Politecnico di Torino

DIPARTIMENTO DI SCIENZE MATEMATICHE APPLICATE

CORSO DI LAUREA IN INGEGNERIA MATEMATICA

TESI DI LAUREA MAGISTRALE

EDTECH: elaborazione di percorsi formativi disciplinari attraverso l'utilizzo del Machine Learning e di Tecniche Statistiche

Relatore:
Prof. Alfredo Benso

Laureando:
Federico Grasso

Anno Accademico 2024/2025

Indice

1 Introduzione

1.1 Contesto

L'avvento delle tecnologie educative *(EDTECH)* ha rivoluzionato i metodi tradizionali di insegnamento e apprendimento, ponendo le basi per un'istruzione sempre più personalizzata e interattiva. La presente tesi si colloca in questo contesto di innovazione, proponendo un approccio metodologico che integra strumenti avanzati di intelligenza artificiale e tecniche di analisi quantitativa per migliorare i percorsi di apprendimento.

La ricerca si concentra sull'elaborazione dei dati provenienti da **Alatin**[1], una piattaforma per l'apprendimento del latino che offre risorse per studenti di diversi livelli. Alatin fornisce informazioni preziose, come esercizi svolti, errori ricorrenti, tempi di svolgimento e preferenze di navigazione. Questi dati, opportunamente analizzati, costituiscono una base ideale per sviluppare modelli di apprendimento automatico e applicare tecniche statistiche avanzate.

Verranno introdotte, specialmente nel primo capitolo, tutte le conoscenze di base necessarie per comprendere appieno ogni aspetto di questo lavoro. Tra di esse figurano nozioni quali il **Principio di Pareto**, per l'identificazione degli argomenti consigliati ad ogni studente, **test di ipotesi** utili per la verifica del lavoro svolto o nel contesto del clustering, particolare attenzione sarà data all'algoritmo **K-Means**, che permette di raggruppare gli studenti in cluster omogenei.

Nel secondo capitolo verrà presentato il funzionamento dell'algoritmo sviluppato attraverso il linguaggio Python per l'elaborazione di questo progetto attraverso diversi pseudocodici volti a fornire una comprensione chiara e accessibile del flusso logico dell'algoritmo e delle sue interazioni con i dati.

Infine, il terzo capitolo sarà dedicato all'analisi e alla validazione empirica dei risultati. Verranno condotti diversi test utilizzando dati aggiornati e più recenti, al fine di valutare l'efficacia delle metodologie proposte evidenziando i vantaggi e le eventuali criticità del sistema.

1.2 Obiettivi

Questo lavoro si propone di sviluppare e analizzare modelli in grado di supportare gli studenti nel loro percorso di apprendimento, fornendo suggerimenti personalizzati per migliorare la loro preparazione. A tal fine, vengono implementati due algoritmi capaci di generare raccomandazioni mirate, con l'obiettivo di ottimizzare l'efficacia dello studio e adattarlo alle esigenze specifiche di ciascun individuo.

Aspetto fondamentale della ricerca è valutare in che misura l'adesione degli studenti ai suggerimenti generati dai due algoritmi incida sulle loro performance accademiche. Per questo motivo, viene condotta un'analisi quantitativa basata su tecniche statistiche, con l'obiettivo di misurare l'effettivo impatto delle raccomandazioni sul miglioramento del rendimento e sull'acquisizione delle conoscenze. Inoltre, la ricerca intende confrontare i due modelli per individuare quello più efficace nel fornire suggerimenti coerenti con il livello di preparazione di ogni studente. L'analisi dei risultati, supportata dalla rappresentazione grafica dei dati, consentirà di identificare eventuali pattern di apprendimento e comprendere meglio le dinamiche che regolano la personalizzazione del percorso formativo.

Per concludere, questo lavoro mira a fornire una base metodologica per futuri sviluppi nell'ambito dell'apprendimento adattivo. L'integrazione di nuove strategie e parametri avanzati potrebbe migliorare ulteriormente l'accuratezza dei suggerimenti, contribuendo alla creazione di modelli di insegnamento sempre più efficaci e personalizzati.

2 Conoscenze Preliminari

2.1 Alatin

Alatin è un software per la somministrazione adattiva e la valutazione degli esercizi di latino. Il suo funzionamento è molto semplice: una volta che un'unità/argomento è assegnato allo studente, per completarlo lo studente deve concludere tutti i suoi livelli di difficoltà chiamati nella piattaforma "lezioni pratiche". Per concludere una lezione pratica, lo studente deve rispondere correttamente al primo tentativo a 12 domande. Se la risposta fornita è corretta, il software fa procedere, altrimenti:

1. avvisa che la risposta fornita è sbagliata;
2. elenca tutte le risposte corrette a quella domanda;
3. offre anche l'opportunità di controllare la teoria;
4. aggiunge una nuova domanda, diversa dalle precedenti, e la somministra aggiungendola in coda alla batteria iniziale di 12 domande;
5. quando ha ottenuto le 12 risposte corrette al primo tentativo aggiunge in coda tutte le domande che hanno generato in prima battuta una risposta errata

Si può ragionevolmente affermare che Alatin sia efficace, poiché gli studenti e le studentesse, spinti dal desiderio di completare più rapidamente i compiti, sono incentivati a impegnarsi attivamente e, grazie a questo impegno, apprendono in modo più profondo. Il docente non riceve un report dettagliato sulle risposte a ogni singola domanda, poiché ciascuno studente svolge un compito personalizzato sia in termini di quantità che di tipologia delle domande. Tuttavia, il sistema restituisce alcune informazioni chiave:

- **Il completamento dell'argomento**: indipendentemente dal numero di errori commessi, il software considera due studenti che hanno terminato un'unità come aventi risposto correttamente al primo tentativo allo stesso numero di domande.

- **Gli esercizi più frequentemente sbagliati**: queste informazioni sono utili per supportare il docente nell'organizzare una correzione formativa e mirata, volta a migliorare il processo di apprendimento.

Inoltre, il sistema fornisce una percentuale di errore individuale per ogni studente, che assolve una funzione specifica:

- **Individuare se un elevato tasso di errore è condiviso dall'intera classe**: questo indica la necessità di riprendere l'argomento perché non è stato assimilato correttamente.
- **Evidenziare discrepanze nel rendimento di uno studente rispetto al proprio andamento abituale**: questo permette al docente di intervenire tempestivamente per comprendere e affrontare le difficoltà riscontrate.

Alatin è stato pubblicato per la prima volta nel 2015 e oggi ha un numero stabile di circa 20-22 mila studenti ogni anno, che generano circa un milione di nuove risposte ogni mese.

2.2 Fondamenti di Statistica

2.2.1 *Modello di Pareto o Principio dell' 80/20*

Il **Modello di Pareto**, noto anche come **regola dell'80/20**, è un principio statistico che suggerisce che, in molti contesti, una piccola parte delle cause genera la maggior parte degli effetti. Ad esempio, in economia, il 20% delle persone potrebbe possedere l'80% della ricchezza, o in un'azienda il 20% dei prodotti potrebbe rappresentare l'80% delle vendite. Questo modello è ampiamente utilizzato in diversi campi per identificare le priorità e ottimizzare gli sforzi: concentrarsi su quel 20% che genera il massimo impatto può portare a risultati significativi. Tuttavia, è importante notare che il rapporto 80/20 non è una regola rigida ma una linea guida, e può variare in base al contesto.

2.2.2 *Quantili*

I **quantili** sono valori che dividono un insieme di dati ordinati in intervalli uguali, offrendo un modo per descrivere la distribuzione di un dataset. Ad

esempio, i **quartili** dividono i dati in quattro parti uguali, mentre i **decili** li suddividono in dieci parti e i **percentili** in cento parti. Il concetto è molto utile per capire dove si trovano determinati valori rispetto al resto del dataset. Ad esempio, se uno studente è al 90° percentile in un test, significa che ha ottenuto un punteggio superiore al 90% degli altri partecipanti.

Esistono due tipi principali di quantili:

- **Quantili empirici**: calcolati direttamente dai dati osservati. Ad esempio, prendendo un dataset ordinato, il mediano (o 50° percentile) è semplicemente il valore al centro.
- **Quantili teorici**: derivati da una distribuzione teorica, come la distribuzione normale. In questo caso, i quantili si calcolano in base a una formula matematica che dipende dalla forma della distribuzione.

I quantili empirici sono utili per analisi pratiche di dataset reali, mentre quelli teorici vengono spesso usati per confrontare dati osservati con modelli teorici o per calcolare probabilità in situazioni predeterminate. Questo concetto è essenziale per interpretare le distribuzioni dei dati e confrontare posizioni relative tra diversi dataset o individui.

2.2.3 *Regressione Lineare Semplice*

La **regressione lineare semplice** è un metodo statistico utilizzato per analizzare la relazione tra due variabili:

- Una **variabile indipendente** (X), che è il fattore che supponiamo influenzi l'altra.
- Una **variabile dipendente** (Y), che è il valore che vogliamo prevedere o spiegare.

L'obiettivo della regressione lineare semplice è trovare una retta che meglio rappresenti la relazione tra queste due variabili. Questa retta è descritta dall'equazione:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Dove:

- β_0 è l'intercetta, cioè il valore assunto da Y quando $X = 0$.
- β_1 è il coefficiente di regressione, che rappresenta l'impatto di X su Y, ovvero il coefficiente angolare della retta.
- ϵ è l'errore residuo, che rappresenta la differenza tra i valori effettivi di Y e quelli previsti dal modello.

La regressione lineare semplice può esser usata, ad esempio, per capire se il numero di ore di studio (X) influenza i voti in un esame (Y). Il modello ci direbbe come cambiano i voti al variare delle ore di studio.

2.2.4 *Calcolo dei Minimi Quadrati*

Il **metodo dei minimi quadrati** è una tecnica utilizzata per trovare la "migliore" retta nella regressione lineare, cioè quella che si adatta meglio ai dati. La chiave del metodo è minimizzare la somma dei quadrati degli errori (ϵ), che sono le differenze tra i valori osservati (Y) e quelli previsti dal modello ($\hat{Y}$).

In pratica, per ogni punto dei dati, si calcola la distanza tra il valore osservato di Y e il valore che la retta prevede per lo stesso X. Questa distanza viene elevata al quadrato (per evitare che gli errori positivi e negativi si annullino) e poi sommata per tutti i punti. La retta scelta dal metodo dei minimi quadrati è quella che rende questa somma il più piccola possibile, chiamata **Sum Squared Error(SSE)** o **Residual Sum Squared(RSS)**.

Matematicamente, si minimizza:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Dove:

- Y_i è il valore osservato.
- $\hat{Y}_i$ è il valore previsto dalla retta per lo stesso X_i.
- n è il numero totale di osservazioni.

Questo metodo ci permette di calcolare i valori ottimali di β_0 (intercetta) e β_1 (coefficiente di regressione).

Il metodo garantisce che il modello sia il più accurato possibile rispetto ai dati forniti.

2.2.5 *P Value*

Il **P-value** è una misura statistica utilizzata per quantificare la probabilità che i risultati osservati siano dovuti al caso, supponendo che l'ipotesi nulla sia vera. L'ipotesi nulla rappresenta una situazione di "assenza di differenza" tra gruppi o variabili. Ad esempio, in un test statistico per confrontare due gruppi, l'ipotesi nulla potrebbe affermare che non ci sia differenza significativa tra le loro medie. Un valore p basso (di solito inferiore a una soglia predefinita, come 0,05) indica che i risultati osservati sono improbabili se l'ipotesi nulla fosse vera, suggerendo quindi che esiste un effetto o una differenza significativa.

2.2.6 *Test T*

Il **Test T** è un metodo statistico usato per confrontare la media di un gruppo di dati con un valore noto o per confrontare le medie di due gruppi. Ad esempio, si può utilizzare un Test T per verificare se gli studenti che utilizzano un nuovo metodo di apprendimento ottengono punteggi mediamente migliori rispetto a quelli che seguono un metodo tradizionale. Il test genera un valore T, che viene poi confrontato con il P-value per decidere se la differenza osservata è statisticamente significativa. È particolarmente importante nell'ambito della regressione lineare semplice e multipla, dove viene utilizzato per valutare la significatività dei coefficienti del modello, noti come beta (β). Per ogni coefficiente β, il Test T verifica se il valore stimato è significativamente diverso da zero, ovvero se la variabile indipendente (X) ha un impatto significativo su Y. Questo viene fatto calcolando il valore T, che dipende dalla stima del coefficiente β e dalla sua deviazione standard. Un valore T elevato (e un P-value associato basso) indica che il coefficiente è significativo, e quindi la variabile indipendente contribuisce significativamente al modello.

2.2.7 *Indice* R^2

L'**Indice** R^2, chiamato anche coefficiente di determinazione, è una misura che indica quanto bene un modello statistico spiega la variabilità dei dati. La sua formula è

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

dove y_i sono i valori osservati, $\hat{y}_i$ i valori predetti dal modello e $\bar{y}$ la media dei valori osservati. Ha un valore compreso tra 0 e 1: un R^2 vicino a 1 significa che il modello si adatta molto bene ai dati, mentre un valore vicino a 0 indica che il modello spiega poco la variabilità osservata. Ad esempio, un $R^2 = 0.8$ significa che l'80% della variabilità dei dati è spiegata dal modello.

2.2.8 *Indice R^2 aggiustato*

L'R^2 **aggiustato** è una versione modificata del coefficiente di determinazione che, a differenza dell'R^2 standard, penalizza l'aggiunta di variabili non significative. Mentre l'R^2 aumenta con l'aggiunta di qualsiasi predittore, anche se irrilevante, l'R^2 aggiustato considera il numero di osservazioni n e il numero di predittori p, ed è calcolato con la formula

$$R^2_{\text{adj}} = 1 - \frac{(n-1)}{(n-p-1)} \cdot (1 - R^2)$$

Questo indice rimane compreso tra 0 e 1, e aumenta solo se l'aggiunta di nuove variabili migliora davvero la capacità esplicativa del modello. In tal modo, l'R^2 aggiustato consente di confrontare modelli con diversi numeri di predittori in modo più equo, evitando la sovrastima della bontà di adattamento dovuta a variabili superflue.

