

Relazione Data Analytics for Business

Docenti

Prof.ssa Ester Zumpano
Prof. Eugenio Vocaturo
Dott. Ing. Tommaso Ruga

Studenti

Mastroianni Matteo 252422
Cantafio Matteo 257301
Froio Alessia 252226
Rizzo Giuseppe 252419

Anno 2024/2025

Sommario

Questo progetto di ricerca affronta il clustering e l'analisi di attributi facciali dal dataset CelebA, con un focus crescente sulla valutazione della fairness e sull'interpretazione dei modelli. Inizialmente, si esplora l'uso di embedding pre-addestrati da FaceNet e attributi CSV originali, con riduzione dimensionale tramite Principal Component Analysis (PCA), per il clustering K-Means. Successivamente, si adotta un approccio più sofisticato basato sulla metodologia di Anzalone et al., utilizzando un modello pre-addestrato (MobileNetV2 modificata) per generare vettori binari di 37 attributi facciali predetti per ogni immagine. Questi vettori, che rappresentano caratteristiche semantiche più esplicite, sono poi impiegati sia per il clustering, dimostrando una superiore capacità di separazione dei gruppi, sia per un'analisi approfondita della predizione dell'attributo soggettivo "Attractive". Quest'ultima fase investiga l'impatto di diverse strategie di feature selection e di bilanciamento dei dati su classificatori quali Random Forest e Regressione Logistica. E' stato dimostrato che le feature predette semanticamente ricche migliorano significativamente la separabilità dei cluster (Silhouette Score fino a 0.72). Tuttavia, la predizione di "Attractive" mostra una notevole eterogeneità e potenziale bias tra i cluster. Lo studio evidenzia la complessità della modellazione di attributi percettivi e la necessità impellente di un'analisi critica delle fonti di bias, dal dataset ai modelli intermedi e finali.

Indice

1	Introduzione	2
2	Approccio Baseline	3
2.1	Il Dataset CelebA	3
2.2	Pre-elaborazione dei dati all'interno del CSV	4
2.3	Selezione delle features	5
2.4	Metodologia	5
2.4.1	Generazione degli embeddings tramite FaceNet	5
2.4.2	Riduzione della dimensionalità (PCA)	5
2.4.3	Clustering con K-Means, valutazione e visualizzazione	6
2.5	Risultati sperimentali e analisi	7
2.5.1	Clustering su embeddings di FaceNet	7
2.5.2	Clustering su attributi CSV	8
2.6	Discussione dei risultati dell'approccio baseline	10
3	Approccio basato su Anzalone et al.	12
3.1	Approccio basato su Anzalone et al.	15
3.1.1	Clustering su attributi originali del dataset CelebA	15
3.1.2	Replicazione esperimento Anzalone et al. con feature predette	16
3.2	Addestramento e valutazione dei classificatori per l'attributo "Attractive"	17
3.2.1	Analisi SHAP	19
4	Conclusioni e Lavori Futuri	22
	Bibliografia	24

1. Introduzione

L'interpretazione automatica delle caratteristiche facciali umane, e in particolare la valutazione di attributi soggettivi come l'attrattività, rappresenta una frontiera della computer vision densa di implicazioni tecnologiche e socio-culturali. L'avvento di vasti dataset di immagini, quale CelebA [1], e di potenti architetture di deep learning ha aperto nuove prospettive per l'analisi e il raggruppamento di volti su larga scala. Tuttavia, la natura intrinsecamente complessa e spesso ambigua degli attributi percettivi solleva questioni fondamentali relative all'affidabilità, all'equità (fairness) e all'interpretabilità dei modelli computazionali impiegati. La definizione e l'annotazione di tali attributi sono processi intrinsecamente influenzati da fattori culturali e demografici, introducendo potenziali bias che possono essere amplificati o perpetuati dai sistemi di intelligenza artificiale.

Il presente lavoro si inserisce in questo contesto problematico, con l'obiettivo primario di esplorare e confrontare diverse strategie per il clustering di immagini facciali e la successiva analisi predittiva dell'attributo "Attractive". La ricerca si propone di investigare come differenti rappresentazioni delle feature da embedding generici pre-addestrati a vettori di attributi semanticamente esplicativi influenzino la capacità di algoritmi di apprendimento non supervisionato di identificare raggruppamenti significativi all'interno del dataset CelebA.

Un secondo obiettivo cruciale è quello di valutare le performance e il comportamento di modelli di classificazione (Random Forest e Regressione Logistica) addestrati per predire l'attributo "Attractive", utilizzando i cluster precedentemente identificati come base per un'analisi disaggregata. Questa fase mira a comprendere non solo l'accuratezza predittiva globale, ma anche come tale accuratezza vari attraverso i diversi sottogruppi di volti, evidenziando potenziali disparità. In questo sforzo, si esplorano l'impatto della selezione delle feature e di diverse tecniche di bilanciamento del training set.

Infine, una componente fondamentale di questo studio è l'applicazione di tecniche di interpretabilità dei modelli per investigare il contributo delle singole feature facciali alla predizione dell'attributo "Attractive" all'interno dei diversi cluster. Questa analisi si propone di far luce sui meccanismi decisionali dei modelli e di identificare eventuali pattern di bias o dipendenze da feature specifiche che potrebbero variare tra i sottogruppi, offrendo così una prospettiva critica sulle sfide etiche e interpretative che emergono quando si trattano attributi umani complessi con metodi computazionali. Il percorso di ricerca si articola quindi attraverso un approccio baseline, volto a stabilire un punto di riferimento con tecniche di clustering standard, per poi evolvere verso una metodologia più sofisticata, ispirata al lavoro di Anzalone et al. [2], che sfrutta attributi facciali predetti per migliorare sia la qualità del clustering sia la profondità dell'analisi sulla predizione dell'attrattività e sui relativi bias.

2. Approccio Baseline

2.1 Il Dataset CelebA

Il dataset primario utilizzato è CelebA (CelebFaces Attributes Dataset) [1]. Contiene oltre 200.000 immagini di celebrità, ognuna annotata con 40 attributi facciali binari. Per questo progetto, utilizziamo sia le immagini sia le annotazioni degli attributi. Le immagini presentano variazioni significative in posa, espressione, età e angolazioni. Preso atto che il dataset CelebA presenta una marcata disomogeneità nella distribuzione delle features: oltre un terzo di esse risulta estremamente raro (con una frequenza inferiore al 10%), mentre solo poche sono particolarmente diffuse (comparando in oltre il 70% dei casi). Questa disparità ha evidenziato durante l'analisi delle criticità nelle Loss functions comunemente adottate, in particolare nell'addestramento di modelli su istanze poco rappresentate.

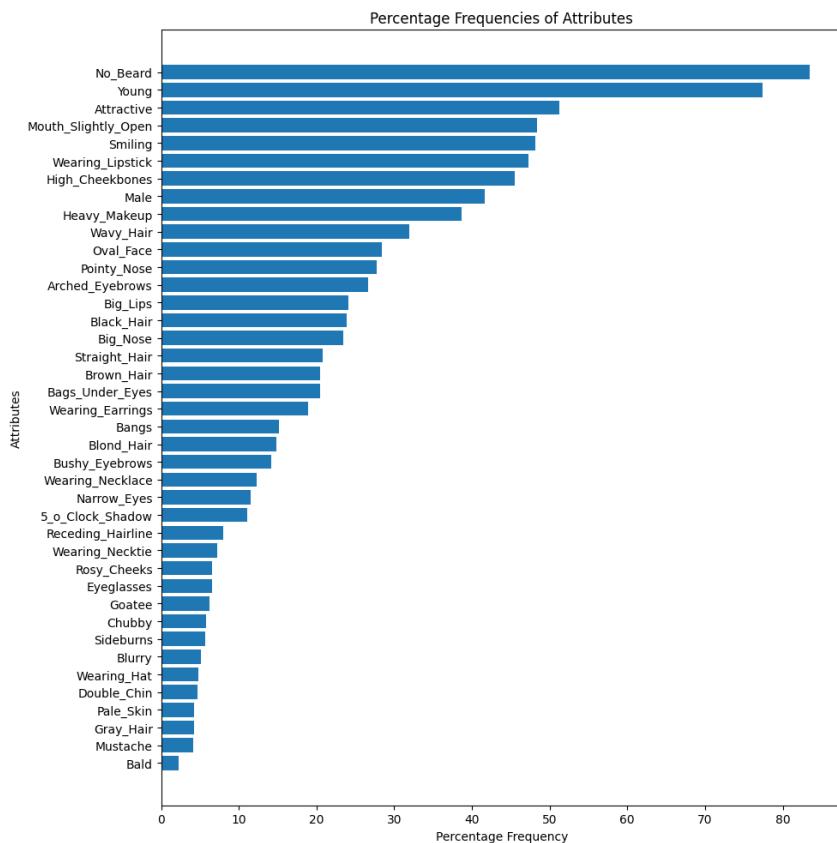


Figura 2.1: Frequenze delle features nel dataset CelebA

2.2 Pre-elaborazione dei dati all'interno del CSV

I dati degli attributi sono forniti in un file CSV ('list_attr_celeba.csv'), dove ogni riga corrisponde a un'immagine e le colonne rappresentano i 40 attributi binari. Questi attributi sono inizialmente codificati come -1 (assente) e 1 (presente).

Per prima cosa, gli attributi vengono trasformati in modo che il valore -1 diventi 0, rappresentando l'assenza di un attributo, mentre il valore 1 rimane invariato, indicando la sua presenza. Questo passaggio crea una rappresentazione puramente binaria (0/1) per tutte le feature. Il dataset risultante contiene 40 colonne di attributi (oltre alla colonna 'image_id') per 202.599 voci, confermando l'assenza di valori mancanti.

Successivamente alla conversione dei valori, è stata generata una matrice di correlazione, visualizzata come heatmap, per investigare le interdipendenze tra le 40 feature binarie. L'analisi di questa matrice è cruciale per comprendere la struttura intrinseca dei dati e per anticipare potenziali sfide nel processo di clustering.

