UNIVERSITÀ DEGLI STUDI DI SALERNO

Corso di Laurea in Informatica
Fondamenti di Intelligenza Artificiale

IGEA

Gruppo di Progetto:

Gennaro Pio Albano (Mat. 0512119547)

Giuseppe Annunziata (Mat. 0512120144)

Alessandro Bonelli (Mat. 0512119640)

Samuele Nacchia (Mat. 0512119128)

Link al repository: https://github.com/jupex69/IGEA

Anno Accademico 2025/2026

Indice

1 Introduzione

Il benessere psicologico degli studenti universitari rappresenta una tematica di crescente rilevanza nel panorama accademico e sanitario. Il percorso universitario, spesso caratterizzato da elevate pressioni performative, transizioni sociali significative e incertezza verso il futuro, costituisce una fase critica che può favorire l'insorgenza di disturbi dell'umore, tra cui la depressione.
Tale disagio, se trascurato, può aggravarsi fino a livelli insostenibili, portando nei casi più drammatici a gesti estremi.

1.1 Sistema attuale

Attualmente, l'individuazione di studenti universitari a rischio di depressione avviene principalmente tramite autosegnalazione o osservazioni indirette da parte di docenti e tutor.
I servizi di supporto psicologico operano in maniera reattiva, intervenendo solo quando il disagio è già esplicitamente manifestato. Non sono presenti strumenti automatici di analisi o predizione basati sui dati, rendendo difficile un'identificazione precoce e sistematica degli studenti potenzialmente vulnerabili.

1.2 Obiettivi

Il sistema IGEA - Intelligent Guide for Emotional Assessment è stato progettato con l'obiettivo di fornire un supporto proattivo nella rilevazione precoce della depressione tra gli studenti universitari, al fine di favorire interventi tempestivi e mirati da parte dell'università e degli psicologi dell'ateneo. IGEA è un alleato nel monitoraggio del benessere psicologico, ma non sostituisce il lavoro degli esperti. Il sistema non ha l'intento di diagnosticare o curare la depressione, ma piuttosto di identificare segnali di rischio che possano indicare uno stato di disagio emotivo o mentale, permettendo una valutazione iniziale della salute psicologica degli studenti.

2 Descrizione agente

Il sistema *IGEA* è modellato come un agente intelligente di tipo **classificatore**, progettato per supportare l'individuazione precoce di potenziali stati di disagio psicologico negli studenti universitari. L'agente analizza le risposte fornite dagli studenti tramite questionari psicologici strutturati e produce una valutazione automatica del rischio di sintomi depressivi.

2.1 Specifica PEAS

Di seguito è riportata la descrizione PEAS dell'ambiente operativo in forma tabellare.

Componente	**Descrizione**
Performance	La misura di performance del sistema si basa sulla capacità dell'agente di distinguere correttamente studenti inclini alla depressione e studenti non inclini, con particolare attenzione all'identificazione accurata della classe Depressione=True.
Environment	L'ambiente consiste negli studenti universitari, i quali completano un questionario psicologico per valutare il loro benessere emotivo.
Actuators	Gli attuatori consistono in un sistema di classificazione che assegna un'etichetta ("Rischio Depressione Rilevato" o "Nessun Rischio Rilevato") e segnala gli studenti identificati come a rischio con una panoramica chiara dei diversi report. Questi attuatori permettono di attivare interventi per supportare gli studenti.
Sensors	I sensori consistono nelle risposte al questionario psicologico fornito dagli studenti, che vengono analizzate dal sistema.

Tabella 1: Specifica PEAS dell'agente IGEA

2.2 Specifiche dell'ambiente

L'ambiente operativo in cui agisce il sistema *IGEA* è costituito dall'insieme dei dati relativi agli studenti universitari, provenienti da questionari autocompilati. Il modello di machine learning interagisce con tale ambiente analizzando i dati disponibili al fine di stimare la presenza o meno del rischio di disagio psicologico.

L'ambiente di IGEA può essere classificato come segue:

- **Parzialmente osservabile**: l'agente non ha accesso diretto allo stato psicologico reale dello studente, ma solo a informazioni indirette e parziali, spesso soggettive e rumorose, come risposte a questionari. Di conseguenza, lo stato dell'ambiente non è completamente osservabile.
- **Non deterministico**: la relazione tra i dati osservabili e il reale stato emotivo dello studente non è deterministica. A parità di input possono corrispondere stati psicologici differenti, a causa di fattori esterni non completamente modellabili e dell'elevata variabilità individuale.

- **Episodico**: ogni valutazione prodotta dal sistema è indipendente dalle precedenti.
  Il modello analizza le osservazioni raccolte in un determinato istante temporale senza mantenere uno stato interno che tenga traccia delle valutazioni passate. Ciascun episodio corrisponde pertanto a una singola compilazione del questionario, e la decisione non influenza né dipende da episodi futuri.
- **Statico**: durante l'elaborazione dei dati relativi a una singola compilazione del questionario, l'ambiente non cambia: le percezioni fornite all'agente e lo stato interno considerato restano invariati fino al termine dell'analisi. Eventuali variazioni nello stato emotivo dello studente o nei dati disponibili si verificano solo tra episodi distinti, e non influenzano il processo decisionale in corso.
- **Discreto**: poiché il sistema produce esclusivamente una classificazione binaria e opera su percezioni, stati e azioni discreti
- **Singolo agente**: il sistema IGEA opera come agente singolo e non interagisce direttamente con altri agenti intelligenti. Esso fornisce supporto decisionale a figure umane quali psicologi e servizi di supporto universitari, che rimangono responsabili degli interventi finali.

