Progetto di Intelligenza Artificiale

Impatto del benessere psicologico degli studenti sul loro rendimento accademico

Corso di Laurea Magistrale in Informatica

A.A. 2023/24

Nome del team

BiblioTeam

Membri del team

Delia Cavalca - 0001126401
delia.cavalca@studio.unibo.it

Claudia Brunetti - 0001128544
claudia.brunetti4@studio.unibo.it

Indice

1 Introduzione

1.1 Descrizione del Problema

La salute mentale degli studenti universitari è una questione di crescente importanza, considerando gli elevati livelli di stress e pressione che molti affrontano durante il percorso accademico. Identificare i fattori chiave che influenzano il benessere psicologico degli studenti è fondamentale per migliorare la loro esperienza universitaria.
L'analisi dell'andamento accademico degli studenti potrebbe rivelare correlazioni e pattern che riflettono l'impatto delle loro condizioni sociali e psicologiche.
È importante analizzare e comprendere come queste variabili abbiano influenza sull'andamento accademico degli studenti per poter adottare strategie mirate di supporto e intervento.

L'obiettivo primario di questo progetto è quello di capire l'impatto del benessere psicologico degli studenti sul loro rendimento accademico.
Attraverso un'analisi dei dati relativi alle prestazioni accademiche e alle condizioni psicologiche degli studenti, si vogliono identificare correlazioni tra salute mentale e successo o fallimento accademico. Questo consentirà di sviluppare interventi mirati per migliorare il benessere complessivo degli studenti.
I benefici di una soluzione efficace sono molteplici: miglioramento del rendimento accademico, riduzione del tasso di abbandono degli studi, promozione del benessere psicologico e creazione di un ambiente universitario più sano e inclusivo.

1.2 Soluzione Proposta

1.2.1 Approccio alla soluzione

Per affrontare la sfida relativa all’identificazione della correlazione tra le caratteristiche degli studenti e il loro andamento accademico, adotteremo un approccio sperimentale che integra l'analisi dei dati raccolti attraverso un questionario somministrato a studenti di differenti corsi di laurea dell'Università degli Studi di Bologna.

Il metodo sperimentale prevede una prima descrizione e rappresentazione dei dati. La rappresentazione dei dati degli studenti prevede due gruppi di informazioni: le informazioni relative al percorso accademico e quelle relative alla loro condizione psicologica. Questo approccio consentirà di valutare come queste due dimensioni influenzino reciprocamente il rendimento degli studenti.

Per quanto riguarda il percorso accademico, si considerano i dati relativi a:
- **corso di studi (CDL)**;
- **anno del corso (ANNO)**;
- **voto di diploma (VOTO DIPLOMA)**, riflette l'attitudine allo studio;
- **numero di esami superati (NUMERO ESAMI)**, rappresentano il raggiungimento di un obiettivo;
- **numero di esami non superati (NON SUPERATI)**, il mancato superamento di un esame è un evento che potrebbe causare una rottura nel percorso dello studente, come perdita di motivazione, aumento di frustrazione o abbandono del percorso;
- **numero di esami risostenuti (RISOSTENUTI)** per un miglioramento del voto;
- **media dei voti (MEDIA VOTI),** potrebbe avere forte impatto su soddisfazione e motivazione;

Per quanto riguarda la condizione psicologica, si esaminano i dati relativi a:
- **soddisfazione (SODD)** in relazione al corso di studi scelto (uno studente soddisfatto è maggiormente motivato, impegnato e ottimista; ciò può contribuire a un migliore equilibrio emotivo e mentale);
- **motivazione (MOT_INTR, MOT_ID, MOT_ES)** allo studio, con distinzione tra intrinseca, identificata, estrinseca (la mancanza di motivazione può portare a frustrazione, insoddisfazione e stress);
- **autogestione (LOC_INT)**, sicurezza e abilità nella gestione del proprio percorso di studi (gli studenti maggiormente sicuri tendono ad essere meno stressati e più resilienti di fronte alle difficoltà);
- **interazione con i docenti (IIS_INT_DOC)**, influenza su obiettivi, atteggiamenti e aspirazioni (possono favorire un ambiente stimolante, contribuendo a maggiore soddisfazione e motivazione);
- **attenzione dei docenti (IIS_ATT)** riguardo lo sviluppo degli studenti e dell'insegnamento;
- **interazione con altri studenti (IIS_PARI)**, rapporti sociali (possono promuovere il senso di appartenenza, il supporto reciproco e la condivisione di esperienze, elementi fondamentali per il benessere mentale);
- **sviluppo accademico e intellettuale (IIS_SVACC)**, influenza dell'esperienza accademica sullo sviluppo intellettuale (ha impatto sulla soddisfazione dello studente);
- **impegno verso l'istituzione e gli obiettivi (IIS_IMP1, IIS_IMP2)**, gradimento verso l'università e impegno nel raggiungere un certo obiettivo;
- **insoddisfazione (RIC_IMP)** in relazione al corso di studi scelto (può causare smarrimento e stress);
- **regolazione emotiva (REAG_TEAM)**, capacità di gestire situazioni difficili, controllo dei sentimenti (importante per affrontare lo stress accademico e le sfide personali);
- **capacità di lavorare in team (TEAM)**;

- **autoefficacia accademica (AUT_ACC)**, sicurezza nelle proprie capacità e nel riuscire ad organizzare ed eseguire azioni per raggiungere gli obiettivi prefissati (influenza la motivazione, la fiducia e il benessere complessivo);
- **procrastinazione (RISP_SCA)**, incapacità nel rispettare obiettivi prefissati, può indicare scarsa motivazione e concentrazione (può avere impatto sulla salute mentale, causando stress, ansia e sensi di colpa);
- **chiarezza di carriera (CH_CARR)**, consapevolezza rispetto alle proprie aspirazioni, interessi e obiettivi professionali futuri (l'incertezza o l'insicurezza riguardo al futuro professionale possono generare ansia, stress e frustrazione, influenzando negativamente la motivazione e l'autostima);
- **ricerca di carriera (RIC_CARR)**, conoscenza e pensiero critico relativo alle opportunità di lavoro presenti sul mercato (la mancanza di opportunità soddisfacenti o la difficoltà nel trovare un lavoro adatto può causare frustrazione e ridurre l'autostima);
- **strategie di miglioramento (TASK_CRA)**, capacità di esplorare nuovi metodi, adattare i compiti agli interessi e capacità personali, dare priorità alle attività di studio più adatte (influenzano soddisfazione e motivazione);
- **autoefficacia di carriera (AUT_CAR)**, capacità di definire obiettivi professionali e step per raggiungerli terminato il percorso accademico (la mancanza di fiducia nelle proprie capacità può causare frustrazione e ridurre l'autostima, influenzando negativamente la motivazione);

Oltre a tali informazioni, si considerano anche il genere e l’età di ciascuno studente.

Per analizzare i dati si è scelto di utilizzare la tecnica del clustering. Questo approccio mira a separare gli studenti in gruppi distinti sulla base delle loro caratteristiche, consentendo di individuare pattern e tendenze.

Inizialmente, i dati sono stati suddivisi in base all’anno di corso frequentato dagli studenti. Questa suddivisione ci ha consentito di esaminare le dinamiche nei differenti anni di studio. L’analisi di clustering è stata quindi svolta separatamente su ciascuna ripartizione.

L’intenzione è esplorare le caratteristiche rilevanti e rappresentative di ciascun cluster. Sono stati condotti diversi test di clustering per ciascun gruppo di studenti, escludendo man mano differenti features per valutare l’impatto di determinate caratteristiche sulla formazione dei cluster.