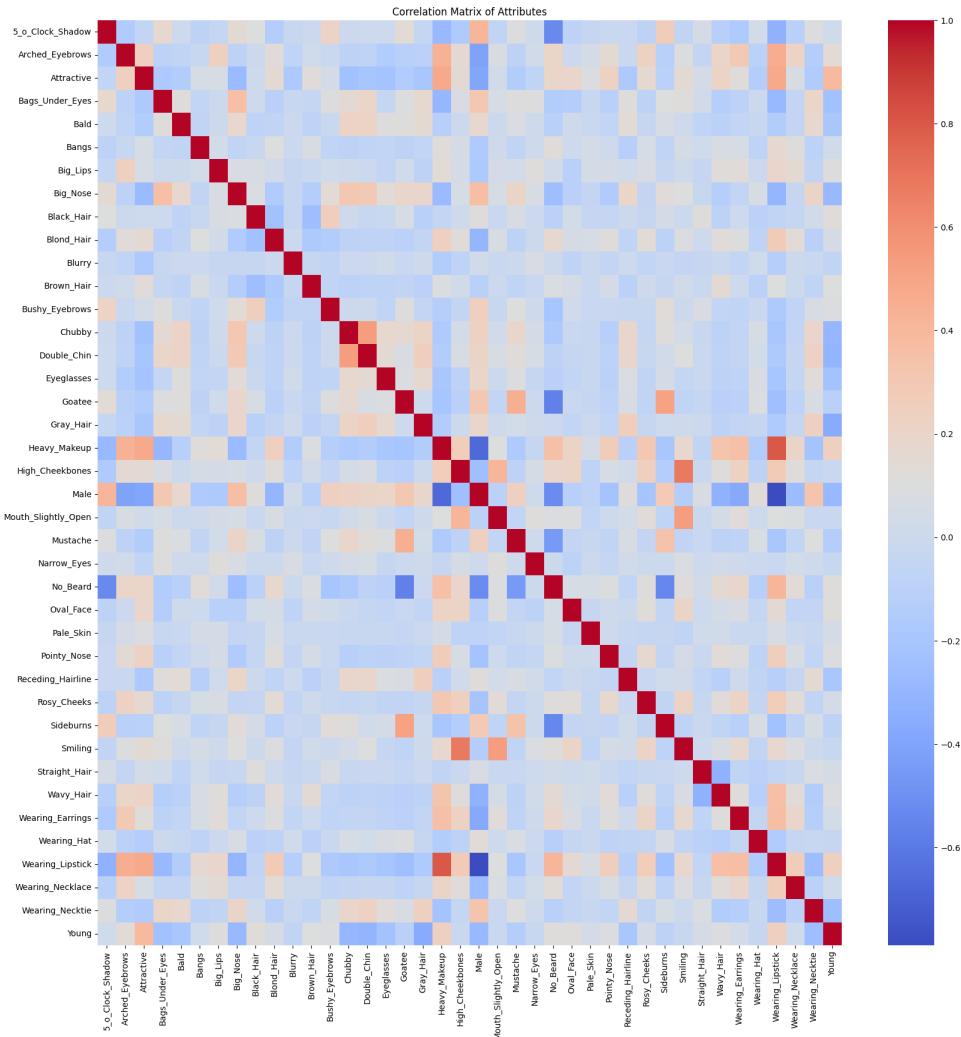


Figura 2.2: Matrice di Correlazione (Heatmap) degli attributi del dataset CelebA.

Dall'osservazione della Figura 2.2, emergono alcuni pattern significativi. Sebbene la maggior parte delle coppie di attributi mostrino una correlazione bassa o moderata, indicando una certa indipendenza, si notano anche alcune correlazioni più marcate (sia positive che negative). Ad esempio, è possibile osservare una correlazione positiva tra attributi come "Baffi" (Mustache) e

"Barba" (No_Beard, in forma negativa se No_Beard è 0), oppure una correlazione negativa tra "Capelli Biondi" (Blond_Hair) e "Capelli Neri" (Black_Hair).

Tuttavia, un aspetto fondamentale che emerge dalla heatmap, e che si allinea con la disomogeneità distributiva menzionata in precedenza (Sezione 2.2), è la generale **assenza di strutture di correlazione estremamente forti e pervasive**.

2.3 Selezione delle features

Per le features binarie, la varianza è calcolata come $p(1 - p)$, dove p rappresenta la proporzione di una delle due modalità. Un valore vicino a 0.25 indica una feature bilanciata, mentre un valore prossimo a 0 segnala una forte asimmetria. Le features con varianza molto bassa risultano poco utili ai fini del clustering, in quanto scarsamente discriminanti. Per questo motivo, è stata applicata una soglia di varianza per escludere tali feature.

La soglia iniziale scelta è stata $0.05 \times (1 - 0.05) \approx 0.0475$. Questo filtro ha ridotto il numero di attributi da 40 a 34, escludendo la colonna *image_id*.

Non sono state rilevate feature altamente correlate (ad esempio con correlazione > 0.90) tra i 34 attributi rimanenti. Di conseguenza, non si è resa necessaria un'ulteriore riduzione basata sulla correlazione. Il DataFrame finale, denominato *attr*, contiene quindi *image_id* e 34 attributi binari processati.

2.4 Metodologia

2.4.1 Generazione degli embeddings tramite FaceNet

Per estrarre caratteristiche visive di elevata qualità dai volti presenti nelle immagini, si è fatto ricorso a un modello FaceNet pre-addestrato, che è stato eseguito su un terminale in locale, a differenza di quasi tutte le altre operazioni che sono state eseguite sul notebook Colab. Questo modello, basato su una rete neurale convoluzionale profonda, è stato adottato con l'obiettivo di apprendere una mappatura dalle immagini facciali a uno spazio euclideo compatto, in cui le distanze tra i punti riflettono in maniera diretta il grado di somiglianza tra i volti corrispondenti. Abbiamo deciso di adottare la versione del modello disponibile su Kaggle Hub [3]. Le immagini sono state preliminarmente sottoposte a un processo di pre-elaborazione che ha previsto il loro ridimensionamento alla risoluzione di 160 x 160 pixel, la conversione al formato RGB e la normalizzazione dei valori dei pixel affinché presentassero media nulla e varianza unitaria, conformemente ai requisiti di input del modello.

Una volta completata questa fase di preparazione, le immagini sono state processate in batch dal modello FaceNet, il quale ha generato per ciascuna di esse un vettore di embedding di 128 dimensioni. In totale, sono stati così ottenuti embedding per 202.599 immagini, che sono stati successivamente archiviati per essere impiegati nelle fasi di clustering e analisi.

2.4.2 Riduzione della dimensionalità (PCA)

La Principal Component Analysis (PCA) è stata impiegata come tecnica fondamentale per la riduzione della dimensionalità, applicata distintamente sia agli embedding generati da FaceNet sia agli attributi CSV precedentemente pre-elaborati.

Per quanto riguarda gli embedding di FaceNet, originariamente a 128 dimensioni, questi sono stati inizialmente sottoposti a una standardizzazione Z-score. Successivamente, l'applicazione

della PCA ha permesso di investigare come la contrazione dello spazio delle feature influenzasse le prestazioni del clustering. La selezione del numero ottimale di componenti principali è stata guidata dall'analisi della varianza spiegata cumulativa, con un'attenzione particolare al "gomito" dello scree plot, visibile in Figura 2.3 e al raggiungimento di una soglia di varianza desiderata, come ad esempio il 95%. In questo contesto, sono stati condotti esperimenti variando il numero di componenti (ad esempio, utilizzando 12, 6 o 4 componenti), osservando che, per illustrare, 32 componenti erano in grado di catturare circa il 95.3% della varianza totale degli embedding standardizzati.

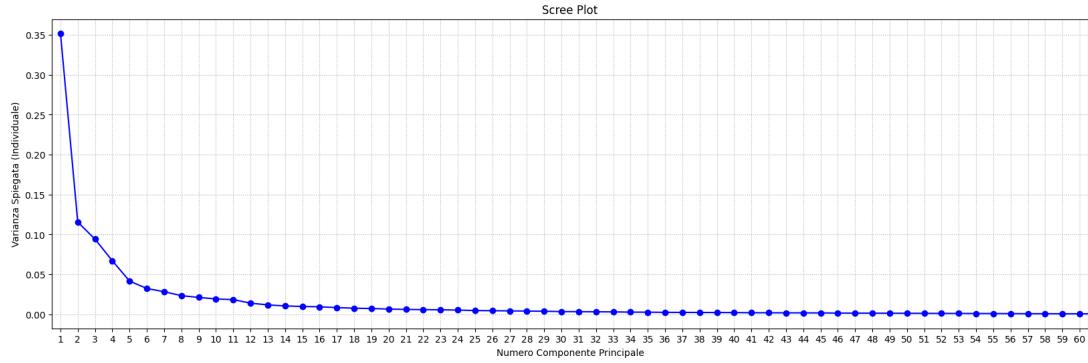


Figura 2.3: Scree-Plot per il file degli embeddings

Analogamente, la PCA è stata applicata anche ai 34 attributi binari derivanti dalla pre-elaborazione dei dati CSV, come descritto nella Sezione 2.3. Anche in questo caso, l'obiettivo era ridurre la dimensionalità mantenendo al contempo l'informazione più saliente. A titolo esemplificativo, si è riscontrato che 28 componenti principali riuscivano a spiegare circa il 95.7% della varianza intrinseca nei dati degli attributi. Parallelamente, sono state esplorate configurazioni con un numero inferiore di componenti, ovvero 7, 4 e 2 componenti principali, per valutare l'impatto di una riduzione più drastica sulla successiva fase di clustering.

