Analisi e Predizione dei Dati di Cittadinanza in Gran Bretagna

1. Introduzione

Il progetto affronta il task di analisi e predizione dei dati di cittadinanza in Gran Bretagna basandosi su un dataset contenente informazioni relative alle richieste e concessioni di cittadinanza nel paese. L'obiettivo principale è stato quello di costruire un sistema che, partendo dai dati grezzi, potesse:

- Preprocessare e pulire i dati
- Addestrare e confrontare diversi modelli di machine learning
- Valutare le prestazioni dei modelli utilizzando metriche standard

L'obiettivo del progetto è di addestrare i modelli per predire se un cittadino di un paese europeo potrà ottenere la cittadinanza della Gran Bretagna basandosi sulle cittadinanze concesse annualmente, l'influenza del paese di provenienza e di fattori come età, sesso e motivo della richiesta.

2. Dataset Utilizzato

Il dataset è costituito da un file Excel disponibile sul sito del governo inglese (https://www.gov.uk/government/statistical-data-sets/immigration-system-statistics-data-tables#settlement).
Dal dataset sono presi in considerazione i due fogli di lavoro ("*Data - Cit_D01*" e "*Data - Cit_D02*") contenenti informazioni relative agli anni, dati personali dei richiedenti la cittadinanza, tipo di applicazioni e numero di concessioni di cittadinanza. Il primo foglio riguarda le richieste mentre il secondo le concessioni.
Sono stati applicati i seguenti filtri e trasformazioni per semplificare il dataset e analizzare solamente i dati dei richiedenti provenienti da paesi Europei:

- Filtro temporale: mantenuti i dati dal 2014 in poi.
- Regioni selezionate: incluse solo le regioni europee
- Unione dei dataset: i due fogli sono stati uniti sulla base delle colonne "Year" e "Nationality".
- Selezione delle colonne rilevanti: mantenute solo le colonne di interesse come "Applications", "Application type", "Sex", "Age" e "Grants"
- Imputazione dei dati mancanti: sostituiti i valori mancanti utilizzando la strategia della moda.

In seguito, sono rappresentati graficamente i dati del dataset utilizzato:

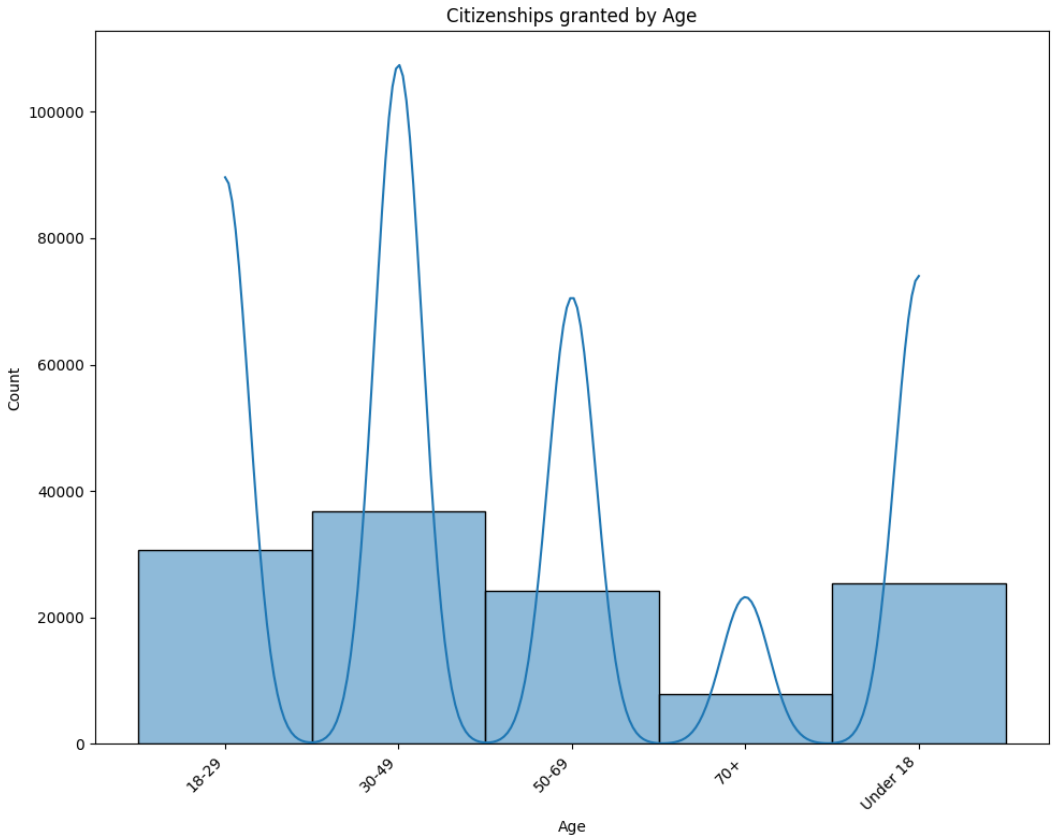

Figura 1: Cittadinanze concesse a seconda dell'età

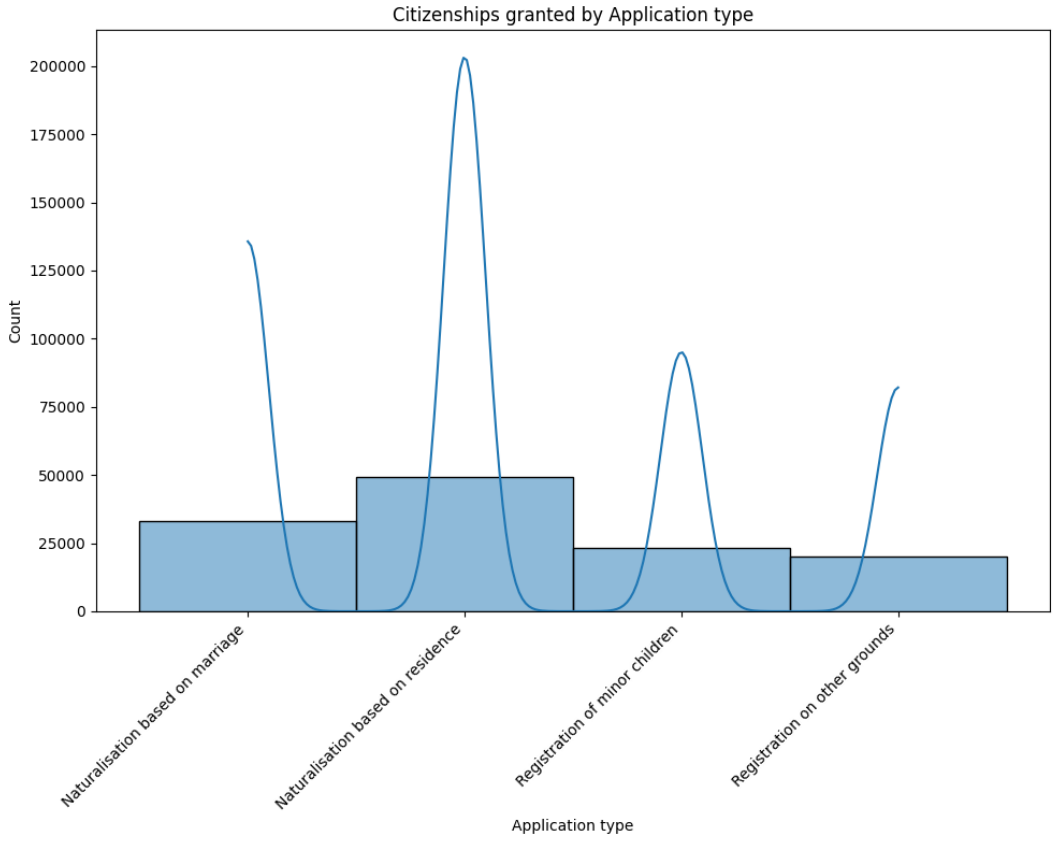

Figura 2: Cittadinanze concesse a seconda della nazionalità

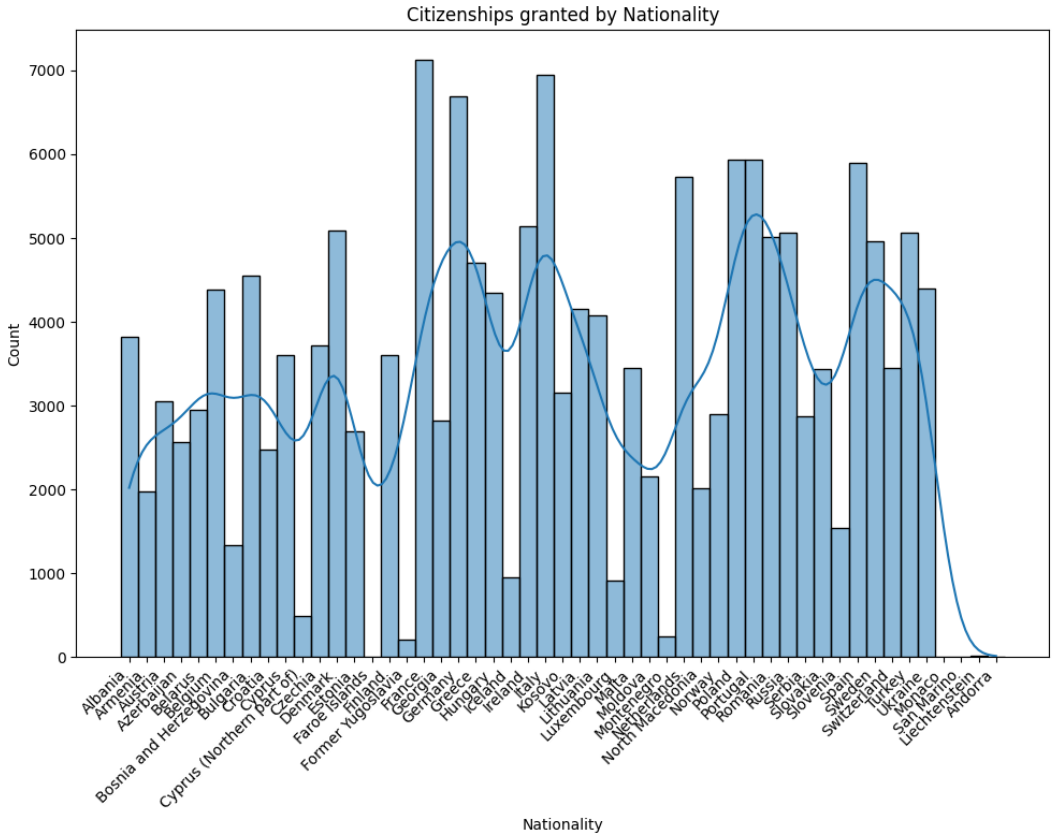

Figura 3: Cittadinanze concesse a seconda della nazionalità

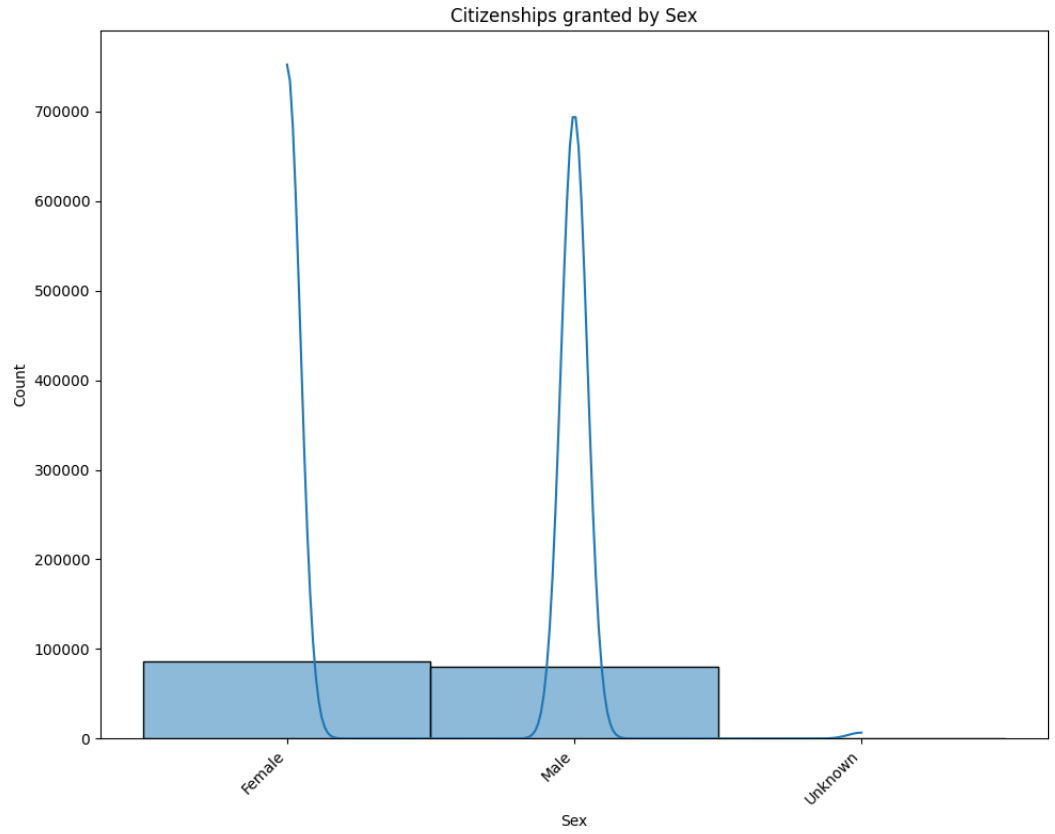

Figura 4: Cittadinanze concesse a seconda del sesso

3. Scelte Progettuali

3.1. Preprocessing

Il modulo di preprocessing ha implementato:

- La lettura e il filtraggio del dataset per anno e paese di provenienza
- La generazione di grafici esplorativi per visualizzare la distribuzione delle variabili chiave e avere un quadro iniziale della struttura dei dati
- La standardizzazione dei dati per garantire l'omogeneità delle scale, tramite *StandardScaler*

Per minimizzare l'impatto dei dati mancanti mantenendo la coerenza del dataset sono stati sostituiti i valori mancanti utilizzando la strategia della moda, adatta per variabili categoriche, dove la scelta del valore più frequente riduce il rischio di introdurre distorsioni significative. Non richiede supposizioni sulla distribuzione dei dati, rendendola robusta per variabili non numeriche.

Per garantire l'omogeneità delle scale delle variabili, è stata applicata la standardizzazione utilizzando la classe *StandardScaler* della libreria *sklearn.preprocessing*. Questo processo trasforma i dati in modo che ogni variabile abbia una media pari a 0 e una deviazione standard pari a 1 (normalizzazione). La formula applicata è:

$$x' = \frac{x - \bar{x}}{\sigma}$$ Dove $\begin{cases} \bar{x} \text{ è la media della variabile x;} \\ \sigma \text{ è la sua deviazione standard.} \end{cases}$

La standardizzazione è stata eseguita separatamente per il set di training e per quello di test. Lo scaler è stato adattato, con il metodo *fit,* solo sui dati di training, calcolandone la media e la deviazione standard, e successivamente applicato ai dati di test (con il metodo *transform*) per evitare problemi di data leakage garantendo la correttezza dell'intero processo.

3.2. Modelli di Machine Learning