Una particolare attenzione è stata posta rispetto alla feature relativa al numero di esami sostenuti, in modo da valutare se le caratteristiche comuni identificate in un certo cluster corrispondono a un determinato numero di esami sostenuti dagli studenti del gruppo stesso.

Questo approccio ha fornito indicazioni sulla possibile correlazione tra benessere psicologico e andamento accademico.

1.2.2 Sfide informatiche affrontate

Prima di condurre l'analisi vera e propria è stata necessaria una fase di pre-elaborazione dei dati, al fine di garantirne qualità e coerenza.
Di fondamentale importanza è stata la gestione dei valori mancanti, dei valori incoerenti e la normalizzazione del formato dei valori.

In seguito, sono stati analizzati differenti algoritmi di clustering.
La principale difficoltà riscontrata è stata l'individuare metodologie in grado di ottenere cluster ben distinti. Tale problema è causato dalla somiglianza dei dati fra loro. Di conseguenza, è stato importante definire criteri specifici per l'analisi e la valutazione dei cluster al fine di identificare le caratteristiche più rilevanti di ciascuno.

1.2.3 Divisione dei compiti nel gruppo

Il progetto è stato portato a termine tramite una serie di step (definizione obiettivo, definizione rappresentazione dei dati, preparazione dei dati, studio di algoritmi di clustering, applicazione algoritmi scelti e analisi dei risultati).
Ad ogni step, i compiti sono stati distribuiti in modo equo fra i membri del gruppo.
Ognuno ha potuto lavorare in modo individuale sui propri compiti e ci siamo regolarmente confrontate per condividere idee, discutere dei progressi e valutare come procedere.

1.2.4 Risultati ottenuti in sintesi

I risultati dell'analisi dei dati mediante clustering hanno rivelato una mancanza di distinzione significativa tra i cluster identificati. È emersa una correlazione interessante tra soddisfazione e motivazione degli studenti. Tuttavia, è stato difficile dimostrare un impatto significativo della condizione psicologica sull'andamento accademico. Questi risultati mostrano la necessità di ulteriori ricerche per comprendere al meglio le dinamiche sottostanti al successo degli studenti e l'importanza del loro benessere psicologico.

2 Metodo Proposto

2.1 Scelta della Soluzione

Si è deciso di analizzare i risultati relativi a sei tecniche di clustering differenti, in modo da avere un maggior numero di considerazioni da confrontare e poter definire conclusioni maggiormente precise.

Ogni algoritmo è stato valutato attraverso il coefficiente di silhouette, un indicatore di coesione dei cluster ottenuti, utile per confrontare e valutare diversi algoritmi e determinare il numero ottimale di cluster in un dataset.
Il coefficiente di silhouette calcola quanto un punto è simile ai punti all'interno dello stesso cluster rispetto a quelli in altri cluster. Viene calcolato per ciascun punto nel dataset e può assumere valori compresi tra -1 e 1. Se il risultato è vicino a 1 allora i punti sono stati assegnati ai cluster corretti, se è vicino a 0 significa che i punti si trovano vicino al limite tra due cluster, se è vicino a -1 i punti potrebbero essere stati assegnati ai cluster sbagliati.
In generale, un coefficiente di silhouette più alto indica una migliore qualità di clustering, con cluster compatti e ben separati.

Di seguito vengono descritti gli algoritmi di clustering utilizzati.

2.1.1 K-MEANS Clustering

L'algoritmo K-MEANS clusterizza cercando di separare i campioni in n gruppi con varianza uguale, minimizzando il criterio di inerzia. L'inerzia è una misura che indica quanto i cluster sono internamente coerenti.
Ogni cluster può essere rappresentato attraverso il corrispettivo "centroide", ovvero il campione corrispondente alla media dei campioni nel cluster.
Tale algoritmo richiede che il numero di cluster k desiderati sia specificato. Analizzando i risultati e il coefficiente di silhouette per diversi valori di k, si stabilisce che il valore ottimale di cluster è k = 2.

Parametri dell'algoritmo:

num_clusters	2

2.1.2 DBSCAN Clustering

L'algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) considera i cluster come aree di alta densità separate da aree di bassa densità. Il concetto di densità viene espresso attraverso i parametri min_samples ed eps.
Un campione viene definito come campione centrale se esistono almeno min_samples altri campioni entro una distanza eps, definiti come vicini del campione

centrale. Un cluster è quindi un insieme di campioni centrali che può essere costruito prendendo ricorsivamente un campione centrale, trovando tutti i suoi vicini, e così via.
Un qualsiasi campione che non è un campione centrale, e che è almeno a eps distanza da qualsiasi campione centrale, è considerato un outlier.
È importante definire in modo appropriato il valore del parametro eps: se troppo piccolo la maggior parte dei dati non verrà clusterizzata, se troppo grande i cluster vicini vengono fusi in uno unico. Per questo, è stato analizzato il grafico delle distanze dei vicini più prossimi per individuare tale valore.
Scelti i parametri dell'algoritmo, è stato controllato il valore relativo al numero di outlier, per evitare che la clusterizzazione coinvolgesse solo un numero ristretto di dati.

Parametri dell'algoritmo:

min_samples	30
eps	5

2.1.3 Agglomerative Hierarchical Clustering

Il clustering gerarchico è una famiglia di algoritmi che costruiscono cluster nidificati unendo o dividendo man mano determinati cluster. La gerarchia di cluster può essere rappresentata tramite un albero.
Nello specifico, l'algoritmo Agglomerative Hierarchical Clustering utilizza un approccio bottom-up: ogni osservazione è inizialmente in un proprio cluster e i cluster vengono successivamente uniti. Il criterio di linkage dell'algoritmo determina la metrica utilizzata come strategia di fusione. Si è scelto di utilizzare linkage = ward, ovvero la strategia che minimizza la somma delle differenze quadratiche all'interno di tutti i cluster (minimizza la varianza), in quanto permette di ottenere risultati di clusterizzazione migliori rispetto alle altre.
Analizzando il dendrogramma gerarchico, si è stabilito che il valore ottimale di cluster da analizzare è 2. I dati sono infatti molto simili e vicini fra loro, considerare un numero più elevato di cluster risulterebbe poco significativo.

Parametri dell'algoritmo:

num_clusters	2
linkage	ward

2.1.4 UMAP Clustering

UMAP (Uniform Manifold Approximation and Projection) è un algoritmo di riduzione della dimensionalità utilizzato per proiettare i dati in uno spazio a dimensionalità

inferiore, preservando la struttura e le relazioni dei dati originali. Questo algoritmo è noto per la sua capacità di mantenere sia la struttura globale che quella locale dei dati, rendendolo una scelta preferita rispetto al t-SNE. All'interno dell'algoritmo UMAP, è stato utilizzato anche il Elbow method e l'algoritmo KMeans. Il primo è stato impiegato per determinare il numero ottimale di cluster da creare. Questo viene identificato nel punto in cui il valore della metrica smette di diminuire rapidamente, fornendo una guida chiara sulla struttura dei dati; nel nostro caso il numero ottimale risulta essere 2. Successivamente, è stato applicato l'algoritmo KMeans per suddividere gli elementi nei cluster. I centroidi di questi cluster sono stati poi selezionati come punti di riferimento.

