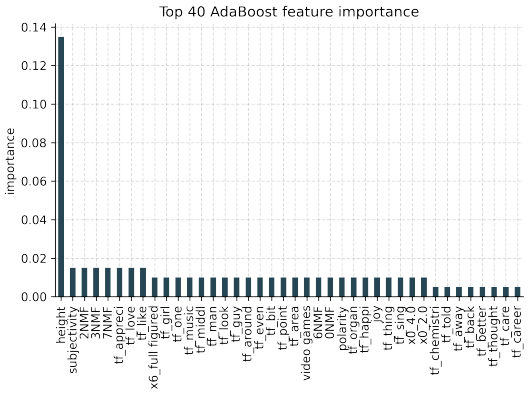


Figura 17: Top 40 feature importance in AdaBoost

Per provare ad avere una interpretazione di quello che fa il modello abbiamo estratto anche in questo caso le feature importance per AdaBoost, visibili in figura 17. Possiamo notare che le feature più importanti in generale sono simili a quelle della Random Forest, anche se, stranamente, sembrano comparire le feature estratte con la NNF.

Clustering with selected features

Clustering

Dopo aver visualizzato la feature importance di AdaBoost abbiamo deciso di utilizzare le feature estratte dal classificatore per fare un'ultima volta DBSCAN. In figura 18 è possibile vedere un grafico sommario di questo step. Anche riducendo la dimensionalità a 130 feature l'algoritmo di clustering trova un unico cluster con una proporzione paritaria fra maschi e femmine.

Support Vector Machine

Abbiamo provato ad allenare un Support Vector Classifier utilizzando le feature più importanti secondo AdaBoost. Per il SVC abbiamo normalizzato le variabili diversamente da quanto fatto con random forest e AdaBoost. Utilizzando i parametri di default abbiamo raggiunto performance simili alla random forest (accuratezza e f1–score dell'89%). Per provare a migliorare l'algoritmo abbiamo provato anche in questo caso a fare una gridsearch cross–validation provando diversi parametri di C (`1, 1e-2, 1e-3, 1e-4`) e diversi parametri di gamma (`scale, auto, 0.5, 0.33`). I risultati della gridsearch sono stati i parametri standard quindi 1 per la C e "auto" per gamma, raggiungendo quindi accuratezza e F1–score detti prima. In tabella 6 si trovano le metriche complete e in figura 15c la confusion matrix.

Voting Classifier

In un ultimo tentativo, al fine di migliorare la nostra classificazione, abbiamo provato ad usare un ensamble dei modelli usati in precedenza ma allenati sulle feature più importanti estratte con AdaBoost e utilizzando i migliori parametri trovati con le gridsearch precedenti. Il classificatore così importato ha raggiunto l'accuratezza del 91% e un f1–score del 91%.

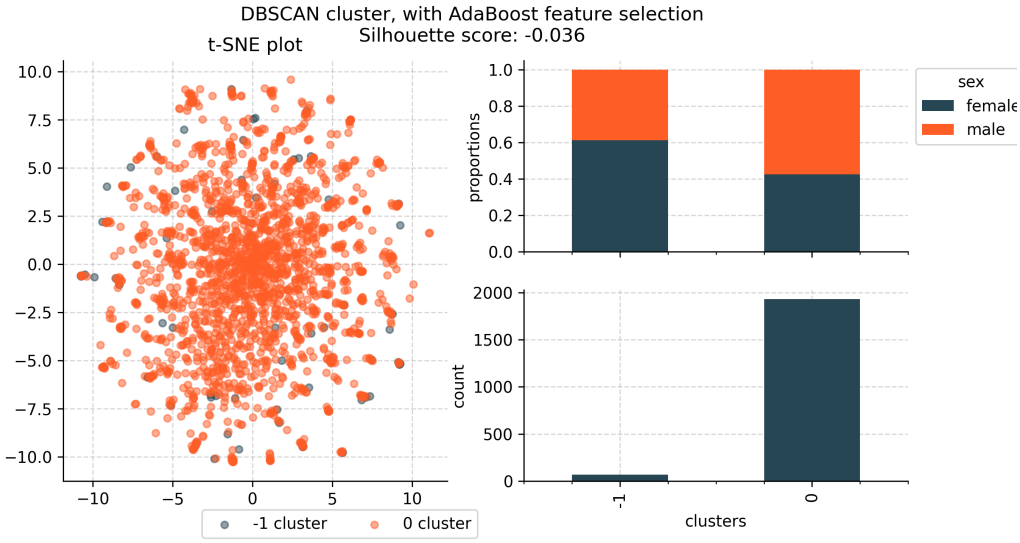


Figura 18: Density Based clustering utilizzando le features con importance non zero di AdaBoost