Sono stati addestrati quattro modelli:

- Regressione Logistica: per prevedere la probabilità che un'osservazione appartenga a una determinata classe.
- Random Forest: basato su alberi decisionali. Costruisce molti alberi decisionali indipendenti su sottoinsiemi casuali dei dati e ne aggrega i risultati.
- Hist Gradient Boosting: versione ottimizzata del Gradient boosting, crea sequenzialmente alberi decisionali, ogni albero cerca di correggere gli errori commessi dagli alberi precedenti. Divide le variabili in "bin" (istogrammi) per velocizzare i calcoli e migliorare l'efficienza. È stato preferito l'hist gradient boosting a causa di vincoli di efficienza dei dispositivi usati per l'addestramento.
- Decision Tree: basato su struttura ad albero, composto da nodi che dividono i dati in base a condizioni su una caratteristica. Ogni foglia rappresenta un output.

I modelli sono stati addestrati su un set di training (80% dei dati) e testati su un set di test (20%).

3.3. Valutazione

Le prestazioni dei modelli sono state valutate utilizzando le seguenti metriche:

- Accuracy: percentuale di previsioni corrette
- Precision: accuratezza delle previsioni positive;
- Recall: copertura delle previsioni positive;
- F1 Score: media armonica tra precision e recall.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Risultati dell'Analisi

4.1. Prestazioni dei Modelli

I risultati ottenuti sono stati salvati in formato tabellare e visivamente rappresentati tramite grafici. Esempio di risultati:

```
Dataframe risultante:
                          Accuracy  Precision    Recall  F1 Score
Logistic Regression       0.236433   0.143973  0.236433  0.159412
Random Forest             0.236761   0.056056  0.236761  0.090649
Hist Gradient Boosting    0.074943   0.242389  0.074943  0.010647
Decision Tree             0.236970   0.073980  0.236970  0.091623
```

Sono stati generati istogrammi e grafici a barre per visualizzare le metriche di prestazione per ogni modello e il confronto fra modelli.

I grafici sono stati salvati in cartelle specifiche per facilitare la consultazione.