Parametri dell'algoritmo:

num_clusters	2

2.1.5 SOM Clustering

Il Self-Organizing Map (SOM) è un algoritmo di clustering non supervisionato che organizza i dati in modo topologico su una griglia di neuroni, identificando le Best Matching Units (BMU) per ogni dato e utilizzandole per formare cluster. La BMU è il neurone della griglia che ha i pesi più simili al dato considerato. Utilizzando le BMU, la SOM suddivide i dati in cluster, dove ogni cluster è formato dai dati che sono stati associati alla stessa BMU. Questo processo consente di raggruppare i dati in base alle loro somiglianze, senza la necessità di etichette o supervisione esterna. Una volta formati i cluster, è possibile analizzare i centroidi di ciascun cluster. È importante notare che l'algoritmo SOM restituirà rappresentazioni, cluster e centroidi diversi in base ai dati utilizzati per l'addestramento e per il test. Questa variazione dipende dalla casualità introdotta nella suddivisione dei dati in set di addestramento e di test. La divisione casuale dei dati è fondamentale per valutare l'affidabilità e la generalizzazione dell'algoritmo rispetto a differenti sottoinsiemi di dati. Nell'esperimento, più precisamente utilizzeremo l'algoritmo minisom visto il numero ristretto di dati.

2.1.6 GMM Clustering

Il Gaussian Mixture Model (GMM) è un algoritmo di clustering che si basa sull'idea che i dati siano generati da una combinazione di diverse distribuzioni gaussiane nello spazio delle features. Il GMM identifica la struttura nascosta nei dati modellando questa miscela di gaussiane sovrapposte e assegnando i punti ai cluster in base alla probabilità di appartenenza a ciascuna gaussiana. Questo processo avviene attraverso quattro fasi principali: inizializzazione dei parametri, E-step, M-step e convergenza. Durante l'assegnazione dei cluster, ogni punto viene associato al cluster con la probabilità di appartenenza più alta, calcolata utilizzando la densità di probabilità delle gaussiane associate a ciascun cluster. Questo approccio flessibile

consente al GMM di catturare la struttura complessa dei dati, consentendo la sovrapposizione delle gaussiane e l'assegnazione flessibile dei punti ai cluster in base alla loro distribuzione di probabilità. Successivamente, per valutare ulteriormente la distribuzione dei dati all'interno dei cluster, vengono calcolate le statistiche sui centroidi.

Parametri dell’algoritmo:

num_clusters	4

3 Risultati Sperimentali

3.1 Dimostrazione e Tecnologie

I diversi algoritmi di clustering presentati sono stati applicati ai dataset degli studenti dei diversi anni di studi.
L'obiettivo è l'individuazione di caratteristiche rilevanti e rappresentative di ogni cluster, al fine di delineare una correlazione tra benessere psicologico e rendimento accademico.
Per fare ciò, ottenuta la clusterizzazione di certi dati, sono stati confrontati i valori delle features nei diversi cluster calcolandone per ciascuna media e varianza. Sono state considerate maggiormente identificative quelle features con media differente nei due cluster e varianza bassa.
Nello specifico, il calcolo effettuato per valutare se una certa feature risulta essere rappresentativa dei cluster è il seguente: data la media e la varianza della feature nei due cluster differenti, si controlla se vi è una sovrapposizione fra l'intorno della media nel primo cluster con l'intorno della media nel secondo cluster. Se non vi è sovrapposizione, tale caratteristica è identificativa dei cluster.
Vista la scarsa variabilità dei dati e conseguente bassa capacità di clusterizzazione, si è deciso di considerare rappresentative anche quelle features con una piccola sovrapposizione, oltre ad analizzare i risultati ottenuti nel loro insieme.

I risultati ottenuti dai differenti algoritmi sono stati confrontati fra loro per delineare al meglio determinate conclusioni.

Per lo svolgimento del progetto abbiamo fatto affidamento su tecnologie specifiche per garantire la riproducibilità e la coerenza dei risultati.
Per quanto riguarda l'ambiente di sviluppo, abbiamo utilizzato Python nella versione 3.11.0. Per l'implementazione degli algoritmi di clustering, abbiamo impiegato la libreria scikit-learn nella versione 1.4.0, la libreria minisom nella versione 2.3.1 e umap-learn nella versione 0.5.5. Queste librerie sono ampiamente riconosciute per la loro versatilità e robustezza nel campo del machine learning e del data mining, offrendo una vasta gamma di algoritmi e funzionalità per l'analisi dei dati.

3.2 Risultati

3.2.1 Risultati K-MEANS Clustering

Parametro di sovrapposizione accettata utilizzato

soglia	0.5

Analisi studenti del primo anno
In un primo test non è stata esclusa nessuna feature dal dataset.

Tale configurazione non ha permesso di identificare in modo significativo caratteristiche rappresentative dei cluster individuati.
In seguito, si è cercato di fare clustering escludendo alcune feature considerate meno rilevanti rispetto agli obiettivi, come età, corso di studi e genere.
Da tale analisi è emerso che i due cluster individuati sono maggiormente differenziati rispetto all'indice di sviluppo accademico e intellettuale e rispetto all'indice di impegno verso l'istituzione e gli obiettivi. Inoltre, risultano significativi gli indicatori di motivazione intrinseca e motivazione identificata.
Il primo cluster, rappresenta studenti meno soddisfatti del proprio sviluppo intellettuale e della propria esperienza accademica, non sicuri di aver fatto la decisione giusta nel frequentare tale università. Tali soggetti sono allo stesso tempo meno motivati.
Il secondo cluster individua, invece, soggetti con alta soddisfazione accademica e che hanno una forte spinta nel voler conoscere nuove cose (motivazione intrinseca) e ritengono molto importante concludere i propri studi per poter avere maggiori possibilità professionali (motivazione identificata).
Inoltre, è possibile notare come gli studenti del secondo cluster abbiano sostenuto in media un maggior numero di esami, non ne abbiano superati in media un numero inferiore e abbiano dovuto risostenere meno esami.
Tali differenze relative agli esami sono però ridotte e i relativi dati sono abbastanza variabili.

Analisi studenti del secondo anno
Analizzando i cluster individuati sugli studenti del secondo anno, ancora una volta emergono come caratteristiche rilevanti lo sviluppo accademico e intellettuale e l'impegno verso l'istituzione e gli obiettivi, oltre all'indicatore di riconsiderazione dell'impegno.
Il gruppo di studenti meno soddisfatti dal punto di vista della scelta accademica e del proprio sviluppo intellettuale risultano essere maggiormente propensi a cambiare università e professione futura.
Ancora una volta, emerge che gli studenti maggiormente soddisfatti del proprio percorso hanno sostenuto in media un numero più elevato di esami.

Analisi studenti del terzo anno
Anche per gli studenti del terzo anno, risultano maggiormente rappresentative dei cluster individuati le caratteristiche che esprimono la soddisfazione degli studenti.
In corrispondenza del gruppo di studenti maggiormente soddisfatti, è possibile notare maggiore motivazione e migliori risultati nel percorso accademico.

In generale, risulta quindi emergere una correlazione tra soddisfazione e motivazione, che potrebbe avere influenza sul numero di esami sostenuti e in generale sull'andamento del percorso accademico.

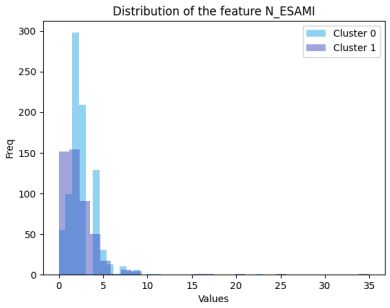
Tuttavia, indipendentemente dall'anno di studi, le differenze osservate tra i due cluster sono piuttosto lievi. I valori medi dei relativi indicatori mostrano infatti scostamenti ridotti. Questo risultato è ulteriormente dimostrato dal coefficiente di silhouette restituito dall'algoritmo, sempre intorno a 0.10. Tale valore indica che i cluster identificati mostrano una distinzione poco marcata fra loro. Pertanto, nonostante siano emerse tendenze interessanti riguardo la soddisfazione, motivazione e impegno accademico degli studenti, occorre considerare ulteriori analisi per comprendere meglio le dinamiche presenti nel campione di studenti considerato.

