MENTALLY STABILITY OF THE PERSON

Autori del progetto:

Giacomo Visciotti
Simone Verrengia
Giuseppe Pio del Vecchio
Liliana Gilca

1. Introduzione

Il progetto in oggetto si propone di sviluppare un modello predittivo per la classificazione del rischio di depressione tra gli adulti, utilizzando un dataset sintetico generato da un modello di deep learning, che è stato addestrato su un dataset originale denominato *Depression Survey/Dataset for Analysis*. L'obiettivo principale è costruire un sistema che possa individuare, attraverso variabili demografiche e comportamentali, i soggetti maggiormente a rischio di sviluppare una condizione depressiva.

Il dataset è stato acquisito dalla competizione *Mentally Stability of the Person* su Kaggle, il quale raccoglie informazioni anonime fornite da adulti di età compresa tra i 18 e i 60 anni. Le informazioni includono variabili relative a caratteristiche sociodemografiche e comportamentali, tra cui: età, genere, livello di istruzione, soddisfazione lavorativa o accademica, abitudini quotidiane e la presenza di una storia familiare di disturbi psichiatrici.

La variabile target del dataset è di tipo binario, dove "1" indica la presenza di un rischio di depressione e "0" l'assenza di tale rischio. L'approccio predittivo si concentra sull'analisi dei pattern nelle variabili indipendenti (come l'età, il livello di istruzione, e le abitudini di vita) per determinare una probabilità di rischio associata alla depressione, al fine di supportare la diagnosi precoce e l'intervento tempestivo in ambito clinico o sociale.

2. Pre-elaborazione dei Dati

Nel processo di pulizia del dataset sono state effettuate una serie di operazioni per preparare i dati all'analisi predittiva e migliorare la qualità del modello. La prima operazione cruciale riguarda la mappatura della durata del sonno. Molte delle variabili nel dataset contengono informazioni testuali sulle ore di sonno, come "6-7 hours" o "less than 5 hours", che possono risultare difficili da trattare per un modello di machine learning. Per risolvere questo problema, è stata implementata una funzione che trasforma queste informazioni testuali in valori numerici, consentendo una più facile manipolazione dei dati e prevenendo errori legati a formati non omogenei o imprecisioni nelle risposte.

Per quanto riguarda la gestione dei dati mancanti per le variabili legate alla professione, i valori mancanti sono stati imputati tenendo conto del ruolo dell'individuo, "Studente" o "Professionista". All'interno di questa categorizzazione, per i professionisti sono stati impostati a 0 i valori relativi alle colonne accademiche e per i studenti sono stati impostati a 0 i valori nelle colonne lavorative. Per le

variabili numeriche, è stata utilizzata la mediana per sostituire i valori mancanti, in quanto questa misura è meno influenzata da eventuali outliers rispetto alla media. Le variabili categoriche sono state imputate utilizzando la moda all'interno di ciascun sottogruppo pertinente, per garantire che la sostituzione mantenesse una coerenza logica con i dati disponibili.

La codifica delle variabili categoriche è un altro passo fondamentale. In questo caso, le variabili categoriche come il sesso sono state trasformate in variabili numeriche, con "Male" mappato a 1 e "Female" a 0. Analogamente è stato effettuato per le variabili professione lavorativa e le abitudini alimentari, sono state mappate su categorie numeriche generali per facilitare l'input nel modello di machine learning. Ad esempio, la professione è stata codificata in categorie come "IT/Tech", "Marketing". Le abitudini alimentari sono state classificate come "Unhealthy", "Moderate", e "Healthy".

Nel dataset, la variabile City è stata trasformata in una nuova variabile Region, che raggruppa le città in aree geografiche più ampie. Le città sono state mappate a categorie regionali specifiche, come ad esempio: “North-East India”, “North-West India”, “West-Gujarat”, “West-Maharashtra”, “Central India”, “East India”,”South India”. In seguito, la variabile Region viene sostituita dalla variabile Region_Encoded applicando il Label Encoder.

In modo analogo viene applicato la stessa struttura logica alle variabili Profession e Degree, ovvero vengono sostituite con le corrispettive variabili, Professional_Group_Encoded e Degree_Group_Encoded, applicando il LabelEncoder.

Un altro passo significativo riguarda la rimozione delle variabili problematiche. Per ottimizzare la qualità del modello predittivo, è stata eseguita una valutazione delle variabili attraverso il calcolo del Variance Inflation Factor (VIF). Questo processo ha permesso di identificare eventuali variabili con elevata multicollinearità. Tuttavia, l'analisi non ha evidenziato la presenza di variabili con problematiche significative di multicollinearità, indicando che tutte le variabili selezionate contribuiscono in modo distintivo al modello senza causare interferenze tra di loro.

Una volta completate tutte le operazioni di pulizia e trasformazione, le variabili non necessarie come ID e Name sono state eliminate per ricavare un file CSV pronto per essere utilizzato in analisi successive e per l'addestramento del modello predittivo.

3. Esplorazione del Dataset e Rappresentazioni Grafiche delle Variabili

L’analisi esplorativa del dataset ha come l'obiettivo di esaminare le distribuzioni tra le principali variabili socio-demografiche e comportamentali, al fine di preparare i dati per il modello predittivo.