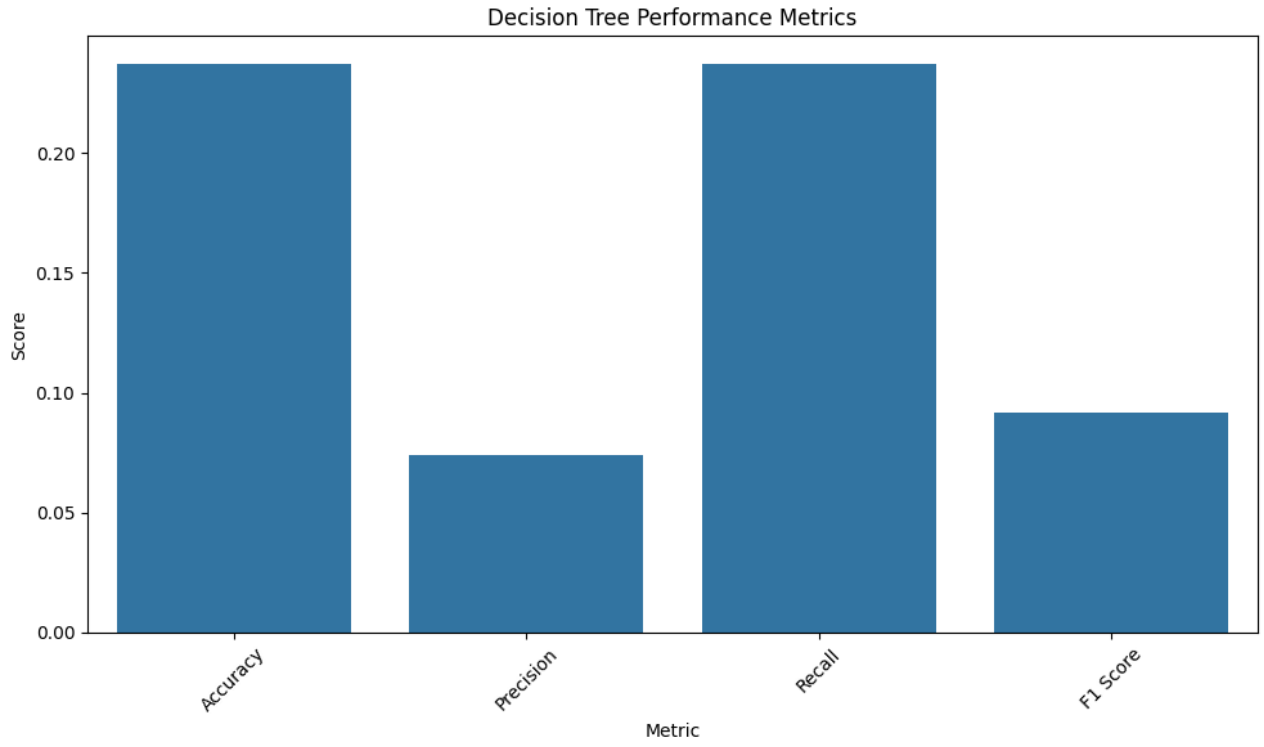

Figura 5: Performance per la metrica decision tree

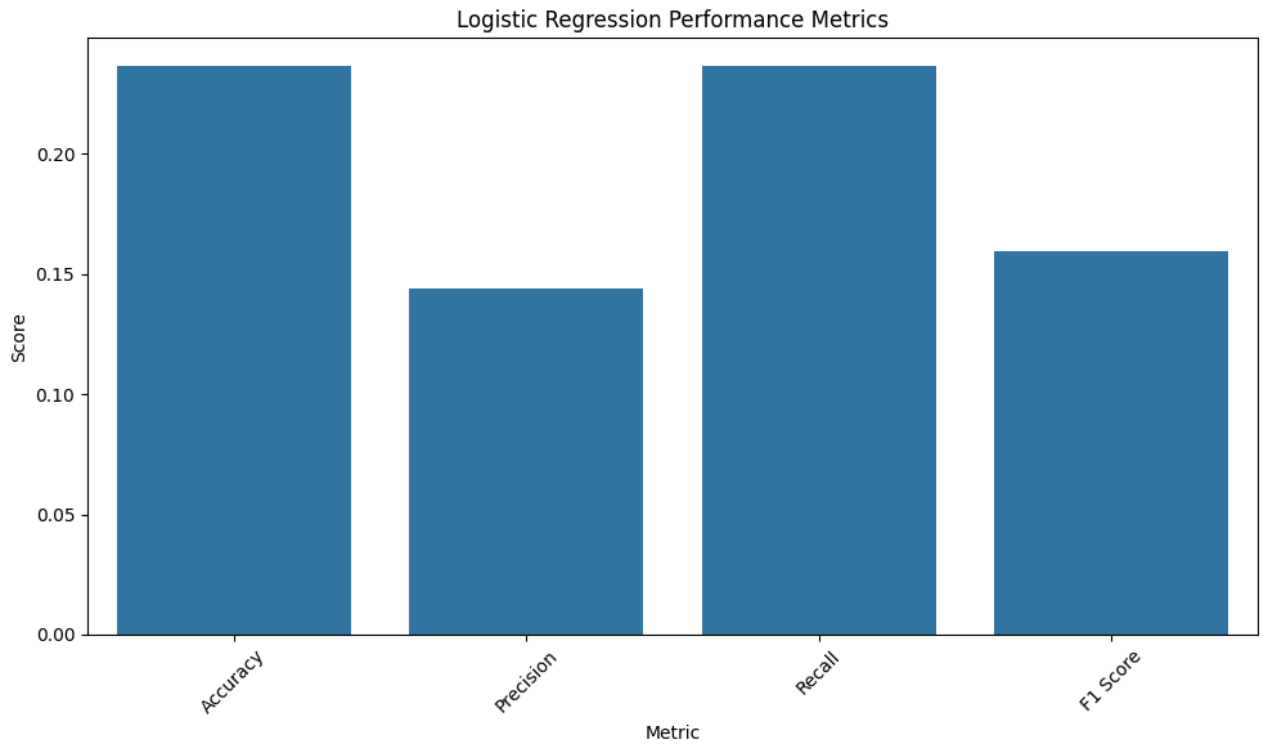

Figura 6: Performance per la metrica hist gradient boosting

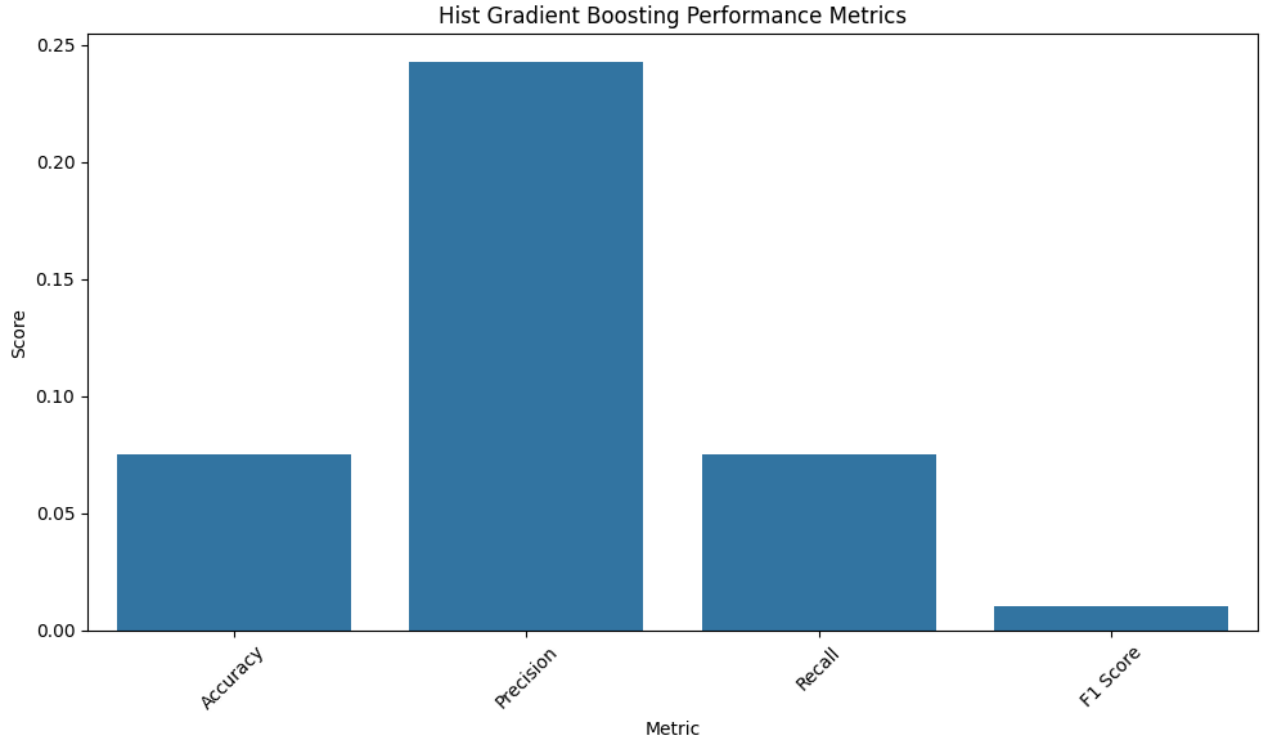

Figura 7: Performance per la metrica logistic regression

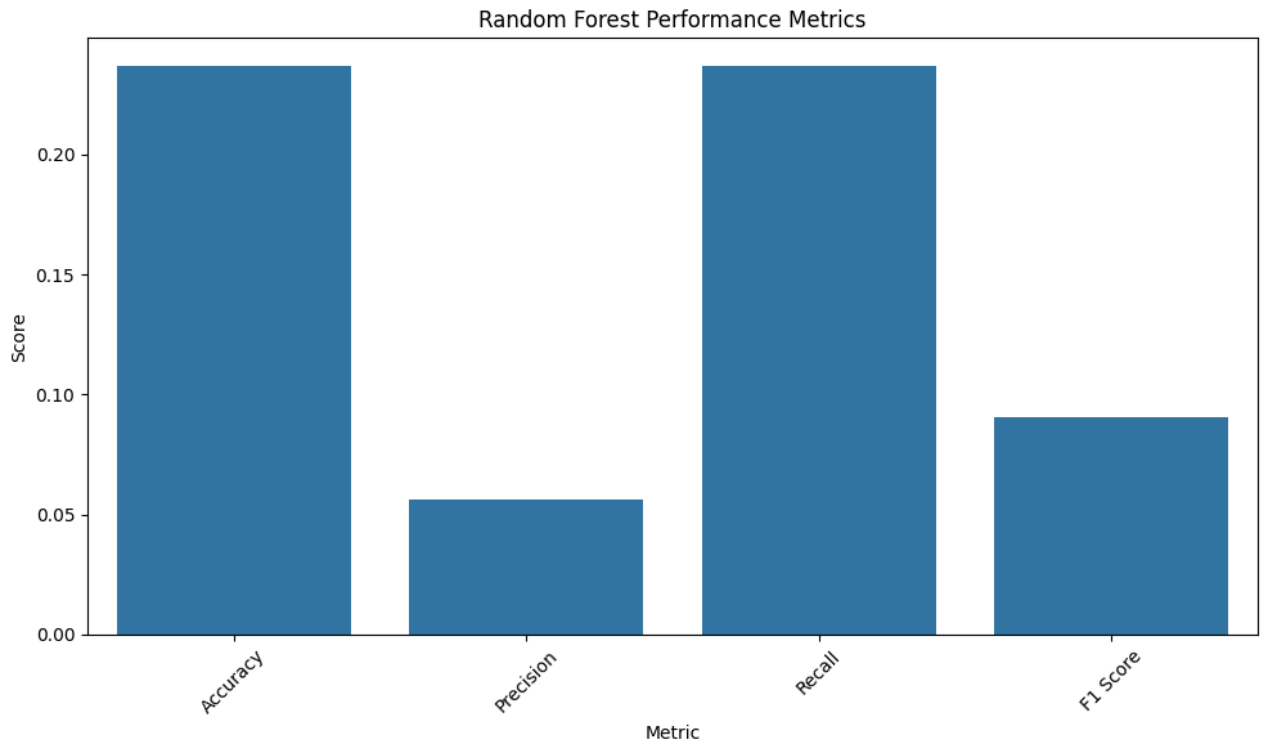