Tabella risultati K-MEANS Clustering

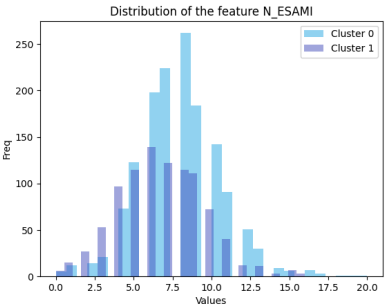
studenti	features maggiormente rilevanti	features rilevanti (meno)
1 anno	IIS_SVACC, IIS_IMP1	MOT_INTR, MOT_ID
2 anno	IIS_SVACC, IIS_IMP1	RIC_IMP
3 anno	IIS_SVACC	RIC_IMP

Grafico K-MEANS Clustering (risultati test 3):
analisi distribuzione N_ESAMI nei differenti cluster

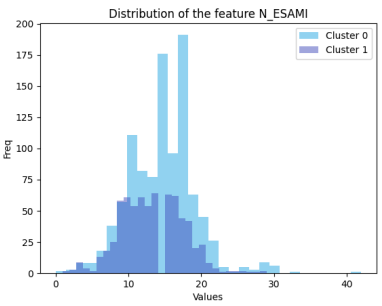
studenti 1 anno



studenti 2 anno



studenti 3 anno



3.2.2 Risultati DBSCAN Clustering

Tale algoritmo, applicato ai tre gruppi di studenti, è in grado di individuare un solo cluster in ciascun gruppo. Siccome i dati sono molto simili fra loro, con una densità abbastanza uniforme nel dataset, DBSCAN non è in grado di individuare cluster distinti.
Per questo, i risultati ottenuti sono stati analizzati in modo differente. Si è cercato di individuare quali fossero le caratteristiche più rappresentative del cluster, ovvero quelle con bassa varianza.

Parametro per considerare una feature rilevante

varianza max	0.5

Analisi studenti del primo anno
Dai risultati, emerge che la maggior parte degli studenti del primo anno ha elevata motivazione intrinseca, ritiene che l'università sia una opportunità importante per le proprie future possibilità professionali, non ha avuto una particolare interazione coi docenti, è abbastanza soddisfatto della propria scelta accademica e dell'impatto della stessa sulla propria crescita intellettuale, non ritiene di voler cambiare percorso universitario, ha una buona regolazione emotiva e capacità organizzative in merito al proprio studio.

Analisi studenti del secondo anno
Gli studenti del secondo anno risultano avere elevata motivazione intrinseca, sono molto soddisfatti del proprio livello di sviluppo intellettuale e dell'esperienza accademica, si impegnano molto nel raggiungere i propri obiettivi accademici, non vogliono cambiare percorso di studi e hanno una buona regolazione emotiva.

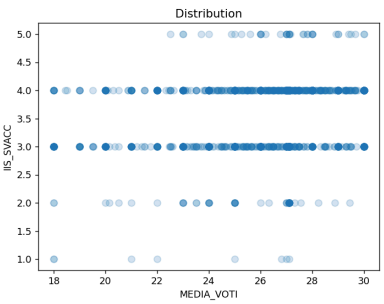
Analisi studenti del terzo anno
Le caratteristiche che si distinguono per essere maggiormente rappresentative degli studenti del terzo anno sono quelle che ne esprimono la soddisfazione. Gli studenti sono soddisfatti del proprio livello di sviluppo intellettuale e ritengono molto importante l'impegno verso l'università e gli obiettivi.

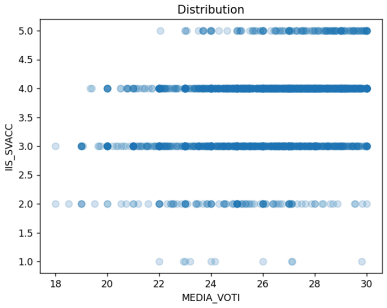
In nessun gruppo di studenti è stato possibile individuare come identificative caratteristiche relative all'andamento del percorso accademico. Il numero di esami superati e la media dei voti risulta essere molto variabile.
Questa osservazione riflette differenze dal punto di vista del progresso accademico all'interno di ciascun gruppo. Non è quindi detto che esista una forte correlazione fra condizione psicologica (motivazione e soddisfazione) e risultati accademici.

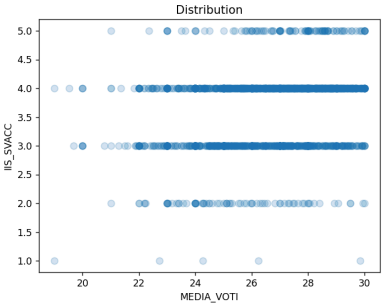
Tabella risultati DBSCAN Clustering

studenti	features maggiormente rilevanti
1 anno	IIS_SVACC, IIS_IMP1, AUT_CAR, CH_CARR, IIS_ATT, IIS_INTDOC, REGOLAZIONE_EMOTIVA
2 anno	IIS_SVACC, IIS_IMP1, AUT_CAR, MOT_INTR, REGOLAZIONE_EMOTIVA
3 anno	IIS_SVACC, IIS_IMP2

Grafico DBSCAN Clustering (risultati test 3):
analisi correlazione fra IIS_SVACC e MEDIA_VOTI
studenti 1 anno studenti 2 anno studenti 3 anno







3.2.3 Risultati Agglomerative Hierarchical Clustering

Con tale algoritmo è stato difficile individuare caratteristiche rappresentative dei cluster, i cluster individuati risultano essere poco distinguibili fra loro.
Si è deciso di accettare un maggior livello di sovrapposizione fra i valori delle features per tentare un'analisi.

Parametro di sovrapposizione accettata utilizzato

soglia	1

Analisi studenti del primo anno
Risultano essere maggiormente significative le caratteristiche che identificano la soddisfazione degli studenti e il numero di esami risostenuti.
Gli studenti maggiormente soddisfatti della propria carriera accademica hanno risostenuto in media un numero inferiore di esami.
Si nota, in generale, una correlazione fra andamento accademico e soddisfazione, che non è però possibile dimostrare a causa dell'elevata varianza dei dati nei cluster.

Analisi studenti del secondo anno
Anche in questo caso i risultati ottenuti sono poco significativi a causa dell'elevata varianza degli indicatori.
Si nota come gli studenti maggiormente soddisfatti e motivati hanno una media dei voti più alta e hanno sostenuto un maggior numero di esami.
Tuttavia, i due cluster risultano essere molto simili e non è possibile dimostrare una concreta correlazione fra andamento accademico e condizione psicologica dei soggetti.

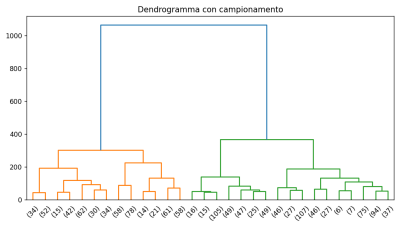
Analisi studenti del terzo anno
Negli studenti del terzo anno emerge una correlazione tra motivazione intrinseca e soddisfazione dal punto di vista dello sviluppo accademico e intellettuale. Gli studenti maggiormente soddisfatti del proprio percorso sono più motivati.
Nonostante emergano queste caratteristiche, i due cluster risultano essere molto simili, la correlazione è da analizzare in modo accurato per capire se effettivamente presente nella realtà.