2.2.9 *Test ANOVA per modelli annidati*

Il **test ANOVA** per modelli annidati è una procedura statistica usata per confrontare due modelli, uno più complesso (esteso) e uno più semplice (base), per determinare se l'aggiunta di variabili o parametri nel modello completo migliora significativamente la capacità di spiegare i dati.

Due modelli si definiscono **annidati** quando il modello semplice è una versione ridotta del modello complesso. In altre parole, il modello semplice è ottenuto togliendo alcune variabili o restrizioni dal modello complesso. Ad esempio:

- **Modello semplice**:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- **Modello complesso**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Qui, il modello semplice ignora la variabile X_2, mentre il modello complesso la include.

Il test confronta i due modelli per vedere se l'aggiunta di variabili o parametri nel modello complesso porta a un miglioramento significativo nella capacità di predire o spiegare Y. Questo viene fatto analizzando la **varianza residua**, ovvero quanto errore rimane nei due modelli.

1. **Calcolo della devianza**: La devianza è una misura dell'errore residuo, cioè quanto i dati osservati differiscono da quelli previsti dal modello.
   - Il modello semplice ha una devianza maggiore, perché spiega meno.
   - Il modello complesso ha una devianza minore, perché spiega meglio.
2. **Confronto delle devianze**: Si calcola la differenza tra le devianze dei due modelli. Se questa differenza è grande, significa che il modello complesso spiega significativamente di più rispetto al modello semplice.
3. **F-test**: La differenza delle devianze viene trasformata in un valore F, che viene confrontato con una distribuzione F per stabilire se è significativa. Un valore F alto (e un P-value basso) indica che il modello complesso è significativamente migliore.

Questa metodologia è particolarmente utile per garantire la parsimonia del modello, evitando di includere variabili superflue e mantenendo il modello il più semplice possibile senza perdere in capacità esplicativa. Pertanto, il test ANOVA per modelli annidati rappresenta uno strumento fondamentale per la costruzione e la valutazione di modelli statistici robusti ed efficaci.

2.2.10 *QQ Plot*

Il **QQ Plot** (Quantile-Quantile Plot) è uno strumento grafico utilizzato per verificare se un insieme di dati segue una determinata distribuzione teorica, molto spesso la distribuzione normale. Questo grafico è composto da:

- Sull'asse x, i **quantili della distribuzione teorica**.
- Sull'asse y, i **quantili dei dati osservati**.

L'idea principale è confrontare i valori attesi dalla distribuzione teorica con quelli effettivamente osservati nei dati. Se i dati seguono bene la distribuzione teorica, i punti del QQ Plot si allineano lungo una diagonale, indicando una corrispondenza tra i quantili. Deviazioni dalla linea diagonale possono indicare:

- **Coda larga o corta**: se i punti si discostano dalla linea alle estremità, può significare che i dati hanno code più lunghe (o più corte) rispetto alla distribuzione teorica.
- **Asimmetria**: un pattern curvo può indicare che i dati sono asimmetrici rispetto alla distribuzione teorica.

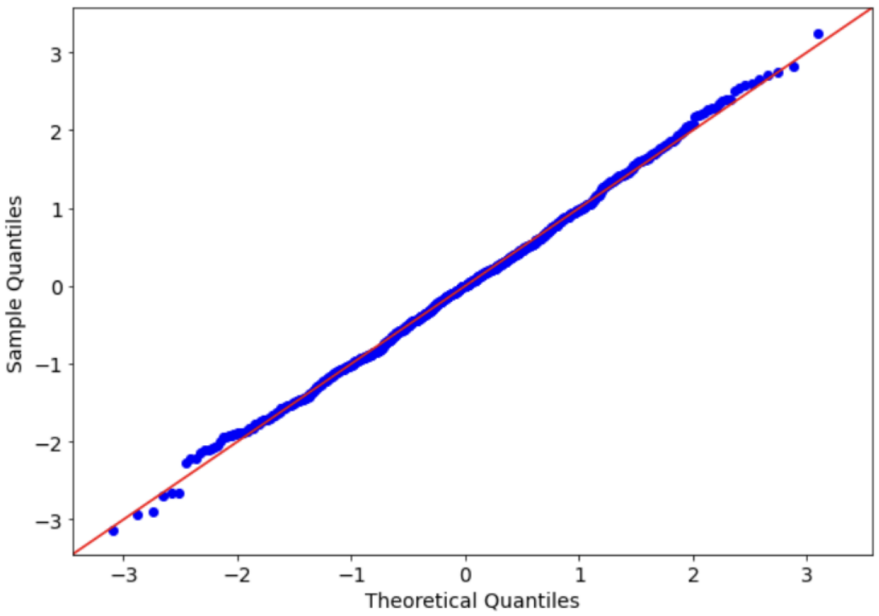


Figura 1: Esempio Di QQ-Plot

Ad esempio, nel contesto della regressione lineare, il QQ Plot viene spesso utilizzato per verificare se i residui seguono una distribuzione normale, un'assunzione chiave per l'affidabilità di molti test statistici. Se i residui non seguono la normalità, potrebbe essere necessario trasformare i dati o scegliere un modello alternativo.

In pratica, il QQ Plot è essenziale per diagnosticare la normalità dei dati e individuare eventuali anomalie, come outlier, che potrebbero influenzare l'analisi statistica.

2.2.11 *Residuals vs Fitted Plot*

Il **Residuals vs Fitted Plot** è un grafico diagnostico usato per valutare la bontà di adattamento di un modello statistico, in particolare nella regressione lineare. Questo grafico rappresenta:

- Sull'asse x, i **valori previsti** (*fitted values*) dal modello.
- Sull'asse y, i **residui** (cioè la differenza tra i valori osservati e quelli previsti).

L'obiettivo principale è verificare alcune assunzioni fondamentali della regressione:

1. **Linearità**: i residui dovrebbero essere distribuiti in modo casuale attorno alla linea dello zero, senza mostrare alcun pattern sistematico. Un pattern, come una curva o una forma a ventaglio, potrebbe indicare che il modello non cattura correttamente la relazione tra variabile indipendente e dipendente.
2. **Omoschedasticità**: la varianza dei residui dovrebbe essere costante lungo l'intera gamma di valori previsti. Se i residui diventano progressivamente più grandi o più piccoli (un fenomeno noto come eteroschedasticità), ciò potrebbe indicare problemi nel modello.
3. **Indipendenza**: i residui non dovrebbero essere correlati tra loro. Una correlazione tra i residui potrebbe suggerire che ci sono variabili non incluse nel modello che influenzano i dati.

Un esempio pratico è quello di un modello di regressione per prevedere i punteggi scolastici in funzione delle ore di studio. Se il Residuals vs Fitted Plot mostra un pattern evidente (ad esempio, una forma a "U"), potrebbe significare che esiste una relazione non lineare tra le due variabili che il modello lineare non riesce a catturare.

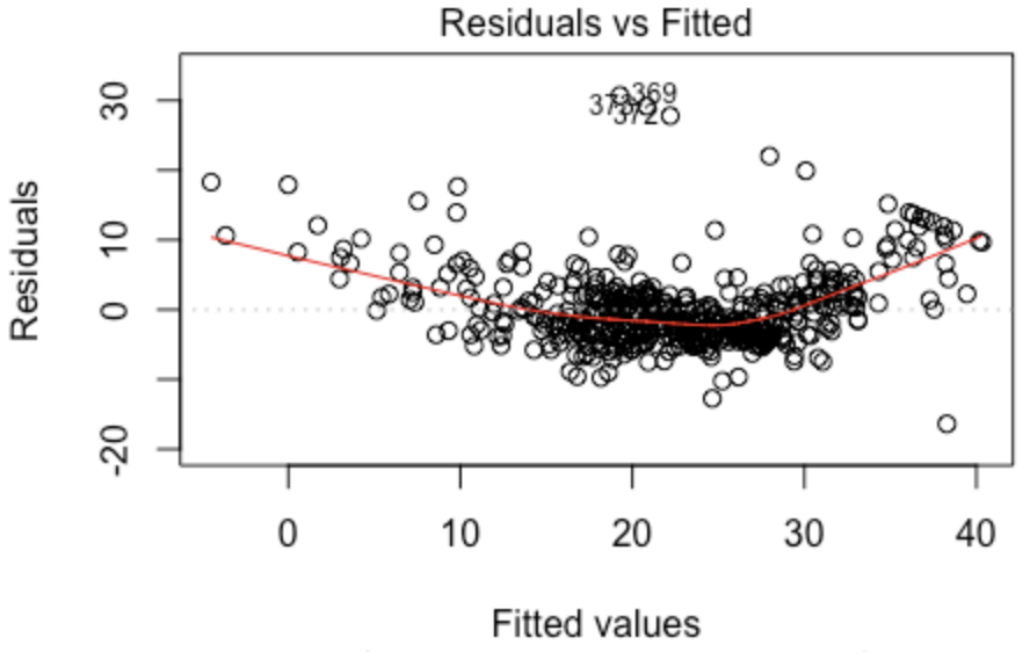


Figura 2: Esempio Di ResidualVsFitted-Plot

In sintesi, questo grafico è cruciale per diagnosticare problemi strutturali nel modello e suggerisce possibili miglioramenti, come aggiungere variabili o trasformare i dati.

2.2.12 *Scale-Location Plot*

Il **Scale-Location Plot**, noto anche come **Spread-Location Plot**, è un grafico diagnostico utilizzato per valutare l'assunzione di omoschedasticità (varianza costante dei residui) in un modello di regressione. Questo grafico rappresenta:

- Sull'asse x, i **valori previsti** (*fitted values*).
- Sull'asse y, la radice quadrata del valore assoluto dei residui standardizzati.

L'obiettivo del grafico è verificare se la dispersione dei residui è costante su tutta la gamma dei valori previsti. Un pattern uniforme dei punti suggerisce che l'assunzione di varianza costante è rispettata, mentre:

- Un aumento o diminuzione sistematica della dispersione (ad esempio, una forma a cono o ventaglio) indica **eteroschedasticità**, ovvero che la varianza dei residui cambia al variare dei valori previsti.

- Una dispersione irregolare potrebbe suggerire la presenza di outlier o errori di modellazione.

Ad esempio, supponiamo di analizzare i salari in funzione degli anni di esperienza. Se i salari mostrano una variabilità maggiore per valori più alti (cioè, differenze più grandi tra i salari elevati rispetto a quelli bassi), il Scale-Location Plot potrebbe mostrare una forma a ventaglio, suggerendo che il modello non tiene conto di questa variabilità.

Il Scale-Location Plot è utile non solo per individuare problemi di eteroschedasticità, ma anche per diagnosticare la necessità di trasformazioni, come il logaritmo dei dati o l'uso di un modello diverso. Questo grafico aiuta a garantire che i residui siano distribuiti in modo appropriato, migliorando la validità delle inferenze statistiche basate sul modello.

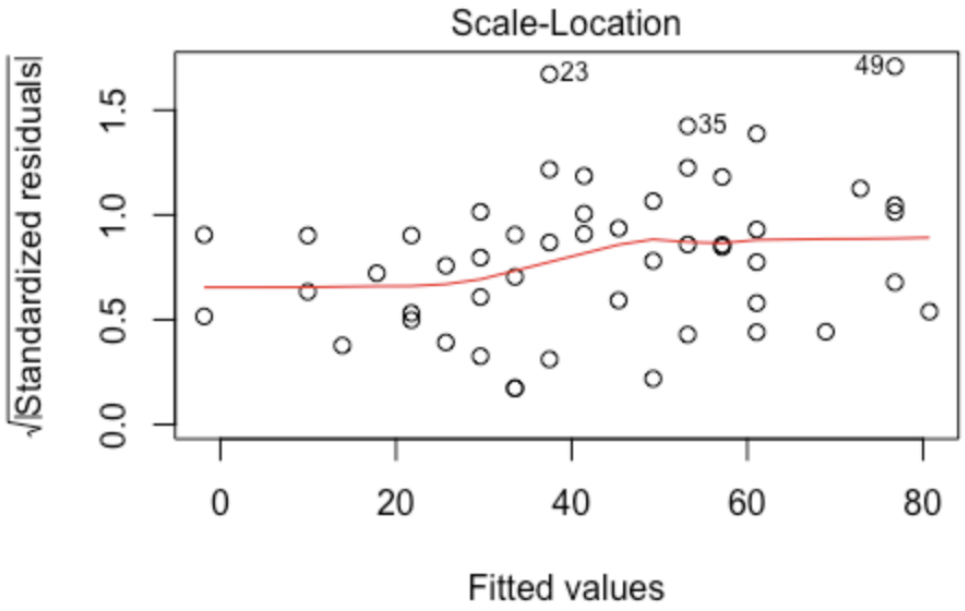


Figura 3: Esempio Di Scale-Location Plot

2.3 Tecniche di Machine Learning

2.3.1 *K-Means Algorithm*

L'**algoritmo K-Means** è un metodo di clustering molto utile per dividere un insieme di dati in gruppi, chiamati **cluster**, in modo che i punti all'interno

dello stesso cluster siano più simili tra loro rispetto ai punti di altri cluster. In pratica, il suo scopo è quello di raggruppare i dati in K cluster, cercando di minimizzare la distanza tra i punti di uno stesso cluster e il loro centro, detto **centroide**.