3 Individuazione dataset

Nella fase iniziale del progetto, sono state valutate due diverse strategie per l'acquisizione dei dati necessari all'addestramento del modello di Machine Learning:

1. **Creare** un dataset da zero, sottoponendo un questionario ad un campione di studenti;
2. **Cercare** sulla rete un dataset già formato, e adeguarlo alle nostre esigenze;

Dopo un'analisi comparativa, la prima opzione è stata scartata a causa di limitazioni metodologiche critiche. In primo luogo, la raccolta di un numero di campioni statisticamente significativo avrebbe richiesto tempi eccessivamente lunghi.

Tuttavia, l'ostacolo principale è rappresentato dalla **mancanza di un esperto di dominio**.
Per addestrare un modello di classificazione accurato, ogni record deve essere etichettato correttamente (es. "Depresso: Sì/No"). Senza la supervisione di un professionista in grado di valutare clinicamente le risposte dei candidati, il dataset prodotto sarebbe risultato privo di validità scientifica, rendendo il modello inaffidabile.

Di conseguenza, si è deciso di procedere con la seconda soluzione, individuando sulla piattaforma **Kaggle** il dataset: *"Student Depression Dataset"*

3.1 Punti di forza del dataset

La scelta è ricaduta su questo archivio per i seguenti motivi:

- **Varietà dei parametri:** Include fattori determinanti come la pressione accademica, la soddisfazione nello studio, le ore di sonno e la storia clinica familiare.
- **Ampiezza dei dati:** Le migliaia di osservazioni forniscono una base solida per l'addestramento, permettendo all'algoritmo di riconoscere pattern complessi con maggiore precisione.

3.2 Limiti del dataset

Tale dataset presenta i seguenti limiti:

- **Provenienza culturale**: Il dataset proviene da studenti indiani, e le esperienze psicologiche e comportamentali potrebbero differire da quelle degli studenti italiani, influenzando i risultati.
- **Fattori socio-culturali e accademici**: Sebbene il dataset descriva variabili psicologiche e comportamentali legate alla vita universitaria (che non sono specifiche di un singolo contesto nazionale), le dinamiche socio-culturali e accademiche in India potrebbero comunque influenzare la rilevazione del rischio di depressione.
- **Validità limitata**: Senza una validazione su dati italiani, il modello potrebbe non riflettere accuratamente il rischio di depressione tra gli studenti italiani.

Sebbene il dataset quindi sia stato raccolto su studenti indiani, le informazioni disponibili descrivono aspetti psicologici e comportamentali legati alla vita universitaria che non sono specifici di un singolo contesto nazionale. Il modello viene pertanto utilizzato come studio preliminare per valutare la fattibilità di un sistema predittivo del rischio di depressione, riconoscendo che una validazione su dati italiani reali sarebbe necessaria per un impiego operativo.

4 Data Understanding

Prima di procedere con le fasi di preparazione dei dati, è stata svolta una fase di *data understanding*, con l'obiettivo di comprendere la struttura del dataset, il significato delle variabili e la loro coerenza con il dominio applicativo del sistema.

In questa fase sono state analizzate tutte le colonne del dataset, identificandone il ruolo e il tipo di informazione rappresentata. In particolare, le variabili sono state suddivise nelle seguenti categorie:

- **Variabili demografiche**: *Gender*, *Age*;
- **Variabili psicologiche e autovalutative**: *Academic Pressure*, *Work Pressure*, *Study Satisfaction*, *Job Satisfaction*, *Have you ever had suicidal thoughts?*, *Family History of Mental Illness*, *Financial Stress*;
- **Variabili di contesto accademico e personale**: *Degree*, *Profession*, *Work/Study Hours*, *CGPA*, *Sleep Duration*, *Dietary Habits*, *City*.

È stata inoltre individuata la variabile target del problema di classificazione, identificata nella colonna *Depression*, che rappresenta se lo studente è a rischio di depressione.

Descrizione semantica delle variabili La Tabella 2 riporta una breve descrizione semantica delle variabili presenti nel dataset, al fine di chiarirne il significato e il ruolo nel contesto del problema affrontato.

Variabile	Descrizione semantica
Gender	Genere dichiarato dallo studente.
Age	Età dello studente espressa in anni.
Academic Pressure	Livello di pressione percepita legata alle attività accademiche.
Work Pressure	Livello di pressione percepita legata alle attività lavorative.
Study Satisfaction	Livello di soddisfazione dello studente rispetto allo studio.
Job Satisfaction	Livello di soddisfazione rispetto ad attività lavorative.
CGPA	Media accademica di tutti i voti ottenuti nel percorso di studi.
Work/Study Hours	Numero medio di ore dedicate quotidianamente a studio e lavoro.
Sleep Duration	Durata media del sonno giornaliero dichiarata.
Dietary Habits	Indicatore delle abitudini alimentari dichiarate.
Degree	Livello di istruzione dichiarato dallo studente.
Financial Stress	Livello di stress percepito legato alla situazione finanziaria.
Have you ever had suicidal thoughts?	Indicatore della presenza di pensieri suicidari auto-riferiti.
Family History of Mental Illness	Presenza di precedenti familiari di disturbi mentali.
Profession	Stato occupazionale dichiarato (nel dataset: Student).
City	Città di residenza dichiarata dallo studente.
Depression	Variabile target che indica la presenza o assenza di depressione.

Tabella 2: Descrizione semantica delle variabili del dataset

Questa fase di data understanding ha consentito di validare la coerenza del dataset rispetto all'obiettivo del sistema e di impostare correttamente l'analisi preliminare e le successive fasi di preparazione dei dati, senza introdurre modifiche al dataset originale.