Tabella risultati Agglomerative Hierarchical Clustering

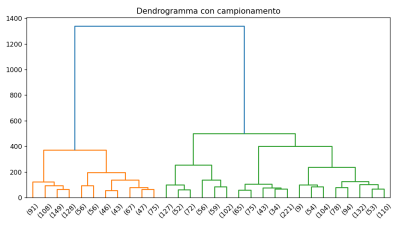
studenti	features maggiormente rilevanti	features rilevanti (meno)
1 anno	-	IIS_SVACC, IIS_IMP1, IIS_ATT, RISOSTENUTI
2 anno	-	IIS_IMP1
3 anno	-	IIS_SVACC, MOT_INTR

Grafico Hierarchical Clustering (risultati test 3)

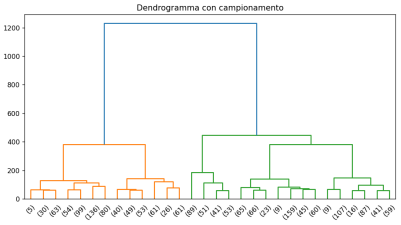
studenti 1 anno



studenti 2 anno



studenti 3 anno



3.2.4 Risultati UMAP Clustering

Durante l'analisi condotta su studenti del primo, secondo e terzo anno, sono stati eseguiti tre test nei quali progressivamente sono state escluse alcune feature del dataset ritenute meno significative.
Nel primo test sono state considerate tutte le features, mentre nel secondo caso sono state escluse l'età, il genere e il corso di laurea. Nella terza analisi, oltre alle suddette feature, è stato escluso anche il numero di esami superati.
In tutti e tre i test, per l'identificazione delle features rilevanti, è stato necessario accettare un maggior livello di sovrapposizione fra i valori delle features (valore soglia pari a 1) per ottenere dati osservabili. Il numero di cluster identificati in tutti e tre i test per ogni anno è stato pari a due.

Nonostante non siano emerse caratteristiche di impatto significativo sulla distinzione dei due cluster, sono state analizzate quelle features con una sovrapposizione maggiormente ridotta.

Parametro di sovrapposizione accettata utilizzato

soglia	1

Analisi studenti del primo anno
Per il primo anno le principali feature rappresentative, comuni a tutti e tre gli esperimenti, sono il numero di esami risostenuti, lo sviluppo accademico e intellettuale e l'impegno verso l'istituzione e gli obiettivi.
Nel secondo e terzo test è emersa un'altra caratteristica distintiva: l'attenzione dei docenti verso gli studenti.
Tali distinzioni indicano che un primo cluster identifica studenti maggiormente soddisfatti dal punto di vista del proprio percorso accademico e sviluppo intellettuale, con maggiore interazioni positive con i docenti e un minor numero di esami risostenuti.

Analisi studenti del secondo anno
Nel secondo anno, la principale caratteristica di distinzione comune a tutti e tre gli esperimenti è l'impegno verso l'istituzione e gli obiettivi. Nel primo test, è emersa anche come caratteristica distintiva l'attenzione dei docenti.
Tuttavia, i cluster sono estremamente simili fra loro, è difficile trarre conclusioni in merito a determinate correlazioni fra caratteristiche degli studenti.

Analisi studenti del terzo anno
Nel terzo anno, la feature che risulta essere maggiormente significativa è lo sviluppo accademico e intellettuale. Nel primo test, emergono come indici distintivi anche la motivazione e l'impegno verso l'istituzione e gli obiettivi.
Analizzando i risultati si osserva come studenti maggiormente motivati siano anche maggiormente soddisfatti del proprio sviluppo intellettuale e del proprio percorso accademico in generale. Tuttavia, non si riesce a individuare una correlazione di tali caratteristiche con particolari indicatori dell’andamento accademico degli studenti.

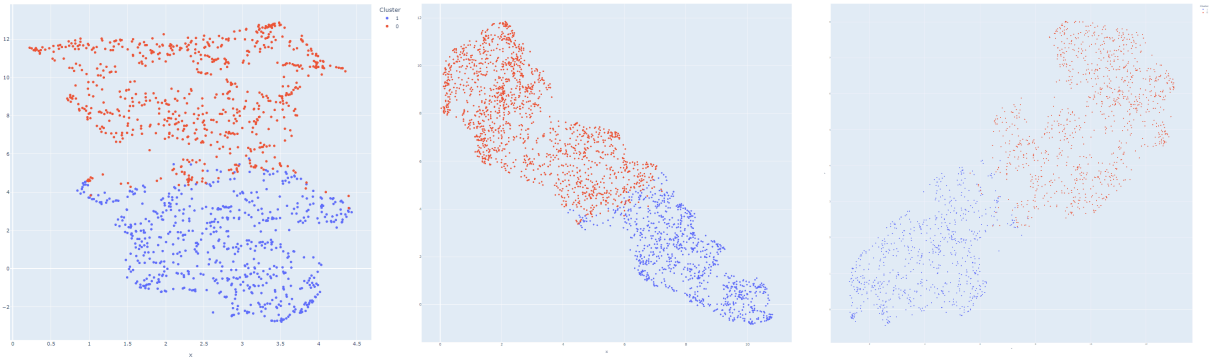
I risultati ottenuti sono da analizzare considerando che è stato utilizzato come valore soglia per l’identificazione delle features rappresentative un numero molto elevato rispetto al dominio di valori delle features stesse.
L’uso di un valore così alto suggerisce che le caratteristiche dei cluster non sono completamente distinte, vi sono sovrapposizioni abbastanza importanti.
In conclusione, è stato riscontrato che l'impegno verso l'istituzione e gli obiettivi, sembrano essere più stabili e rilevanti nei diversi anni.

Tabella risultati UMAP Clustering

studenti	features maggiormente rilevanti	features rilevanti (meno)
1 anno	-	IIS_SVACC, IIS_IMP1, RISOSTENUTI
2 anno	-	IIS_IMP1
3 anno	-	IIS_SVACC

Grafico UMAP Clustering (risultati test 3)

studenti 1 anno studenti 2 anno studenti 3 anno



3.2.5 Risultati SOM Clustering

Durante l'analisi condotta su studenti del primo, secondo e terzo anno, sono stati eseguiti tre test nei quali progressivamente sono state escluse alcune feature del dataset ritenute meno significative, come nel caso precedente.
Con tale algoritmo è stato possibile ritenere come significativo un valore soglia più basso (pari a 0.5), per quanto riguarda l'analisi delle sovrapposizioni fra caratteristiche nei cluster. Nel terzo esperimento, per tutti e tre gli anni, tale valore è stato impostato però a 1 in modo tale da ottenere dei dati osservabili, evidenziando maggiori sovrapposizioni.

Parametro di sovrapposizione accettata utilizzato

soglia	0.5 - nel test 1 e nel test 2
soglia	1 - nel test 3

Analisi studenti del primo anno
Nel primo test sono state identificate le seguenti feature rappresentative: il numero di esami non superati, la capacità di saper lavorare in team, la motivazione identificativa ed estrinseca.
Tuttavia, escludendo l'età, il corso di laurea e il genere, ritroviamo rispetto al test precedente solo l'indice equivalente agli esami non superati; troviamo invece:

l’autogestione, la regolazione emotiva, la soddisfazione, le interazioni con docenti, l’attenzione dei docenti, lo sviluppo accademico e intellettuale, l'impegno verso l'istituzione e gli obiettivi, l’autoefficacia accademica, la procrastinazione e le strategie di miglioramento.
Dunque risultano significative buona parte delle features definite per la rappresentazione della condizione psicologica.
Escludendo anche l’età, troviamo come identificative solo: motivazione identificata, l’attenzione dei docenti, lo sviluppo accademico e intellettuale, l’impegno verso l’istituzione e gli obiettivi, l’insoddisfazione, la regolazione emotiva e la capacità di lavorare in team.
Nonostante tale algoritmo sia stato in grado di individuare cluster con caratteristiche significative, emerge una strana correlazione fra motivazione e soddisfazione: studenti che hanno espresso maggiore motivazione risultano avere un livello di soddisfazione leggermente inferiore.
Emerge un'importante relazione fra le caratteristiche degli studenti: gli studenti che hanno espresso di avere migliori relazioni sociali con altri studenti o docenti, risultano essere maggiormente soddisfatti dal punto di vista del percorso accademico.