Il funzionamento è abbastanza intuitivo:

1. **Scelta del numero di cluster (K)**: inizialmente, si decide quanti gruppi (K) si vogliono creare.
2. **Inizializzazione dei centri (*centroidi*)**: si scelgono K punti iniziali, chiamati *centroidi*, che rappresentano il centro di ogni cluster. Questi punti possono essere scelti casualmente o tra i dati stessi.
3. **Assegnazione ai cluster**: per ogni punto del dataset, si calcola la distanza da ciascun centroide (ad esempio, usando la distanza euclidea). Ogni punto viene assegnato al cluster del centroide più vicino.
4. **Aggiornamento dei centroidi**: una volta assegnati i punti ai cluster, si ricalcolano i centroidi spostandoli nella posizione media dei punti assegnati a ciascun cluster.
5. **Iterazione**: i passaggi 3 e 4 vengono ripetuti fino a quando i centroidi non cambiano più significativamente o fino a raggiungere un numero massimo di iterazioni.

Un' illustrazione più implicita della logica da seguire è la seguente:

Questo algoritmo è particolarmente efficace per dataset di grandi dimensioni, caratterizzati da dati ben definiti e chiaramente separati. Tuttavia, può risultare meno performante quando i cluster presentano forme irregolari o dimensioni significativamente diverse tra loro. Di seguito, un'illustrazione di come viene applicato il K-Means ad un set di dati:

2.3.2 *Cosine Similarity*

La **Cosine Similarity** è una misura matematica utilizzata per determinare quanto due vettori siano orientati nella stessa direzione nello spazio. In ambito statistico e di machine learning, è ampiamente impiegata per calcolare la somiglianza tra due oggetti rappresentati come vettori numerici, indipendentemente dalla loro magnitudine.

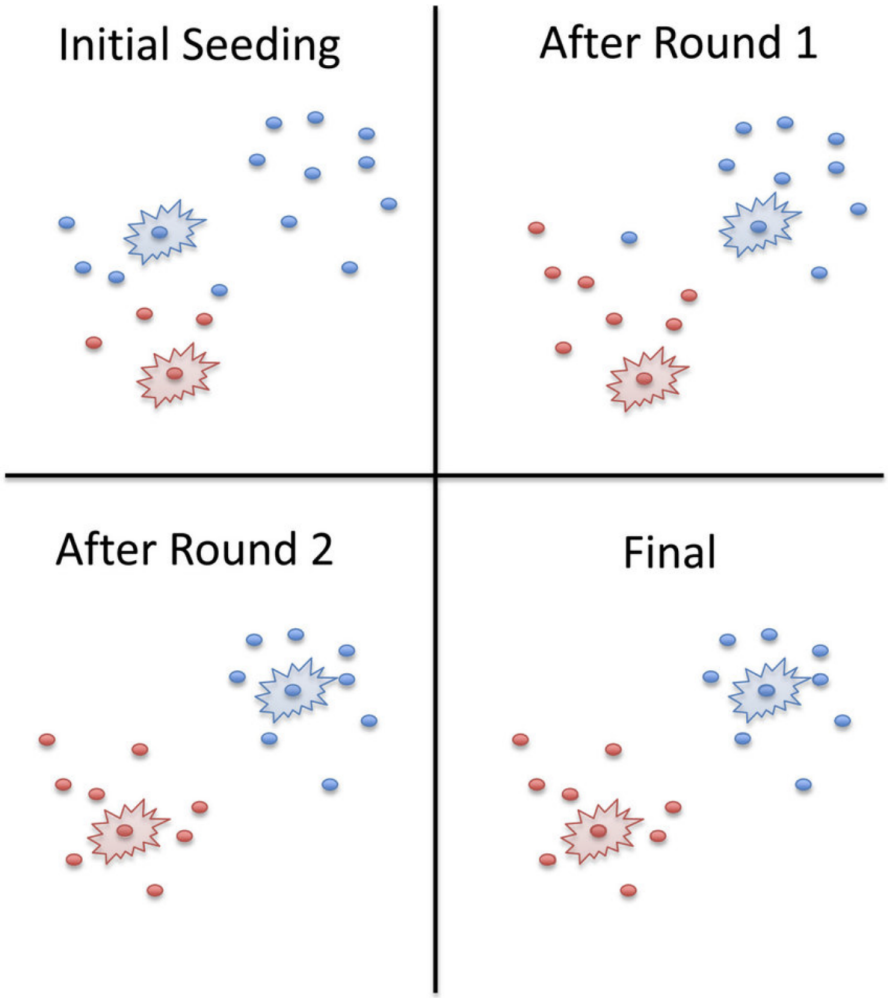


Figura 4: Steps del K-Means

La formula per calcolare la Cosine Similarity tra due vettori $\mathbf{A}$ e $\mathbf{B}$ è la seguente:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Dove:

- $\mathbf{A} \cdot \mathbf{B}$ è il prodotto scalare tra i vettori.
- $\|\mathbf{A}\|$ e $\|\mathbf{B}\|$ sono le norme (o lunghezze) dei vettori.
- A_i e B_i sono le componenti dei vettori.

Il valore risultante varia tra -1 e 1:

- Un valore di 1 indica che i due vettori puntano esattamente nella stessa direzione (massima somiglianza).
- Un valore di 0 indica che i vettori sono ortogonali e non correlati.
- Un valore di -1 indica che i vettori puntano in direzioni opposte.

La Cosine Similarity è particolarmente utile in contesti dove la direzione dei dati è più importante rispetto alla loro magnitudine, ovvero della lunghezza dei loro vettori nello spazio. In termini matematici, la magnitudine di un vettore **A** si calcola come la sua norma euclidea:

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} A_i^2}$$

2.3.3 *Item Based Method*

Il **metodo Item Based Collaborative Filtering**[3] è una tecnica utilizzata per la generazione di raccomandazioni, che si concentra sull'analisi delle relazioni tra oggetti (o "item") piuttosto che tra utenti. Questo approccio lavora sulla matrice utente-oggetto, calcolando la somiglianza tra gli oggetti sulla base delle valutazioni o interazioni degli utenti. L'obiettivo è prevedere l'interesse di un utente verso un nuovo oggetto, sfruttando la similarità con oggetti che l'utente ha già valutato positivamente.

Il funzionamento del metodo si basa su alcune fasi principali:

1. **Calcolo della similarità tra gli oggetti:** Si analizzano solo gli utenti che hanno valutato entrambi gli oggetti, al fine di misurare quanto questi siano simili tra loro.
    - **Cosine Similarity:** Misura l'angolo tra i vettori di valutazioni, fornendo un'indicazione di quanto due oggetti siano orientati nella stessa direzione.
    - **Correlazione di Pearson:** Valuta la relazione lineare tra le valutazioni degli utenti, correggendo eventuali differenze di scala individuali.

2. **Selezione degli oggetti simili:** Una volta calcolata la similarità, si identificano per ciascun oggetto gli k oggetti più simili, che saranno utilizzati per generare raccomandazioni personalizzate.

Le predizioni vengono prodotte combinando le valutazioni di un utente sugli oggetti simili, utilizzando tecniche come la somma pesata, dove le valutazioni sono moltiplicate per la similarità tra gli oggetti, o modelli di regressione, che stimano la valutazione prevista con maggiore precisione. Questo approccio è particolarmente utile per dataset di grandi dimensioni, poiché le somiglianze tra oggetti possono essere precomputate, migliorando l'efficienza del sistema in tempo reale.

Nel contesto educativo, il metodo Item-Item Based può essere applicato per supportare percorsi di apprendimento personalizzati. Ad esempio, dopo che uno studente ha completato con successo un esercizio sull'analisi logica, il sistema potrebbe suggerire esercizi correlati sulla sintassi o sull'analisi del periodo, identificati come simili in base alle prestazioni di altri studenti. In questo modo, il metodo consente di costruire percorsi di studio progressivi e mirati, basandosi su dati reali. Grazie alla sua scalabilità e alla capacità di adattarsi a dataset ampi e sparsi, questa tecnica rappresenta una soluzione efficace per progettare sistemi di raccomandazione che migliorano l'esperienza di apprendimento individuale.

3 Algoritmi di raccomandazioni e verifica dell'implementazione

3.1 Dataset di lavoro

Nell'ambito del presente lavoro di ricerca, i dataset, forniti dalla piattaforma Alatin, rappresentano una risorsa fondamentale per sviluppare un sistema di raccomandazione personalizzato e adattivo, in grado di supportare gli studenti nel loro percorso di apprendimento. I Dataset, denominati cnr_elementi_21.csv, cnr_prerequisiti_21.csv e cnr_risposte_21.csv, cronologicamente riempiti a partire dall'anno 2015 fino ad oggi, contengono dati relativi agli esercizi disponibili, agli argomenti didattici e alle interazioni degli studenti con tali esercizi.

Questi dati sono stati utilizzati per diversi scopi quali analizzare il comportamento degli studenti durante lo svolgimento degli esercizi e identificare le loro aree di difficoltà, oppure elaborare percorsi di apprendimento personalizzati, rispettando i prerequisiti didattici e ottimizzando il livello di difficoltà degli esercizi in base alle competenze dello studente.

L'analisi combinata di questi dataset, integrata con tecniche di Machine Learning e strumenti statistici avanzati, ha permesso di creare una base metodologica solida per il design e lo sviluppo di soluzioni innovative nel campo dell'apprendimento digitale (EdTech), rendendo i processi educativi più efficaci e personalizzati.

Nel dettaglio, mostriamo un esempio delle prime righe di ognuno, specificando per ognuno le colonne di maggior rilevanza per la nostra ricerca. È bene prima di tutto chiarire che non tutte le colonne a disposizione sono state sfruttate per la nostra analisi, pertanto, di ogni blocco dati ci limiteremo a descrivere solo quelle fondamentali per il nostro lavoro:

3.1.1 cnr_risposte_21.csv

Questo dataset raccoglie le interazioni degli studenti con gli esercizi e registra le loro risposte. Ogni riga rappresenta un tentativo di risposta di uno studente a un esercizio:

- `owner_id`: identificativo univoco dello studente.

```
owner_id,esercizio_id,elementi_pks,stato,created
468974,60282,"{23496,23497}",C,2018-06-28 08:39:56.771055+00
559607,36301,"{4254,4255}",C,2018-06-12 14:47:45.744796+00
559607,35950,"{4020,4021}",S,2018-06-12 14:47:51.178605+00
559607,36265,"{4200,4201,4202}",C,2018-06-12 14:47:18.949106+00
559607,35926,"{3985,3986}",C,2018-06-12 14:49:08.404095+00
559607,35926,"{3987,3988}",S,2018-06-12 14:49:33.341238+00
559607,35924,"{3973,3974}",C,2018-06-12 14:49:56.979388+00
559607,35924,"{3975,3980}",C,2018-06-12 14:50:12.653854+00
559607,36314,"{4272,4273}",S,2018-06-12 14:50:44.859927+00
559607,35950,"{4020,4021}",C,2018-06-12 14:51:12.163092+00
565470,59977,"{23042,23043}",C,2018-06-13 08:30:50.575469+00
559607,36274,"{4239,4240,4241}",C,2018-06-12 14:51:07.161294+00
```

Figura 5: Dati presenti in cnr_risposte_21.csv

- `esercizio_id`: ID dell'esercizio al quale lo studente ha risposto.
- `elementi_pks`: coppia di informazioni (o coppia di id consecutivi) relativamente alla risposta corretta da dare e la teoria spiegata dietro ad essa
- `stato`: stato della risposta, con valori come "C" = corretta, "S"= sbagliata e "A" = assente.
- `created`: Anno-Mese-Giorno e Ora in cui l'interazione con la piattaforma è avvenuta per tale esercizio.

Questo dataset è fondamentale per tracciare le attività degli studenti, comprendere le loro performance e identificare eventuali difficoltà o lacune. I dati temporali permettono anche di analizzare la progressione dello studente nel tempo.

3.1.2 cnr_elementi_21.csv

Questo dataset contiene informazioni relative agli esercizi e argomenti disponibili per gli studenti. Ogni riga rappresenta fa riferimento ad un esercizio specifico, contraddistinto dalla propria difficoltà e facente capo ad uno specifico argomento:

- `esercizio_id`: ID specifico associato all'esercizio.

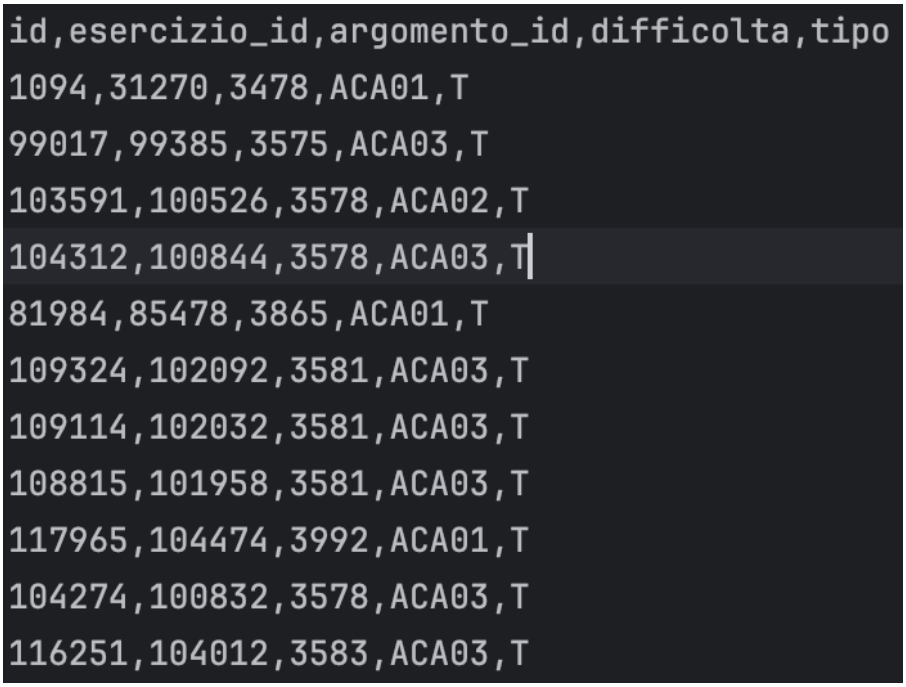


Figura 6: Dati presenti in cnr_elementi_21.csv

- `argomento_id`: identificativo dell'argomento a cui l'esercizio appartiene. Questo campo permette di associare un esercizio a un argomento didattico specifico.
- `difficolta`: indica il livello di difficoltà dell'esercizio, categorizzato tramite codici (es. ACA01, ACA02, ACA03, ACA04).