Figura 8: Performance per la metrica random forest

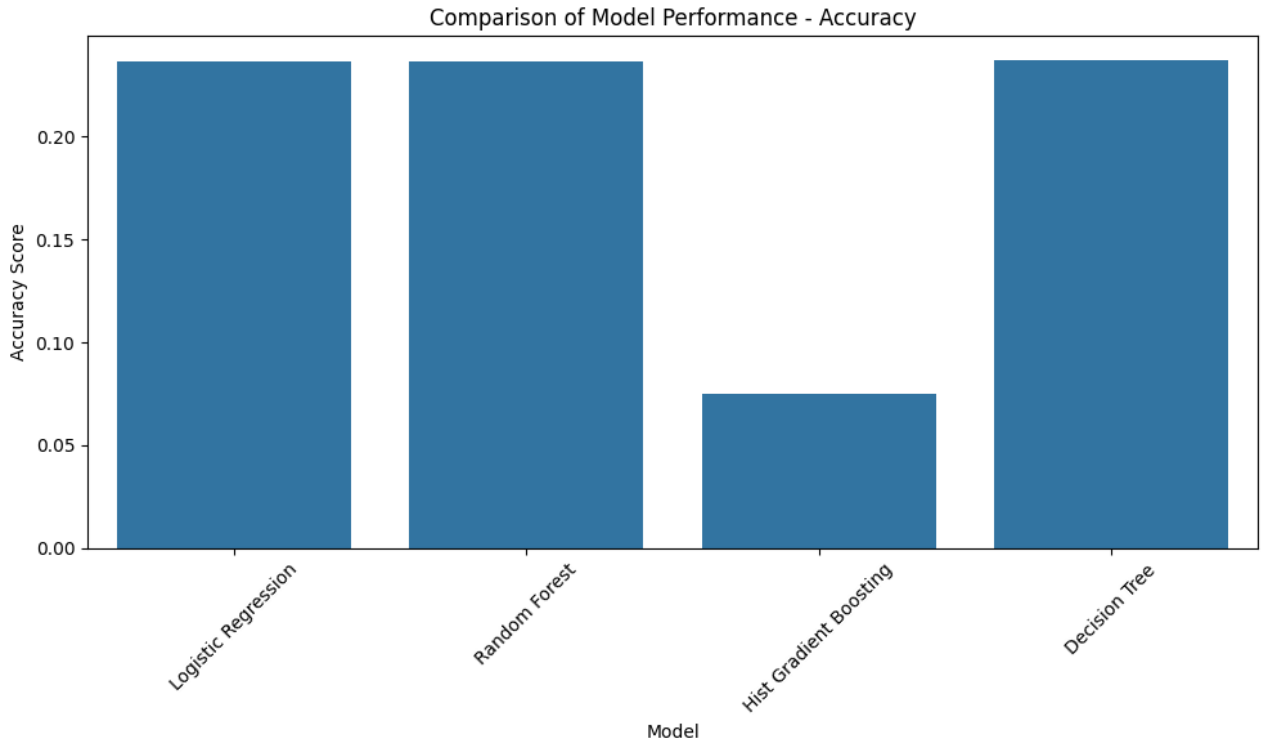

Figura 9: Confronto fra metriche per l'accuracy

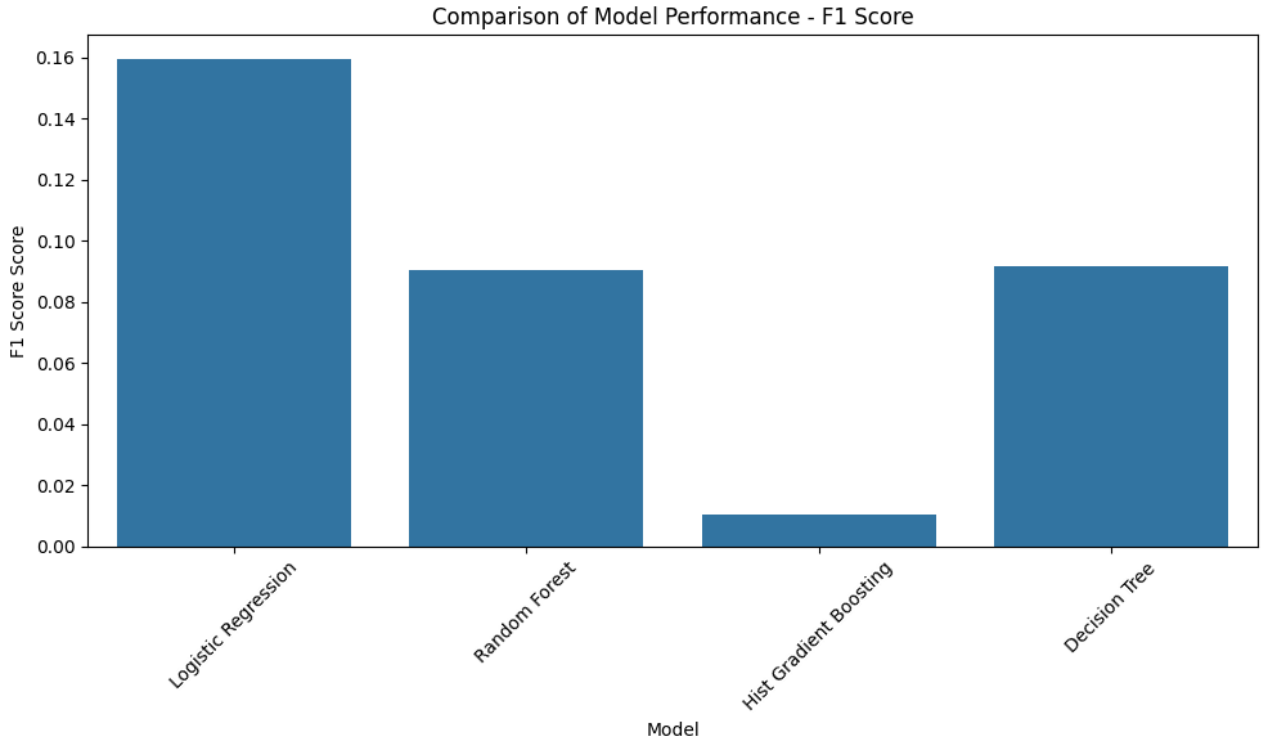

Figura 10: Confronto fra metriche per il f1 score

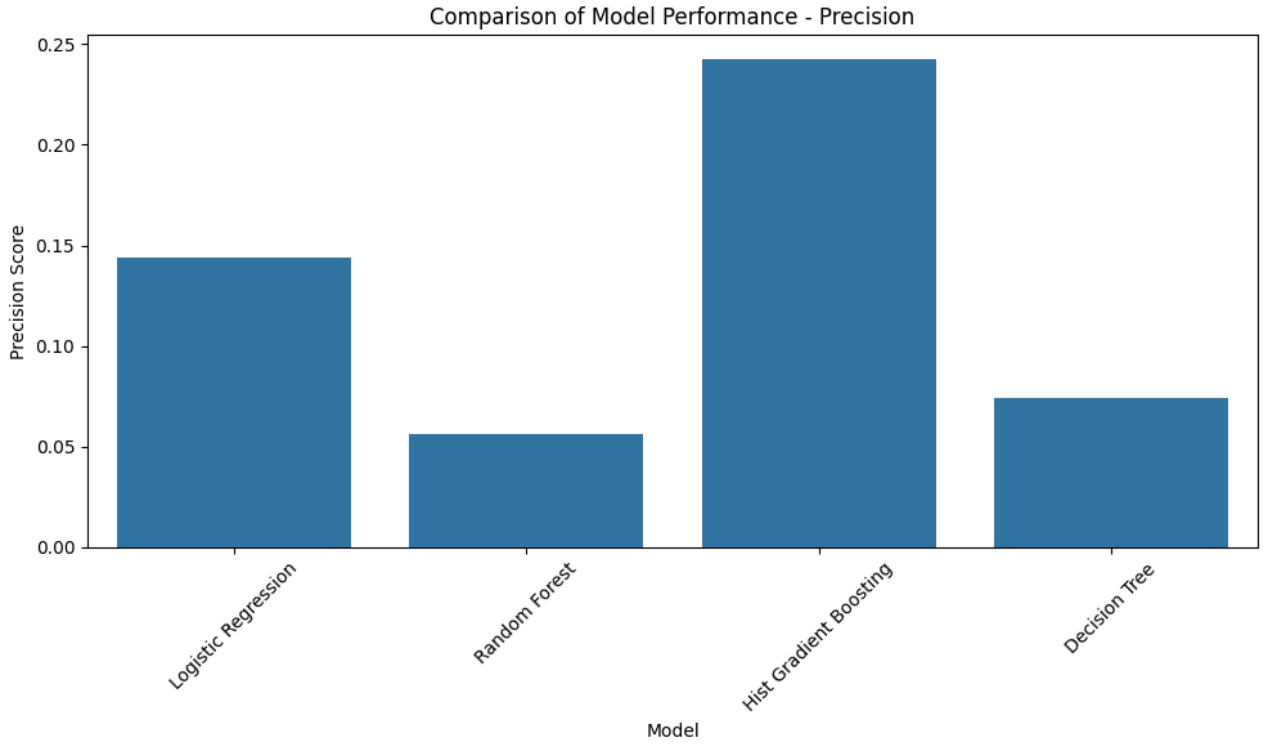

Figura 11: Confronto fra metriche per la precision

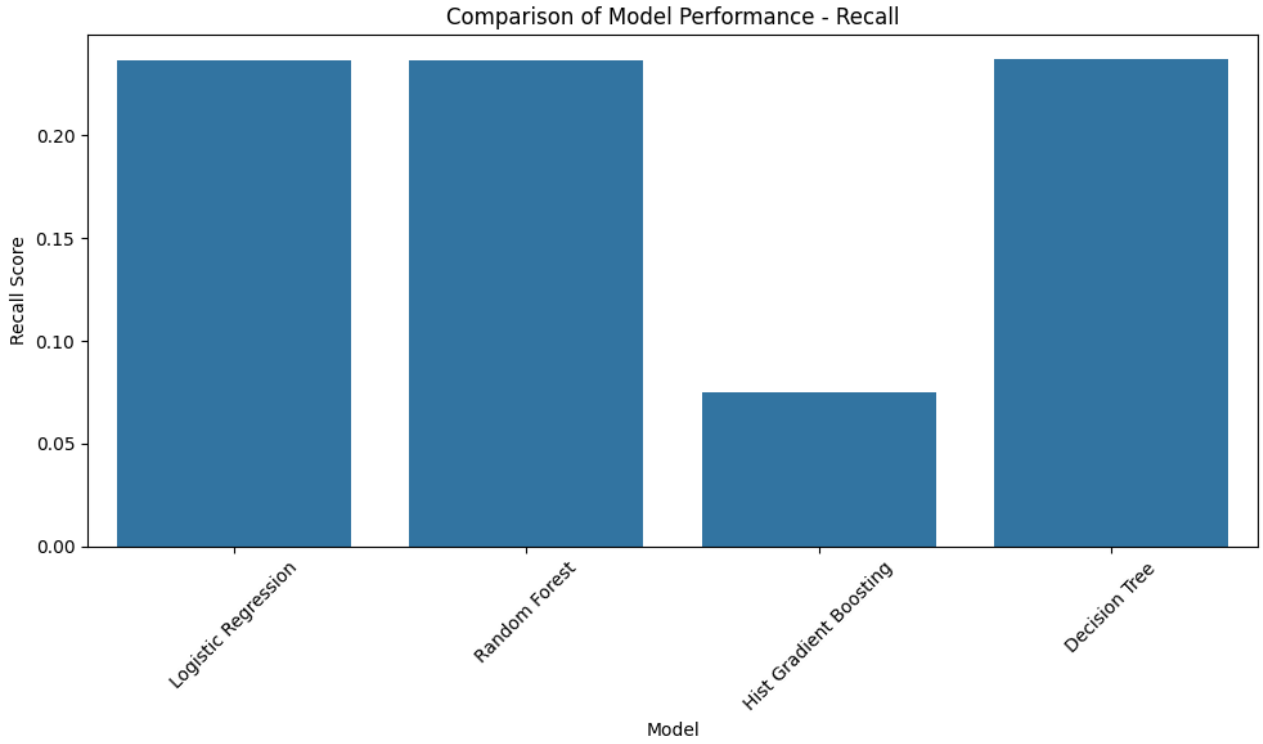