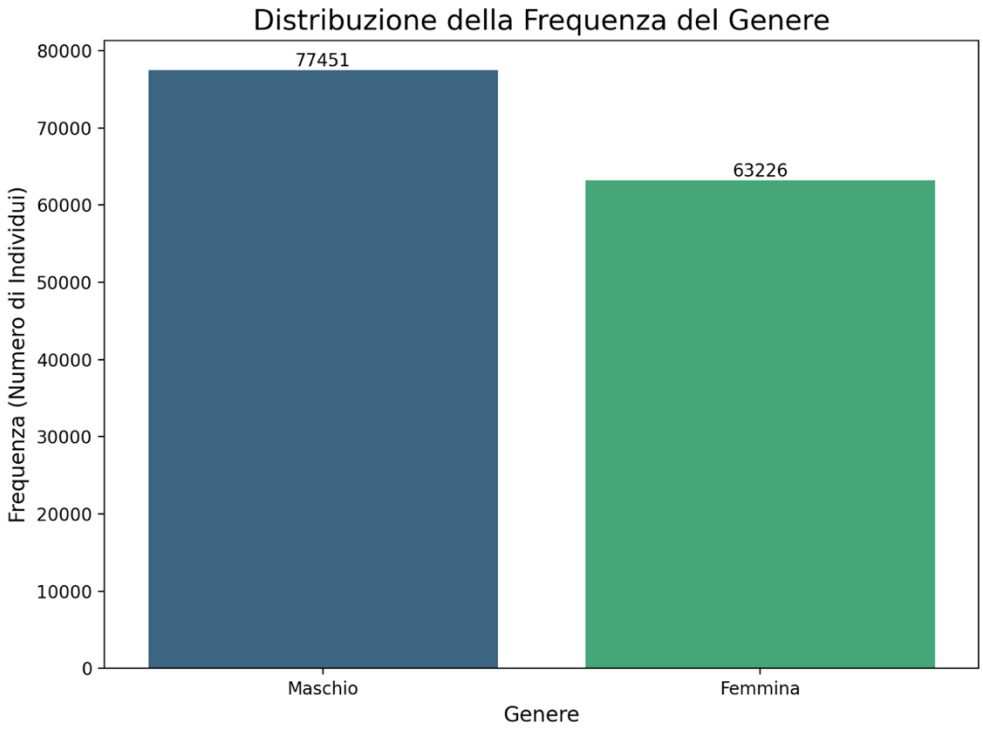

Figura1: Distribuzione della Frequenza del Genere

La distribuzione della frequenza del genere offre una prima panoramica del dataset, evidenziando un moderato sbilanciamento tra i due gruppi. In particolare, gli individui di sesso maschile sono presenti in misura leggermente maggiore rispetto a quelli di sesso femminile.

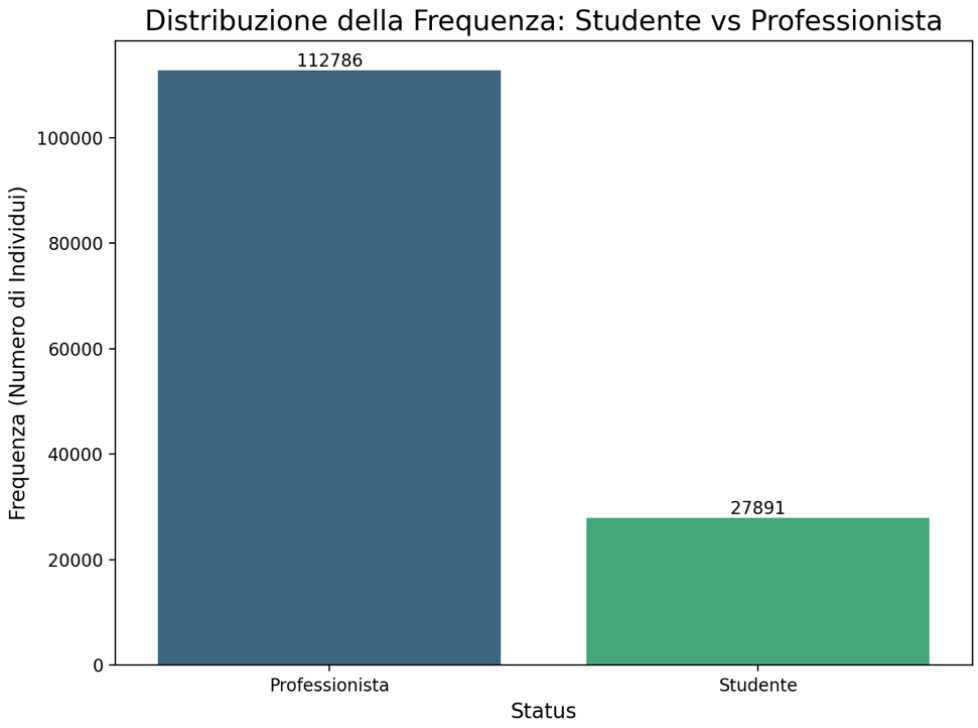

Figura 2: Distribuzione della Frequenza Studenti e Professionista

La distribuzione della frequenza tra studenti e professionisti mostra uno squilibrio evidente nella composizione del campione in relazione allo status professionale. Un numero significativamente maggiore di rispondenti si identifica come professionista rispetto agli studenti. Questo sbilanciamento potrebbe riflettere una rappresentazione più alta di individui con uno sfondo lavorativo stabile rispetto a coloro che sono ancora in formazione accademica. La differenza riflette il fatto che i professionisti potrebbero avere esperienze e stili di vita differenti rispetto agli studenti, il che potrebbe comportare differenze nei fattori di rischio o nei comportamenti legati alla depressione.