Questo dataset fornisce la base per comprendere quali esercizi sono disponibili per ciascun argomento e per identificare esercizi più o meno difficili in base al livello dello studente.

3.1.3 cnr_prerequisiti_21.csv

Questo dataset elenca informazioni sugli argomenti didattici, inclusi i loro prerequisiti. Ogni riga rappresenta un argomento e i dati associati:

- `id`: identificativo univoco dell'argomento.
- `nome`: nome descrittivo dell'argomento, che fornisce una comprensione immediata del contenuto trattato (es. "Ablativo assoluto").

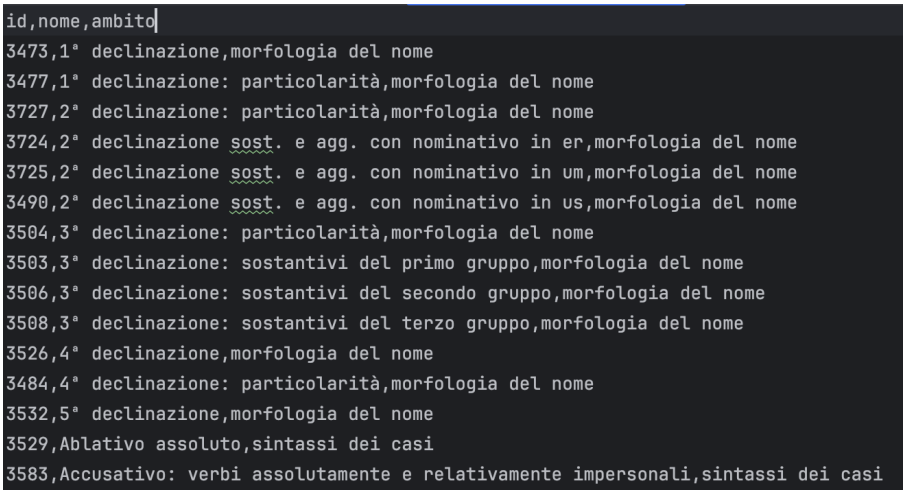


Figura 7: Dati presenti in cnr_prerequisiti_21.csv

- `ambito`: Specifica l'ambito didattico a cui appartiene l'argomento, come "morfologia del nome" o "sintassi dei casi". Questo campo è importante per organizzare gli argomenti in macro-categorie e per costruire percorsi di studio mirati.

Questo dataset è essenziale per comprendere il contesto e la struttura degli argomenti. Può essere utilizzato per definire percorsi di apprendimento logici, dove i prerequisiti sono rispettati e ogni argomento si basa su conoscenze precedenti.

3.2 Creazione delle Matrici di Performance

Il dataset `cnr_risposte_21.csv` rappresenta la fonte principale di informazioni a nostra disposizione, poiché, come descritto nella sezione precedente, fornisce una chiara evidenza dell'interazione tra ciascuno studente e la piattaforma. Data la disponibilità dei dati dal 2015 ad oggi, l'obiettivo è quello di costruire, per ogni anno, una matrice di performance. Questa matrice sarà strutturata in modo tale che:

- Ogni riga corrisponda a uno studente in ingresso per quell'anno.
- Ogni colonna rappresenti un argomento affrontato dallo studente durante l'anno in esame.

- Ogni cella rappresenti la somma delle performance ottenute dallo studente specifico per un determinato argomento.

Relativamente al punteggio associato a ciascun argomento svolto, è necessario definire in prima istanza il tipo di logica implementata nell'assegnazione dei punteggi e sfruttarla successivamente nella creazione delle matrici.

3.2.1 Definizione somma pesata dei punteggi

Come detto in precedenza, l'intero programma di latino, fornito dalla piattaforma Alatin, viene suddiviso in 5 ambiti:

1. 'morfologia del nome';
2. 'sintassi dei casi';
3. 'morfologia del verbo';
4. 'sintassi del periodo';
5. 'competenze miste'.

Ad ogni ambito didattico sono associati diversi argomenti, e ciascun argomento include numerosi esercizi che possono essere classificati in base al loro livello di difficoltà secondo le seguenti categorie:

1. ACA01;
2. ACA02;
3. ACA03;
4. ACA04.

Questi livelli di difficoltà, ordinati in modo crescente, consentono di determinare il punteggio da associare a ciascun esercizio. Utilizzando le stringhe sopra elencate, è possibile definire una funzione di mapping che associa a ogni stringa un punteggio specifico, basandosi sulle ultime due cifre contenute in essa. In questo modo, ogni esercizio viene automaticamente classificato e valorizzato secondo i seguenti punteggi:

1. ACA01 $\rightarrow$ 1.0;
2. ACA02 $\rightarrow$ 2.0;
3. ACA03 $\rightarrow$ 3.0;
4. ACA04 $\rightarrow$ 4.0.

Grazie a questa classificazione, se uno studente si esercita su un determinato argomento completando esercizi di diverse difficoltà, il suo "punteggio finale" associato ad un determinato argomento sarà calcolato come la somma pesata dei punteggi relativi agli esercizi svolti correttamente. Tuttavia, gli esercizi non svolti sono esclusi da questa logica, determinando di conseguenza un azzeramento automatico delle performance ad essi associate. Al contrario, durante la fase di valutazione, un ragionamento analogo, ma applicato agli esercizi errati, viene utilizzato per penalizzare la performance complessiva: questi ultimi, infatti, comportano una riduzione del punteggio nella somma dei pesi relativi a uno o più argomenti. Questo approccio consente di differenziare i punteggi ottenuti dagli studenti, anche per lo stesso argomento proposto da Alatin, riflettendo così la variabilità delle competenze e delle performance individuali.

3.3 Clusterizzazione degli studenti in base alle performance

Per ciascun ambito, utilizzando una matrice di performance dedicata, il codice applica l'algoritmo *K-Means* per classificare gli studenti in tre gruppi distinti:

- **Carenti**: studenti con performance inferiori rispetto alla media complessiva;
- **Intermedi**: studenti con performance mediocri o nella media;
- **Migliori**: studenti con performance ottimali rispetto agli altri.

Nel dettaglio, un possibile pseudo codice che rimarca quanto svolto dall'algoritmo *K-Means* risulta essere il seguente:

Algorithm 1 Pseudocodice per il raggruppamento degli studenti basato su performance

Input:

- Informazioni sugli argomenti suddivisi in ambiti tematici.
- Dati sulle performance degli studenti in diversi anni.
- Numero di gruppi da creare (k).

Output: Raggruppamenti degli studenti per ciascun ambito e per ogni anno.

Passo 1: Organizzazione degli argomenti per ambito
Suddividere gli argomenti in gruppi basati sull'ambito tematico a cui appartengono.

Passo 2: Inizializzazione
Preparare una struttura per memorizzare i risultati dei gruppi di studenti, organizzati per anno e per ambito.

Passo 3: Analisi per ciascun anno
foreach *anno* **do**

Passo 3.1: Aggregazione delle performance per ambito
Per ogni gruppo di argomenti appartenenti a uno stesso ambito:

- Identificare i dati disponibili per gli argomenti appartenenti al gruppo.
- Calcolare un punteggio complessivo per ciascuno studente, considerando le performance sugli argomenti di quel gruppo.

Passo 3.2: Creazione dei gruppi di studenti
Per ciascun gruppo tematico:

- Utilizzare un algoritmo di raggruppamento (*clustering*) per suddividere gli studenti in gruppi in base ai loro punteggi complessivi.
- Se i dati disponibili non sono sufficienti, assegnare a tutti gli studenti un gruppo predefinito.

Passo 3.3: Memorizzazione dei risultati
Salvare i gruppi di studenti formati per ciascun ambito e per l'anno analizzato.

end

Passo 4: Restituzione dei risultati
Restituire i gruppi di studenti creati per ciascun anno e ciascun ambito, fornendo una visione organizzata delle loro performance.

3.4 Metodi di suggerimento degli argomenti per ogni studente

I metodi sviluppati hanno l'obiettivo di individuare per ciascuno studente gli argomenti su cui concentrare l'attenzione per migliorare le proprie competenze, fornendo al contempo una chiara visione dei contenuti necessari per arricchire il proprio bagaglio culturale. Di seguito vengono presentati due esempi di algoritmi progettati specificamente a tale scopo.

3.4.1 Metodo di Pareto

Il metodo di Pareto segue 2 implementazioni parallele al fine di suggerire gli argomenti "migliori" sui quali ogni studente può far affidamento per migliorare le proprie capacità. Di seguito una rappresentazione dettagliata degli step da seguire per tale logica:

▷ **Determinazione degli studenti migliori**
Dopo aver creato, per ciascun ambito, tre cluster che suddividono gli studenti nelle categorie precedentemente indicate (Carenti, Intermedi e Migliori), l'algoritmo identifica gli studenti migliori all'interno di ogni cluster calcolando il 90° percentile. È fondamentale sottolineare che, una volta individuati gli studenti migliori, si procede a una verifica per assicurarsi che in ogni cluster siano presenti almeno due studenti appartenenti a questa categoria. Le ragioni di tale scelta saranno approfondite successivamente.

▷ **Determinazione degli argomenti migliori**
Per "argomenti migliori" si intendono quei temi già affrontati dagli studenti migliori di ciascun cluster, sui quali l'algoritmo si basa per fornire suggerimenti personalizzati a ogni studente. Il principio sottostante è simile a quello scolastico, secondo cui, per prepararsi efficacemente a un'interrogazione o a una verifica, è spesso utile ricevere consigli da chi possiede una conoscenza più approfondita della materia. Seguendo questa logica, l'algoritmo applica il principio dell'80/20 per individuare gli argomenti chiave di ciascuno studente migliore, i quali possono costituire un riferimento anche per gli altri studenti, inclusi sé stessi, che a loro volta si affidano ad altri pari di alto livello. Questo spiega la necessità, già evidenziata in precedenza, di garantire la presenza di almeno due studenti migliori per ciascun cluster.

In dettaglio, l'algoritmo analizza la matrice delle performance, identifica lo studente in esame e determina la categoria di appartenenza. Successivamente, individua gli studenti migliori all'interno di tale categoria e, per ciascuno di essi, esamina il numero di argomenti affrontati. A questo punto, viene calcolata una somma pesata decrescente basata sugli argomenti in cui ogni studente migliore ha ottenuto le performance più elevate. L'aggregazione dei pesi avviene progressivamente, privilegiando gli argomenti con maggiore rilevanza percentuale, fino

al raggiungimento della soglia dell'80%. L'elenco risultante comprende quindi quel 20% degli argomenti che, nel complesso, ha contribuito a generare l'80% degli effetti positivi sulle performance degli studenti.

Esempio Applicativo

Supponiamo che l'algoritmo debba suggerire a **Marco** quali argomenti studiare, basandosi sulle performance di **Bob**, lo studente migliore nella sua categoria. Gli argomenti studiati da Bob e il loro impatto sulle sue performance complessive sono riportati nella seguente tabella:

Argomento	**Peso relativo sulle performance di Bob**
Congiuntivo indipendente e dipendente	25%
Periodo ipotetico	22%
Declinazioni dei sostantivi	18%
Costruzioni con il gerundio e il gerundivo	15%
Versioni di Cesare	12%
Metri poetici in Orazio	8%

Tabella 1: Peso degli argomenti studiati da Bob sulle sue performance

L'obiettivo dell'algoritmo è identificare circa gli argomenti che hanno generato **l'80% dell'effetto complessivo**, ovvero il miglioramento di Bob.

1. Calcolo dell'effetto cumulativo

L'algoritmo somma i pesi in ordine decrescente fino a raggiungere l'**80% degli effetti complessivi**:

Argomento	**Peso relativo**	**Effetto cumulativo**
Congiuntivo indipendente e dipendente	25%	25%
Periodo ipotetico	+22%	47%
Declinazioni dei sostantivi	+18%	65%
Costruzioni con il gerundio e il gerundivo	+15%	80%
Versioni di Cesare	+12%	92%
Metri poetici in Orazio	+8%	100%

Tabella 2: Calcolo dell'effetto cumulativo degli argomenti studiati da Bob

A questo punto, abbiamo raggiunto l'**80% degli effetti complessivi** attraverso i primi **quattro argomenti**. Questi rappresentano gli **"argomenti**

migliori" da suggerire a Marco.

2. Selezione degli argomenti migliori
Gli argomenti selezionati, secondo il principio dell'80/20, sono:

- Congiuntivo indipendente e dipendente
- Periodo ipotetico
- Declinazioni dei sostantivi
- Costruzioni con il gerundio e il gerundivo

Gli argomenti *Versioni di Cesare* (12%) e *Metri poetici in Orazio* (8%) **non vengono selezionati**, poiché il loro impatto sulle performance di Bob è minore e non rientra tra gli argomenti con il maggiore effetto.

3. Risultato
Ora l'algoritmo suggerirà a Marco di concentrarsi su questi quattro argomenti, poiché sono quelli che, nel caso di Bob, hanno contribuito in modo determinante al suo successo.

▷ Filtro degli argomenti migliori
È naturale pensare che tanti studenti migliori possono consigliare ad uno studente qualunque lo stesso argomento più volte. Per tali ragioni, ottenuto l'elenco definitivo degli argomenti suggeriti, l'algoritmo applica un filtro per definire in modo univoco ogni singolo argomento consigliato.

▷ Selezione degli argomenti da suggerire
A questo punto, non è garantito che tutti gli argomenti migliori vengano effettivamente suggeriti. Affinché un argomento non venga consigliato a uno studente, devono verificarsi due determinate condizioni. In particolare, lo studente in questione non deve aver già raggiunto almeno il **60%** delle performance totali associate a quell'argomento (dove il 60% si riferisce al punteggio massimo ottenibile sommando tutti gli esercizi relativi a tale argomento).