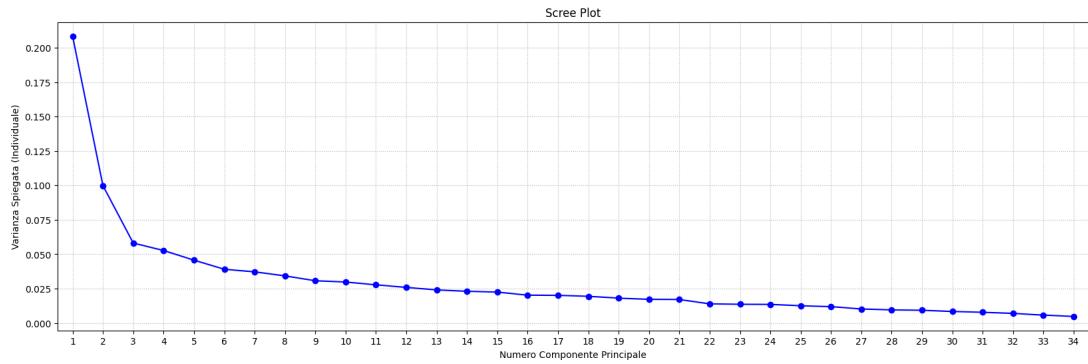


Figura 2.4: Scree-Plot per il file degli attributi presenti nel CSV

2.4.3 Clustering con K-Means, valutazione e visualizzazione

Successivamente è stata condotta un'analisi di clustering mediante l'algoritmo K-Means, applicato a differenti rappresentazioni delle caratteristiche facciali. In particolare, l'algoritmo è stato impiegato sia sugli embedding grezzi generati dal modello FaceNet, composti da vettori a 128 dimensioni, sia su versioni di tali embedding ridotte dimensionalmente tramite Principal Component Analysis (PCA), con un numero di componenti variabile ad esempio, 12, 6 o 4.

Analogamente, anche gli attributi derivanti da file CSV sono stati sottoposti a una riduzione dimensionale mediante PCA, considerando diverse configurazioni, quali 28, 7, 4 e 2 componenti principali.

In tutti gli esperimenti condotti, il numero di cluster K è stato generalmente fissato a quattro, e si è fatto uso di un valore deterministico di ‘random_state’ per garantire la riproducibilità dei risultati. Le assegnazioni di cluster per ciascuna immagine sono state memorizzate per le successive fasi di valutazione e analisi.

Per quanto concerne la valutazione della qualità del clustering, si è adottato il Silhouette Score quale indicatore quantitativo. Tale metrica esprime il grado di coesione interna di un cluster rispetto alla separazione dagli altri, fornendo valori compresi tra -1 e +1: punteggi prossimi a 1 indicano una netta distinzione tra i cluster e una buona coerenza interna. Il calcolo del Silhouette Score è stato frequentemente effettuato su un campione rappresentativo di circa 20.000 punti, al fine di ottimizzare l’efficienza computazionale.

A supporto dell’analisi qualitativa, si è ricorso inoltre alla tecnica di visualizzazione t-Distributed Stochastic Neighbor Embedding (t-SNE), utile per proiettare i dati ad alta dimensionalità siano essi embedding grezzi o ridotti tramite PCA in uno spazio bidimensionale. I punti ottenuti da tale proiezione sono stati colorati in base al cluster di appartenenza, generando scatter plot che facilitano l’interpretazione visiva della separazione tra i gruppi. Anche in questo caso, per contenere il costo computazionale, l’analisi è stata condotta su un campione di 20.000 osservazioni.

Infine, al fine di offrire una comprensione visiva più immediata delle caratteristiche distintive di ciascun cluster, per ogni gruppo identificato è stata generata una griglia di immagini campione. Questa rappresentazione ha permesso di cogliere, a colpo d’occhio, le peculiarità visive comuni all’interno di ciascun insieme, fornendo un ulteriore elemento di interpretazione qualitativa dei risultati ottenuti.

2.5 Risultati sperimentali e analisi

Questa sezione presenta i risultati ottenuti applicando il clustering K-Means ($K = 4$) ai diversi set di feature elaborati. Per valutare quantitativamente la qualità dei cluster, abbiamo calcolato il Silhouette score su campioni di 20.000 punti per ridurre i tempi di calcolo. La Figura 2.8 riassume i Silhouette score in funzione del numero di componenti PCA utilizzati per i due set di feature: FaceNet Embeddings e CSV Attributes.

2.5.1 Clustering su embeddings di FaceNet

Il clustering sugli embedding grezzi di FaceNet (128 dimensioni) ha prodotto un Silhouette score di 0.1624. La visualizzazione t-SNE in uno spazio a 2 dimensioni non ha mostrato la presenza di separazione tra i clusters, bensì una considerevole sovrapposizione, riflettendo quindi la complessità dello spazio degli embedding grezzi.

Applicando la PCA per ridurre la dimensionalità degli embedding, abbiamo osservato un lieve miglioramento progressivo del Silhouette score. Come illustrato nella Figura 2.5, con 12 componenti PCA il punteggio sale a 0.1809, con 6 componenti raggiunge 0.2220, e con 4 componenti si ottiene il punteggio più alto per questo set di feature: 0.2582. Questo suggerisce che la PCA aiuta a catturare le dimensioni più discriminanti per il clustering in questo spazio di feature, ma allo stesso tempo non permette di separare in maniera chiara i dati, come è possibile vedere in maniera qualitativa anche dalla Figura 2.6.

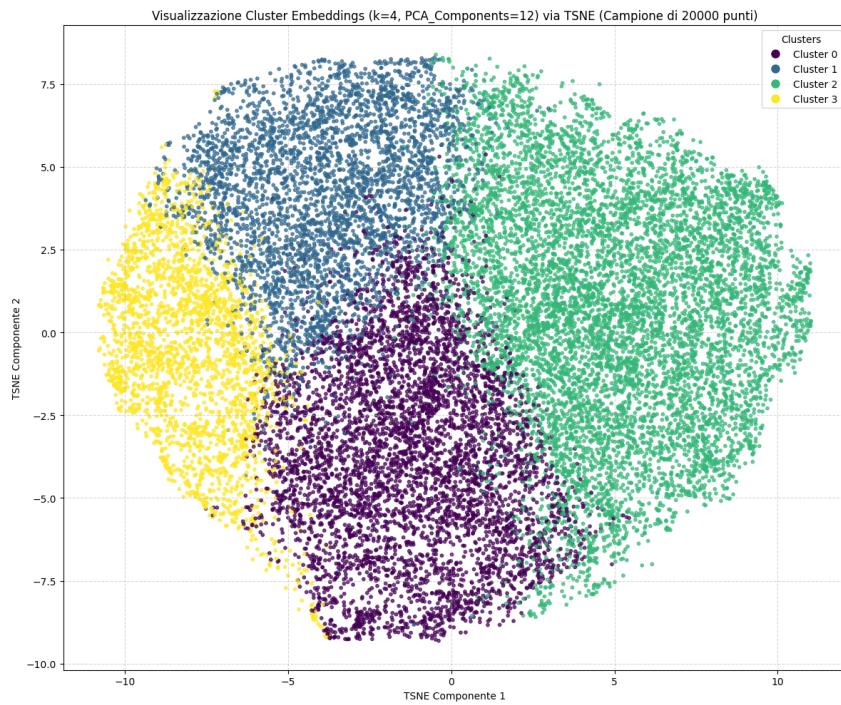


Figura 2.5: Clustering K-Means su Embeddings con PCA a 12 componenti e visualizzazione TSNE

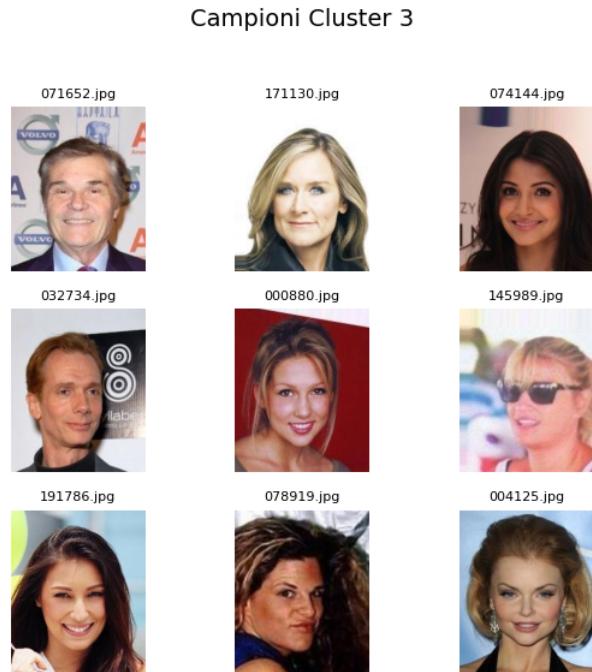


Figura 2.6: Campione di immagini provenienti dal terzo Cluster ottenuto con PCA a 4 componenti sugli embeddings

2.5.2 Clustering su attributi CSV

Il clustering basato sui 34 attributi CSV dopo la feature selection ha mostrato risultati significativamente influenzati dalla riduzione di dimensionalità tramite PCA. Inizialmente, con 28 com-

ponenti PCA (che spiegano circa il 95.7% della varianza), il Silhouette score è relativamente basso (0.1059).

Tuttavia, riducendo ulteriormente il numero di componenti, il Silhouette score aumenta marcatamente (Figura 2.8). Con 7 componenti raggiunge 0.2112, con 4 componenti 0.3090, e il punteggio più alto in assoluto tra tutti gli esperimenti è ottenuto con soli 2 componenti PCA (0.5116). Questo incremento suggerisce che, per gli attributi CSV, una forte riduzione della dimensionalità può isolare le variazioni più rilevanti per il raggruppamento in K-Means, o che il Silhouette score è particolarmente sensibile alla dimensionalità molto bassa in questo caso. Visivamente, lo scatter plot t-SNE per 2 componenti PCA degli attributi (2.7) ha mostrato una chiara separazione dei cluster, il che è coerente con l'aumento del Silhouette score in uno spazio 2D.