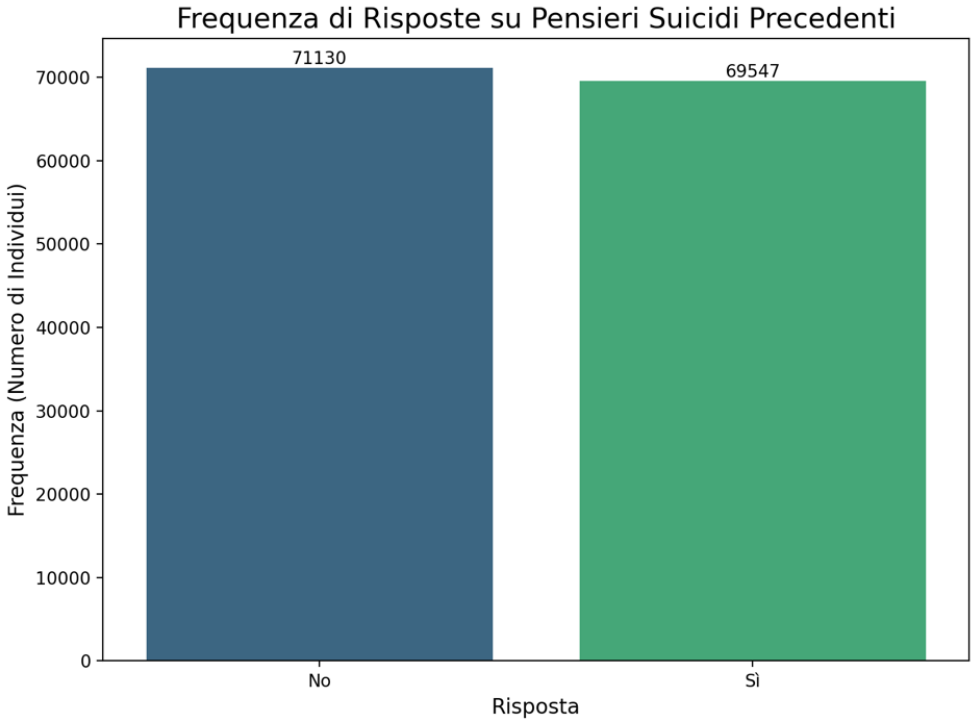

Figura 3: Distribuzione della Frequenza di Risposte su Pensieri Suicidi

Il grafico nella *figura 3* che rappresenta la frequenza delle risposte alla domanda relativa a pensieri suicidi precedenti è distribuito in modo equilibrato e ha un potenziale significativo come indicatore del rischio di depressione.

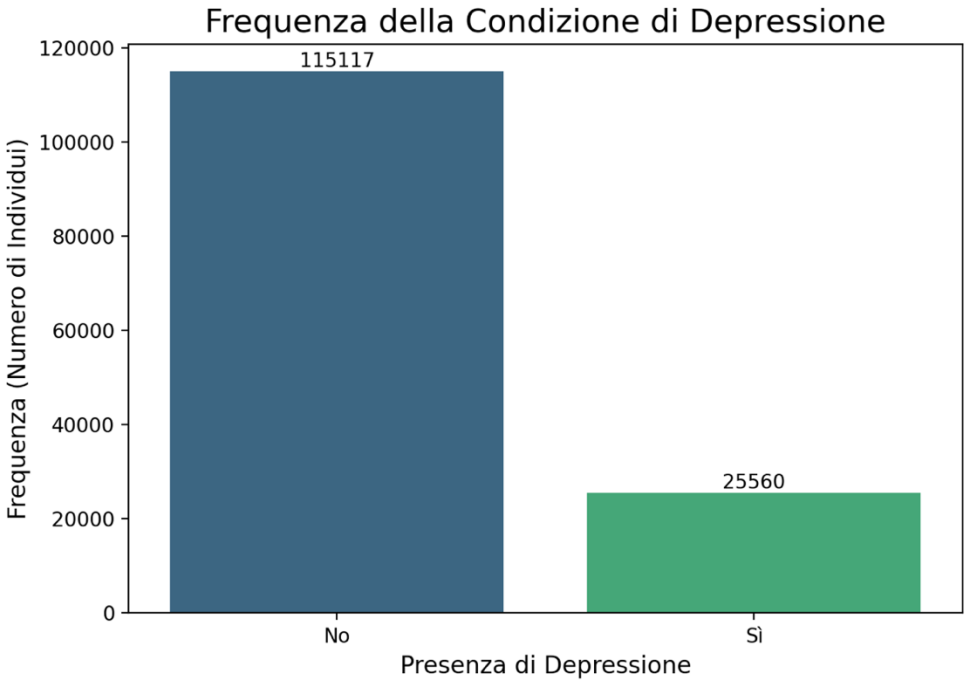

Figura 4: Distribuzione della Frequenza della variabile Target Depressione

La variabile target è fortemente sbilanciata, con una netta predominanza di non depressi. Questo sbilanciamento potrà influenzare la performance dei modelli di classificazione, favorendo la classe predominante.

