

Predicting Hiring Decisions in Recruitment Data

Il dataset (preso e modificato da kaggle) si trova al link <https://bit.ly/4ckMxMd> e contiene dati relativi ad alcuni candidati da assumere in un'azienda. Si vuole predire se un candidato verrà assunto (HiringDecision).

1. Quante sono le istanze contenute nel dataset? Il dataset è completo (cioè per ogni istanza sono sempre specificati tutti i valori di ogni attributo)? Il dataset è bilanciato rispetto alla classe da predire?
2. **Caricare il dataset**, eliminare eventuali attributi inutili (motivare la scelta), eliminare eventuali istanze con valori nulli, dividere il dataset in train (75%) e test (25%), preservando le proporzioni delle classi del target.
3. Valutare le performance sia sul dataset train sia sul dataset test del modello **SGDClassifier**, tenendo in considerazione F1-score e la confusion matrix.
4. **Analisi della fairness del modello**: valutare, con i dati del test set e rispetto al modello SGDClassifier, se la probabilità di predire 0 è la stessa per uomini (0) e donne (1). Il modello ha le stesse performance sul dataset degli uomini e sul dataset delle donne? Calcolando la metrica `demographic_parity_ratio` della libreria `fairlearn`, è possibile stabilire che il modello rispetta la “parità demografica”? Eliminare l'attributo Gender e valutare se le performance del modello ottenute negli uomini sono le stesse ottenute nelle donne.
5. **Analisi della feature importance**: applicare la tecnica Permutation Feature Importance (PFI) considerando il modello SGDClassifier per trovare le feature più importanti. Per garantire una certa stabilità dei risultati, applicare 10 permutazioni diverse ad ogni feature. Visualizzare attraverso un boxplot per ogni feature le differenze di F1-score in ogni permutazione.
6. **Analisi della feature importance**: verificare se i risultati ottenuti al punto precedente sono confermati dalla tecnica Leave-One-Covariate-Out (ovvero escludo una feature alla volta e vedo come variano le performance del modello).

** Prima delle predizioni, trasformare tutte le feature del dataset tranne Gender usando lo **StandardScaler** (puoi usare il `ColumnTransformer`).