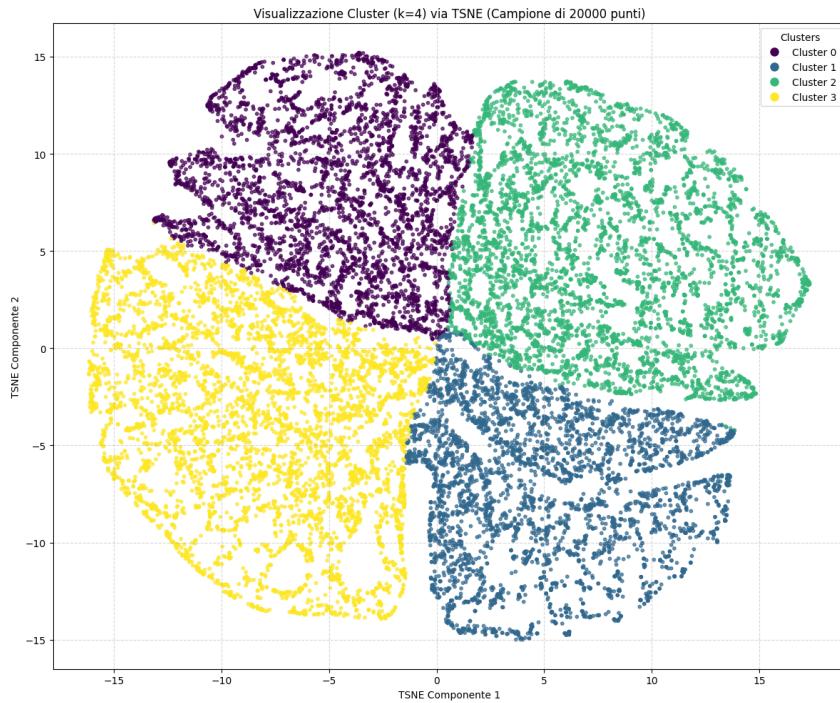


Figura 2.7: Clustering K-Means su CSV con PCA a 2 componenti e visualizzazione TSNE

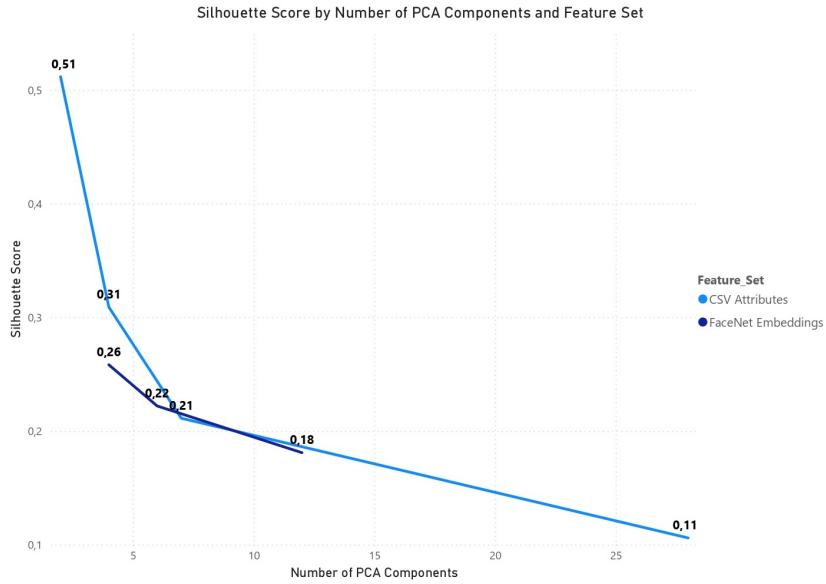


Figura 2.8: Grafico a linee del Silhouette Score in funzione del numero di Componenti PCA per Feature Set tramite Power BI

2.6 Discussione dei risultati dell'approccio baseline

L'analisi degli esperimenti di clustering K-Means ($K = 4$) condotti sui dati CelebA rivela che gli approcci iniziali, basati sia su embedding generici di FaceNet sia sugli attributi CSV pre-elaborati, non hanno prodotto raggruppamenti ottimali. Questa sezione esplora le possibili cause di tali risultati.

Una prima criticità emerge dalla natura stessa del dataset CelebA. Come osservato nell'analisi preliminare (Figura 2.1), molti attributi facciali presentano una distribuzione fortemente sbilanciata, con alcune feature estremamente rare e altre quasi onnipresenti. Questa disomogeneità intrinseca può rendere difficile per algoritmi come K-Means, che si basano su misure di distanza, identificare cluster ben bilanciati e significativi, poiché le feature rare potrebbero avere un impatto limitato sulla definizione delle similarità tra campioni.

Inoltre, l'analisi della matrice di correlazione degli attributi CSV (Figura 2.2) ha mostrato una generale assenza di forti e pervasive strutture di correlazione. Sebbene esistano alcune correlazioni logiche (es. tra baffi e barba), la mancanza di blocchi di feature altamente co-ocorrenti o mutuamente esclusive implica che gli attributi non si raggruppano naturalmente in insiemi distinti. Di conseguenza, K-Means fatica a trovare partizioni chiare basate sulla similarità complessiva delle feature, portando a cluster con confini sfumati.

Per quanto riguarda gli embedding di FaceNet, il Silhouette score iniziale di 0.1624 riflette la difficoltà di separare i cluster. L'applicazione della PCA (Figura 2.8) ha portato a un leggero miglioramento (fino a 0.2582 con 4 componenti), suggerendo che la riduzione del rumore e la focalizzazione sulle dimensioni a maggiore varianza possono essere parzialmente utili. Tuttavia, le visualizzazioni t-SNE (es. Figura 2.5) hanno costantemente mostrato una notevole sovrapposizione tra i cluster, indipendentemente dal numero di componenti PCA. Questo suggerisce che un modello FaceNet **pre-addestrato per il riconoscimento facciale generico** potrebbe non generare embedding ottimali per il compito specifico di raggruppare le immagini in base a una combinazione sottile di attributi facciali, che non erano l'obiettivo primario del suo addestramento. Le feature apprese potrebbero essere troppo orientate alla discriminazione dell'identità piuttosto che alla similarità basata su attributi multipli.

Relativamente agli attributi CSV, i risultati del clustering sono stati particolarmente scarsi quando si è tentato di preservare gran parte della varianza originale tramite PCA (Silhouette score di 0.1059 con 28 componenti). È interessante notare come una drastica riduzione dimensionale (fino a 2 componenti PCA) abbia portato al Silhouette score più alto (0.5116, Figura 2.8). Se da un lato questo indica una buona separazione in uno spazio a bassissima dimensionalità, dall’altro solleva interrogativi sull’interpretabilità e la robustezza di cluster definiti da così poche componenti aggregate, che potrebbero aver perso molta dell’informazione originale dei 34 attributi. L’incremento potrebbe anche riflettere una peculiarità del Silhouette score in spazi a dimensionalità estremamente ridotta piuttosto che una vera scoperta di raggruppamenti profondamente significativi nei dati originali.

In sintesi, gli esperimenti iniziali evidenziano che né gli embedding generici di FaceNet né gli attributi CSV grezzi o con PCA ad alta dimensionalità conducono a cluster ottimali con K-Means. I Silhouette score sono generalmente rimasti modesti, indicando che i cluster formati potrebbero non essere perfettamente sferici o ben separati, o che $K = 4$ potrebbe non essere il numero ottimale di cluster per queste rappresentazioni. L’ispezione visiva delle immagini campione dai cluster (omessa qui per brevità ma presente nel notebook originale) aiuta a comprendere il significato semantico di questi raggruppamenti, ma la loro chiarezza e utilità pratica rimangono limitate.

Queste osservazioni preliminari motivano l’esplorazione di un approccio più sofisticato per la generazione di feature, ripreso dal lavoro di Anzalone et al. [2]. L’idea è di utilizzare il transfer learning per addestrare una rete neurale specificamente sulla predizione degli attributi facciali del dataset CelebA. L’ipotesi è che gli embedding estratti da una tale rete, ottimizzata per comprendere gli attributi, possano fornire una base più robusta e significativa per il successivo clustering, e potenzialmente per un’analisi più approfondita dei bias, rispetto agli embedding generici di FaceNet o agli attributi CSV direttamente utilizzati. Questo costituirà la base per la fase successiva della nostra analisi.

3. Approccio basato su Anzalone et al.

Dopo un'esplorazione preliminare del dataset e una prima fase di clustering basata sugli embedding facciali ottenuti attraverso l'impiego del modello FaceNet, si è deciso di adottare un approccio metodologico più raffinato e semanticamente significativo. Tale approccio mira a rappresentare i volti non più mediante vettori astratti, bensì attraverso l'esplicitazione dei loro attributi facciali, con l'obiettivo di migliorare l'interpretabilità e la coerenza semantica dei risultati. A tal fine, si è fatto riferimento al lavoro di Anzalone et al. [2], intitolato "Transfer Learning for Facial Attributes Prediction and Clustering", che propone una strategia efficace per la predizione e il raggruppamento dei volti basata su attributi facciali esplicativi. In particolare, gli autori hanno sviluppato un modello di classificazione multi-etichetta basato sull'architettura MobileNetV2, opportunamente modificata mediante l'aggiunta di uno strato denso (dense layer) composto da 37 neuroni, ognuno dei quali attivato da una funzione sigmoide. Ciascun neurone restituisce un valore continuo compreso tra 0 e 1, interpretato come la probabilità stimata della presenza dell'attributo facciale corrispondente.

La funzione di attivazione sigmoide utilizzata è definita come segue:

$$\hat{y}_j = \sigma(z_j) = \frac{1}{1 + e^{-z_j}}$$

dove \hat{y}_j rappresenta la probabilità stimata per l'attributo j , e z_j è l'input lineare del neurone. Per trasformare tali probabilità in predizioni binarie, si applica una soglia a 0.5 secondo la seguente formula:

$$\hat{y}_j = [round(\hat{y}_1), round(\hat{y}_2), \dots, round(\hat{y}_{37})]$$

Ogni immagine viene dunque rappresentata da un vettore binario di 37 dimensioni, in cui ciascun elemento indica la presenza (1) o l'assenza (0) di un determinato attributo, come ad esempio Smiling, Male, Eyeglasses, Blond_Hair, ecc.

Grazie alla pubblicazione dei pesi del modello ottimizzato di Anzalone et al.(file nominato weights-FC37-MobileNetV2-0.92.hdf5), è stato reso possibile l'utilizzo diretto della rete pre-addestrata per effettuare predizioni su nuovi dati, evitando così il processo di addestramento, notoriamente dispendioso dal punto di vista computazionale. Considerata l'estensione del dataset CelebA, che comprende oltre 200.000 immagini, ciò ha rappresentato un vantaggio considerevole. Il passaggio successivo consiste nell'applicazione su un terminale in locale del modello pre-addestrato all'intero dataset, ottenendo per ciascuna immagine un vettore binario rappresentante gli attributi facciali. È fondamentale sottolineare che tali predizioni non costituiscono embedding generici per il riconoscimento facciale (come quelli ottenuti con FaceNet), bensì sono l'output diretto di una rete neurale addestrata specificamente per il compito di classificazione multi-etichetta degli attributi facciali.