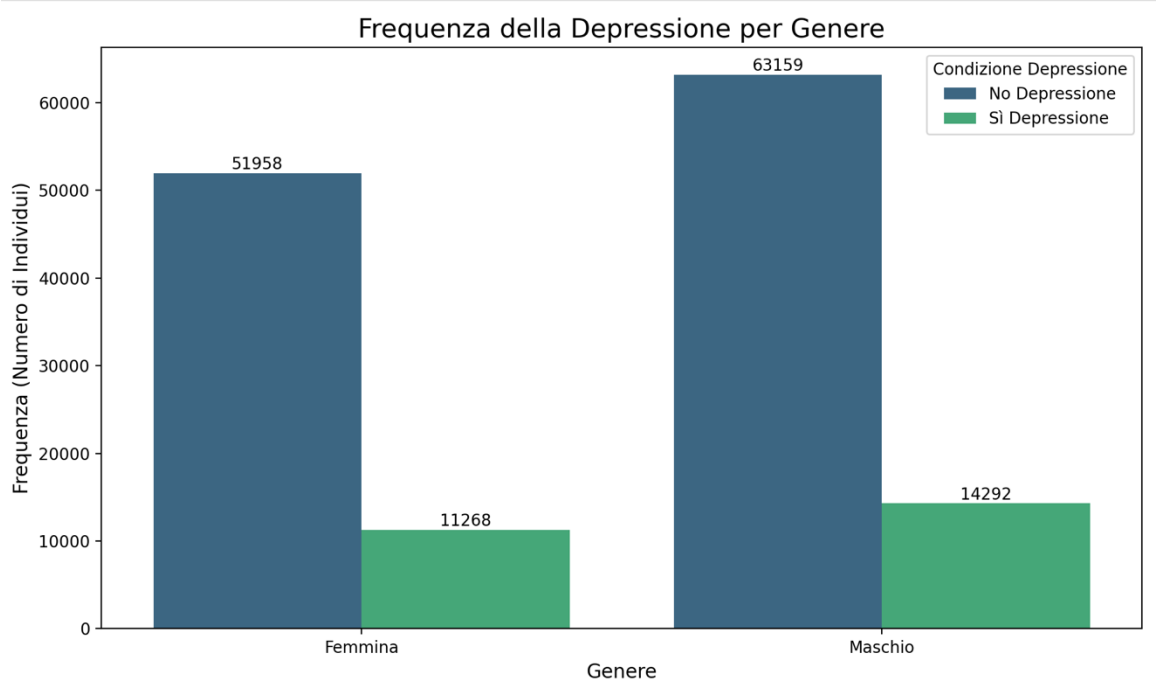

Figura 5: Distribuzione della Frequenza della Depressione per Genere

La *figura 5* evidenzia come la proporzione tra individui con e senza depressione appare abbastanza simile nei due gruppi, suggerendo che il genere potrebbe non essere un fattore discriminante primario per la condizione depressiva nel dataset, almeno in termini di frequenza relativa.

Depression Gender	0	1
0 M	82.18%	17.82%
1 F	81.54%	18.45%

Tabella 1: Frequenze Relative della Depressione per Genere

Dalla seguente *tabella 1* è evidente che le differenze tra i generi sono minime, con una leggera prevalenza di casi di depressione tra i maschi.

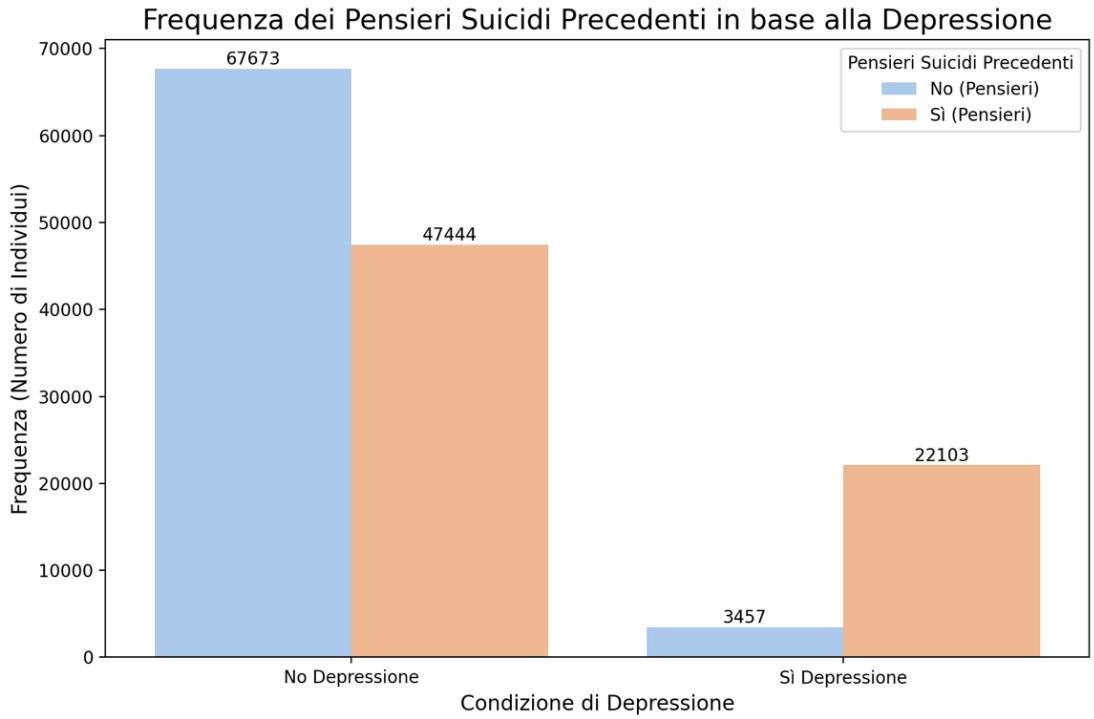

Figura 6: Distribuzione della frequenza dei pensieri suicidi in base alla depressione

Il grafico evidenzia una forte associazione tra la presenza di pensieri suicidi precedenti e la condizione di depressione. Tra gli individui con depressione, la maggior parte ha dichiarato di aver avuto pensieri suicidi, mentre solo una piccola parte non li ha avuti. Al contrario, tra coloro che non presentano depressione, sebbene prevalgano gli individui senza pensieri suicidi, una quota significativa ha comunque riferito di averne avuti.