4.1 Analisi preliminare del dataset

L'analisi preliminare del dataset per comprenderne la struttura e le principali caratteristiche statistiche si è concentrata su:

- controllo del bilanciamento dei dati.
- calcolo della dipendenza delle feature numeriche e categoriche dalla variabile target *Depression*;
- analisi delle distribuzioni delle feature numeriche e categoriche per identificare variabili sbilanciate o a bassa variabilità.
- identificazione di possibili data leakage.

Bilanciamento dei dati La Tabella 3 riporta la distribuzione delle osservazioni tra le due classi della variabile target *Depression*, da cui si osserva un moderato sbilanciamento a favore della classe positiva.

Tabella 3: Distribuzione delle classi della variabile target *Depression*

Classe	**Numero di elementi**	**Percentuale (%)**
Depressi (1)	16336	58.55
Non depressi (0)	11565	41.45
Totale	27901	100.00

Dipendenza dalla variabile target Per le feature numeriche, è stata calcolata la correlazione di Pearson con la variabile target. Per le feature categoriche, la dipendenza è stata stimata come la massima differenza tra le percentuali di studenti depressi e non depressi all'interno delle categorie di ciascuna variabile. In altre parole, indica quanto la presenza di una determinata categoria sia associata allo stato di depressione.

I risultati principali sono i seguenti:

Correlazione delle feature numeriche con la target *Depression*:

- Academic Pressure: 0.475
- Age: -0.226
- Work/Study Hours: 0.209
- Study Satisfaction: -0.168
- CGPA: 0.022
- Job Satisfaction: -0.003
- Work Pressure: -0.003
- id: 0.001

Correlazione delle feature categoriche con la target *Depression*:

- City: 1.000
- Profession: 1.000
- Financial Stress: 0.626
- Have you ever had suicidal thoughts ?: 0.581
- Degree: 0.415
- Dietary Habits: 0.415
- Sleep Duration: 0.290
- Family History of Mental Illness: 0.225
- Gender: 0.173

Analisi delle distribuzioni Per ciascuna feature, oltre a valutare la correlazione o la dipendenza dalla target, è stata analizzata la distribuzione dei valori:

- **Feature numeriche**: sono state calcolate media, deviazione standard, skewness e kurtosis. Feature con skew > 1, kurtosis > 5 o deviazione standard molto bassa (<0.01) sono state considerate *anomale*, poiché potrebbero non fornire informazioni significative al modello.
- **Feature categoriche**: sono state valutate la numerosità delle categorie e la distribuzione percentuale. Variabili con una sola categoria o con una categoria dominante ($>90\%$) sono considerate sbilanciate e poco informative.

Data leakage Al fine di verificare la presenza di potenziali fenomeni di data leakage, è stata analizzata separatamente la feature *Have you ever had suicidal thoughts ?* tramite una tabella di contingenza normalizzata rispetto alla variabile target *Depression*. I risultati mostrano una dipendenza estremamente marcata tra le due variabili, suggerendo che la feature possa fornire informazione quasi deterministica sulla classe target.

Tabella 4: Distribuzione percentuale della variabile *Depression* in funzione della feature *Have you ever had suicidal thoughts ?*

Have you ever had suicidal thoughts ?	**Depression = 0 (%)**	**Depression = 1 (%)**
No	76.8	23.2
Yes	21.0	79.0

5 Data preparation

Al fine di analizzare in modo più approfondito l'impatto delle scelte di preparazione dei dati sul comportamento e sulle prestazioni dei modelli di classificazione, sono state definite due pipeline alternative di preparation. L'obiettivo non è individuare una pipeline intrinsecamente migliore dell'altra, ma valutare come differenti strategie di selezione e trasformazione delle feature influenzino i risultati del sistema proposto.

Le due pipeline rappresentano approcci complementari alla preparazione del dataset.

La prima adotta una strategia basata sulla rimozione delle variabili potenzialmente ridondanti, poco informative o fortemente dipendenti dal contesto di origine del dataset. Questa pipeline consente di valutare le prestazioni dei modelli in una configurazione minimale, riducendo il rischio di bias e di dipendenza da specifiche assunzioni sul dominio.

La seconda pipeline, invece, mantiene alcune variabili di contesto accademico, introducendo trasformazioni controllate volte a migliorarne l'interpretabilità e la trasferibilità nel contesto universitario di riferimento. In particolare, tale pipeline consente di analizzare l'effetto di una rappresentazione più informata dei dati, pur senza introdurre assunzioni forti o modifiche che alterino il significato originale delle informazioni.

L'adozione di due pipeline permette quindi di:

- analizzare la sensibilità dei modelli alla presenza o assenza di specifiche feature;
- confrontare le prestazioni dei modelli a parità di algoritmo, isolando l'effetto della preparazione dei dati;
- ottenere una valutazione sperimentale più solida e meno dipendente da una singola configurazione del dataset.

Per garantire un confronto corretto e controllato, gli stessi algoritmi di classificazione sono stati applicati a entrambe le pipeline. In questo modo, è possibile attribuire eventuali differenze nei risultati esclusivamente alle diverse strategie di preparazione dei dati, mantenendo invariata la componente di apprendimento.

5.1 Data cleaning

La fase di *data cleaning* è finalizzata a rimuovere elementi potenzialmente dannosi per il processo di apprendimento del modello, con l'obiettivo di migliorare la qualità e l'affidabilità complessiva del dataset. Sono state eliminate la feature

- *id*: attributo identificativo degli studenti, irrilevante per la predizione del rischio di depressione
- *Have you ever had suicidal thoughts?*: considerata la natura semantica della feature, che rappresenta un sintomo clinico direttamente associato alla depressione e il rischio di performance artificialemnte gonfiate, la variabile è stata identificata come leaky predictor ed esclusa dalla fase di modellazione.

