

Analisi Dataset Assessing Mathematics Learning in Higher Education (AMLHE)

Intelligenza Artificiale II

Allegra Berardi (2116628)

Sofia Bruno (2146543)

Simone Manunza (2156783)

Il Dataset AMLHE

Il dataset (AMLHE) raccoglie i risultati di problemi matematici di studenti della scuola superiore provenienti da diversi paesi, Europei e non.



Dimensione: 9546 istanze di tipo numerico, categorico, binario e testuale



Feature:

- Student ID: identificativo del singolo studente
- Student Country: paese di provenienza
- Question ID: identificativo della domanda
- Question level: difficoltà della domanda (Basic/Advanced)
- Topic: macro-area del problema
- Subtopic: sotto-argomento specifico dei topic
- Keywords: parole chiave associate ai topic



Target: Type of Answer (1 = corretta, 0 = errata)

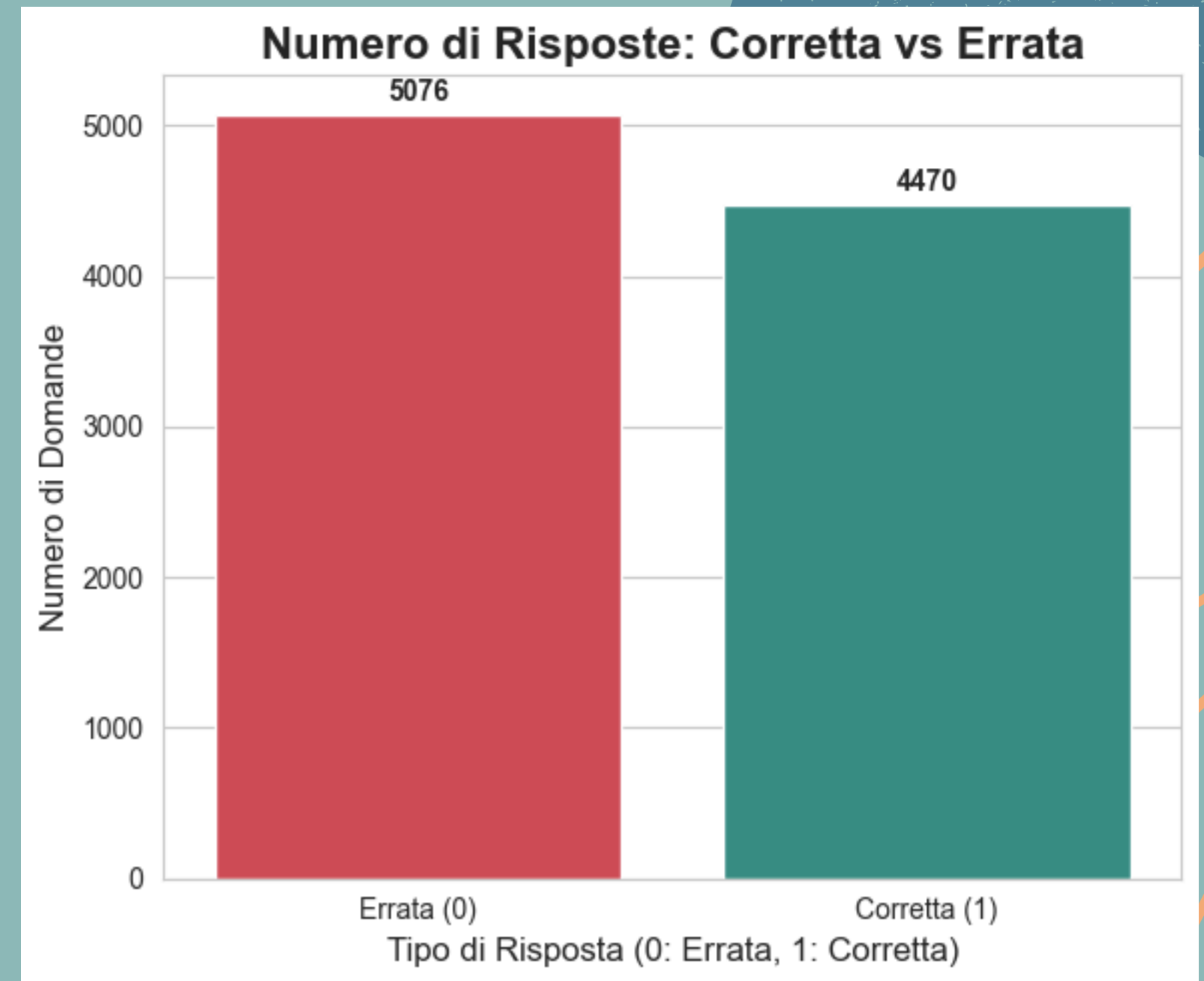


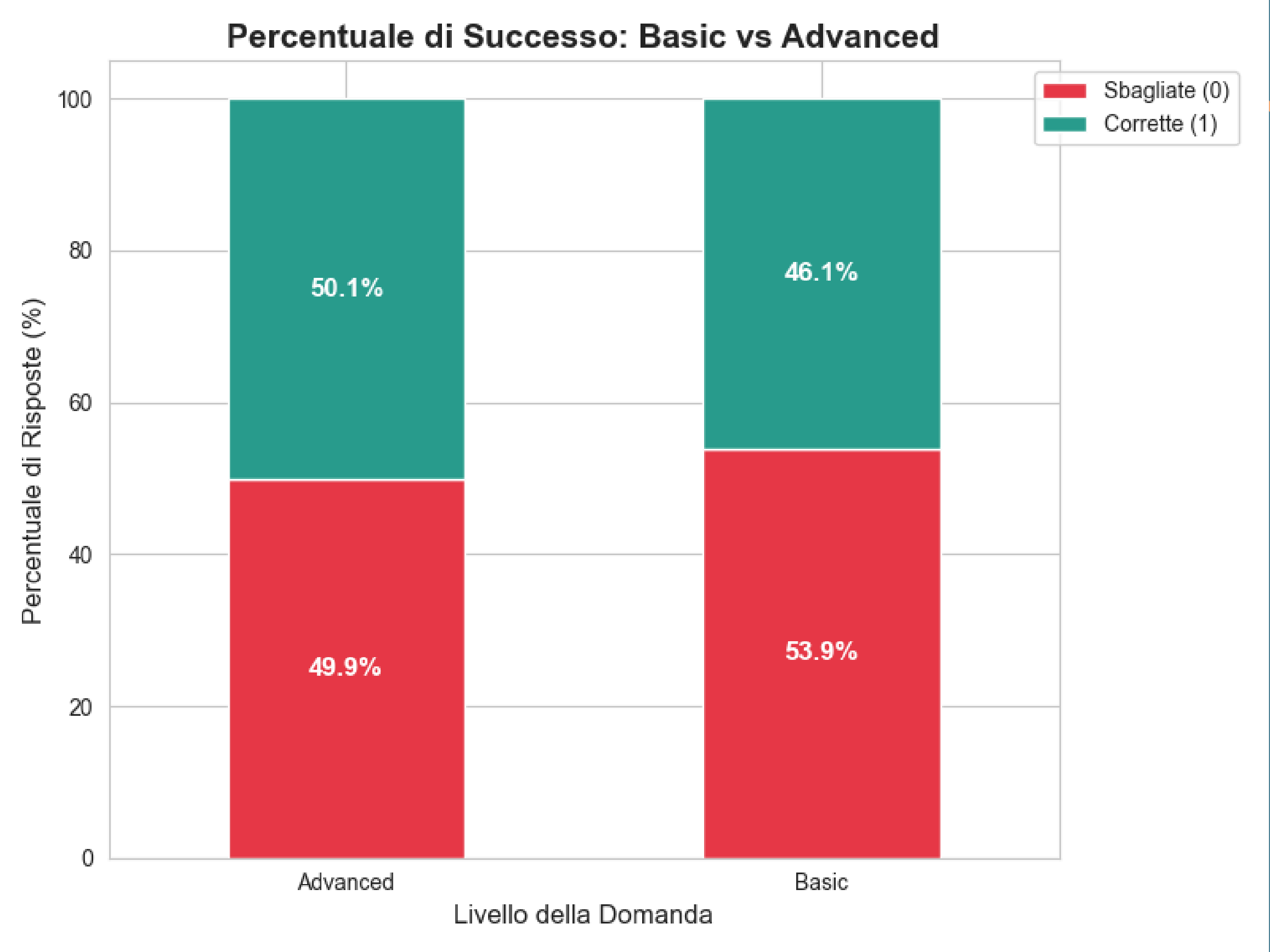
Obiettivo del modello: identificare i pattern che influenzano il successo degli studenti in ambito matematico.