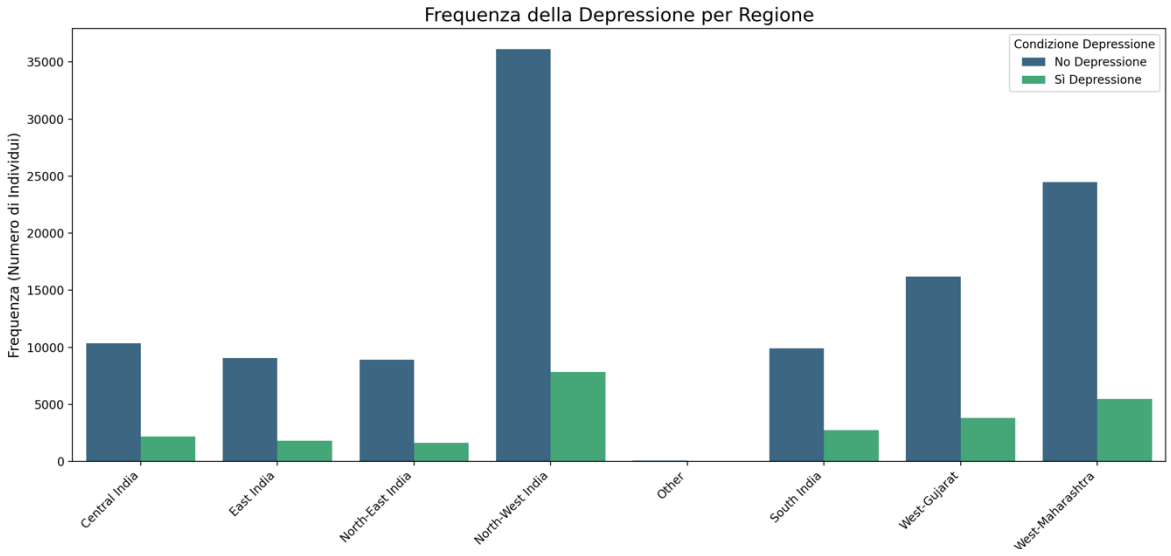

Figura 7: Distribuzione della frequenza del target rispetto alle varie regioni

Il grafico mostra la distribuzione della condizione di depressione suddivisa per regione geografica. La North-West India e la West-Maharashtra presentano il maggior numero di individui, con un'elevata incidenza sia di casi con che senza depressione. La regione North-West India mostra il numero più alto di persone con depressione in valori assoluti. Al contrario, regioni come North-East India, East India e Central India hanno numeri inferiori sia per i casi depressivi che non.

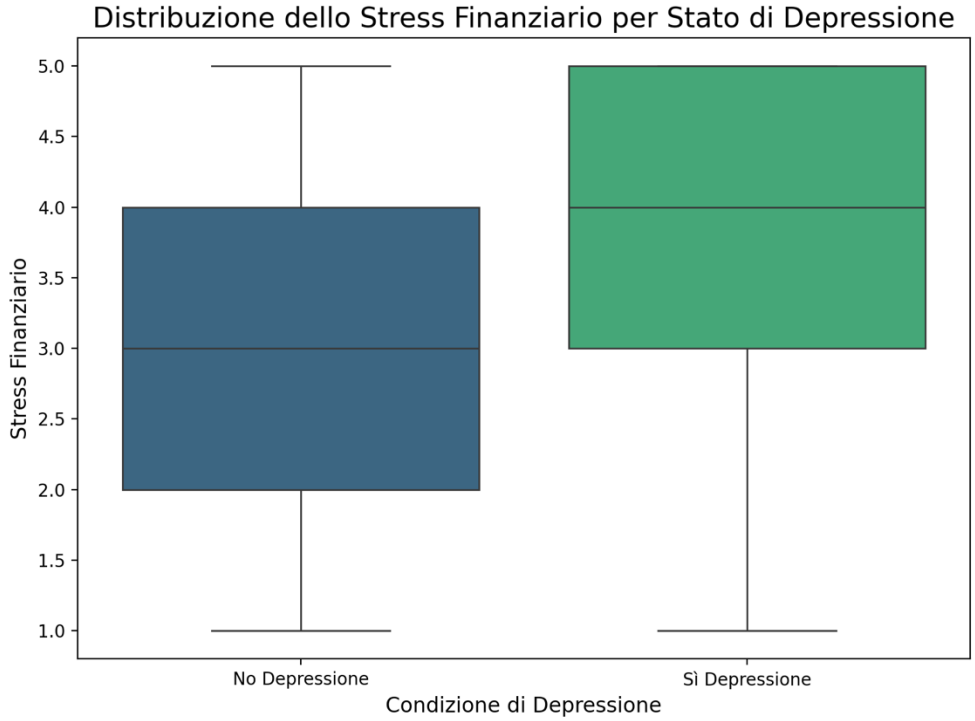

Figura 8: Distribuzione dello stress finanziario per stato di depressione

Il grafico nella *figura 8* mostra la distribuzione dello stress finanziario in relazione alla condizione di depressione. L'analisi visiva rivela che le persone con depressione tendono ad avere livelli di stress finanziario più elevati. La mediana dello stress finanziario è più alta tra chi è depresso (4) rispetto a chi non lo è (3). Inoltre, l'intervallo interquartile (IQR) per il gruppo con depressione si sposta verso l'alto, suggerendo una tendenza generale a livelli di stress più alti tra questi individui.