Inoltre, l'argomento non verrà suggerito se la differenza tra la performance dello studente migliore e quella dello studente corrente risulta negativa o nulla,

ovvero se quest'ultimo ha ottenuto un risultato superiore allo studente migliore. In tal caso, non vi sarebbe alcuna ragione per consigliare tale argomento.

3.4.2 Metodo Item-Item Based

Il metodo *Item-Item Based Collaborative Filtering* utilizzato in questo lavoro si basa su un approccio di filtraggio collaborativo che sfrutta la similarità tra argomenti per prevedere le performance future di uno studente su argomenti non ancora affrontati.

▷ **Descrizione del metodo**

L'obiettivo del modello è identificare gli argomenti sui quali uno studente potrebbe ottenere buoni risultati, sfruttando:

- La **similarità coseno** tra gli argomenti;
- Le **performance medie** degli studenti appartenenti allo stesso cluster;
- Una **normalizzazione** che evita di dare troppo peso a singoli argomenti molto simili.

Per ciascuno studente, il modello segue i seguenti passi:

1. Calcolo della *similarità coseno* tra gli argomenti.
2. Identificazione degli argomenti **già svolti** dallo studente.
3. Determinazione delle **performance medie del cluster** dello studente sugli argomenti già svolti.
4. Stima della performance sugli argomenti **non ancora svolti** utilizzando la formula:

$$\hat{P}_i = \frac{\sum_{j \in \text{ArgSvolti}} \text{sim}(i,j) \cdot \text{MediaCluster}(j)}{\sum_{j \in \text{ArgSvolti}} |\text{sim}(i,j)|} \in [0,1] \tag{1}$$

dove:

- $\hat{P}_i$ è la performance prevista per l'argomento i;
- j rappresenta un argomento già svolto dallo studente;

- $\text{sim}(i, j)$ è la similarità coseno tra gli argomenti i e j;
- $\text{MediaCluster}(j)$ è la performance media del cluster dello studente sull'argomento j.

▷ **Selezione degli argomenti da suggerire**

Gli argomenti con una performance prevista superiore a una certa soglia, determinata come il *quantile* 0.60 della distribuzione delle predizioni, vengono suggeriti allo studente. Il risultato finale è un elenco personalizzato di argomenti su cui lo studente ha una maggiore probabilità di successo.

Esempio Applicativo

Supponiamo che **Marco** abbia studiato i seguenti argomenti:

- Periodo ipotetico
- Declinazioni dei sostantivi
- Versioni di Cesare
- Metri poetici in Orazio

Utilizzando la formula del modello, possiamo calcolare la probabilità di successo sugli argomenti non ancora svolti. Supponiamo che tra gli argomenti rimanenti vi siano:

Argomento	**Performance prevista ($\hat{P}$)**
Congiuntivo indipendente e dipendente	0.72
Costruzioni perifrastiche	0.55
Forme irregolari del participio	0.49
Complementi di luogo e moto	0.53

Tabella 3: Performance prevista sugli argomenti non ancora svolti da Marco

Dopo l'applicazione della formula 1, il modello restituisce che tra gli argomenti non ancora svolti, quello su cui Marco ha la maggiore probabilità di ottenere buoni risultati è il **Congiuntivo indipendente e dipendente**. Pertanto, il sistema suggerisce a Marco di concentrarsi su questo argomento per massimizzare il suo apprendimento.

3.5 Verifica dei metodi adoperati

Visti i metodi precedentemente implementati per la ricerca e selezione dei suggerimenti da proporre a ciascuno studente, sorge spontanea la necessità di valutare se tali suggerimenti abbiano un impatto positivo o negativo sulle loro performance. Il problema centrale, dunque, consiste nell'individuare una relazione tra i suggerimenti generati dai due algoritmi impiegati (**Pareto** e **Item-Based**) e il miglioramento registrato dagli studenti nel corso degli anni.

Per supportare questa analisi, facciamo ricorso a tecniche statistiche già introdotte nel primo capitolo di questo lavoro e introduciamo due nuove variabili di riferimento: il **tasso di adesione** e la **variazione**.

3.5.1 Il Tasso di adesione

Il **tasso di adesione** rappresenta il rapporto tra il numero di argomenti consigliati dall'algoritmo (sulla base di uno dei due metodi adottati) su cui lo studente ha effettivamente deciso di esercitarsi e il numero totale di argomenti suggeriti. In formula:

$$T_{Adesione} = \frac{N_{\text{argomenti seguiti}}}{N_{\text{argomenti consigliati}}} \in [0, 1] \tag{2}$$

dove:

- $N_{\text{argomenti seguiti}}$ indica il numero di argomenti consigliati su cui lo studente si è effettivamente esercitato;
- $N_{\text{argomenti consigliati}}$ rappresenta il numero totale di argomenti suggeriti dall'algoritmo.

Chiaramente un tasso di adesione prossimo a 1 indica un'elevata fiducia dello studente nei suggerimenti forniti dall'algoritmo, poiché implica che la maggior parte degli argomenti consigliati siano stati effettivamente seguiti. Viceversa, un valore prossimo allo 0 implica poca affidabilità nei confronti dell'algoritmo.

3.5.2 La Variazione

Dal momento che i dati a disposizione consentono di analizzare le performance ottenute da ciascuno studente nel corso degli anni e che, attraverso l'uso

delle matrici di performance precedentemente introdotte, è possibile ottenere una visione d'insieme di tale andamento e risulta immediato determinare la performance marginale di ogni studente nell'anno corrente. Questa può essere calcolata sommando i punteggi ottenuti in ciascun argomento di ogni ambito, così come indicato dalla matrice di performance.

Ricordiamo che i punteggi associati ad ogni argomento sono dei valori reali pertanto anche la somma di essi scaturisce un valore reale.

Esempio Applicativo

Studente	Periodo ipotetico	Versioni di Cesare	Metri poetici in Orazio	Performance totale
Marco	6	-3	4	7
Lucia	8	2	-1	9
Davide	5	-4	-3	-2
Sara	7	-9	2	0

Tabella 4: Performance degli studenti nei diversi argomenti e performance totale

A tal proposito, ripetendo lo stesso ragionamento per ogni matrice di performance a nostra disposizione e andando a interessarci unicamente dell'ultima colonna nell'esempio mostrata, la **variazione** rappresenta la differenza tra la performance ottenuta da uno studente nell'anno successivo e quella registrata nell'anno corrente. Essa fornisce un'indicazione quantitativa dell'andamento delle sue competenze, permettendo di distinguere tra:

- **Miglioramento** se la variazione assume un valore **positivo**.
- **Peggioramento** se la variazione assume un valore **negativo**.
- **Invariato** se la variazione è **zero**.

Questo parametro è utile per analizzare l'evoluzione del rendimento dello studente nel tempo e valutare l'efficacia dei suggerimenti proposti dagli algoritmi.

3.5.3 Il modello di Regressione Lineare

Sulla base dei presupposti precedentemente esposti, possiamo considerare un modello di regressione lineare semplice in cui il **tasso di adesione**, calcolato

per tutti gli studenti, costituisce l'insieme dei valori della variabile indipendente. Allo stesso modo, le diverse **variazioni** rappresentano le variabili dipendenti.

L'obiettivo è determinare in che misura il tasso di adesione influenzi significativamente le variazioni di performance. Per questo motivo, risulta necessario applicare un modello di regressione lineare al fine di stimare valori che descrivano in modo accurato la relazione esistente tra le due variabili.

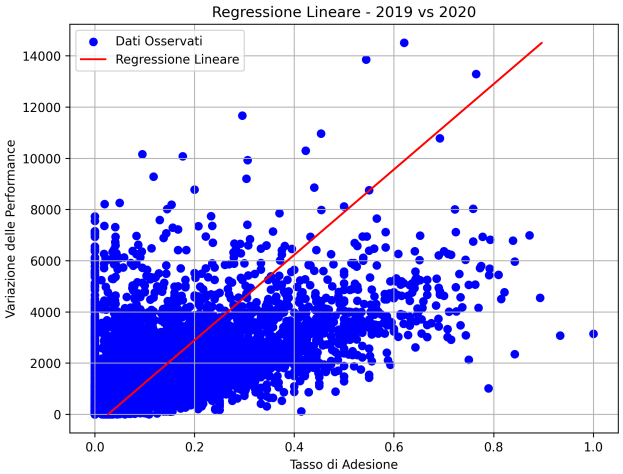
4 Applicazione e Test sui Modelli

In questo capitolo ci dedichiamo alla visualizzazione e alla fase di test per analizzare a fondo la relazione esistente tra le variabili, distinguendo due diverse tipologie di rappresentazione in base al metodo adottato. Per rafforzare ulteriormente la nostra analisi, è stata condotta un'indagine più approfondita sulla **frequenza di studio media** di ciascuno studente, calcolata su intervalli di due anni. L'obiettivo è valutare in che misura un maggiore impegno nello studio influenzi il miglioramento individuale e con quale grado di significatività. Per questo motivo, accanto ai modelli di base, sono stati sviluppati **modelli estesi**, nei quali, oltre al **tasso di adesione**, viene inclusa la **frequenza di studio** come variabile predittiva aggiuntiva. Il confronto tra i due approcci verrà effettuato attraverso il **test ANOVA**, che consentirà di determinare quale modello descriva meglio la relazione tra le variabili analizzate.

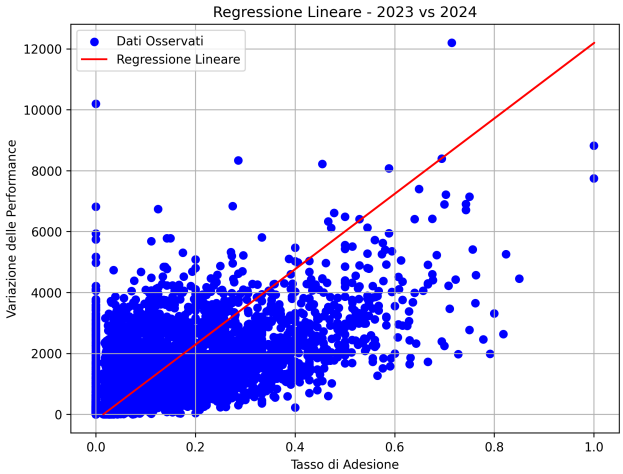
4.1 Risultati ottenuti attraverso il Metodo di Pareto

Analisi modelli base

Attraverso i grafici presentati, è possibile osservare come i suggerimenti forniti influenzino il miglioramento di ciascuno studente nel corso degli anni. Di seguito sono riportate alcune rappresentazioni grafiche relative alle diverse annualità.



(a) Analisi Regressione Lineare anni 2019-2020



(b) Analisi Regressione Lineare anni 2023-2024

Figura 8: Confronto tra le analisi di regressione lineare per gli anni 2019-2020 e 2023-2024

Dove possiamo osservare i seguenti risultati ottenuti:

Parametro	**Valore**
DataFrame di riferimento	2019-2020
β_0 (intercetta)	0.0268
β_1 (coefficiente)	0.0001
t-statistic per β_1	69.7256
P-value per β_1	$\leq$ 2e-10
R^2 (coefficiente di determinazione)	0.3977
Conclusione	C'è evidenza statistica per rifiutare l'ipotesi nulla. β_1 è significativo.
Considerazione sul modello	Il modello spiega poca variabilità dei dati (R^2 basso).

Tabella 5: Risultati del modello di regressione lineare per la coppia di anni 2019-2020 (8a)

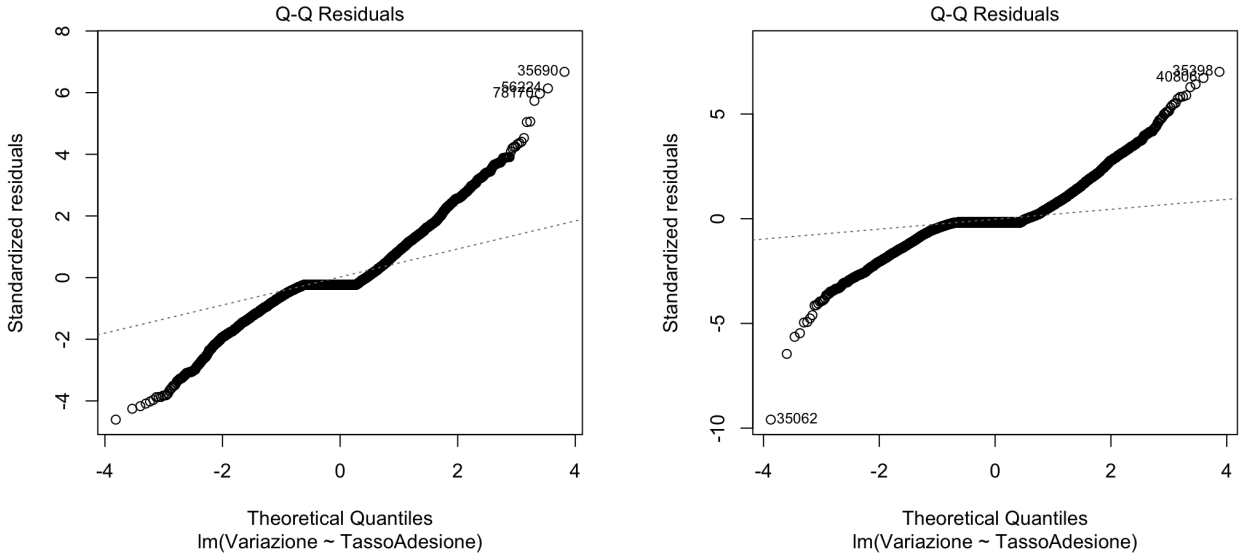
Parametro	**Valore**
DataFrame di riferimento	2023-2024
β_0 (intercetta)	0.0153
β_1 (coefficiente)	0.0001
t-statistic per β_1	96.5178
P-value per β_1	$\leq$ 2e-10
R^2 (coefficiente di determinazione)	0.4995
Conclusione	C'è evidenza statistica per rifiutare l'ipotesi nulla. β_1 è significativo.
Considerazione sul modello	Il modello spiega una parte moderata della variabilità dei dati.