Il modello in questione è stato addestrato mediante l'utilizzo della funzione di perdita *Cosine Proximity*, particolarmente adatta a scenari multi-etichetta con output sparsi. Essa è formalmente definita come:

$$L = -\frac{\mathbf{y} \cdot \hat{\mathbf{y}}}{\|\mathbf{y}\|_2 \cdot \|\hat{\mathbf{y}}\|_2}$$

dove:

- \mathbf{y} è il vettore binario delle etichette reali (ground truth);
- $\hat{\mathbf{y}}$ è il vettore delle probabilità predette dalla rete.

Questa funzione ha l'obiettivo di massimizzare l'allineamento tra le direzioni dei due vettori, trattandoli come punti nello stesso spazio vettoriale e penalizzando la dissimilitudine angolare tra predizione e realtà.

A differenza di quanto effettuato dagli autori originali, i quali hanno applicato il proprio modello a un campione casuale di 2.000 immagini estratte dall'intera collezione del dataset CelebA (senza quindi un esplicito riferimento alle partizioni di training, validation o test ufficiali per questa specifica valutazione), nel presente lavoro si è deciso di estendere le predizioni a tutte le immagini disponibili nel dataset CelebA. È importante notare che, poiché il modello di Anzalone et al. è stato sottoposto a *fine-tuning* utilizzando la partizione di training di CelebA, l'applicazione dei suoi pesi all'intero dataset implica che le nostre predizioni includono anche le immagini che il modello ha "visto" durante la sua fase di addestramento. Di conseguenza, le metriche di performance (accuratezza e tasso di errore) calcolate sull'intero dataset potrebbero risultare più ottimistiche rispetto a una valutazione condotta su un set di test completamente inedito per il modello. Tuttavia, poiché il nostro obiettivo primario era ottenere un vettore di attributi per ogni immagine da utilizzare nelle successive fasi di clustering, e non una valutazione rigorosa della capacità di generalizzazione del modello di Anzalone su dati unseen, questa scelta è stata ritenuta accettabile. Il nostro scopo era verificare la corretta applicazione dei pesi e la coerenza generale delle predizioni.

Le metriche di performance ottenute, in termini di accuratezza e tasso di errore per ogni feature, come visualizzato nelle figure sottostanti (Figura 3.2: "Risultati Anzalone" e Figura 3.1: "Nostrti Risultati"), si sono rivelate sostanzialmente in linea con i risultati riportati da Anzalone et al. Questo suggerisce una corretta implementazione nell'utilizzo dei pesi forniti e una comparabilità generale nel comportamento del modello, pur tenendo conto della potenziale influenza positiva sulle metriche dovuta all'inclusione dei dati di training nella nostra valutazione.

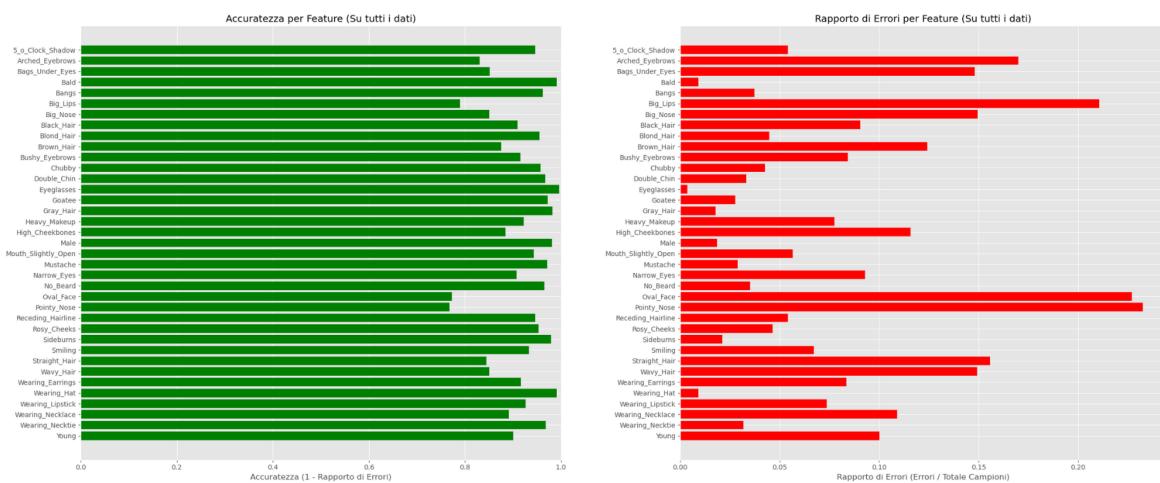


Figura 3.1: Risultati ottenuti in questo lavoro seguendo il metodo di Anzalone

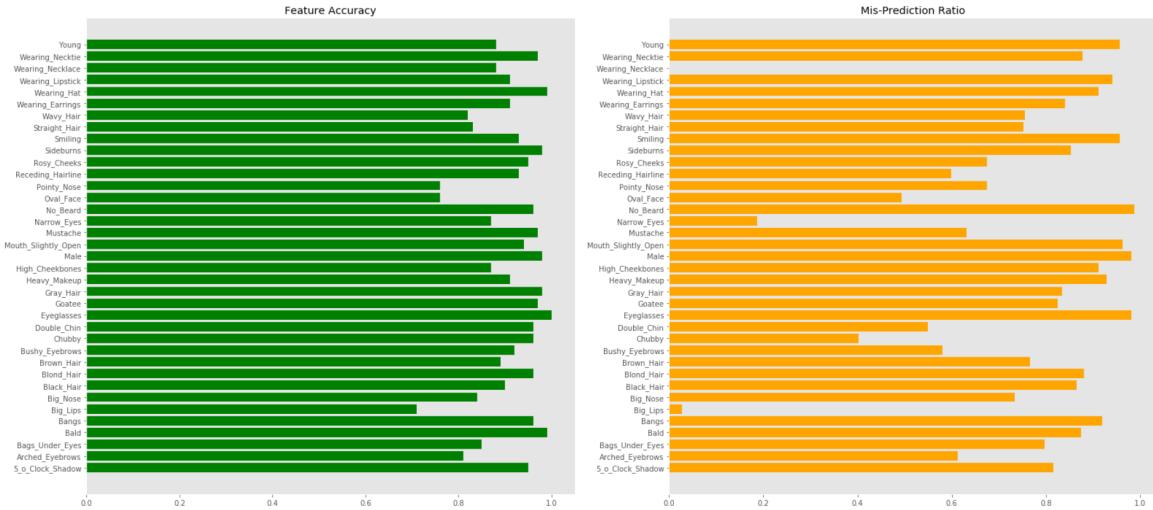


Figura 3.2: Risultati ottenuti da Anzalone

I vettori binari così ottenuti costituiscono la base per la successiva fase del progetto, dedicata all’analisi di clustering. Quest’ultima sarà orientata a raggruppare i volti in funzione della similarità semantica tra i rispettivi attributi facciali, con un duplice obiettivo: da un lato, esplorare la struttura latente del dataset in relazione alle caratteristiche facciali esplicite; dall’altro, indagare l’eventuale presenza di bias nei dati o nelle predizioni del modello (e quindi nelle feature usate per il clustering), valutando la fairness dei risultati rispetto a specifici gruppi demografici o tratti distintivi.

3.1 Approccio basato su Anzalone et al.

Dopo aver esplorato gli approcci di clustering baseline, l'indagine si è concentrata sull'approfondimento dell'impatto della qualità e della natura semantica delle feature sulla formazione dei cluster. Specificamente, si è voluto testare in modo comparativo l'efficacia di utilizzare attributi facciali predetti da un modello specializzato, seguendo la metodologia di Anzalone et al. [2], mettendola a confronto con l'utilizzo degli attributi originali forniti dal dataset CelebA. L'obiettivo primario di questa serie di esperimenti è stato quindi quello di isolare e quantificare il contributo apportato dalle feature predette da un modello addestrato sugli attributi, rispetto all'utilizzo diretto delle etichette ground truth, mantenendo per quanto possibile costanti le altre condizioni sperimentalistiche, come la logica di selezione delle feature e l'algoritmo di clustering. In tutti gli esperimenti descritti in questa sezione, è stato impiegato l'algoritmo K-Means, e per permettere un confronto diretto con i risultati e la granularità proposti da Anzalone et al., il numero di cluster K è stato fissato a 15.

3.1.1 Clustering su attributi originali del dataset CelebA

Come primo passo per stabilire un termine di paragone basato sui dati grezzi, sono stati condotti esperimenti di clustering utilizzando direttamente gli attributi originali del dataset CelebA, applicando due diverse strategie di selezione delle feature.

Attributi originali con esclusione di feature percettive

Inizialmente, per mitigare l'influenza di attributi potenzialmente troppo soggettivi o legati a qualità dell'immagine piuttosto che a caratteristiche intrinseche del volto, è stato condotto un esperimento **rimuovendo le feature "Attractive", "Pale-Skin" e "Blurry"** dal set completo dei 40 attributi originali. Questo ha portato a un dataset con 37 attributi.

Su questo insieme di 37 feature, è stato applicato l'algoritmo K-Means ($K = 15$). Il **Silhouette Score**, calcolato su un campione di 20.000 punti, per questa configurazione è risultato essere **0.1117**. Tale punteggio indica una qualità di clustering molto bassa, suggerendo che l'ampio numero di feature originali, anche epurate da quelle più marcatamente percettive, non facilita la formazione di gruppi distinti e coesi.

Attributi originali selezionati secondo la logica di Anzalone et al.

Per un confronto più diretto con l'approccio di Anzalone et al., un secondo esperimento sugli **attributi originali** ha previsto la selezione di un sottoinsieme di sole 10 feature, corrispondenti concettualmente a quelle da loro identificate come particolarmente rilevanti da Anzalone et al.. Le feature originali (ground truth) considerate sono state: 'Wearing_Lipstick', 'Smiling', 'No_Beard', 'Heavy_Makeup', 'Bald', 'Male', 'Young', 'Eyeglasses', 'Blond_Hair', 'Wearing_Hat'.