Figura 12: Confronto fra metriche per il recall

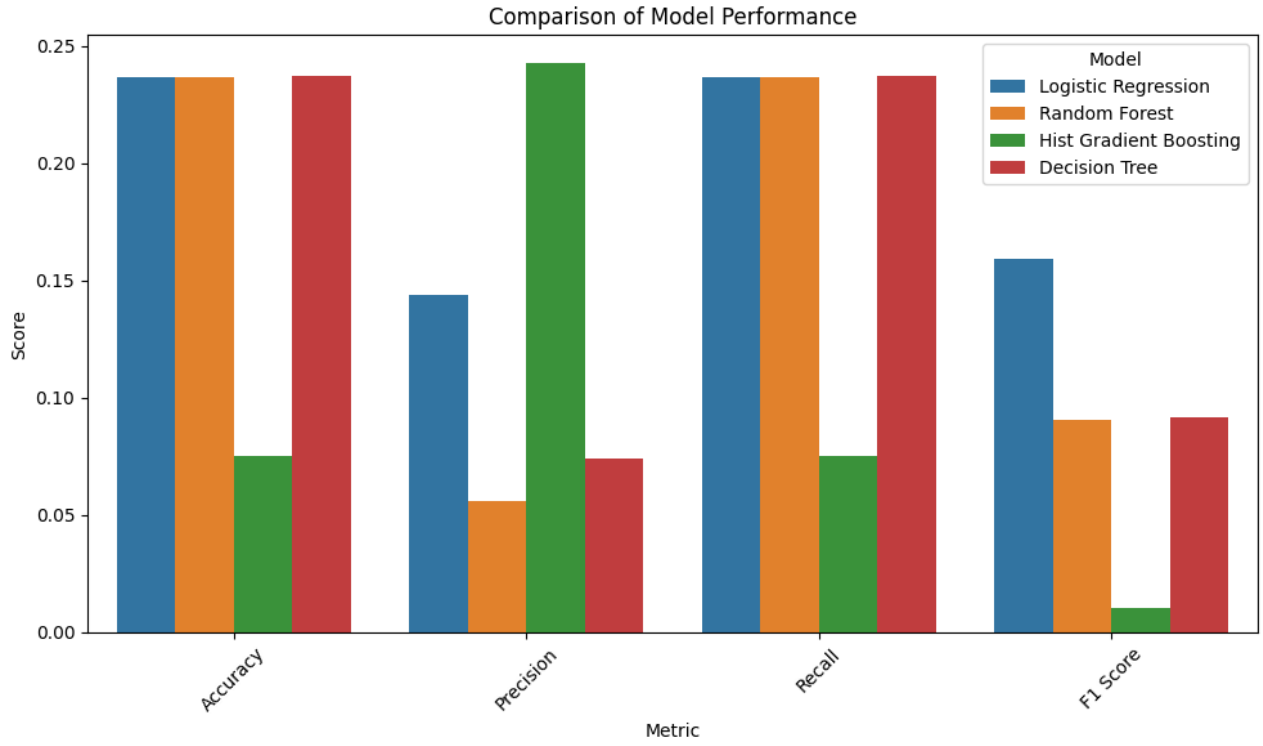

Figura 13: Confronto fra i modelli

5. Conclusioni

Il sistema sviluppato ha permesso di:

- Automatizzare il preprocessing e la pulizia dei dati;
- Confrontare le prestazioni di modelli diversi;
- Fornire una base per analisi future sulla concessione di cittadinanza.

Il modello più performante si è rivelato il **Random Forest**, che ha mostrato una buona capacità predittiva mantenendo un equilibrio tra precision e recall. Simile è il risultato del Decision Tree. Tuttavia, il **Logistic Regression** si è dimostrato più interpretabile, utile per comprendere le relazioni tra variabili, in quanto ha l'F1 score più alto, il che indica un equilibrio relativamente migliore tra precisione e recall rispetto agli altri modelli, sebbene tutte le performance siano basse in termini assoluti).
Hist Gradient Boosting, infine, ha la migliore precision, ma il basso recall e F1 score indicano che non è un modello performante complessivamente.

Sommario