Tabella 6: Risultati del modello di regressione lineare per la coppia di anni 2023-2024 (8b)

I risultati delle analisi di regressione per il 2019-2020 e il 2023-2024 mostrano una relazione statisticamente significativa tra il **tasso di adesione** e la **variazione della performance**, come confermato dai **P-value** molto bassi ($\leq 2e-10$) e dalle elevate **t-statistic** per β_1.

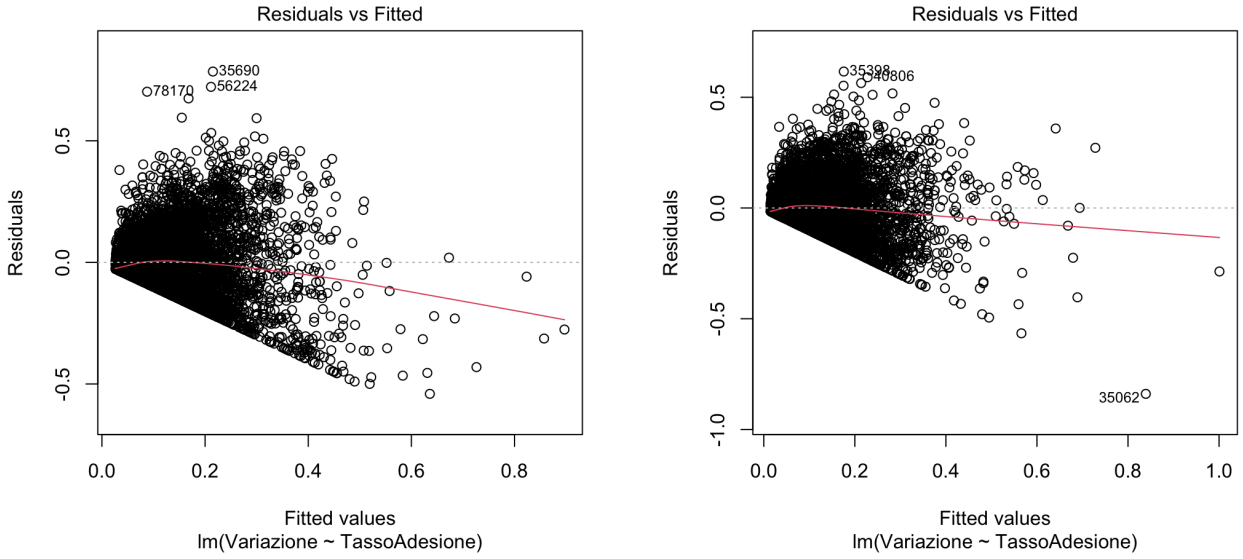
Tuttavia, confrontando i coefficienti di determinazione $\mathbf{R^2}$, si nota un incremento dal **39.77%** nel 2019-2020 al **49.95%** nel 2023-2024, segnalando un miglioramento nel potere esplicativo del modello. Ciò suggerisce una maggiore coerenza tra adesione ai suggerimenti e miglioramento della performance, probabilmente dovuta a strategie più efficaci o a una maggiore consapevolezza degli studenti. Tuttavia, il valore di R^2 indica che altri fattori influenzano la performance, rendendo necessaria un'analisi più approfondita con modelli più complessi o ulteriori variabili predittive.

Passiamo ora all'analisi dei grafici relativi a **QQ-Plot**, **Residuals vs Fitted-Plot** e **Scale-Location Plot** per entrambi gli anni.



(a) Analisi QQ-Plot anni 2019-2020 (b) Analisi QQ-Plot anni 2023-2024

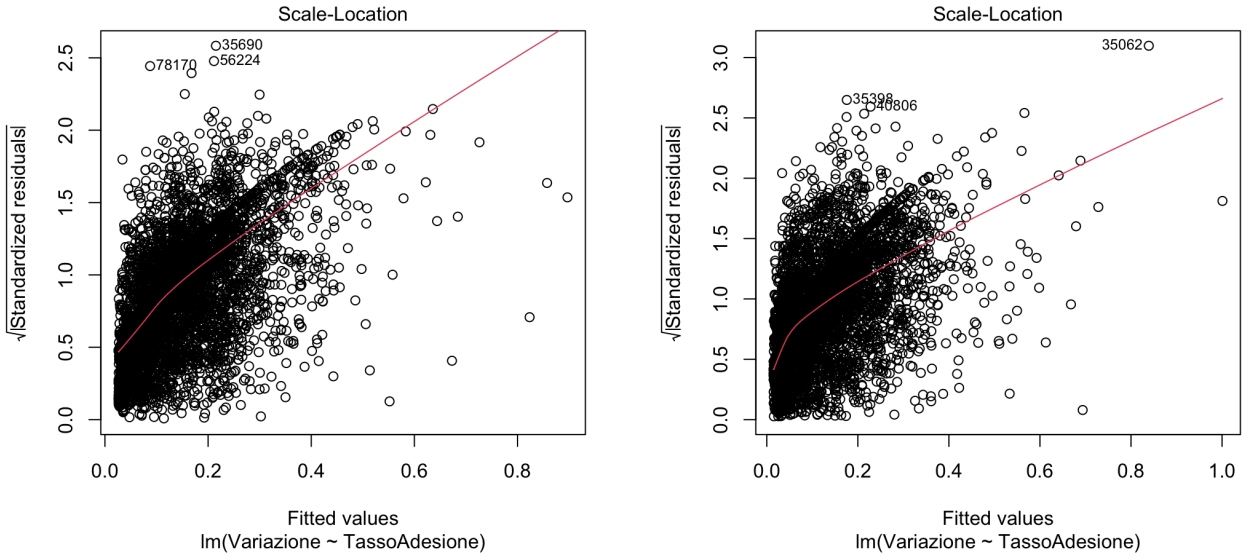
Figura 9: Confronto tra le analisi dei QQ-Plot per gli anni 2019-2020 e 2023-2024



(a) Analisi ResidualvsFitted Plot anni 2019-2020 (b) Analisi ResidualvsFitted Plot anni 2023-2024

Figura 10: Confronto tra le analisi dei ResidualvsFitted Plot per gli anni 2019-2020 e 2023-2024

L'analisi dei QQ-Plot mostra una deviazione significativa dai quantili teorici nelle code della distribuzione, indicando la presenza di valori anomali e una



(a) Analisi Scale Location Plot anni 2019-2020

(b) Analisi Scale Location Plot anni 2023-2024

Figura 11: Confronto tra le analisi dei Scale Location Plot per gli anni 2019-2020 e 2023-2024

possibile violazione dell'ipotesi di normalità dei residui. Tuttavia, nella parte centrale, la distribuzione sembra allinearsi ragionevolmente alla normalità, suggerendo che eventuali deviazioni potrebbero derivare da una limitata porzione dei dati. Nei Residuals vs Fitted Plot, i residui non risultano distribuiti in modo completamente casuale attorno allo zero, ma mostrano una lieve struttura, suggerendo che il modello potrebbe non aver catturato appieno la relazione tra il tasso di adesione e la variazione della performance. Inoltre, si osserva un incremento della dispersione dei residui per valori più elevati del tasso di adesione, evidenziando una possibile eteroschedasticità, ossia una varianza non costante degli errori. Tale fenomeno è confermato dagli Scale-Location Plot, nei quali si nota una tendenza crescente della curva rossa, a indicare che la varianza dei residui tende ad aumentare per valori più elevati della variabile indipendente. Tuttavia, nel modello relativo agli anni 2023-2024, si osserva una riduzione di questa tendenza rispetto agli anni 2019-2020, suggerendo un leggero miglioramento nell'adattamento del modello ai dati più recenti.

Analisi Test ANOVA

Visualizziamo adesso i risultati riportati attraverso il Test ANOVA andando

ad aggiungere le **Frequenze di Studio** come nuova variabile indipendente nel modello esteso:

Parametro	**Valore**
Modello Esteso	Frequenze di Studio
R^2 (coefficiente di determinazione)	0.4074
Test ANOVA	
RSS Modello Base	101.8178
RSS Modello Esteso	100.1816
Delta RSS	1.6362
Gradi di libertà modello	1
Gradi di libertà residuo	7361
Statistica F	120.2198
P-value	$\leq 2e-10$
R^2 **e** R^2 **Aggiustato**	
R^2 Modello Base	0.3977
R^2 Aggiustato Modello Base	0.3976
R^2 Modello Esteso	0.4074
R^2 Aggiustato Modello Esteso	0.4072
Conclusione	Il modello esteso fornisce un miglioramento lieve rispetto al modello base.

Tabella 7: Risultati del test ANOVA per il modello esteso con Frequenze di Studio anni 2019-2020

Parametro	**Valore**
Modello Esteso	Frequenze di Studio
R^2 (coefficiente di determinazione)	0.5075
Test ANOVA	
RSS Modello Base	72.0021
RSS Modello Esteso	70.8428
Delta RSS	1.1592
Gradi di libertà modello	1
Gradi di libertà residuo	9334
Statistica F	152.7369
P-value	$\leq 2e-10$
R^2 **e** R^2 **Aggiustato**	
R^2 Modello Base	0.4995
R^2 Aggiustato Modello Base	0.4994
R^2 Modello Esteso	0.5075
R^2 Aggiustato Modello Esteso	0.5074
Conclusione	Il modello esteso fornisce un miglioramento lieve rispetto al modello base.

Tabella 8: Risultati del test ANOVA per il modello esteso con Frequenze di Studio anni 2023-2024

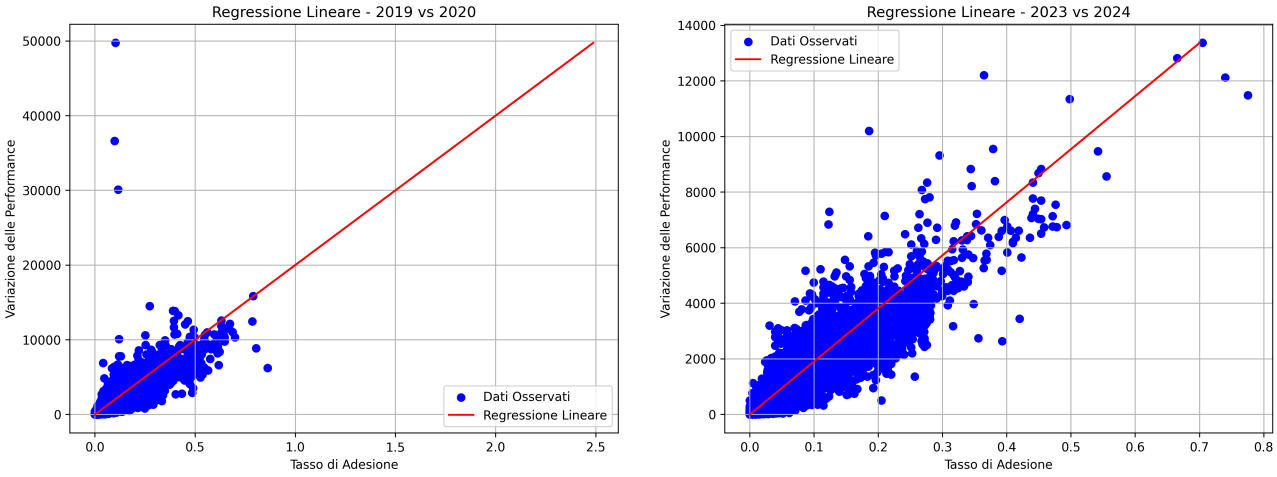
Dai risultati ottenuti possiamo notare che il coefficiente di determinazione R^2 passa da **0.3977** a **0.4074** per gli anni **2019-2020** e da **0.4995** a **0.5075** per gli anni **2023-2024**, indicando un lieve miglioramento nella capacità esplicativa del modello. La **statistica F** assume valori elevati in entrambi i casi (**120.2198**

per 2019-2020 e **152.7369** per 2023-2024), con un **P-value** significativamente basso ($\leq 2e-10$), suggerendo che l'inclusione della nuova variabile ha un impatto statisticamente significativo sulla spiegazione della **variazione della performance**. Nonostante ciò, l'incremento di R^2 rimane modesto, suggerendo che, sebbene la frequenza di studio abbia un certo effetto, il modello esteso fornisce un miglioramento rispetto al modello base, ma in misura contenuta.

4.2 Risultati ottenuti attraverso il Metodo Item Based

Analisi modelli base

Ora riproponiamo le stesse analisi viste precedentemente ma sfruttando il **metodo Item Based**. Anche per quest'ultimo faremo poi un confronto applicando le frequenze di studio ad ogni studente e andando ad osservare i risultati proposti nel test ANOVA. Partiamo, come prima, dalla regressione lineare delle stesse coppie di anni.



(a) Analisi Regressione Lineare anni 2019-2020

(b) Analisi Regressione Lineare anni 2023-2024

Figura 12: Confronto tra le analisi di regressione lineare per gli anni 2019-2020 e 2023-2024

Come notiamo, il confronto tra le analisi di regressione lineare per gli anni **2019-2020** e **2023-2024** evidenzia una relazione positiva tra il **tasso di adesione** e la **variazione della performance**. In entrambi i grafici, la retta di regressione (in rosso) indica una tendenza crescente, suggerendo che un maggiore tasso di adesione ai suggerimenti è associato a un miglioramento delle performance degli studenti. Eppure, si osservano differenze significative tra i

due periodi. Nel **2019-2020**, la dispersione dei dati è maggiore e sono presenti diversi valori estremi, che possono influenzare la stima della relazione. Nel **2023-2024**, i dati risultano più concentrati attorno alla retta di regressione, suggerendo una maggiore coerenza del modello nel descrivere il fenomeno. Di seguito, i risultati nei test applicati:

Parametro	**Valore**
DataFrame di riferimento	2019-2020
β_0 (intercetta)	0.0016
β_1 (coefficiente)	0.0000
t-statistic per β_1	796.1366
p-value per β_1	$\leq$ 2e-10
R^2 (coefficiente di determinazione)	0.8605
Conclusione	C'è evidenza statistica per rifiutare l'ipotesi nulla. β_1 è significativo.
Considerazione sul modello	Il modello spiega gran parte della variabilità dei dati (R^2 elevato).