Applicando K-Means ($K = 15$) a questo dataset ridotto di 10 attributi originali, il **Silhouette Score** è salito notevolmente a **0.6568**. Questo dimostra che una selezione mirata basata su una logica semantica può migliorare drasticamente la qualità del clustering anche quando si utilizzano gli attributi originali, isolando le informazioni più discriminanti.

3.1.2 Replicazione esperimento Anzalone et al. con feature predette

Avendo stabilito delle baseline con gli attributi originali, l'esperimento successivo ha mirato a replicare pienamente l'approccio di Anzalone et al., utilizzando questa volta le *predizioni* del loro modello specializzato. Come descritto nel capitolo precedente, sono stati generati i vettori binari dei 37 attributi facciali per ogni immagine. Da queste predizioni, sono state estratte le stesse 10 feature concettualmente rilevanti utilizzate nell'esperimento precedente (*anzalone_features* nel notebook): 'Wearing_Lipstick_pred', 'Smiling_pred', 'No_Beard_pred', 'Heavy_Makeup_pred', 'Bald_pred', 'Male_pred', 'Young_pred', 'Eyeglasses_pred', 'Blond_Hair_pred', 'Wearing_Hat_pred'.

L'algoritmo K-Means ($K = 15$) è stato quindi applicato a questo dataset di 10 feature *predette*. Su un campione di 20.000 punti, questa configurazione ha prodotto il **Silhouette Score più elevato tra tutti gli esperimenti di clustering: 0.7229**.

L'ispezione qualitativa delle immagini campione in Figura 3.3 ha ulteriormente confermato la coerenza semantica dei raggruppamenti.

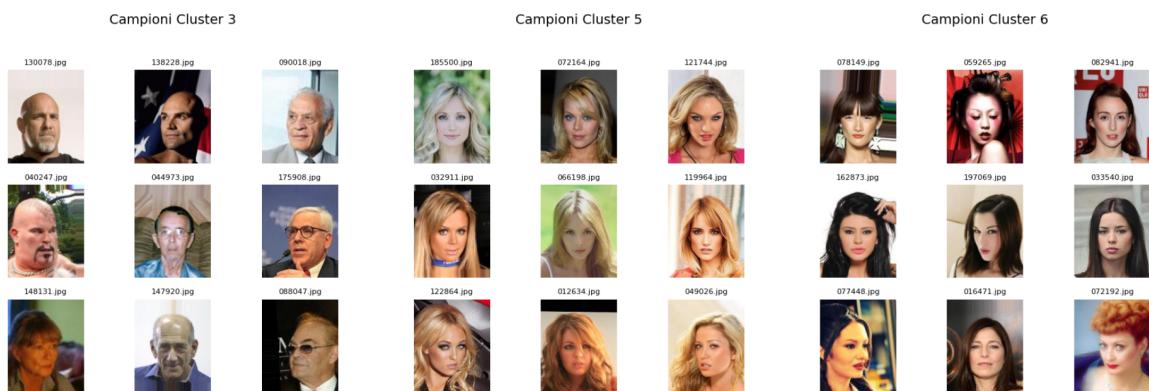


Figura 3.3: Campioni di immagini provenienti da tre Clusters ottenuti dalle 10 feature predette selezionate da Anzalone et al.

Il confronto diretto dei Silhouette Score ottenuti in questi tre esperimenti con $K = 15$ è particolarmente illuminante:

- Attributi Originali (37 feature, percettive escluse): **0.1117**
- Attributi Originali (10 feature selezionate da Anzalone et al.): **0.6568**
- Attributi Predetti (10 feature predette da Anzalone at al.): **0.7229**

Questi risultati dimostrano in modo conclusivo che l'approccio di utilizzare feature predette da un modello specializzato per gli attributi facciali (Anzalone et al.) fornisce una base dati moderatamente più strutturata ed idonea per il clustering K-Means, portando a raggruppamenti di qualità superiore. Anche una selezione oculata degli attributi originali, basata sulla stessa logica semantica, offre un notevole miglioramento rispetto all'uso indifferenziato di un set più ampio di feature originali. L'incremento di performance passando dalle 10 feature originali selezionate (0.6568) alle 10 feature predette (0.7229) suggerisce inoltre che il modello di Anzalone non solo identifica attributi rilevanti, ma le sue predizioni binarizzate (ottenute da output sigmoidi)

potrebbero aver regolarizzato o "pulito" le rappresentazioni delle feature, rendendole più consistenti e quindi più adatte a formare cluster distinti. Data la superiorità dimostrata, i 15 cluster ottenuti replicando l'approccio di Anzalone con le feature *predette* sono stati scelti come base per la successiva analisi approfondita della predizione dell'attributo "Attractive" e dei relativi bias, come discusso nella Sezione 3.2.

3.2 Addestramento e valutazione dei classificatori per l'attributo "Attractive"

Successivamente, è stata intrapresa un'analisi focalizzata sulla predizione dell'attributo "Attractive", utilizzando come ground truth i valori originali forniti dal dataset CelebA. Questa fase sperimentale è stata condotta in tre esperimenti per esplorare l'impatto di diverse configurazioni di feature e strategie di bilanciamento del training set.

In tutti gli esperimenti effettuati sono stati utilizzati due modelli di classificazione, ovvero un modello di Regressione Logistica ed un modello Random Forest. Per entrambi i modelli, è stata eseguita una fase di tuning degli iperparametri tramite *GridSearchCV* con Cross Validation impostata su 3 fold (*CV_FOLDS* = 3) ottimizzando la metrica *AUC-ROC*.

Primo Esperimento di Classificazione: All'interno del primo esperimento entrambi questi modelli iniziali hanno utilizzato come input le 10 feature facciali predette tramite l'applicazione del modello di Anzalone et al. sull'intero dataset (*anzalone_features*). Per il modello **Random Forest**, lo spazio di ricerca degli iperparametri prevedeva una griglia così formata:

- *n_estimators*: [100, 200, 300]
- *max_depth*: [10, 20, 30, None]
- *min_samples_leaf*: [1, 2, 4, 6]
- *min_samples_split*: [2, 5, 10]

Il modello ottimale ha presentato i seguenti iperparametri: *n_estimators*: 100, *max_depth*: 10, *min_samples_leaf*: 2, *min_samples_split*: 5.

Per la **Regressione Logistica**, gli iperparametri sottoposti a tuning erano:

- *logreg_C* (parametro di regolarizzazione): [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]
- *logreg_penalty*: ['l1', 'l2']

Gli iperparametri ottimali identificati per la Regressione Logistica sono stati: *C*: 0.01, *penalty*: l1.

In questo esperimento per mitigare l'impatto dello sbilanciamento della classe target "Attractive" all'interno dei singoli cluster, l'intero dataset è stato bilanciato tramite una tecnica di "undersampling per cluster", dove l'undersampling veniva applicato in modo da bilanciare la proporzione della feature *Attractive* **all'interno di ogni cluster prima dello split train/test**. Dopo l'addestramento dei modelli con gli iperparametri ottimali, le loro performance (in termini di accuratezza, F1-score per entrambe le classi, AUC-ROC, True Positive Rate - TPR e False Positive Rate - FPR per la classe "Attractive") sono state valutate in modo disaggregato, ovvero

separatamente per ciascuno dei 15 cluster precedentemente identificati attraverso il clustering K-Means sugli attributi di Anzalone.

La valutazione per cluster ha rivelato una variabilità nelle performance. Ad esempio, per il Random Forest, l'AUC media sui cluster è stata di circa 0.606, con F1-score medio per la classe "Attractive" (f1_attractive_1) di circa 0.694 e un'accuratezza media di 0.594. Per la Regressione Logistica, l'AUC media è risultata circa 0.597, con un F1-score medio per la classe "Attractive" di circa 0.558 e un'accuratezza media di 0.571. Queste metriche disaggregate hanno permesso di identificare i cluster su cui i modelli generalizzavano meglio o peggio.

Secondo Esperimento di Classificazione: Considerando queste prime osservazioni, per esplorare ulteriormente l'impatto della selezione delle feature sui risultati finali, è stato condotto un secondo esperimento. In questa iterazione, è stato costruito un **nuovo set esteso di 20 feature (new_feature_set)**.

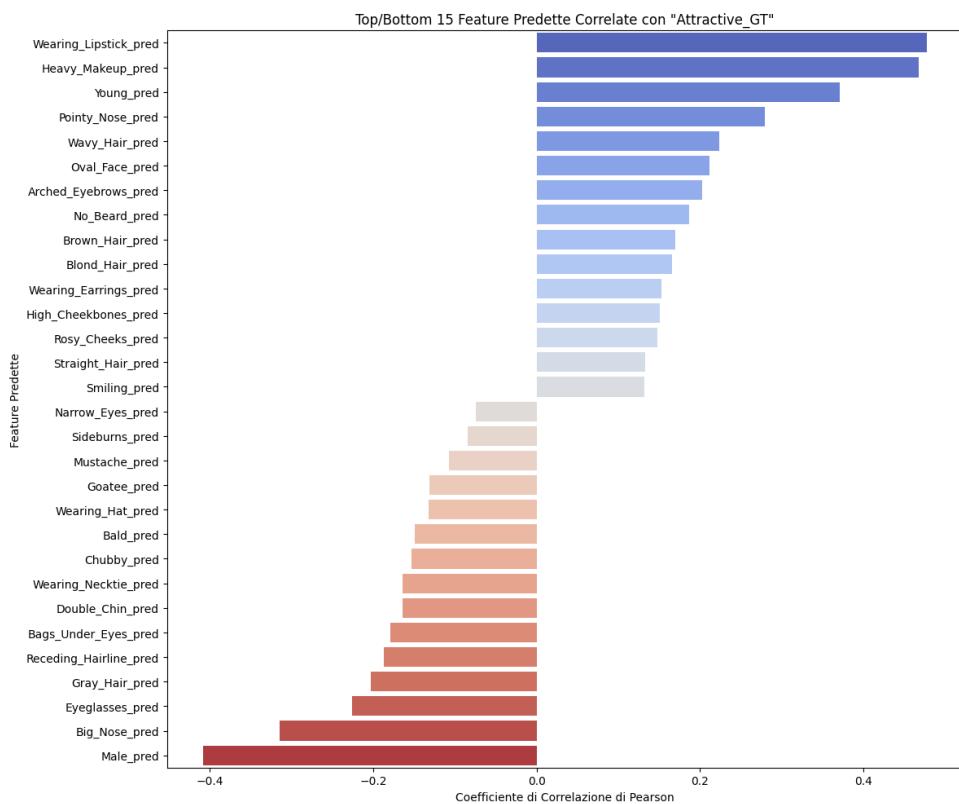


Figura 3.4: 15 feature con correlazione più alta in modulo rispetto alla classe target Attractive (Ground Truth).

