Analisi e Predizione della

Employee Retention

Modelli di machine learning per il turnover aziendale

Raffaele Manzo | ***05121 16484***

Introduzione

- Problema: Il turnover dei dipendenti è un problema costoso per le aziende.
- Obiettivo: costruire un modello predittivo in grado di determinare se un dipendente dovrebbe *rimanere* nell'organizzazione o *lasciarla*.

Colonna	Descrizione
Education	Qualifiche educative, inclusi titolo di studio, istituzione e campo di studio
JoiningYear	Anno di assunzione, indica la durata del servizio
City	Città di base o lavoro del dipendente
PaymentTier	Classificazione in diverse fasce salariali
Age	Età del dipendente, fornisce informazioni demografiche
Gender	Identità di genere, utile per analisi di diversità
EverBenched	Indica se il dipendente è stato temporaneamente senza lavoro assegnato
ExperienceInCurrentDomain	Anni di esperienza nel dominio attuale
LeaveOrNot	Variabile target: 0 (rimane), 1 (lascia)

Caratteristiche del Dataset

- Righe: **5653**
- Colonne: **9**

- Variabili principali:
  - **Età** (Numerica)
  - **Genere** (Categoriale)
  - **Istruzione** (Categoriale)
  - **Anni di Esperienza** (Numerica)
  - **Fascia Salariale** (Numerica)
  - **Anno di assunzione** (Numerica)
  - **Cittá** (Categoriale)
  - **Temporaneamente inutilizzato** (Numerica)

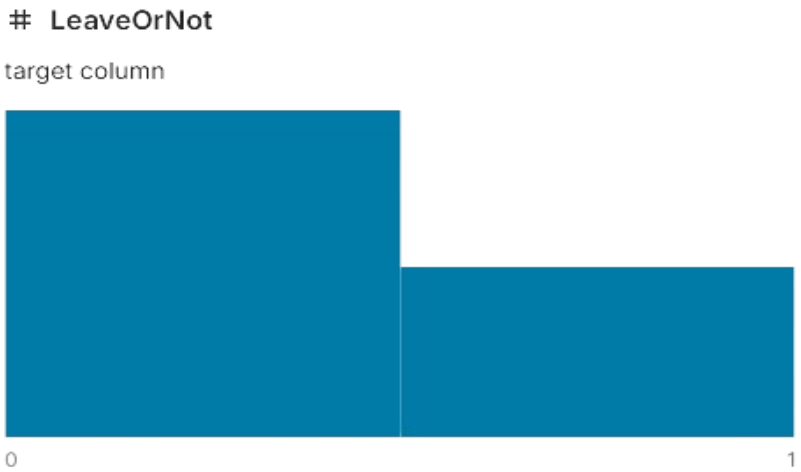
https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset/

Distribuzione della variabile target

Classe 0 (Rimane): 65%

Classe 1 (Lascia): 35%

Dataset sbilanciato: necessità di bilanciamento delle classi.



Preparazione dei dati

Trattamento dei valori mancanti:

Riempimento con Media

Riempimento con Moda

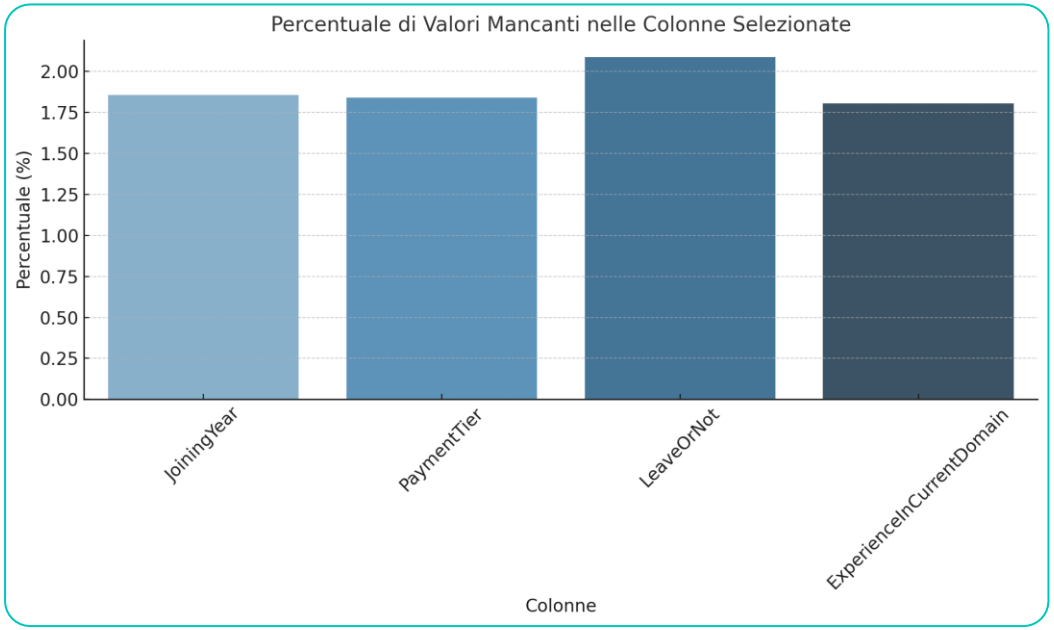
Eliminazione righe critiche

Encoding delle variabili categoriali:

One-Hot Encoding

Label Encoding

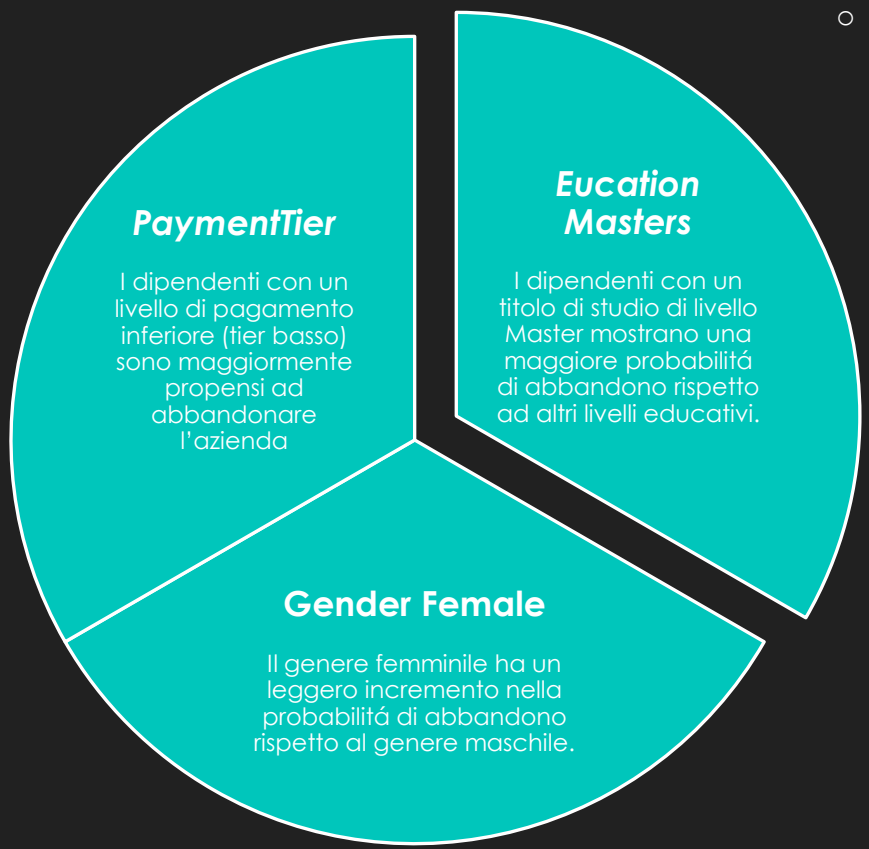
Distribuzione dei valori mancanti:



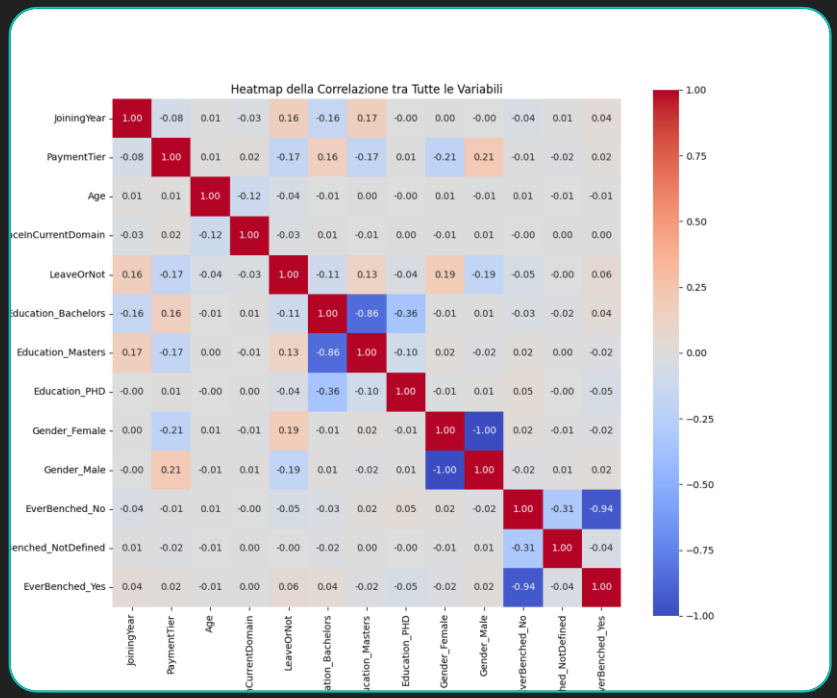
- Education: 108 valori mancanti.
- JoiningYear: 105 valori mancanti.
- City: 115 valori mancanti.
- PaymentTier: 104 valori mancanti.
- Age: 95 valori mancanti.
- Gender: 114 valori mancanti.
- EverBenched: 99 valori mancanti.
- ExperienceInCurrentDomain: 102 valori mancanti.
- LeaveOrNot: 118 valori mancanti.