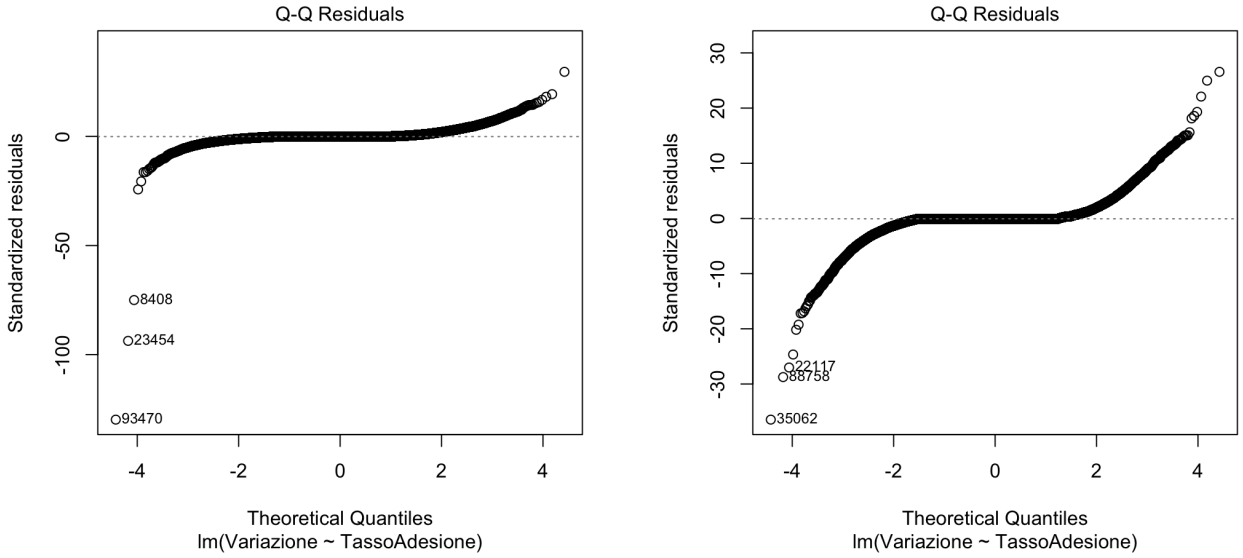
Tabella 9: Risultati del modello di regressione lineare per la coppia di anni 2019-2020

Parametro	**Valore**
DataFrame di riferimento	2023-2024
β_0 (intercetta)	0.0006
β_1 (coefficiente)	0.0001
t-statistic per β_1	960.4921
p-value per β_1	$\leq$ 2e-10
R^2 (coefficiente di determinazione)	0.8998
Conclusione	C'è evidenza statistica per rifiutare l'ipotesi nulla. β_1 è significativo.
Considerazione sul modello	Il modello spiega gran parte della variabilità dei dati (R^2 elevato).

Tabella 10: Risultati del modello di regressione lineare per la coppia di anni 2023-2024

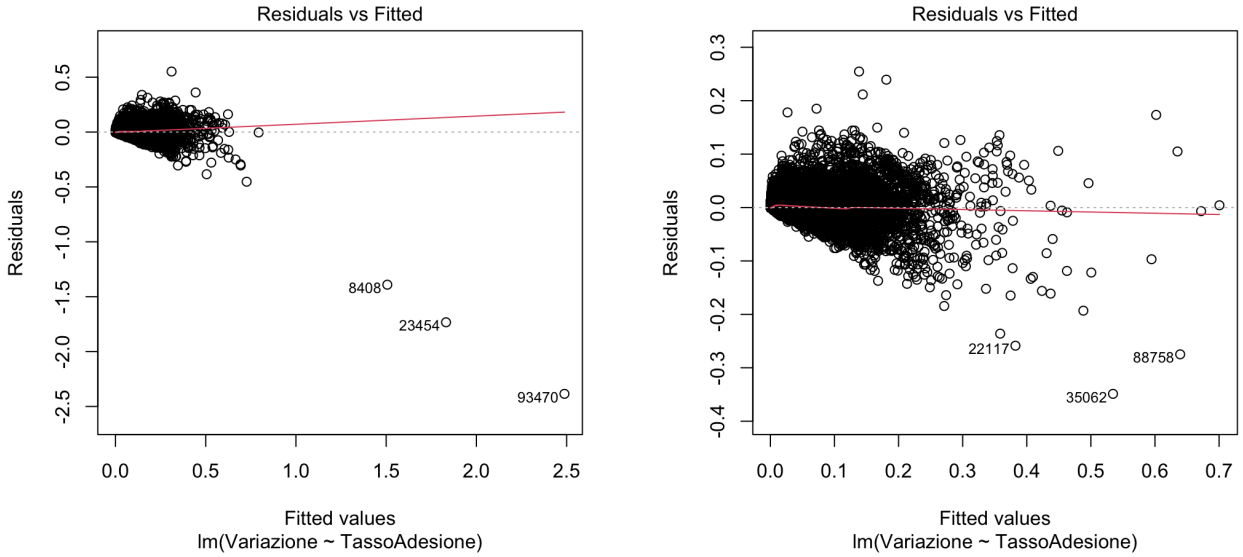
Come si può facilmente notare dalle tabelle, in entrambi i casi, il coefficiente β_1 risulta **statisticamente significativo**, come indicato dal **P-value estremamente basso** ($\leq 2e-10$), confermando ancora una volta l'esistenza di una relazione tra il **tasso di adesione** e la **variazione della performance**. Il valore di R^2 è **elevato in entrambi i modelli** (0.8605 nel 2019-2020 e 0.8998 nel 2023-2024), suggerendo che il modello spiega gran parte della variabilità dei dati. Inoltre, l'aumento di R^2 nel 2023-2024 rispetto al 2019-2020 indica un miglioramento della capacità predittiva del modello nel tempo. Passiamo ora all'analisi dei grafici relativi a **QQ-Plot**, **Residuals vs Fitted-Plot** e **Scale-Location Plot** per entrambi gli anni.

Dai **QQ-Plot**, emerge che in entrambi i periodi i residui mostrano una di-



(a) Analisi QQ-Plot anni 2019-2020 (b) Analisi QQ-Plot anni 2023-2024

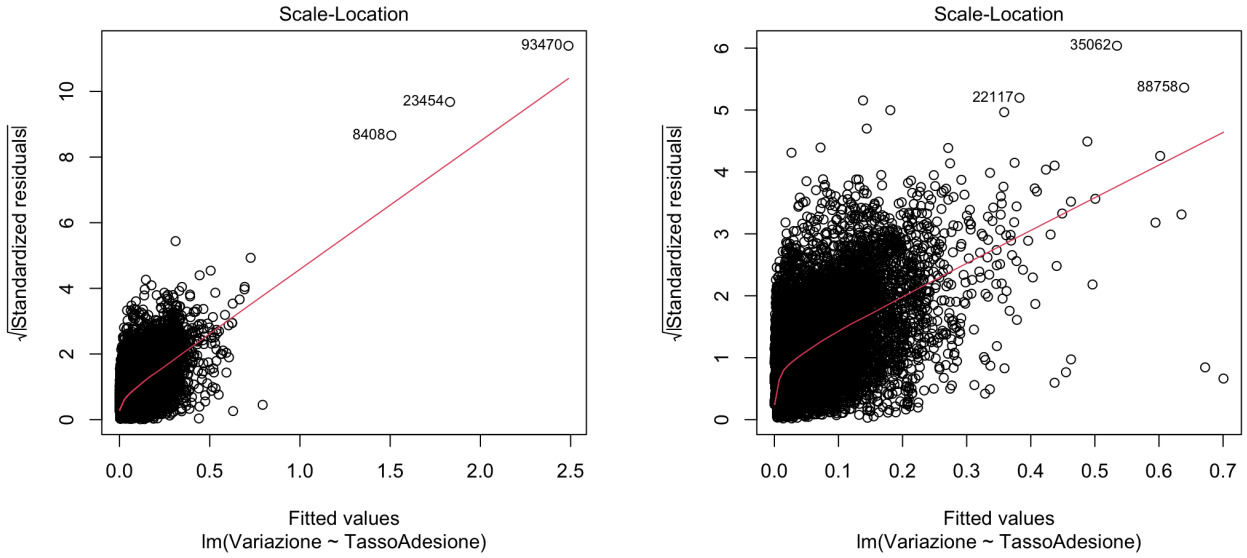
Figura 13: Confronto tra le analisi dei QQ-Plot per gli anni 2019-2020 e 2023-2024



(a) Analisi ResidualvsFitted Plot anni 2019-2020 (b) Analisi ResidualvsFitted Plot anni 2023-2024

Figura 14: Confronto tra le analisi dei ResidualvsFitted Plot per gli anni 2019-2020 e 2023-2024

screta aderenza alla distribuzione normale nella parte centrale, ma presentano deviazioni evidenti nelle code, suggerendo la presenza di valori anomali. Tut-



(a) Analisi Scale Location Plot anni 2019-2020

(b) Analisi Scale Location Plot anni 2023-2024

Figura 15: Confronto tra le analisi dei Scale Location Plot per gli anni 2019-2020 e 2023-2024

tavia, nel **2023-2024** queste deviazioni risultano meno accentuate rispetto al **2019-2020**, indicando un leggero miglioramento nella normalità dei residui. Nei **Residuals vs Fitted Plot**, il modello relativo al **2019-2020** mostra una maggiore dispersione e la presenza di alcuni valori anomali con residui molto distanti dalla linea centrale. Al contrario, nel **2023-2024**, la distribuzione dei residui appare più compatta, suggerendo un adattamento migliore del modello ai dati. Tuttavia, in entrambi i periodi, si notano leggere strutture nei residui che potrebbero indicare una relazione non completamente lineare tra le variabili. Infine, i **Scale-Location Plot** rivelano una chiara tendenza all'**eteroschedasticità** in entrambi i periodi, con un aumento della dispersione dei residui per valori più elevati della variabile indipendente. Questo effetto è più marcato nel **2019-2020**, mentre nel **2023-2024** si osserva una distribuzione più omogenea, suggerendo una lieve riduzione dell'eteroschedasticità.

Analisi Test ANOVA

Anche quì, riportiamo i risultati ottenuti attraverso il Test ANOVA andando ad aggiungere le **Frequenze di Studio** come nuova variabile indipendente nel modello esteso:

Parametro	Valore
Modello Esteso	Frequenze di Studio
R^2 (coefficiente di determinazione)	0.9207
Test ANOVA	
RSS Modello Base	9.4390
RSS Modello Esteso	7.4709
Delta RSS	1.9681
Gradi di libertà modello	1
Gradi di libertà residuo	102780
Statistica F	27075.9496
P-value	$\leq 2e-10$
R^2 **e** R^2 **Aggiustato**	
R^2 Modello Base	0.8998
R^2 Aggiustato Modello Base	0.8998
R^2 Modello Esteso	0.9207
R^2 Aggiustato Modello Esteso	0.9207
Conclusione	Il modello esteso fornisce un miglioramento significativo rispetto al modello base.

Tabella 11: Risultati del test ANOVA per il modello esteso con Frequenze di Studio anni 2019-2020

Parametro	Valore
Modello Esteso	Frequenze di Studio
R^2 (coefficiente di determinazione)	0.9202
Test ANOVA	
RSS Modello Base	35.6989
RSS Modello Esteso	20.4222
Delta RSS	15.2766
Gradi di libertà modello	1
Gradi di libertà residuo	102780
Statistica F	76883.4555
P-value	$\leq 2e-10$
R^2 **e** R^2 **Aggiustato**	
R^2 Modello Base	0.8605
R^2 Aggiustato Modello Base	0.8605
R^2 Modello Esteso	0.9202
R^2 Aggiustato Modello Esteso	0.9202
Conclusione	Il modello esteso fornisce un miglioramento significativo rispetto al modello base.

Tabella 12: Risultati del test ANOVA per il modello esteso con Frequenze di Studio anni 2023-2024

Come si evince, entrambe le analisi mostrano un **miglioramento significativo** del modello esteso rispetto al modello base, come evidenziato dall'aumento dei valori di R^2 e dalla riduzione dell'errore residuo (RSS). In particolare, nella Tabella 11 il coefficiente di determinazione R^2 passa da **0.8998** nel modello base a **0.9207** nel modello esteso, mentre nella Tabella 12 si osserva un incremento da **0.8605** a **0.9202**. Questo indica che l'aggiunta della frequenza di studio contribuisce a spiegare una maggiore quota della variabilità delle performance degli studenti. La **statistica F** assume valori molto elevati in entrambi

i casi (**27075.9496** e **76883.4555**), con un **P-value estremamente basso** ($\leq 2e-10$), confermando la significatività del miglioramento apportato dal modello esteso.

4.3 Confronto dei risultati ottenuti

Giunti a questo punto dell'analisi, è possibile confrontare i risultati ottenuti nei due modelli, evidenziando l'evoluzione dell'indice R^2 nel tempo e calcolandone la media per ciascuno dei metodi proposti.