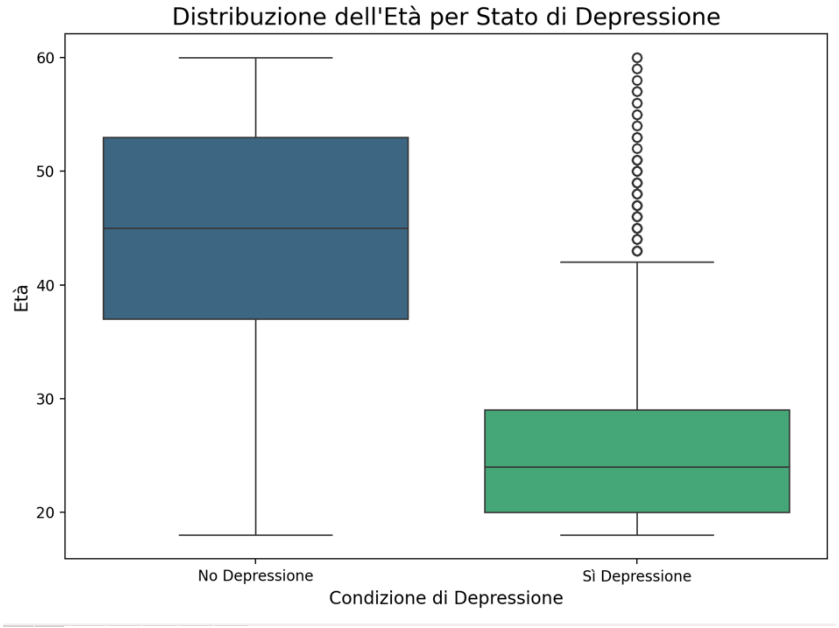

Figura 9: Distribuzione dell'età per stato di depressione

Il grafico rappresenta la distribuzione dell'età in base alla condizione di depressione, è evidente una differenza marcata tra i due gruppi: le persone senza depressione tendono a essere significativamente più anziane con una mediana intorno ai 45 anni, mentre quelle con depressione hanno una mediana molto più bassa intorno ai 25 anni. Inoltre, la distribuzione dei depressi mostra una maggiore presenza di valori anomali verso l'alto, indicando che i casi di depressione tra gli individui più anziani sono meno frequenti, ma non assenti. Tale fenomeno viene confermato dal grafico nella *figura 10*, ovvero che la prevalenza di depressione è maggiore tra gli studenti rispetto ai professionisti, suggerendo una possibile associazione tra lo status e la condizione depressiva.