Sono state selezionate solo le 4 colonne piú significative del dataset

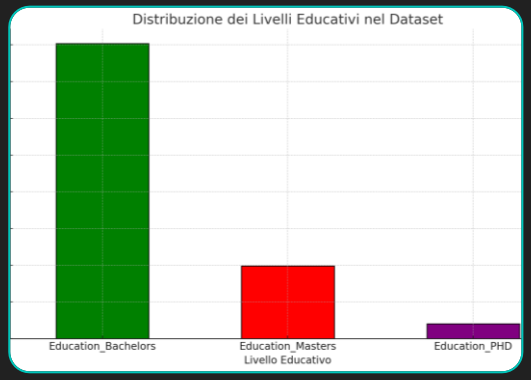
Visualizzazione delle correlazioni:

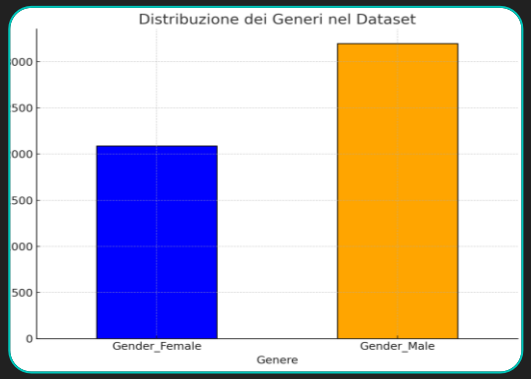


- Dopo aver eseguito l'encoding delle variabili categoriali, é stata effettuata l'analisi della correlazione tra tutte le variabili del dataset.



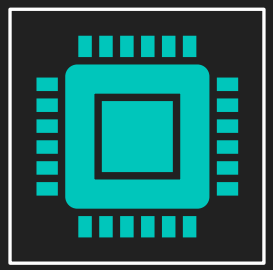
Individuazione potenziali bias





- Lo sbilanciamento tra i livelli educativi implica che le categorie Masters e PhD siano sottorappresentate rispetto a Bachelors. Questo potrebbe distorcere i risultati dei modelli predittivi, riducendone l'affidabilitá
- Il divario tra maschi e femmine nel dataset potrebbe portare a modelli predittivi con prestazioni diseguali per i due generi, enfatizzando caratteristiche prevalenti nei maschi e trascurando quelle delle femmine.

I modelli scelti



Logistic Regression:

Semplice, interpretabile, veloce

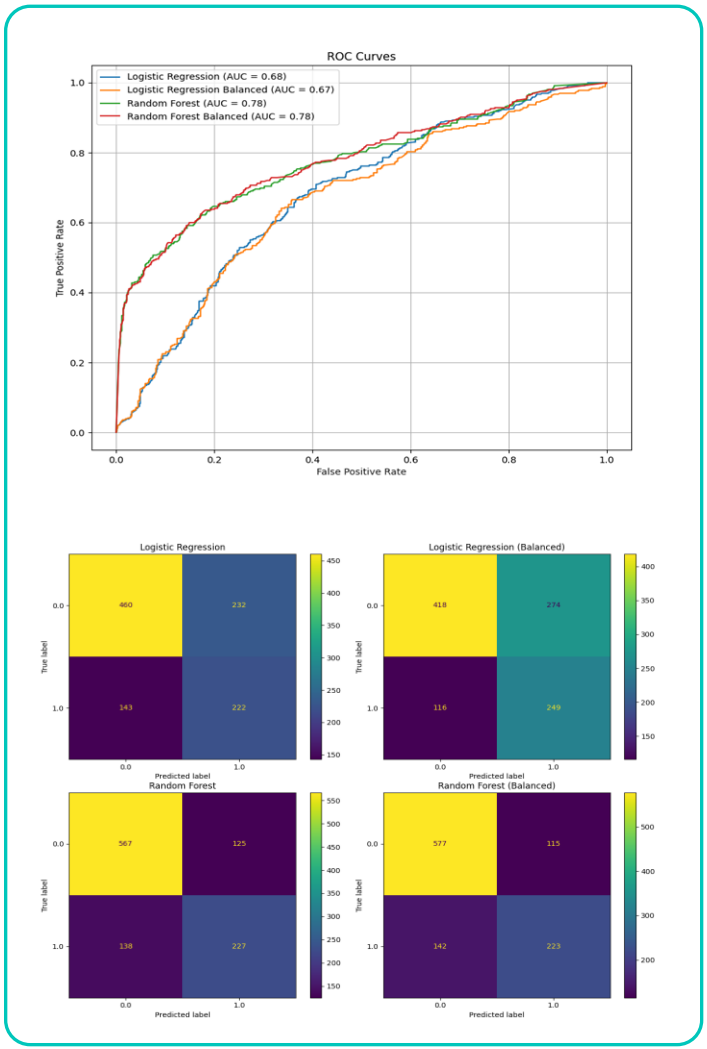
Random Forest:

Robusto, cattura relazioni non lineari

Come valutiamo i modelli

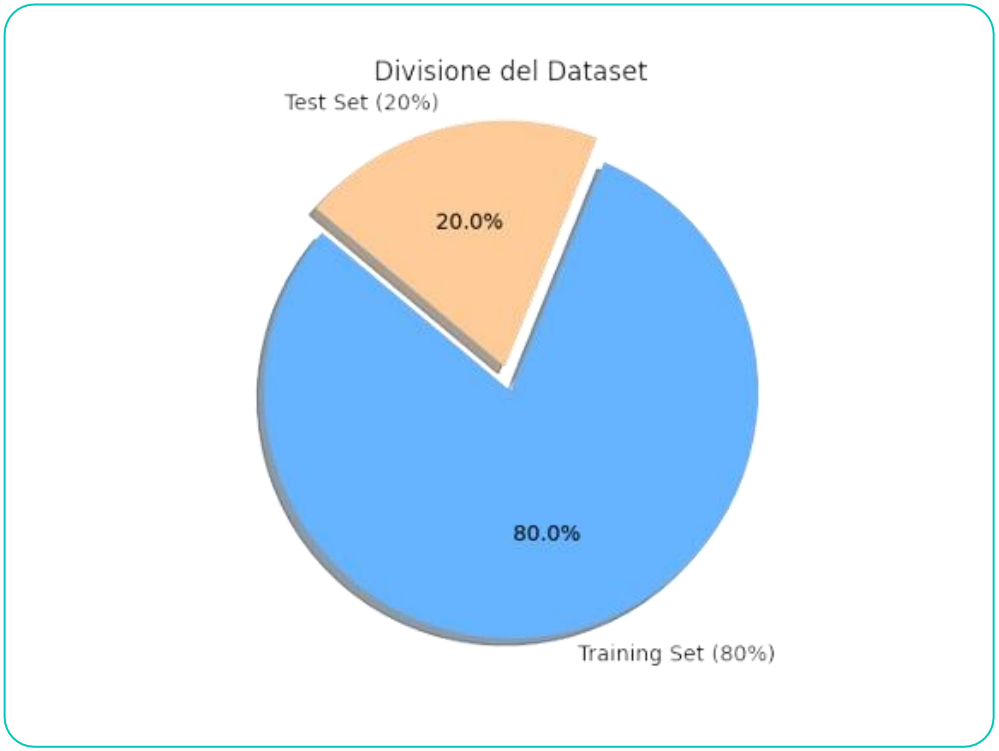
- Metriche principali:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - ROC-AUC

- La curva ROC-AUC e le confusion matrix mostrano i risultati di entrambi i modelli prima e dopo il bilanciamento



Divisione del dataset

- É stata adottata una divisione 80%-20%, riservando l'80% dei dati all'addestramento e il 20% al testing



Confronto dei risultati

Random Forest:

- F1-Score: 63%
- Recall (Classe 1): 61%
- Precision (Classe 1): 66%
- Accuracy: 75%

Logistic Regression:

- F1-Score: 52%
- Recall (Classe 1): 55%
- Precision (Classe 1): 50%
- Accuracy: 65%

- Nota: Random Forest si conferma più performante su tutte le metriche principali.

Sintesi

Random Forest risulta più performante rispetto alla Logistic Regression su tutte le metriche chiave

L'utilizzo del metodo SMOTE per generare dati sintetici ha migliorato il Recall della classe minoritaria e ridotto i falsi negativi, rendendo il modello più affidabile per identificare i dipendenti a rischio.