Analisi studenti del secondo anno
Nel primo test, sono state identificate come feature rappresentative: la strategia di miglioramento, il rapporto docente - studente, l’impegno verso l’istituzione e gli obiettivi, la motivazione, l’interazione con altri studenti, l’attenzione dei docenti, la regolazione emotiva, lo sviluppo accademico e intellettuale, l’autoefficacia accademica e la ricerca di carriera.
Ovvero, troviamo la maggior parte delle features definite per la rappresentazione della condizione psicologica.
Andando ad effettuare l’analisi con l’esclusione delle features, non abbiamo più caratteristiche rappresentative ma solo sovrapposizioni. Nel secondo test, le feature con minore sovrapposizione sono quelle relative all’impegno verso l’istituzione e gli obiettivi, che rivediamo anche nel terzo test insieme all’attenzione dei docenti.

Analisi studenti del terzo anno
Nel primo test, sono state identificate le seguenti feature rappresentative: insoddisfazione, interazione con altri studenti, procrastinazione, l'impegno verso l'istituzione e gli obiettivi, capacità di saper lavorare in team e media dei voti.
Da tali risultati emerge che gli studenti con maggiore impegno verso gli obiettivi, migliori capacità di lavorare in team e migliori relazioni con altri studenti, hanno una media di voti superiore.
Tuttavia, escludendo l’età, il corso di laurea, il genere e anche il numero di esami superati, si nota come l’unica feature che evidenzia una piccola sovrapposizione sia la regolazione emotiva, le altre caratteristiche non sono distinguibili fra i due cluster.

Possiamo notare che le features identificate come rappresentative variano tra i diversi test e anni, indicando che le caratteristiche che influenzano la condizione

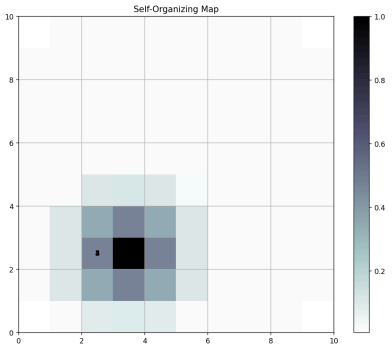
psicologica degli studenti possono cambiare nel tempo e tra i diversi gruppi di studenti. Nonostante le variazioni nelle features identificate come rappresentative, ci sono alcune piccole sovrapposizioni tra i test. Questo suggerisce che alcune caratteristiche possono essere importanti in più contesti o anni, anche se non sono sempre dominanti. Inoltre, possiamo constatare come la regolazione emotiva e l'impegno verso gli obiettivi siano delle costanti in tutti e tre gli anni, sottolineando così il loro peso nella distinzione tra gruppi di studenti.

Tabella risultati SOM Clustering

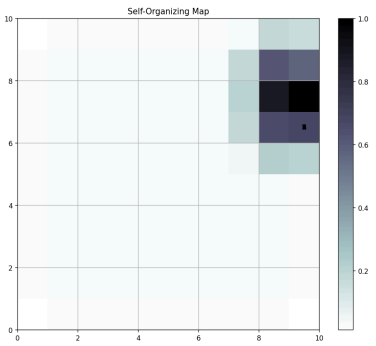
studenti	features maggiormente rilevanti	features rilevanti (meno)
1 anno	molte caratteristiche relative alla condizione psicologica	
2 anno	molte caratteristiche relative alla condizione psicologica	
3 anno	molte caratteristiche relative alla condizione psicologica	

Grafico SOM Clustering (risultati test 3)

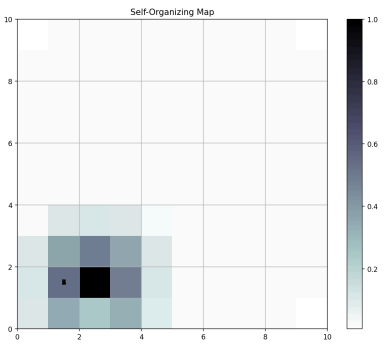
studenti 1 anno



studenti 2 anno



studenti 3 anno



3.2.6 Risultati GMM Clustering

Durante l'analisi condotta su studenti del primo, secondo e terzo anno, sono stati eseguiti tre test nei quali progressivamente sono state escluse alcune feature del database ritenute meno significative.
Nel primo test sono state considerate tutte le features, mentre nel secondo caso sono state escluse l'età, il genere e il corso di laurea. Nella terza analisi, oltre alle suddette features, è stato escluso anche il numero di esami superati.
Il valore soglia per quest'analisi è stato impostato a 0.1 per tutti i test.

Parametro di sovrapposizione accettata utilizzato

soglia	0.1

Analisi studenti del primo anno
Nel primo test, nel quale sono state considerate tutte le features, si è riscontrato che la quasi totalità delle caratteristiche, sia quelle inerenti al percorso accademico sia quelle relative alle condizioni psicologiche degli studenti, hanno contribuito in modo significativo alla definizione dei cluster. Nel secondo e terzo test, dove sono state escluse le features generali, la situazione risulta per lo più invariata.

Analisi studenti del secondo anno
Nel primo test si è riscontrato che la quasi totalità delle caratteristiche, sia a livello generale che inerenti al percorso scolastico e circa la metà delle condizioni psicologiche degli studenti, hanno contribuito in modo significativo alla definizione dei cluster. Nel secondo e terzo test, la situazione risulta per lo più invariata tranne per le condizioni psicologiche, dove si riscontra un incremento delle features che hanno contribuito significativamente alla suddivisione dei cluster.

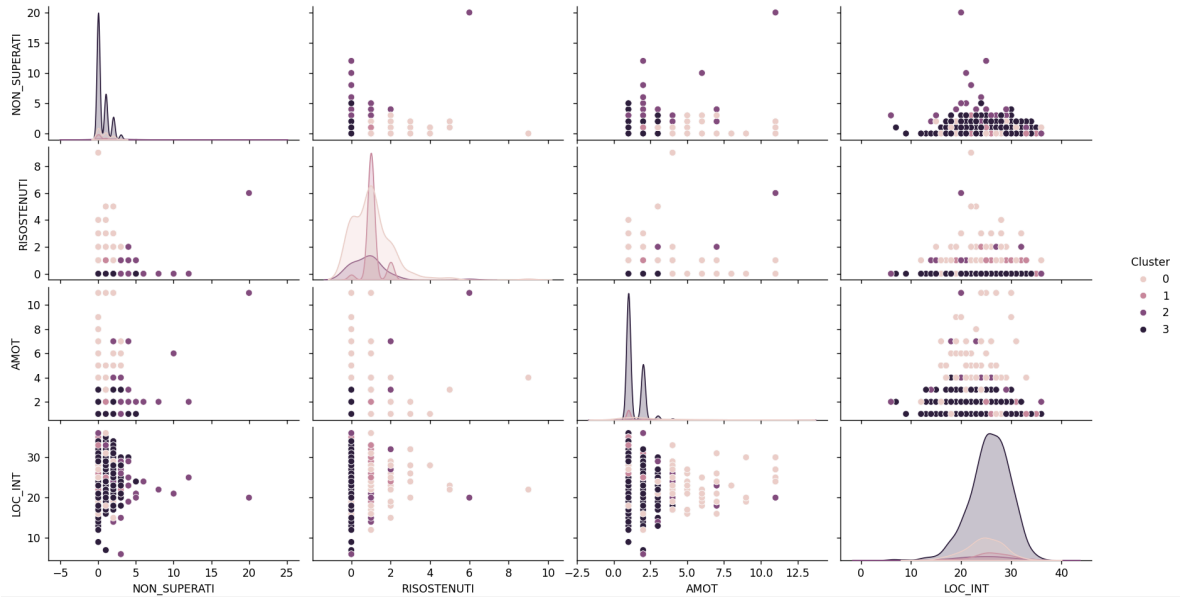
Analisi studenti del terzo anno
Nel primo test la maggior parte delle caratteristiche degli studenti hanno contribuito in modo significativo alla definizione dei cluster. Nel secondo e terzo test, i risultati risultano essere molto simili al caso precedente.