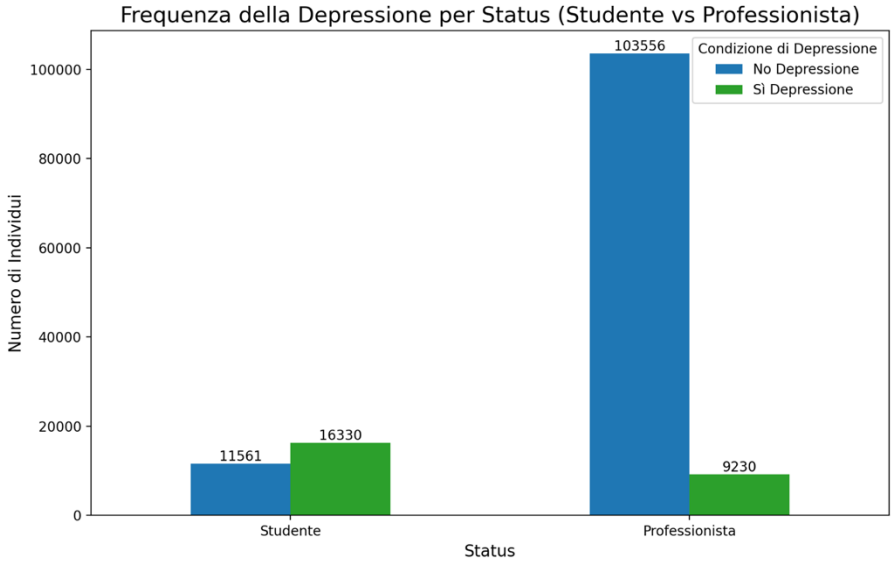

Figura 10: Distribuzione della depressione per status

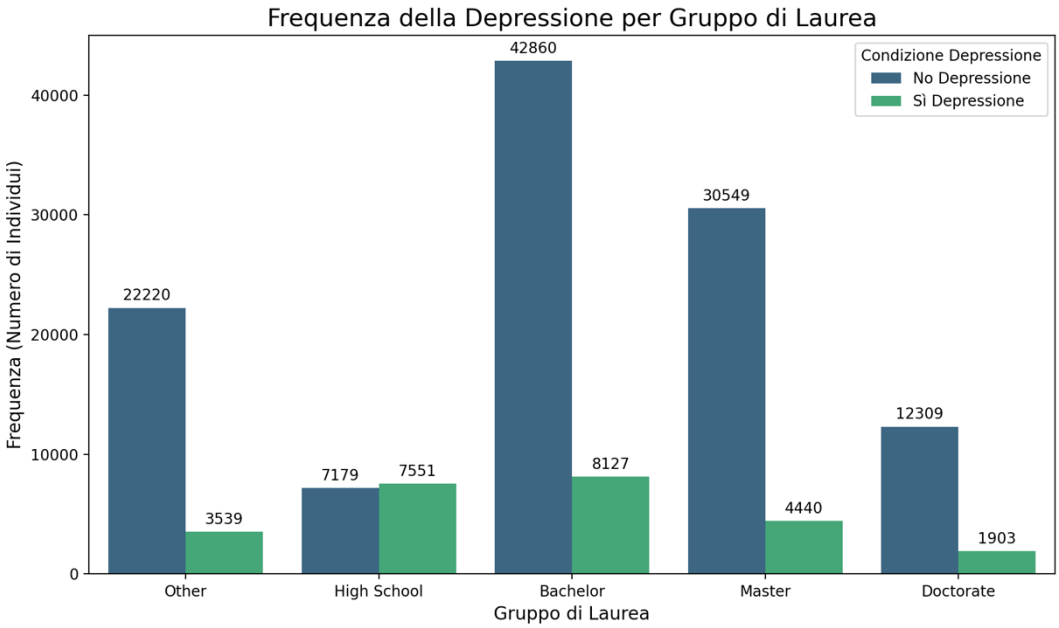

Figura 11: Distribuzione della depressione per grado di istruzione

Il grafico della distribuzione della depressione in base al livello di istruzione è evidente come nei livelli di istruzione più alti (Bachelor, Master, Doctorate), i casi di depressione sono inferiori rispetto a quelli senza depressione. Al contrario, nel gruppo "High School", la frequenza di casi depressivi sono leggermente superiore a quella dei non depressi, indicando una maggiore vulnerabilità.

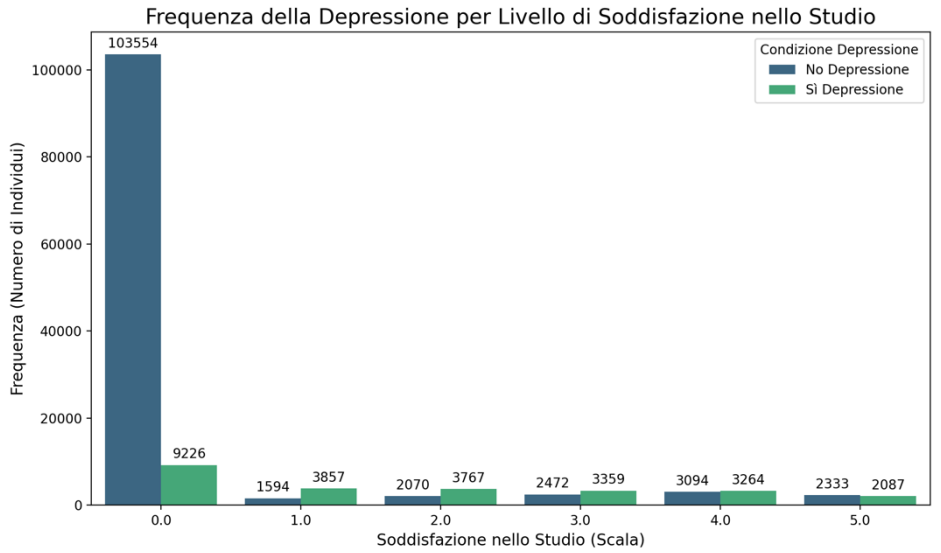

Figura 12: Distribuzione della depressione per soddisfazione nel grado di istruzione

È evidente che la maggioranza degli individui con depressione si concentra nel livello più basso di soddisfazione (0), man mano che il livello di soddisfazione aumenta il numero di casi di depressione diminuisce drasticamente.

4. Addestramento Modello

I dati sono stati suddivisi in due set: il set di addestramento, che rappresenta l'80% del dataset totale, e il set di test, che costituisce il restante 20%. Questa divisione è fondamentale per garantire che il modello venga addestrato su un campione di dati sufficientemente ampio e testato su un altro set di dati per valutarne l'efficacia. È emerso che la distribuzione delle classi nel dataset presenta uno squilibrio significativo tra i soggetti depressi e quelli non depressi. Questo squilibrio potrebbe influenzare negativamente le performance del modello, portando a una maggiore predizione della classe maggioritaria. Perciò, è stato necessario prendere in considerazione questo aspetto durante la fase di costruzione e addestramento del modello, cercando di bilanciare l'impatto delle due classi sulla predizione.

Sono stati testati diversi modelli di machine learning:

- Logistic Regression con class_weight='balanced': Questa variante della regressione logistica permette di adattare automaticamente i pesi delle classi in base alla loro frequenza nel dataset. L'obiettivo di questa tecnica è migliorare la predizione della classe minoritaria, riducendo il bias verso la classe maggioritaria.
- XGBoost utilizzato per confrontare la sua capacità di gestire i dati sbilanciati senza l'uso di tecniche specifiche di bilanciamento.
- XGBoost con SMOTE (Synthetic Minority Over-sampling Technique): per affrontare direttamente il problema dello squilibrio tra le classi è stato utilizzato una tecnica di oversampling che genera nuovi esempi sintetici della classe minoritaria (depressi) a partire dalle osservazioni esistenti. Questa tecnica aiuta a bilanciare il dataset, permettendo al modello di imparare meglio le caratteristiche della classe minoritaria.

Per ottimizzare i parametri del modello XGBoost, è stato utilizzato GridSearchCV per esplorare diverse combinazioni di parametri come il numero di stimatori, la profondità massima degli alberi, il tasso di apprendimento e i pesi delle classi.