Questo set è stato composto dalle 10 feature chiave prese da Anzalone et al., a cui sono state aggiunte ulteriori 5 feature con la più alta correlazione positiva e 5 feature con la più alta correlazione negativa (in valore assoluto) rispetto all'attributo "Attractive" (ground truth), escludendo quelle già presenti nel set di Anzalone (come illustrato in Figura 3.4). Il nuovo set di feature è quindi così composto: ['Arched_Eyebrows_pred', 'Bags_Under_Eyes_pred', 'Bald_pred', 'Big_Nose_pred', 'Blond_Hair_pred', 'Brown_Hair_pred', 'Double_Chin_pred', 'Eyeglasses_pred', 'Gray_Hair_pred', 'Heavy_Makeup_pred', 'Male_pred', 'No_Beard_pred', 'Oval_Face_pred', 'Pointy_Nose_pred', 'Receding_Hairline_pred', 'Smiling_pred', 'Wavy_Hair_pred', 'Wearing_Hat_pred', 'Wearing_Lipstick_pred', 'Young_pred'].

Su questo nuovo set di feature sono stati nuovamente addestrati e ottimizzati i classificatori

Random Forest e Regressione Logistica, utilizzando la stessa strategia di bilanciamento del primo esperimento (undersampling per cluster prima dello split tra training set e test set). Per il modello di Random Forest (*RandomForest_NewFeatures*), gli iperparametri ottimali sono risultati: *n_estimators: 300, max_depth: 30, min_samples_leaf: 4, min_samples_split: 15*. Per il modello di Regressione Logistica(*LogisticRegression_NewFeatures*), gli iperparametri ottimali sono stati: *C: 10, penalty: l1*. Le performance medie ottenute sono state per Random Forest: AUC 0.711, F1 "Attractive" 0.651, accuratezza 0.652. Per Regressione Logistica: AUC 0.691, F1 "Attractive" 0.537, accuratezza 0.613.

Terzo Esperimento di Classificazione: Un’ulteriore modifica metodologica introdotta in questa fase è stata l’adozione di una tecnica di bilanciamento del training set più mirata, denominata "**Robust Sampling**", applicata sempre al *new_feature_set* da 20 feature. A differenza dell’approccio precedente però, questa tecnica prevede lo split train-test sui dati originali di ciascun cluster, seguito dall’undersampling della classe maggioritaria **solo sulla porzione di training del cluster**. Dunque, il *test set* di ciascun cluster è stato mantenuto nella sua forma originale e sbilanciata, al fine di ottenere una valutazione delle performance più realistica e rappresentativa delle condizioni operative reali. Gli iperparametri per i modelli "*RandomForest_RobustSamp*" e "*LogReg_RobustSamp*" sono stati nuovamente ottimizzati. Per Random Forest si è giunti ai medesimi valori del secondo esperimento: *n_estimators: 300, max_depth: 30, min_samples_leaf: 4, min_samples_split: 15*; e per Regressione Logistica: *C: 10, penalty: l1*. Le metriche di performance medie ottenute con questo approccio hanno mostrato per Random Forest: AUC 0.713, F1 "Attractive" 0.583, accuratezza 0.668. Per Regressione Logistica: AUC 0.688, F1 "Attractive" 0.532, accuratezza 0.687. Questo suggerisce che la combinazione del set di feature esteso e della strategia di "Robust Sampling" possa portare a modelli globalmente più performanti e con una migliore capacità di generalizzazione sui dati di test sbilanciati.

Di seguito è riportata una tabella, generata all’interno di una dashboard in Power BI, che sintetizza i risultati dei diversi esperimenti. La tabella mostra le metriche medie calcolate su tutti i cluster, per ciascun modello addestrato in ogni esperimento (cfr. Figura 3.5).

model	Average of auc	Average of accuracy	Average of f1_attractive_0	Average of f1_attractive_1	Average of fpr_attractive_1	Average of tpr_attractive_1
RandomForest_RobustSamp	0.71	0.67	0,62	0.58	0.37	0.68
RandomForest_NewFeatures	0.71	0.65	0.65	0.65	0.36	0.66
RandomForest	0.61	0.59	0.39	0.69	0.73	0.92
LogReg_RobustSamp	0.69	0.69	0.64	0.53	0.29	0.51
LogisticRegression_NewFeatures	0.69	0.61	0.65	0.54	0.28	0.51
LogisticRegression	0.60	0.57	0.41	0.56	0.62	0.76

Figura 3.5: Sommario delle metriche ottenute per gli esperimenti sulla Classificazione - ottenuto tramite PowerBI

3.2.1 Analisi SHAP

In seguito ai risultati dell’ultimo esperimento è stata effettuata un’analisi di interpretabilità tramite la libreria SHAP (SHapley Additive exPlanations). Questa libreria mette a disposizione un metodo di spiegazione dei modelli di apprendimento automatico che si basa sui valori Shapley, un concetto della teoria dei giochi. In tal modo è possibile capire come ogni singola caratteristica di un modello contribuisca alla previsione finale, fornendo un’interpretazione locale e globale del processo decisionale del modello [4]. L’obiettivo era investigare più a fondo l’impatto e il contributo delle singole feature alla predizione dell’attributo "Attractive" all’ind

terno di ciascun cluster, utilizzando i modelli Random Forest (*RandomForest_RobustSamp*) e Regressione Logistica (*LogReg_RobustSamp*) precedentemente ottimizzati.



Figura 3.6: Risultati analisi SHAP per RandomForest_RobustSamp.

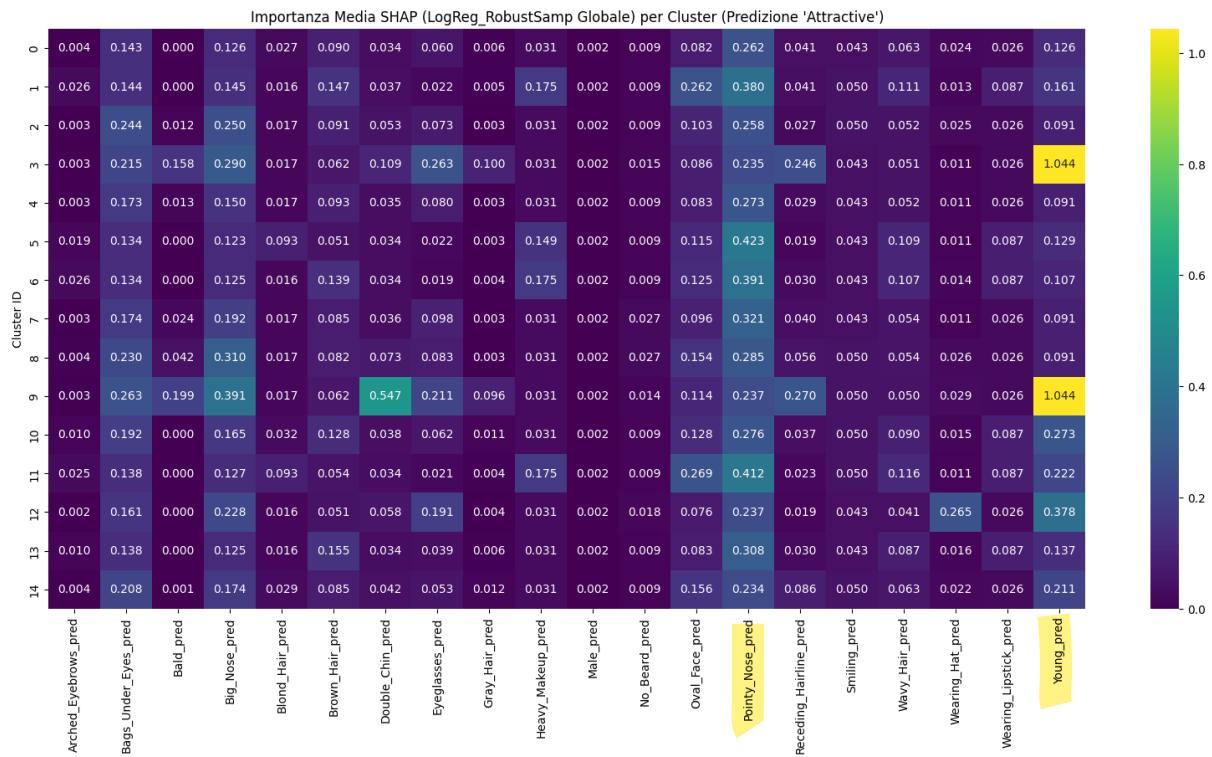


Figura 3.7: Risultati analisi SHAP per LogReg_RobustSamp.

Applicando SHAP a questi modelli per le istanze appartenenti a ciascun cluster (valutate sui rispettivi test set originali e sbilanciati), è stato possibile quantificare l'importanza media assoluta di ciascuna delle 20 feature predittive per ogni cluster. I risultati, visualizzati tramite heatmap che confrontano i valori SHAP medi per feature attraverso i 15 cluster (come mostrato nelle Figure 3.6 e 3.7), hanno confermato una notevole eterogeneità.