Valori nulli Dall'ispezione del dataset non sono emerse osservazioni con valori nulli, pertanto non si è resa necessaria alcuna operazione di imputazione o rimozione di record incompleti.

Osservazioni duplicate Anche in questo caso, non sono state individuate righe duplicate, confermando la coerenza e l'unicità delle istanze presenti nel dataset.

Outlier È stata inoltre condotta un'analisi dei valori anomali (*outlier*) sulle variabili numeriche mediante il metodo dell'*Interquartile Range* (IQR). Tale analisi ha evidenziato la presenza di 22 *outlier* nella variabile *Age*, riferibili a studenti con età significativamente superiore rispetto alla distribuzione centrale, 3 *outlier* nella variabile *Work Pressure*, 6 nella variabile *CGPA* (corrispondenti a valori pari a zero) e 2 nella variabile *Job Satisfaction*.

Per quanto riguarda le variabili categoriche, sono state considerate come valori anomali quelli con una frequenza inferiore all'1% del totale delle osservazioni. In base a questo criterio, sono stati individuati 31 *outlier* nella variabile *Profession*, 18 nella variabile *Sleep Duration*, 12 nella variabile *Dietary Habits* e 3 nella variabile *Financial Stress*.

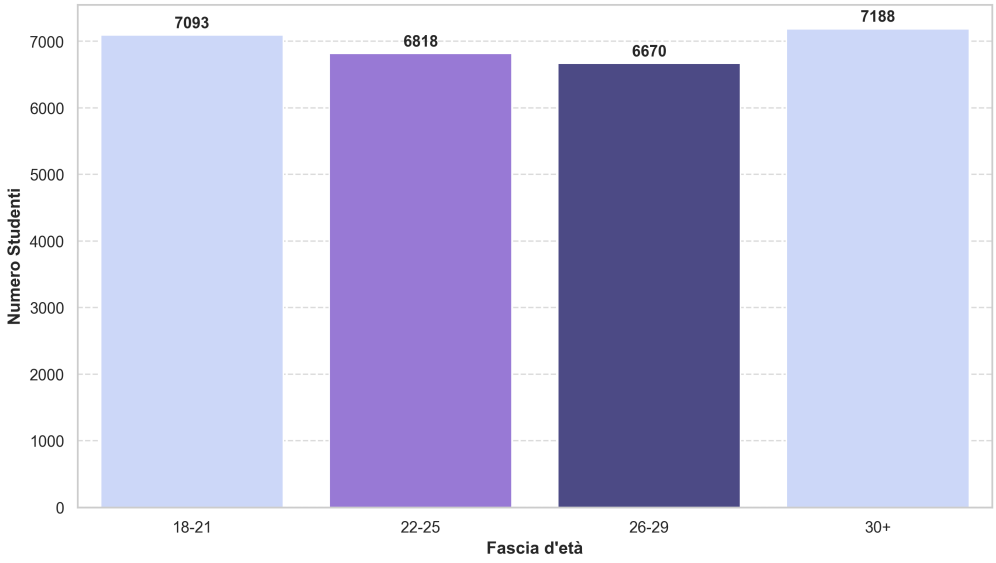
Inoltre, sono state eliminate 35 occorrenze con valore *Others* nella variabile categorica *Degree*, in quanto poco rappresentative e potenzialmente distorsive per l'analisi.

Complessivamente, il processo di identificazione e rimozione dei valori anomali ha comportato l'eliminazione di 132 osservazioni dal dataset, consentendo di ottenere un insieme di dati più coerente e rappresentativo ai fini delle successive analisi statistiche.

5.2 Pipeline 1

5.2.1 Feature engineering

Feature Trasformation La variabile continua *Age* è stata discretizzata nella nuova feature `Age_group` sulla base dell'analisi della sua distribuzione. Tale scelta è motivata dal fatto che la rappresentazione puramente numerica può amplificare il peso di osservazioni con età più elevate, introducendo effetti distorsivi nella stima del rischio. La categorizzazione dell'età consente invece di ridurre l'influenza di valori anomali, preservando al contempo l'informazione semantica associata alle diverse fasi della vita universitaria. Di seguito il grafico rappresenta il numero di osservazioni per ogni classe.



Feature Construction In questa fase, l'analisi si è spostata dalla semplice selezione delle variabili esistenti alla generazione di nuove feature sintetiche (*Derived Features*). L'obiettivo è stato quello di catturare relazioni non lineari e interazioni complesse tra le variabili, basandosi su ipotesi derivanti dalla conoscenza del dominio psicologico (*Domain Knowledge*).

Si è ipotizzato che il rischio di depressione non dipenda unicamente dalla somma lineare dei singoli fattori di stress, ma dalla loro interazione sinergica, aggravata da condizioni fisiologiche precarie come la privazione del sonno. Per modellare questo fenomeno, sono state costruite le seguenti variabili.

Calcolo del Debito di Sonno (Sleep Debt) La variabile originale `Sleep Duration` presenta una relazione inversa rispetto al rischio depressivo (una maggiore durata del sonno è generalmente protettiva). Per allineare la metrica alla direzione del rischio e penalizzare specificamente la carenza di riposo, è stata introdotta la variabile `Sleep_Debt`.

Il debito è calcolato rispetto alla soglia raccomandata di 8 ore giornaliere. La funzione applicata assicura che valori superiori alle 8 ore non generino un debito negativo, ma siano considerati pari a zero.

$$Sleep\_Debt = \max(0, 8 - SleepDuration) \tag{1}$$

Questa trasformazione permette al modello di quantificare linearmente la gravità della privazione del sonno, dove un valore più alto indica una condizione peggiore.