In seguito, vengono riportati i risultati delle analisi introdotte precedentemente. Le metriche utilizzate sono le seguenti:

- Accuracy: Indica la percentuale di previsioni corrette rispetto al totale delle previsioni effettuate dal modello. È una misura generale di quanto il modello sia preciso.
- Precision: Misura quanto il modello è preciso quando prevede la classe positiva. In altre parole, indica quante delle previsioni positive del modello sono effettivamente corrette.
- Recall: Misura quanto il modello è in grado di identificare correttamente i veri positivi. Indica la capacità del modello di non perdere casi positivi.

- F1 Score: È una media bilanciata tra precisione e recall. Viene utilizzato quando è importante avere un buon equilibrio tra le due misure, specialmente in situazioni con classi sbilanciate.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8933	0.646	0.9126	0.7565
XGBoost	0.8977	0.6607	0.8989	0.7616
XGBoost (Optimized)	0.9186	0.7558	0.8159	0.7847

Tabella 2: Metriche di Performance tra i Modelli

L'analisi dei risultati evidenzia che il modello XGBoost ottimizzato presenta le migliori performance in tutte le metriche rispetto agli altri modelli testati. In particolare, l'accuratezza del modello ottimizzato supera sia la Logistic Regression che il XGBoost senza ottimizzazione, dimostrando una maggiore capacità di fare previsioni corrette. La precisione del modello ottimizzato è significativamente migliore rispetto agli altri modelli, indicando una maggiore capacità di predire correttamente la classe positiva.

Il modello XGBoost ottimizzato mostra anche una buona capacità di identificare i veri positivi, anche se la Logistic Regression mostra un valore leggermente superiore per il recall, suggerendo una maggiore sensibilità in alcuni casi. Tuttavia, l'F1 Score del modello ottimizzato risulta il migliore, indicando un buon equilibrio tra precisione e recall, che è fondamentale in contesti con classi sbilanciate.

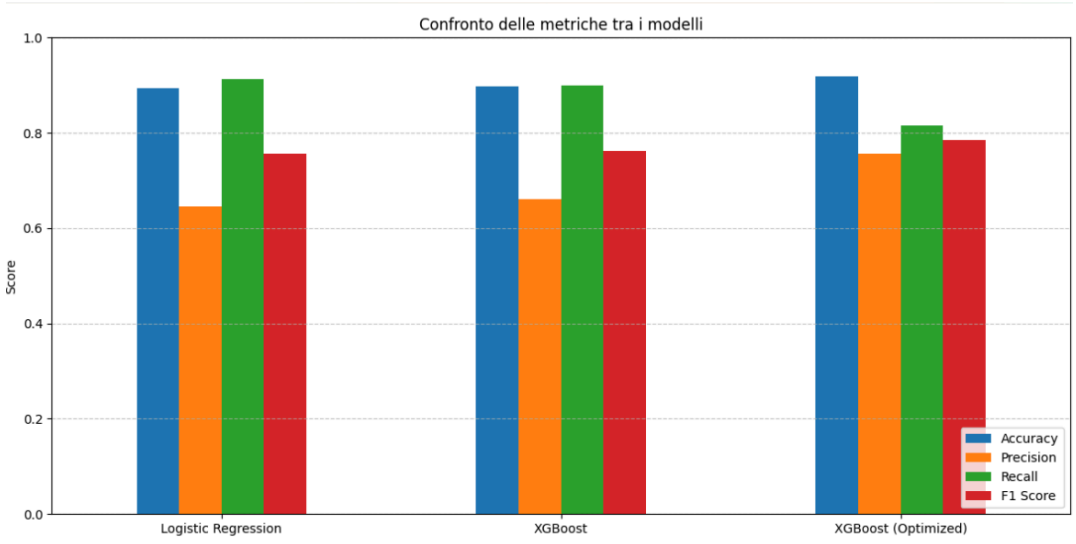

Figura 13: Confronto delle Metriche di Performance tra i Modelli

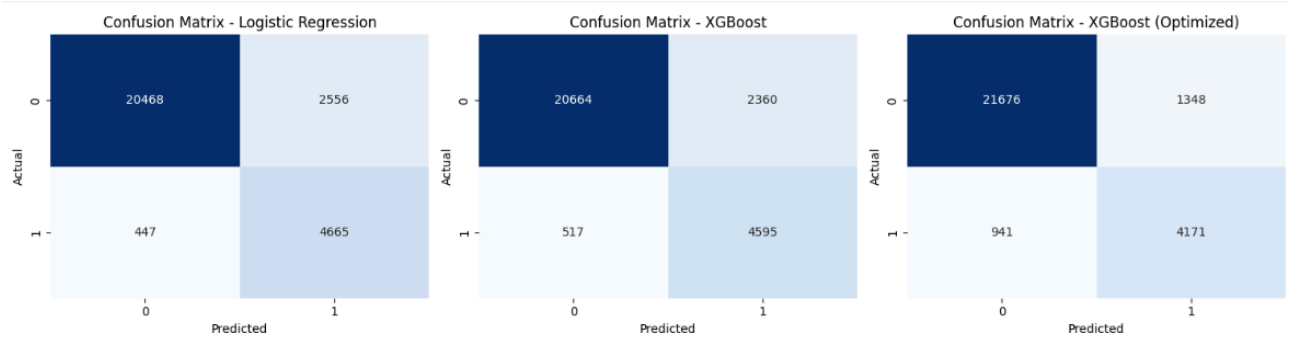

Figura 14: Matrice di confusione per i Modelli

Le matrici di confusione confermano i risultati precedentemente discussi, con il modello ottimizzato che dimostra una buona capacità di distinguere tra le due classi, con un numero maggiore di predizioni corrette rispetto agli altri modelli. In generale, l'XGBoost ottimizzato risulta il modello più robusto e performante per il problema in esame.