Dai risultati emerge che, indipendentemente dall'anno di studio, sia le caratteristiche relative al percorso accademico sia quelle relative alle condizioni psicologiche degli studenti, hanno giocato un ruolo importante nella definizione dei cluster.
Tuttavia, i cluster individuati risultano essere molto simili fra loro. Tali risultati risultano quindi essere poco rilevanti rispetto ai nostri obiettivi di analisi.
Inoltre, i risultati ottenuti sono legati al fatto che l'algoritmo è stato definito in modo da tentare di suddividere i dati in quattro cluster distinti. Questa scelta è stata fatta con l'obiettivo di fornire al modello una flessibilità sufficiente nel rilevare eventuali pattern o sottogruppi presenti nei dati. Tuttavia, i risultati mostrano che l'individuazione di un maggior numero di cluster non porta a conclusioni migliori. Il valore medio di ciascuna feature risulta essere molto simile tra i cluster individuati, quindi nonostante la bassa varianza non è possibile considerare tali caratteristiche come rappresentative. I cluster individuati risultano essere estremamente simili fra loro.

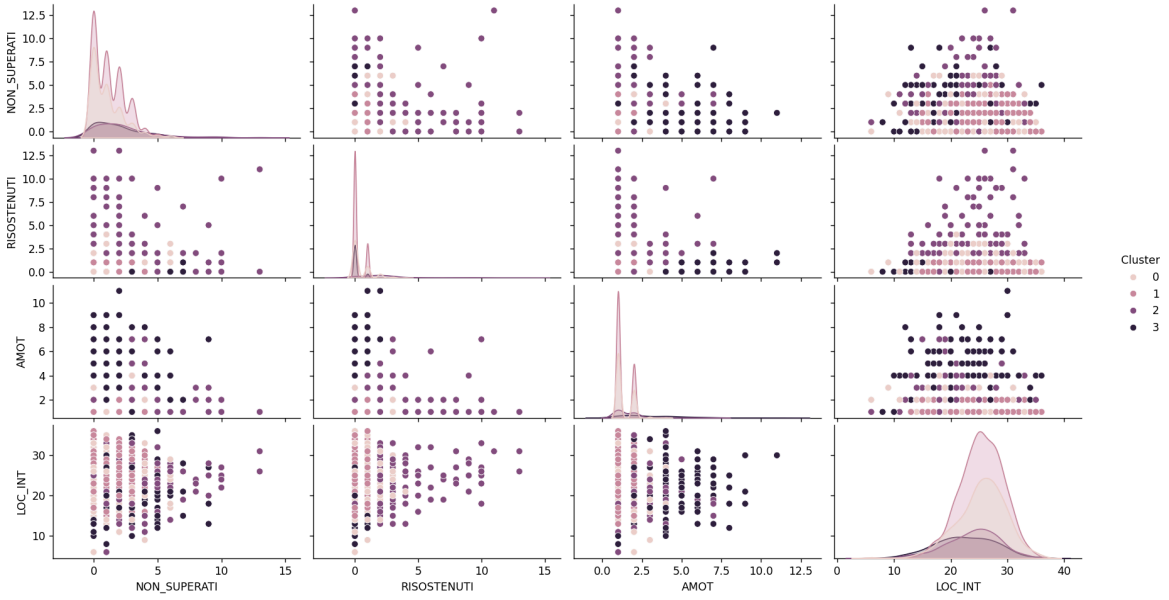
Tabella risultati GMM Clustering

studenti	features maggiormente rilevanti	features rilevanti (meno)
1 anno	molte caratteristiche relative alla condizione psicologica e all'andamento accademico	
2 anno	molte caratteristiche relative alla condizione psicologica e all'andamento accademico	
3 anno	molte caratteristiche relative alla condizione psicologica e all'andamento accademico	

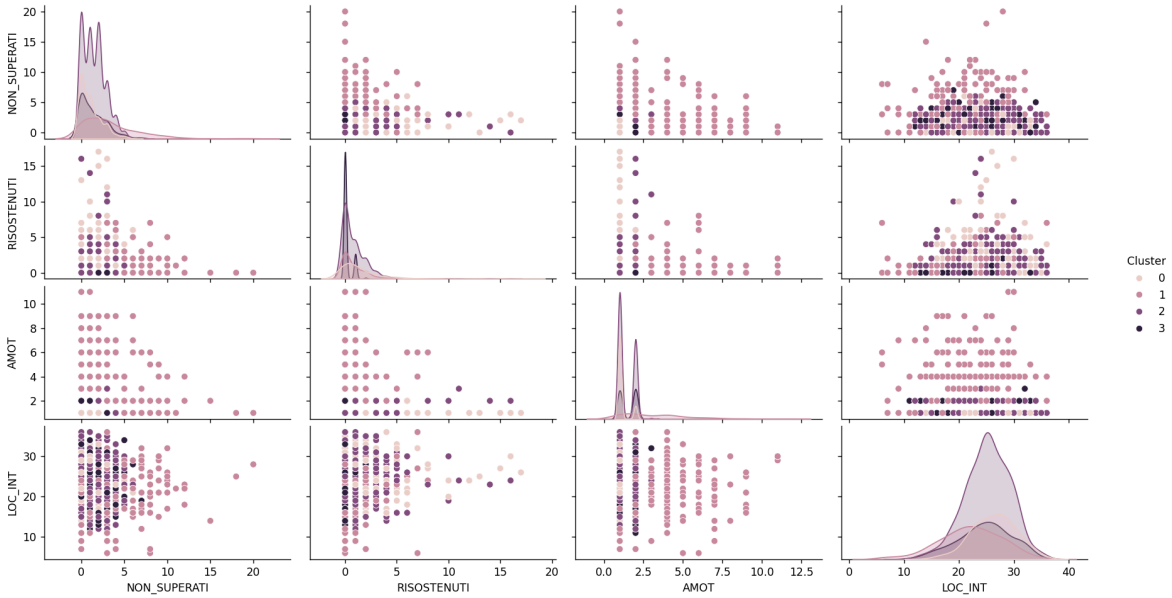
Grafico GMM Clustering (risultati test 3):
analisi correlazione fra alcune features
studenti 1 anno



studenti 2 anno



studenti 3 anno



4 Discussione e Conclusioni

4.1 Discussione dei Risultati

Indipendentemente dal metodo di clustering utilizzato sui dataset di studenti, i risultati ottenuti non hanno mostrato una distinzione significativa tra i cluster identificati.