L'importanza relativa delle diverse feature variava considerevolmente da un cluster all'altro per entrambi i modelli. Ad esempio, feature come "Young_pred" hanno mostrato un impatto positivo e costante nella maggior parte dei cluster, indicando la giovinezza percepita come un fattore determinante per l'attrattività. Al contrario, "Male_pred" ha teso ad avere un'influenza negativa, coerente con le osservazioni precedenti che vedevano questa feature correlata negativamente con la classe target "Attractive". Feature come "Bald_pred" o "Wearing_Hat_pred" sono apparse generalmente poco influenti. Tuttavia, l'importanza di altre feature, come "Pointy_Nose_pred" per Random Forest o "Heavy_Makeup_pred" e "Wearing_Lipstick_pred" per Logistic Regression, è risultata più variabile e significativa solo in specifici sottogruppi.

Queste differenze nell'importanza delle feature suggeriscono che i modelli, pur essendo addestrati globalmente, apprendono e si affidano a diverse combinazioni di segnali visivi a seconda delle caratteristiche predominanti dei volti all'interno di ciascun cluster. Questo non solo riflette la complessità della predizione di un attributo soggettivo, ma sottolinea anche come i bias presenti nel dataset o appresi durante la predizione degli attributi possano manifestarsi in modo differenziato. Ad esempio, cluster con una forte presenza di determinate caratteristiche (es. trucco evidente) potrebbero portare il modello a sovrappesare tali feature per la predizione di "Attractive" in quel contesto specifico.

Nei cluster più difficili e sbilanciati (ad esempio, il Cluster 3), dove i modelli hanno faticato a riconoscere correttamente i casi positivi di "Attractive" nonostante l'undersampling, l'analisi SHAP potrebbe rivelare una forte influenza di feature con correlazione negativa o una dipendenza da segnali deboli, contribuendo a spiegare le basse performance (es. F1-score basso per la classe positiva, pur con AUC accettabile). Questa analisi, resa possibile da SHAP e dalla valutazione per cluster su dati di test sbilanciati, conferma che l'attributo "Attractive" è predetto sulla base di segnali complessi e, non sempre distribuiti o interpretati equamente attraverso i diversi sottogruppi demografici implicitamente rappresentati dai cluster.

4. Conclusioni e Lavori Futuri

Questo progetto ha esplorato l'applicazione del clustering K-Means al dataset CelebA, dimostrando che l'approccio basato su attributi facciali predetti da un modello specializzato, come quello ispirato ad Anzalone et al. [2], è in grado di generare cluster con una separabilità notevolmente superiore (Silhouette Score di 0.72 con K=15) rispetto all'utilizzo di embedding generici estratti da FaceNet (Silhouette Score massimo di 0.26 con K=4 e PCA a 4 componenti) o agli attributi CSV originali (Silhouette Score massimo di 0.65 con K=15 ed il set di feature selezionato da Anzalone et al.). Tale risultato sottolinea l'efficacia di feature semanticamente ricche e mirate per compiti di clustering basati su attributi facciali specifici, confermando che una rappresentazione più esplicita dei tratti distintivi del volto facilita l'identificazione di raggruppamenti più coerenti.

L'analisi successiva, focalizzata sulla predizione dell'attributo "Attractive" (utilizzando le etichette ground truth di CelebA) sui cluster identificati, ha rivelato notevoli variazioni nell'importanza delle feature e nelle performance del modello tra i diversi cluster. Questa eterogeneità, emersa chiaramente nel terzo esperimento con il new_feature_set da 20 attributi e la strategia di "Robust Sampling", indica non solo la complessità intrinseca della predizione di un attributo così soggettivo, ma solleva anche interrogativi cruciali sulla presenza di bias e sulla necessità di strategie di modellazione più consapevoli.

Come evidenziato da Böhlen, Chandola e Salunkhe nel loro studio "Server, server in the cloud. Who is the fairest in the crowd?" (2018) [5], la valutazione algoritmica dell'attrattività è storicamente problematica e intrinsecamente soggetta a bias. Essi sottolineano che l'attributo "Attractive" è particolarmente "illusivo" e meno oggettivamente definibile rispetto ad altri attributi facciali, portando a performance di classificazione generalmente inferiori anche con architetture avanzate. La loro analisi del dataset CelebA, su cui anche il nostro lavoro si basa, rivela che le etichette sono state fornite da un gruppo demograficamente specifico di annotatori (50 partecipanti cinesi tra i 20 e i 30 anni) su immagini di celebrità. Questo processo di annotazione, come discusso da Böhlen et al., introduce potenziali bias culturali e di percezione nelle etichette "ground truth" stesse. Le disparità di performance osservate nel nostro studio, in particolare nel terzo esperimento dove l'AUC media per Random Forest si è attestata a 0.713 e per Regressione Logistica a 0.688, potrebbero quindi non derivare unicamente dai modelli impiegati o dalle feature selezionate, ma essere profondamente radicate nelle caratteristiche intrinseche del dataset e nel processo di etichettatura originale. L'analisi SHAP (Figure 3.6 e 3.7), condotta sui modelli addestrati con il new_feature_set e Robust Sampling, ha ulteriormente corroborato questa visione. Abbiamo osservato che l'importanza relativa delle diverse feature variava considerevolmente da un cluster all'altro per entrambi i modelli. Ad esempio, mentre "Young_pred" ha mostrato un impatto positivo e costante, coerentemente con le osservazioni di Böhlen et al. sulla prevalenza della giovinezza nei canoni di bellezza computazionali (fenomeno da loro definito "computational ageism"), e "Male_pred" ha avuto un'influenza tendenzialmente negativa, altre feature come "Pointy_Nose_pred" o "Heavy_Makeup_pred" hanno mostrato un'importanza variabile e significativa solo in specifici sottogruppi.

Questa variabilità suggerisce che i cluster da noi identificati, pur essendo formati sulla base di attributi predetti dal modello di Anzalone, potrebbero aver implicitamente raggruppato volti per i quali l'etichetta "Attractive" è particolarmente ambigua o riflette le percezioni specifiche e potenzialmente biased degli annotatori originali, piuttosto che un concetto universale di bellezza. La difficoltà dei modelli, emersa anche nel nostro studio, nel riconoscere correttamente i casi positivi di "Attractive" in cluster particolarmente sbilanciati (come il Cluster 3, che ha mostrato F1-score bassi per la classe positiva nonostante un'AUC accettabile), potrebbe essere una manifestazione di questi bias sottostanti e della natura "sfuggente" dell'attributo stesso, come teorizzato da Böhnen et al. L'analisi SHAP, evidenziando la dipendenza da segnali deboli o da feature con correlazione negativa in tali cluster, supporta questa interpretazione.

La strategia di "*Robust Sampling*" adottata (test su dati sbilanciati originali) ha rappresentato un passo verso una valutazione più realistica delle performance dei modelli. Tuttavia, l'interpretazione dei risultati di fairness e delle disparità predittive deve sempre considerare le problematiche intrinseche all'etichettatura di "Attractive" e la potenziale propagazione di bias attraverso le diverse fasi di modellazione: dal dataset originale, al modello di Anzalone per la predizione degli attributi, fino ai nostri classificatori finali. Discernere l'origine precisa di tali bias (dataset, modello intermedio, o classificatore finale) rimane una sfida complessa.

In conclusione, sebbene l'utilizzo di feature derivate da modelli specializzati come quello di Anzalone et al. migliori significativamente la qualità del clustering rispetto ad approcci più generici, la successiva predizione di attributi altamente soggettivi come "Attractive" rimane un compito arduo, profondamente influenzato dalla natura dei dati e dai bias in essi contenuti. L'analisi SHAP si è rivelata uno strumento prezioso per illuminare queste dinamiche complesse e le variazioni di comportamento dei modelli attraverso i cluster.

Per i lavori futuri, sarebbe interessante esplorare tecniche di mitigazione dei bias più avanzate, sia a livello di dati che di modello, e confrontare i risultati con dataset etichettati da gruppi di annotatori più diversificati per investigare ulteriormente l'impatto culturale sulla percezione dell'attrattività. Inoltre, l'integrazione di approcci di "explainable AI" più sofisticati potrebbe fornire insight ancora più profondi sui meccanismi decisionali dei modelli in contesti così complessi e sensibili.

L'uso di modelli pre-addestrati, come il modello Inception pre-addestrato su ImageNet, implica l'ereditare scelte e valori impliciti nel dataset originale di pre-addestramento. Il bias può quindi essere introdotto sia dalla selezione dei dati per il compito specifico sia dalla selezione dell'architettura e dal dataset di pre-addestramento.

Il paper conclude sottolineando l'importanza di investigare la complessità culturale della provenienza e cura dei dati negli algoritmi basati sui big data. Questi algoritmi stanno sempre più formulando giudizi di gusto, un'area precedentemente riservata agli esseri umani e difficile da formalizzare, creando una nuova classe di problemi computazionali-sociali. La binarizzazione dell'attrattività non riflette la complessità delle influenze umane quasi subliminali. L'abilità computazionale nel valutare l'età con precisione rende la giovinezza un proxy accessibile per l'elusivo concetto di bellezza, potenzialmente favorendo una nuova forma di ageismo computazionalmente abilitato

Bibliografia

- [1] Z. Liu, P. Luo, X. Wang e X. Tang. «Large-scale CelebFaces Attributes (CelebA) Dataset». In: *Retrieved from <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>.* 2015.
- [2] L. Anzalone, P. Barra, S. Barra, F. Narducci e M. Nappi. «Transfer Learning for Facial Attributes Prediction and Clustering». In: *Communications in Computer and Information Science*. Vol. 1127. CCIS. Springer, 2019, pp. 101–113. DOI: [10.1007/978-981-15-1301-5_9](https://doi.org/10.1007/978-981-15-1301-5_9).
- [3] F. Ali. *FaceNet - TensorFlow Model*. <https://www.kaggle.com/models/faiqueali/facenet-tensorflow>. Accessed: 2025-05-15. 2023.
- [4] S. M. Lundberg e S.-I. Lee. «A Unified Approach to Interpreting Model Predictions». In: *Advances in Neural Information Processing Systems 30*. A cura di I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774. URL: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
- [5] M. Böhlen, V. Chandola e A. Salunkhe. «Server, server in the cloud. Who is the fairest in the crowd?» In: *CoRR abs/1711.08801* (2017). arXiv: [1711.08801](https://arxiv.org/abs/1711.08801). URL: [http://arxiv.org/abs/1711.08801](https://arxiv.org/abs/1711.08801).