Modellazione dello Stress Amplificato (Stress Amplified) La principale innovazione introdotta nella pipeline è la variabile `Stress_Amplified`. Questa feature nasce dall'ipotesi che la stanchezza cronica agisca come un **moltiplicatore dello stress**: la capacità di resilienza psicologica di uno studente diminuisce drasticamente all'aumentare del debito di sonno.

Invece di considerare `Academic Pressure` e `Financial Stress` come addendi indipendenti, è stato modellato un termine di interazione in cui la pressione esterna totale viene amplificata dal fattore fisiologico del sonno:

$$Pressione\_Totale = AcademicPressure + FinancialStress \tag{2}$$

$$Stress\_Amplified = Pressione\_Totale \times (1 + Sleep\_Debt) \tag{3}$$

L'interpretazione della formula (2) è la seguente:

- Se lo studente è riposato ($Sleep\_Debt \approx 0$), lo stress percepito coincide con la somma delle pressioni reali.
- All'aumentare del debito di sonno, il termine ($1+Sleep\_Debt$) cresce, agendo come coefficiente moltiplicativo. Ad esempio, con 4 ore di debito, il carico di stress viene amplificato di un fattore 5.

Questa operazione di *Feature Engineering* ha permesso di sintetizzare informazioni eterogenee in un unico indicatore ad alto potenziale predittivo, migliorando la capacità del modello di identificare soggetti a rischio che presentano una combinazione critica di stress moderato e grave deprivazione del sonno.

Feature Selection Inizialmente sono state rimosse le feature dipendenti dal contesto di origine del dataset:

- *City*: rappresenta un'informazione geografica specifica del contesto indiano e quindi non generalizzabile agli studenti italiani.
- *Degree*: poichè non relativa al contesto italiano

Successivamente, sono state rimosse le feature numeriche e categoriche con distribuzioni anomale o bassa dipendenza dalla target:

- **Numeriche**: Work Pressure, CGPA, Job Satisfaction
- **Categoriali**: Profession

Sono state inoltre rimosse le feature utilizzate per la creazione di *Stress Amplified*:

- Academic Pressure, Financial Stress, Sleep Duration, Sleep Debt.

Al termine della fase di *feature selection*, il dataset risultante appare coerente, privo di incongruenze significative e adeguato per le successive fasi di trasformazione e addestramento del modello.

5.3 Pipeline 2

La pipeline 2 mantiene le info accademiche ma le rende più trasferibili/leggibili, evitando variabili non trasferibili o ridondanti.

5.3.1 Feature engineering

Feature Transformation - Age Analogalmente alla pipeline 1, la variabile *Age* è stata discretizzata nella variabile *Age_ group*.

Feature Transformation - Degree La variabile *Degree* descrive il livello di istruzione dichiarato dallo studente secondo il sistema educativo del paese di origine del dataset. Poiché tale rappresentazione risulta fortemente dipendente dal contesto accademico indiano e caratterizzata da un'elevata frammentazione delle categorie, è stata effettuata un'operazione di aggregazione in tre macro-categorie, basata esclusivamente sul livello di istruzione e non sul sistema universitario di riferimento.

In particolare, i valori originali sono stati ricondotti alle seguenti categorie:

- **Diploma**: corrispondente al livello di istruzione secondaria superiore (es. *Class 12*);
- **Titolo di primo livello**: comprendente i titoli di livello bachelor;
- **Titolo di secondo livello**: comprendente i titoli di livello master e post-graduate.

Questa trasformazione consente di ridurre la dipendenza dal contesto accademico specifico del dataset, mantenendo al contempo un'informazione rilevante relativa al livello di istruzione dello studente.

Feature Transformation - CGPA La variabile *CGPA* è espressa secondo una scala di valutazione specifica del sistema accademico di origine del dataset, nella quale valori più elevati corrispondono a prestazioni accademiche migliori. In particolare, secondo le specifiche del dataset, un valore di 10 rappresenta il punteggio massimo, mentre un valore di 5 corrisponde alla soglia minima di superamento.

Al fine di rendere i valori maggiormente interpretabili nel contesto universitario italiano, la variabile è stata convertita in una rappresentazione su scala trentesimale. La conversione è stata effettuata mediante una trasformazione lineare, definita a partire dai punti di riferimento della scala ($10 \rightarrow 30$ e $5 \rightarrow 18$). La formula applicata è la seguente:

$$CGPA_{30} = 2.4 \cdot CGPA + 6 \tag{4}$$

Tale trasformazione preserva l'ordine e le proprietà statistiche della variabile originale e non altera la sua capacità predittiva, ma ha esclusivamente una finalità descrittiva e interpretativa.

Nel dataset finale, la variabile originale *CGPA* è stata sostituita dalla nuova rappresentazione *CGPA_ 30*, al fine di evitare ridondanze informative.

Feature Selection Analogamente alla pipeline 1, è stata rimossa la feature **city**: fortemente dipendente dal contesto geografico indiano e quindi non generalizzabile ad altri contesti universitari;

Mentre sono state escluse tra le feature caratterizzate da una bassa dipendenza dalla variabile target e da distribuzioni anomale:

- **Numeriche**: Work Pressure, Job Satisfaction
- **Categoriali**: Profession

5.4 Feature Scaling

Per garantire che tutte le feature abbiano la stessa importanza, si è deciso (per entrambe le Pipeline) di normalizzare le feature numeriche con media 0 e deviazione standard 1 attraverso l'adozione del **Z-score scaling**. Ciò consente di trasformare i dati in una forma che sia uniforme e comparabile, evitando che le variabili con valori estremi o unità diverse abbiano un impatto sproporzionato. Le feature scalate sono:

Tabella 5: Confronto delle feature sottoposte a Scaling nelle due Pipeline

Pipeline 1	**Pipeline 2**
Study Satisfaction	*Study Satisfaction*
Work/Study Hours	*Work/Study Hours*
Stress_ Amplified	*Academic Pressure*
	Financial Stress
	CGPA_ 30

5.5 Encoding

Poiché la maggior parte dei modelli di machine learning opera su numeri si è deciso (per entrambe le Pipeline) di convertire le variabili categoriche in variabili binarie attraverso l'adozione del **One-Hot encoder**. Ciò permette al modello di catturare le differenze tra le categorie senza pregiudizi, trattando ognuna come indipendente e riducendo il rischio di interpretare erroneamente le relazioni tra categorie diverse. Le feature trasformate sono:

Tabella 6: Confronto delle feature sottoposte a Encoding nelle due Pipeline

Pipeline 1	**Pipeline 2**
Gender	*Gender*
Age_ group	*Age_ group*
Dietary Habits	*Dietary Habits*
Family History of Mental Illness	*Family History of Mental Illness*
	Sleep Duration
	Degree_ level

6 Modeling

6.1 Scelta degli Algoritmi

Sono stati selezionati due algoritmi di classificazione appartenenti a paradigmi differenti: la **Logistic Regression** e il **Decision Tree**. La scelta è motivata dalla necessità di bilanciare accuratezza predittiva e trasparenza nell'analisi dei fattori di rischio.

Logistic Regression La Logistic Regression rappresenta il modello *baseline* per eccellenza nei problemi di classificazione binaria. Le ragioni della sua inclusione nel progetto sono le seguenti:

- **Interpretabilità:** Attraverso l'analisi dei coefficienti β, il modello permette di quantificare l'impatto relativo di ogni feature sulla probabilità del target.
- **Natura Probabilistica:** Il modello restituisce un valore continuo nell'intervallo $[0, 1]$. Questo approccio è fondamentale in ambito psicologico per valutare il grado di confidenza della previsione.
- **Ottimizzazione:** Il modello beneficia della standardizzazione delle feature effettuata nella pipeline di modelling, garantendo una convergenza efficiente dell'algoritmo di ottimizzazione.

Decision Tree Il Decision Tree è un algoritmo di apprendimento supervisionato non parametrico che modella le decisioni attraverso una struttura gerarchica ad albero. A differenza della regressione logistica, non assume una relazione lineare tra le variabili.

L'inclusione di questo algoritmo è giustificata dai seguenti punti:

- **Interpretabilità visuale:** La struttura a nodi e rami permette di visualizzare chiaramente le regole decisionali che conducono a una classificazione. In ambito clinico, ciò facilita l'identificazione di soglie critiche nelle risposte ai questionari.
- **Cattura delle interazioni:** Il modello è in grado di identificare automaticamente interazioni complesse tra le variabili senza la necessità di specificarle manualmente in fase di feature engineering.
- **Assenza di assunzioni distribuzionali:** I Decision Tree non richiedono che i dati seguano una distribuzione specifica e risultano poco sensibili alla scala delle feature, offrendo una prospettiva complementare alla Logistic Regression.

6.2 Train-Test Split

Per valutare in modo robusto la capacità di generalizzazione dei modelli e ridurre il rischio di *overfitting*, è stata adottata la metodologia di **k-fold cross validation** con $k = 5$.

Questa tecnica prevede la suddivisione del dataset in k sottoinsiemi (fold) di dimensione approssimativamente uguale. Per ciascuna iterazione, un fold viene utilizzato come insieme di validazione, mentre i restanti $k - 1$ fold costituiscono il training set. Il processo viene ripetuto k volte, consentendo a ciascun sottoinsieme di essere utilizzato esattamente una volta come dati di validazione.

6.3 Addestramento

L'addestramento dei modelli è stato condotto adottando un approccio **pipeline-based**, al fine di garantire una corretta separazione tra i dati utilizzati per l'apprendimento e quelli impiegati per la validazione, evitando fenomeni di *data leakage*.

In particolare, le operazioni di preprocessing (standardizzazione delle feature e codifica delle variabili categoriche) sono state integrate all'interno di una pipeline di *scikit-learn* ed eseguite esclusivamente sui dati di training di ciascun fold durante la cross-validation. Questo approccio assicura che le trasformazioni apprese non siano influenzate dai dati di validazione, preservando la correttezza metodologica del processo di addestramento.

Per entrambi i modelli è stata inoltre adottata una strategia di **bilanciamento delle classi** attraverso l'impostazione del parametro `class_weight = balanced`, al fine di mitigare l'effetto dello sbilanciamento presente nella variabile target e migliorare la capacità dei modelli di individuare correttamente i casi positivi.

Nel caso del **Decision Tree**, l'addestramento è stato preceduto da una fase di **ottimizzazione degli iperparametri** mediante *Grid Search* (integrata nel processo di cross-validation), grazie al quale sono state esplorate diverse configurazioni. Gli iperparametri esplorati sono stati: profondità massima dell'albero, numero minimo di campioni richiesti per effettuare uno split, numero minimo di campioni richiesti per la creazione delle foglie e criterio di impurità adottato. Di seguito le configurazioni ottimali trovate per entrambe le pipeline che risultano essere motlo simili.

Pipeline	Criterion	Max Depth	Min Samples Leaf	Min Samples Split
Pipeline 1	gini	8	14	3
Pipeline 2	entropy	8	14	3

Tabella 7: Configurazioni ottimali degli iperparametri del Decision Tree ottenute tramite Grid Search

Per la **Logistic Regression**, il processo di addestramento è stato condotto fissando un numero contenuto di iterazioni dell'algoritmo di ottimizzazione. In particolare, sono state effettuate **15 iterazioni**, in quanto verifiche empiriche preliminari hanno mostrato la convergenza del modello e l'assenza di variazioni apprezzabili nelle prestazioni e nei coefficienti stimati al crescere del numero di iterazioni. Tale scelta permette di contenere il costo computazionale senza compromettere l'affidabilità del modello.