I cluster sono risultati essere molto simili fra loro, suggerendo una sovrapposizione tra le caratteristiche degli studenti inclusi in ciascuno.
Questo fenomeno potrebbe essere attribuito a diversi fattori. In primo luogo, potrebbe essere dovuto alla natura delle features dei dati. La maggior parte di esse ha un dominio di valori ristretto e presenta una variazione limitata tra le osservazioni. Le differenze tra le istanze nel dataset potrebbero essere troppo sottili per essere rilevate in modo significativo, portando alla formazione di cluster poco distinguibili.
Anche la scelta dei parametri degli algoritmi di clustering e la selezione delle features possono influenzare i risultati ottenuti, per questo sono stati svolti differenti test per massimizzare la capacità di identificare pattern e correlazioni nei dati.

In generale, è stata notata una correlazione interessante fra soddisfazione e motivazione degli studenti. L'esperienza accademica ha un notevole impatto sullo sviluppo intellettuale degli studenti, influenzando la soddisfazione e l'impegno nel raggiungimento degli obiettivi. Tali caratteristiche sono fondamentali per una valutazione della condizione psicologica dei soggetti e per analizzare il loro equilibrio emotivo e mentale.
Analizzando i cluster individuati, è stato possibile osservare che in corrispondenza di maggiore soddisfazione e motivazione vi è un migliore andamento accademico, ovvero lo studente ha in media voti più alti e ha conseguito un maggior numero di esami.
Tuttavia, all'interno di ciascun cluster i valori relativi al numero di esami sostenuti e alla media dei voti mostrano una notevole variabilità. Di conseguenza, non è stato possibile individuare classi di studenti estremamente differenti con correlazioni particolarmente significative tra le caratteristiche analizzate.

L'analisi condotta tramite gli algoritmi SOM (Self-Organizing Map) e GMM (Guassian Mixiture Model) ha prodotto risultati contrastanti rispetto agli altri, portando all'individuazione di un elevato numero di features rappresentative. Tuttavia, confrontando attentamente i cluster individuati, si può concludere che in realtà questi ultimi siano molto simili tra loro. Le features vengono classificate come rilevanti in quanto la varianza è relativamente bassa, tuttavia avendo valore medio molto simile fra i cluster non possiamo ritenerle significative nella clusterizzazione.
Una possibile spiegazione per questi risultati contrastanti potrebbe essere legata al tipo di approccio degli algoritmi. Infatti, mentre Kmeans e DBSCAN tendono a formare cluster basandosi sulla distanza tra i punti nello spazio, SOM e GMM

svolgono una fase di autoapprendimento per identificare pattern nei dati. Questo processo può portare a risultati diversi e a interpretazioni contrastanti delle stesse informazioni. In particolare, l'autoapprendimento potrebbe favorire la scoperta di correlazioni più complesse e non lineari, spiegando così le discrepanze osservate.
Riteniamo che i risultati ottenuti da tali algoritmi siano poco significativi al fine della nostra ricerca.

In conclusione, mentre le nostre ipotesi iniziali suggerivano una forte correlazione tra lo stato di salute mentale degli studenti e il loro andamento accademico, i dati analizzati non permettono di confermare questa relazione in modo definitivo.
Questa discrepanza tra le aspettative e i risultati effettivamente osservati sottolinea l'importanza di ulteriori ricerche per comprendere appieno la complessa interazione tra benessere psicologico degli studenti e il loro rendimento accademico.

4.2 Validità del Metodo

La scelta di applicare la tecnica di clustering ai dati raccolti ha fornito risultati che non rispecchiano totalmente le aspettative iniziali.
Inizialmente ci si aspettava di riuscire ad individuare in modo chiaro gruppi di studenti con caratteristiche simili, mentre l'analisi ha rivelato una notevole sovrapposizione fra i cluster identificati.
Inoltre, è emerso chiaramente quanto i risultati ottenuti siano sensibili alla scelta dei parametri definiti all'interno degli algoritmi. Come evidenziato dall'uso del Gaussian Mixture Model, una non corretta ottimizzazione di tali parametri può portare a risultati poco significativi e persino contraddittori. Questo sottolinea l'importanza di un'attenta selezione e calibrazione dei parametri per garantire la validità e l'affidabilità delle analisi di clustering.
Nonostante questa limitazione, l'analisi dei risultati ci ha permesso di trarre conclusioni preliminari riguardo l'impatto della condizione psicologica degli studenti sul loro rendimento accademico.
Queste considerazioni costituiscono una base per future analisi, che potranno approfondire e confrontare i risultati ottenuti con metodi di analisi e dati differenti.

4.3 Limitazioni e Maturità

Il presente studio presenta alcuni limiti di applicabilità e potenziali bias che possono influenzare l'interpretazione dei risultati.
Uno dei principali limiti riguarda la natura dei dati utilizzati e la loro rappresentatività degli studenti universitari nel complesso. Poiché i dati sono stati raccolti da una singola fonte, potrebbe esserci una mancanza di generalizzabilità dei risultati. Inoltre, è da considerare il fatto che i dati utilizzati derivano da questionari compilati dagli stessi soggetti coinvolti. Ciò potrebbe comportare il rischio di risposte distorte o non completamente sincere, influenzando la precisione e l'attendibilità dei dati raccolti.

Dal punto di vista della maturità tecnologica della soluzione, va considerato che l'applicazione della tecnica di clustering potrebbe essere soggetta a limitazioni legate alla disponibilità di dati di alta qualità e alla complessità delle procedure analitiche. Inoltre, le scelte metodologiche e i parametri selezionati per l'analisi possono influenzare i risultati ottenuti. È importante valutare attentamente queste considerazioni per garantire l'affidabilità e la validità delle conclusioni del presente studio.

4.4 Lavori Futuri

Per estendere ed aumentare la validità della nostra ricerca, vengono proposte di seguito alcune linee guida per avanzare nel progetto.

Suggeriamo di estendere il campione di studenti coinvolti nello studio, al fine di garantire una rappresentatività più ampia e diversificata nelle esperienze accademiche.
Sarebbe utile poter esplorare ulteriori fattori, come il supporto sociale e familiare, la gestione dello stress, il carico di lavoro percepito. Inoltre, potrebbe risultare utile considerare anche l'inclusione di caratteristiche relative a disturbi specifici dell'apprendimento o disturbi fisici e motori. Ciò permetterebbe di esaminare come tali condizioni influenzano le dinamiche accademiche degli studenti.
Ad esempio, i disturbi specifici dell'apprendimento, come la dislessia o la discalculia, possono avere un impatto significativo sulle capacità di apprendimento e sulle strategie di studio degli studenti. Esaminare come queste condizioni interagiscono con le altre variabili psicologiche considerate nello studio potrebbe fornire una prospettiva più completa sull'esperienza accademica degli studenti con tali disturbi.
Allo stesso modo, includere disturbi fisici e motori potrebbe rivelare come le sfide fisiche influenzano il benessere psicologico e le prestazioni accademiche degli studenti.
Si potrebbe inoltre pensare di effettuare una revisione delle domande poste agli utenti nei questionari, cercando di renderle più specifiche o generiche a seconda delle informazioni desiderate.
Inoltre, riteniamo importante integrare le analisi quantitative con metodologie qualitative come interviste e osservazioni sul campo, per poi pensare di confrontare i risultati del clustering con altre fonti di dati e misure psicologiche, consentendo una valutazione più completa e accurata delle correlazioni osservate.
Infine, l'analisi dei dati potrebbe essere svolta con tecniche differenti, al fine di approfondire la comprensione dei risultati.
Sulla base delle considerazioni, potrebbero essere testati determinati interventi mirati al miglioramento del benessere degli studenti e conseguentemente valutati in termini di efficacia. L'analisi nel tempo dei soggetti coinvolti potrebbe essere utile per identificare gli effetti delle eventuali azioni intraprese e il relativo impatto sul successo accademico degli studenti.