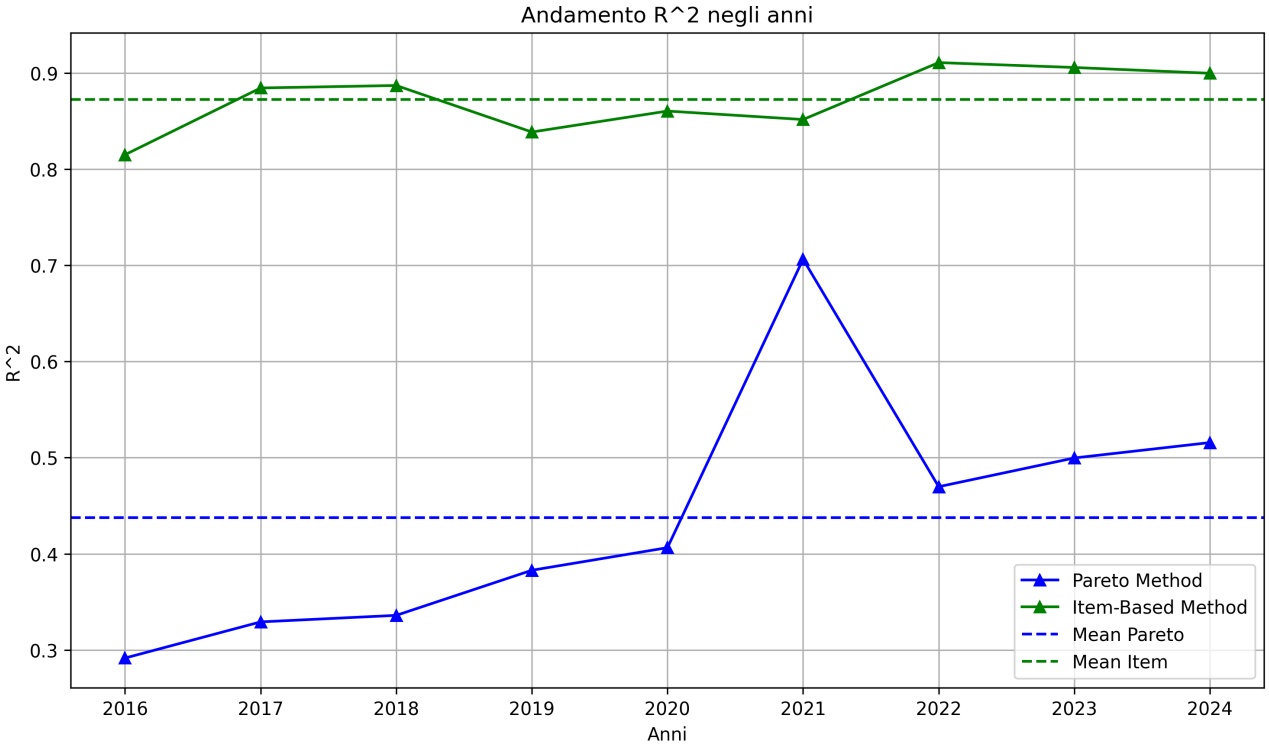


Figura 16: Evoluzione dell'indice R^2 negli anni

Dall'analisi dei grafici emerge che i dati si adattano meglio al secondo metodo, ovvero **Item Based**. Nel primo caso, si osserva un picco dell'indice R^2 nel periodo **2020-2021**, suggerendo un'elevata capacità del secondo metodo nel rappresentare i valori osservati. Questo risultato appare particolarmente affidabile, poiché i dataset disponibili mostrano una forte concentrazione di informazioni derivante dall'elevato numero di studenti che hanno utilizzato la piattaforma Alatin in quel periodo. Tale incremento è strettamente legato al contesto

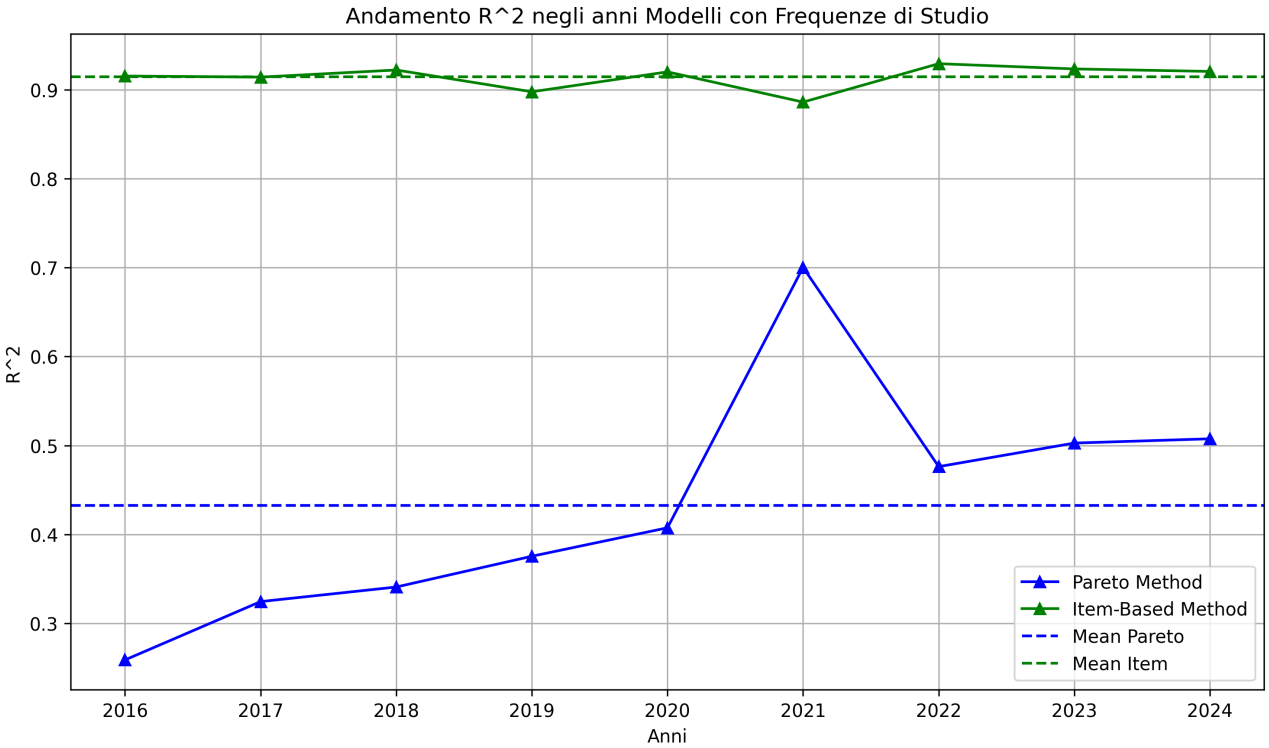


Figura 17: Evoluzione dell'indice R^2 negli anni con Frequenze di Studio

storico: durante la pandemia, l'insegnamento a distanza è diventato una pratica diffusa, influenzando significativamente l'uso della piattaforma. Escludendo quella coppia di anni, il metodo **Pareto** mostra comunque una tendenza crescente rispetto al secondo metodo, che mantiene risultati stabili in un intervallo compreso tra 0.8 e poco oltre 0.9. L'andamento dei modelli base non differisce significativamente da quello dei modelli estesi. Tuttavia, nella Figura 17, si osserva una maggiore concentrazione dei risultati per il metodo **Item Based**, mentre il metodo **Pareto** mostra un'evoluzione più marcata. Questo trend è ulteriormente confermato dal successivo calcolo dell'R^2 aggiustato, suggerendo che l'incremento del numero di variabili predittive potrebbe progressivamente avvicinare i risultati a quelli ottenuti con il metodo precedente.

4.3.1 Considerazioni Finali

Tentando di perfezionare il **metodo di Pareto**, i risultati ottenuti rimangono sostanzialmente invariati. È importante sottolineare che tale metodo richie-

de che i punteggi utilizzati per calcolare le performance degli studenti **non includano penalità**, poiché, dal punto di vista matematico, il processo di generazione dei suggerimenti necessita che i valori delle performance totali di ciascuno studente siano **positivi o nulli** in ogni anno considerato. Un aspetto chiave della differenza tra i due metodi implementati inoltre risiede nel tipo di relazione su cui si basano. Il metodo **Pareto** stabilisce un collegamento tra uno studente generico e un altro studente considerato "migliore" all'interno del proprio cluster, identificando così chi possa fornire suggerimenti più efficaci agli altri. Al contrario, il metodo **Item Based** si basa su una relazione **argomento-argomento**, stimando, in base agli argomenti già affrontati da uno studente, la probabilità che quest'ultimo ottenga buoni risultati su nuovi argomenti. Dai grafici analizzati in precedenza, emerge chiaramente che il metodo **Item Based** si adatta meglio ai dati. Questo risultato può essere interpretato come un'indicazione del fatto che i dati disponibili provengono da una piattaforma multidisciplinare, nella quale **non esiste una relazione diretta tra studenti**, ma piuttosto una correlazione tra gli argomenti studiati. L'analisi condotta suggerisce quindi che, dal punto di vista matematico, il processo di apprendimento degli studenti tende a seguire maggiormente un modello basato sulle connessioni tra argomenti piuttosto che un confronto diretto tra studenti di diversa competenza.

5 Conclusioni e sviluppi futuri

Il progetto sviluppato è stato possibile grazie all'analisi dei dati provenienti dalla piattaforma **Alatin** gestita dall'azienda **Maieutical Labs**. L'obiettivo, come mostrato, è stato quello di individuare metodi capaci di stimare la relazione tra i suggerimenti forniti e il miglioramento delle loro performance. È stata fornita una descrizione dettagliata dell'implementazione dei metodi proposti e del loro ruolo nella generazione dei suggerimenti per ciascuno studente nel corso degli anni. L'analisi è stata realizzata attraverso l'uso di **Python** e **R**, strumenti che hanno permesso di sviluppare il progetto e di visualizzare i dati mediante plot esplicativi. Questi grafici hanno fornito una visione più chiara delle tendenze presenti nei dati, contribuendo alla comprensione dei risultati ottenuti e alla formulazione delle conclusioni finali. Tuttavia, per migliorare ulteriormente la ricerca in questo ambito, sarebbe opportuno approfondire alcuni aspetti del lavoro, con l'obiettivo di ottimizzare i risultati attraverso nuove strategie o metodologie non ancora implementate. Di seguito, vengono riportate alcune possibili direzioni di sviluppo, nella speranza che questa tesi possa rappresentare il punto di partenza per studi più approfonditi.

5.1 Miglioramento della gestione dei cluster

Come discusso in precedenza, per raggruppare gli studenti in base alle loro performance è stato utilizzato l'algoritmo **K-Means**. Questo metodo, nella sua fase iniziale, seleziona arbitrariamente tre centroidi all'interno dei dati disponibili, rappresentati dalle performance degli studenti in un determinato ambito. Eppure, se si considerasse un caso in cui le performance degli studenti fossero **disposte lungo un'unica retta**, risulterebbe complesso ottenere una suddivisione efficace in tre cluster sin dal primo step. Questo accade poiché, se i centroidi iniziali vengono posizionati troppo vicini tra loro, l'algoritmo potrebbe non riuscire a generare raggruppamenti ben distinti. Un'idea immediata per risolvere tale problema potrebbe essere l'adozione dell'algoritmo **DBSCAN**, il quale non dipende dalla disposizione lineare dei dati sulle performance. Tuttavia, anche questa soluzione risulterebbe inefficace per due motivi principali:

- **Numero di cluster generati**: l'algoritmo DBSCAN, per sua natura, tenderebbe a suddividere gli studenti in un numero di cluster superiore a tre.

Questo comprometterebbe la logica adottata, che prevede esattamente tre gruppi nei quali gli studenti vengono classificati come carenti, intermedi e migliori. Tale classificazione è fondamentale per l'analisi basata sul metodo di Pareto, che identifica, all'interno di ciascun cluster, gli studenti con maggiori conoscenze;

- **Numero di studenti esclusi**: nella nostra metodologia, ogni studente deve necessariamente appartenere a un cluster. Tuttavia, a causa dei parametri richiesti dall'algoritmo DBSCAN, è probabile che alcuni studenti con performance sopra la media vengano esclusi da qualsiasi gruppo e classificati come outlier, ovvero rumore nel dataset.

Escludendo dunque l'algoritmo DBSCAN, una possibile alternativa nell'ambito del **K-Means** consiste nel determinare a priori, e non in modo arbitrario, le posizioni iniziali dei tre centroidi. In particolare, tali centroidi potrebbero essere posizionati lungo la retta delle performance corrispondenti ai percentili 25^{o}, 50^{o} e 75^{o}, prima di avviare l'iterazione generale.

5.2 Introduzione di metodi alternativi a Pareto

Nella nostra analisi, il **metodo di Pareto** non prevede l'applicazione di penalità alle performance degli studenti. Per ottenere una visione più realistica del fenomeno e analizzare il comportamento dei dati in un contesto più aderente alla realtà, si potrebbe considerare l'implementazione di un metodo alternativo al precedente. In particolare, un'opzione valida potrebbe essere l'**User-Based Collaborative Filtering** [2], come alternativa al **Metodo di Pareto**. L'approccio **User-Based Collaborative Filtering** si concentra sulle similitudini tra utenti. In pratica, per un utente target vengono individuati altri utenti con gusti simili, e gli articoli che questi ultimi hanno apprezzato vengono proposti come raccomandazioni. A differenza dell'approccio item-based, che guarda alla correlazione tra oggetti, il metodo user-based sfrutta il comportamento collettivo degli utenti per individuare preferenze condivise. Infine, si potrebbe andare ad implementare un approccio ibrido, generalmente definito **Hybrid Recommender System**[4]. In molti casi, questi sistemi utilizzano tecniche di matrix factorization (ad esempio SVD) per combinare le informazioni ottenute dai me-

todi user-based e item-based, sfruttando così i punti di forza di entrambi per ottenere raccomandazioni più accurate.

Lista di figure

Riferimenti bibliografici

[1] Alatin. https://alatin.it. Accessed: 27 February 2025.

[2] CF Pinela. Recommender systems: User-based and item-based collaborative filtering. https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering- 2018. Accessed: 27 February 2025.

[3] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295. ACM, 2001. Accessed: 27 February 2025.

[4] Towards Data Science. Hybrid recommender systems. https://towardsdatascience.com/hybrid-recommender-systems-80a1dd7c2e4f, 2018. Accessed: 27 February 2025.