7 Conclusioni

L'analisi condotta attraverso le due pipeline di preparation e i due algoritmi di classificazione ha permesso di identificare i fattori più influenti nella previsione della depressione studentesca e di selezionare il modello più performante.

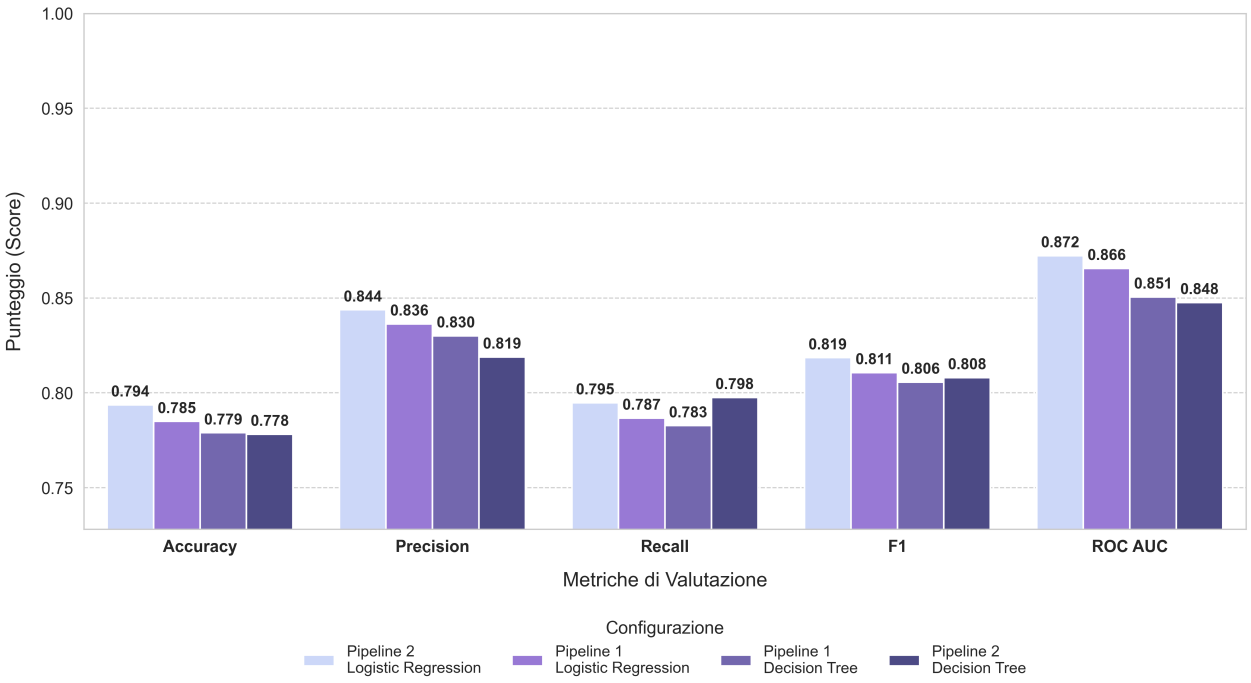
7.1 Metriche di valutazione

Le metriche selezionate per valutare la qualità dei classificatori sono state:

- **Recall:** Misura la capacità del modello di individuare tutti i casi reali di depressione. In ambito medico, è fondamentale massimizzare questa metrica per non ignorare studenti a rischio (pochi falsi negativi).
- **Precision:** Indica l'affidabilità delle previsioni positive. Una precision alta significa che quando il modello prevede "Depressione", è molto probabile che lo studente sia effettivamente depresso (pochi falsi positivi).

7.2 Valutazioni

L'analisi comparativa dei modelli, riassunta nel grafico di performance, evidenzia che la combinazione **Pipeline 2 + Logistic Regression** offre le prestazioni migliori in assoluto, raggiungendo una *Precision* del **84.38%** e una *Recall* di **79.48%**.



7.3 La nostra scelta: Logistic Regression (Pipeline 2)

Alla luce dei risultati ottenuti, il modello selezionato per la fase di deployment è la **Logistic Regression** addestrata sui dati processati dalla **Pipeline 2**. Questa combinazione ha garantito il miglior compromesso tra capacità di generalizzazione e affidabilità nelle predizioni positive.

7.3.1 Analisi dei Coefficienti Beta (β)

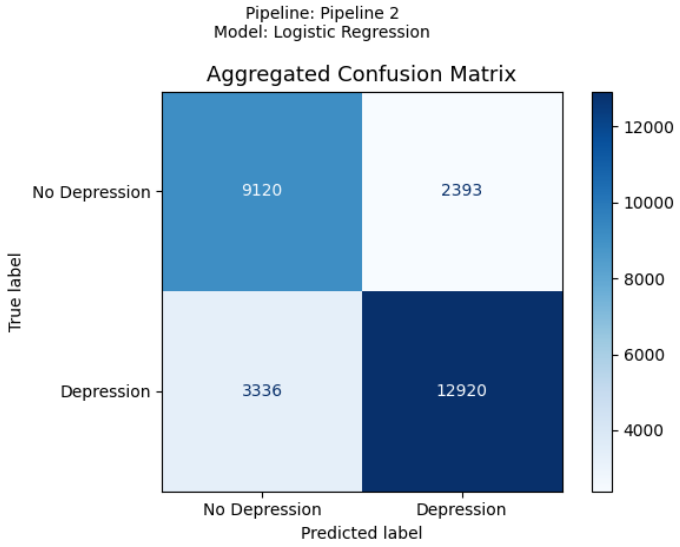
L'analisi dei pesi assegnati dal modello ci permette di comprendere quali variabili influenzano maggiormente la decisione. I coefficienti con valore assoluto più alto indicano i predittori più forti. Di seguito riportiamo le 5 feature più determinanti estratte dal modello finale:

Tabella 8: Top 5 Feature per impatto sulla predizione (Logistic Regression)

Feature	**Coeff. (β)**	**Odds Ratio**
Age_group (30+)	-1.29	0.28
Academic Pressure	+1.16	3.18
Dietary Habits (Unhealthy)	+1.10	3.01
Financial Stress	+0.81	2.26
Dietary Habits (Moderate)	+0.53	1.69

7.3.2 Matrice di Confusione

Per valutare gli errori commessi dal modello, riportiamo la Matrice di Confusione. Essa mostra la distribuzione tra predizioni corrette (sulla diagonale principale) ed errori di classificazione.

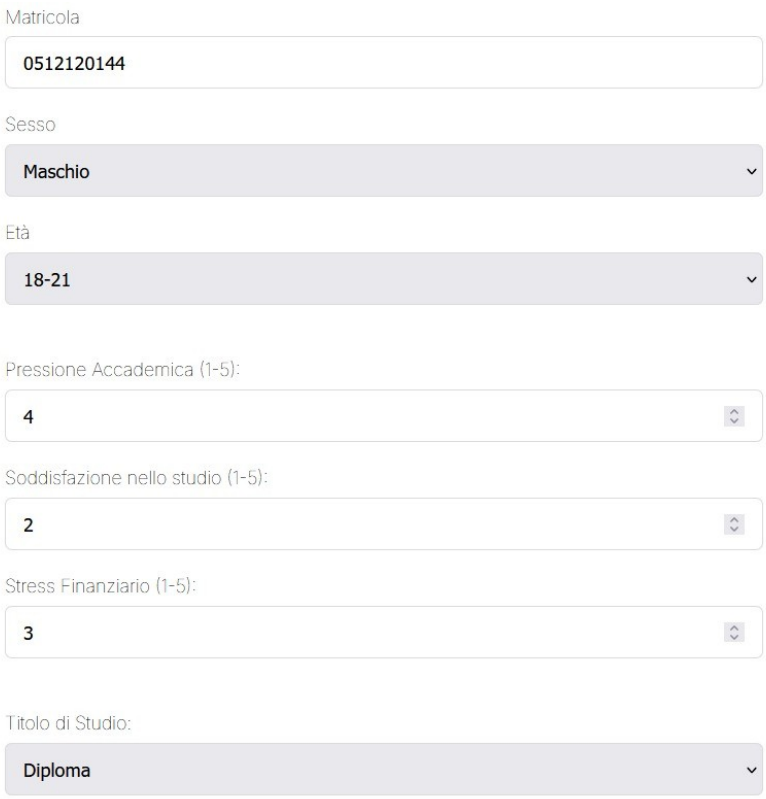


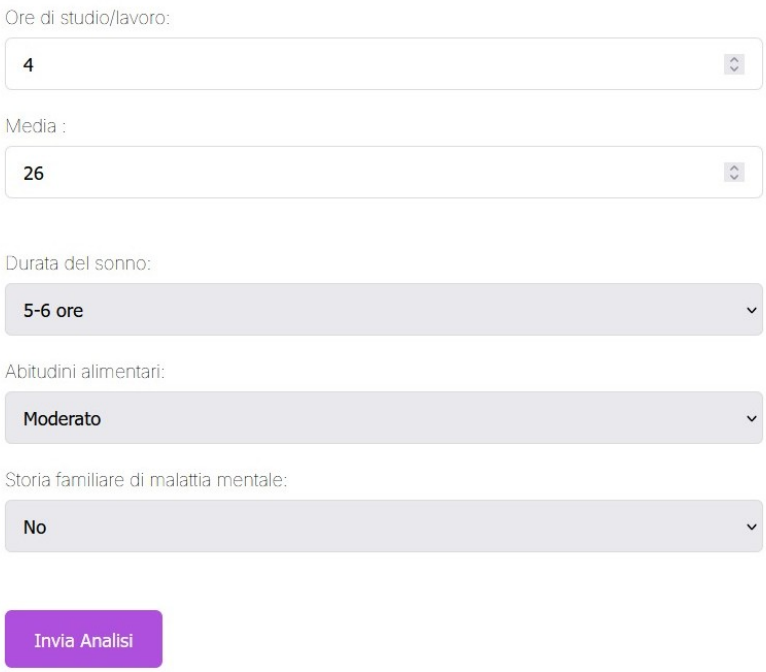
Il modello dimostra una solida capacità di distinguere le classi, minimizzando i Falsi Positivi (alta Precision), aspetto cruciale per evitare di segnalare erroneamente come a rischio studenti che non lo sono.

8 Interfaccia della Demo

Pagina Iniziale La pagina di benvenuto presenta un pulsante per accedere ad un questionario.

Il Questionario di Input L'interfaccia di inserimento dati consiste in un modulo da compilare. In seguito è presente un esempio di compilazione del questionario.





Visualizzazione del Risultato Una volta completata la compilazione e inviato il modulo, il sistema effettua una chiamata al backend dove risiede il modello di classificazione.
Il risultato dell'inferenza viene mostrato in una pagina di output dedicata che presenta:

1. **Esito della Classificazione**.
2. **Componente Visiva:** L'interfaccia utilizza una comunicazione visiva per rendere il responso facilmente interpretabile dall'utente finale.

In seguito è riportato il risultato relativo alla compilazione del form visto in precedenza.