Analisi Esplorativa dei Dati (EDA)

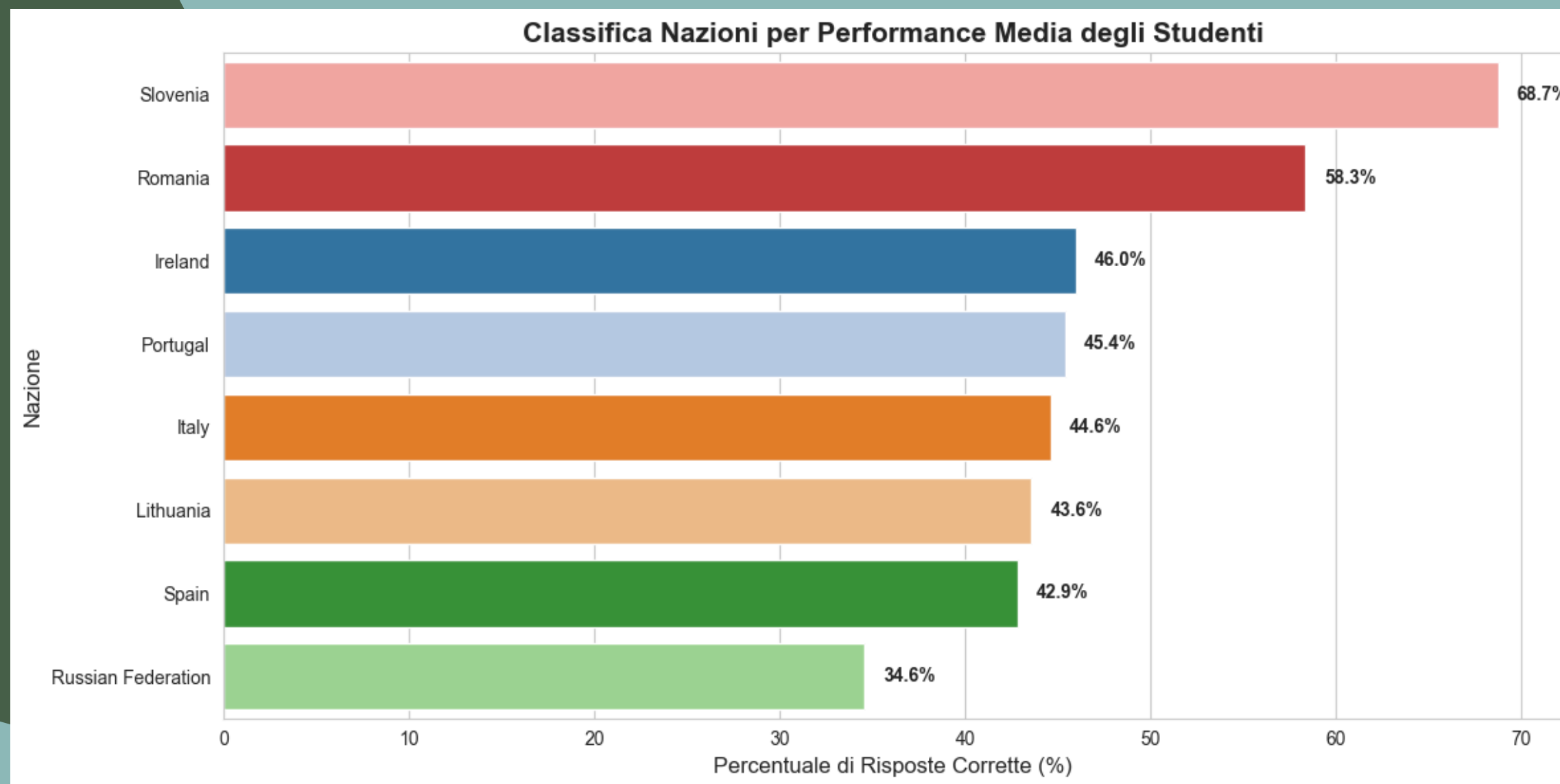
La prima parte dell'EDA affronta un'analisi generale delle domande con i seguenti risultati:

- Numero di domande: 9546
- Percentuale di domande errate: 53,2%
- Percentuale di domande corrette: 46,8%
- Numero di domande "Basic": 7844
- Numero di domande "Advanced": 1702

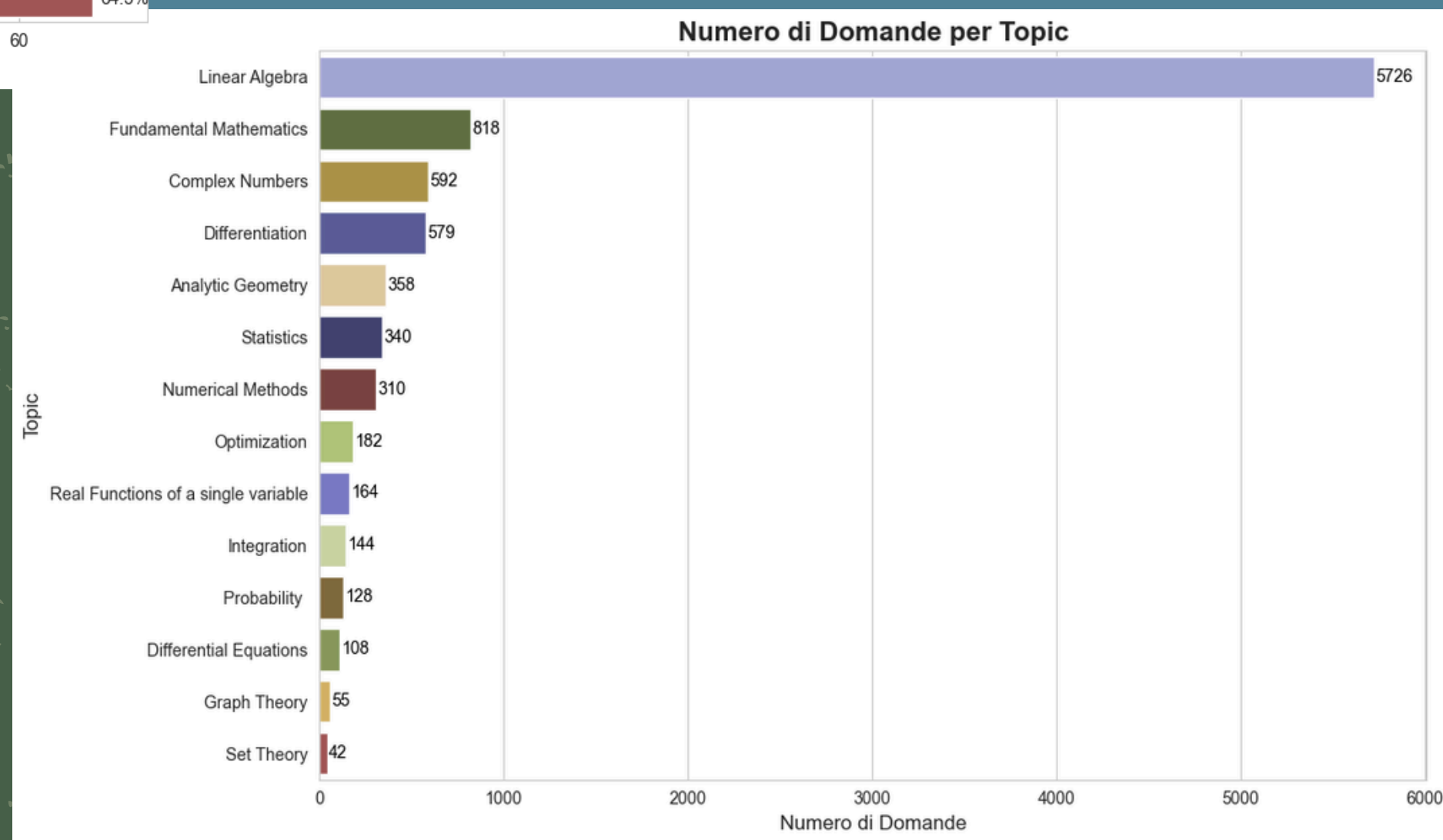
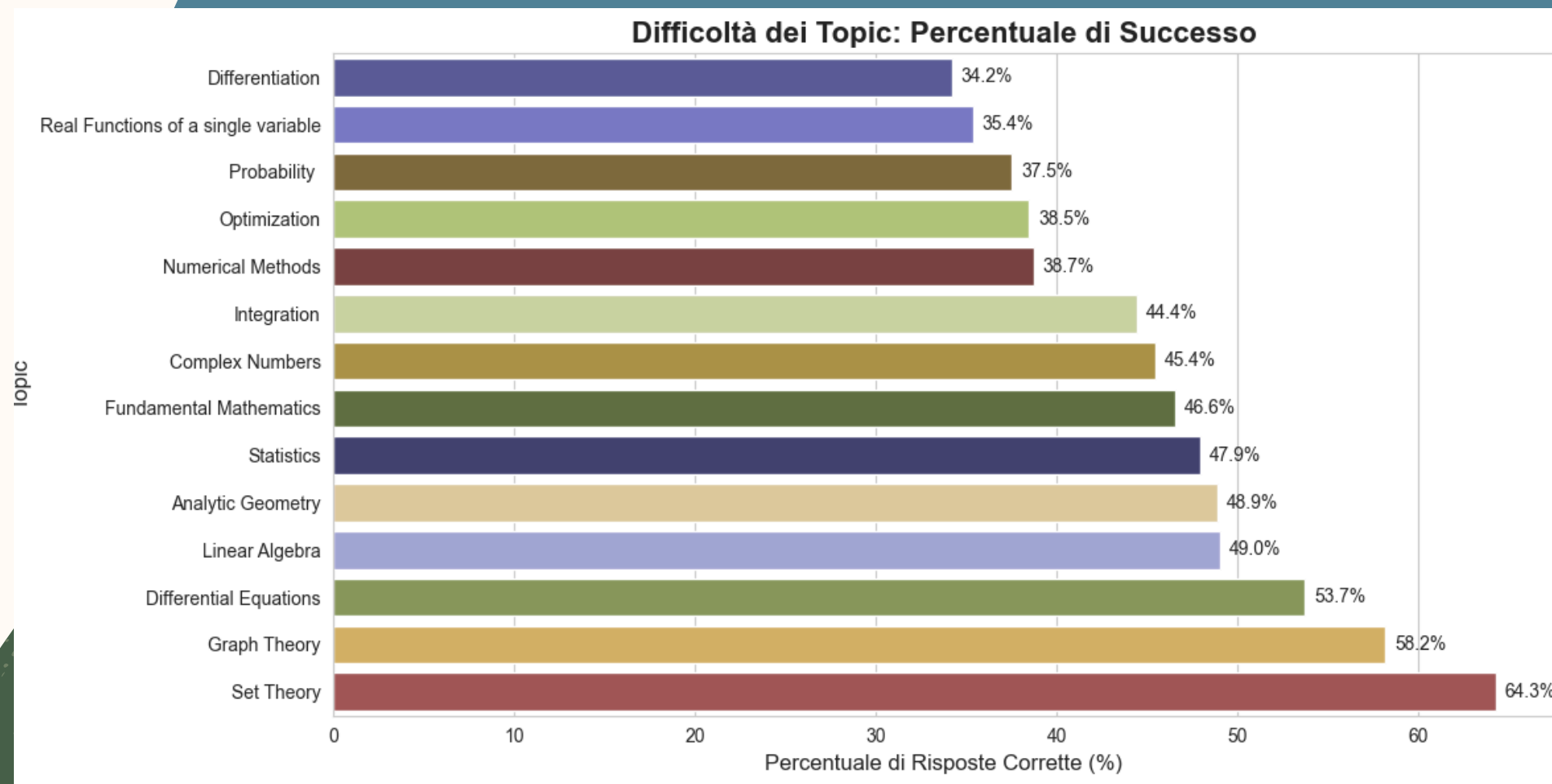




Il grafico mostra le percentuali di risposte corrette per paese. Il paese con il rendimento scolastico migliore è la "Slovenia" con una percentuale di (68,7%), mentre il paese con il rendimento scolastico più scarso è la "Russia" con una percentuale di (34,6%).




L'analisi sul numero di studenti per nazione, dimostra, però, un bias numerico verso nazioni come Portogallo, Lituania e Italia.



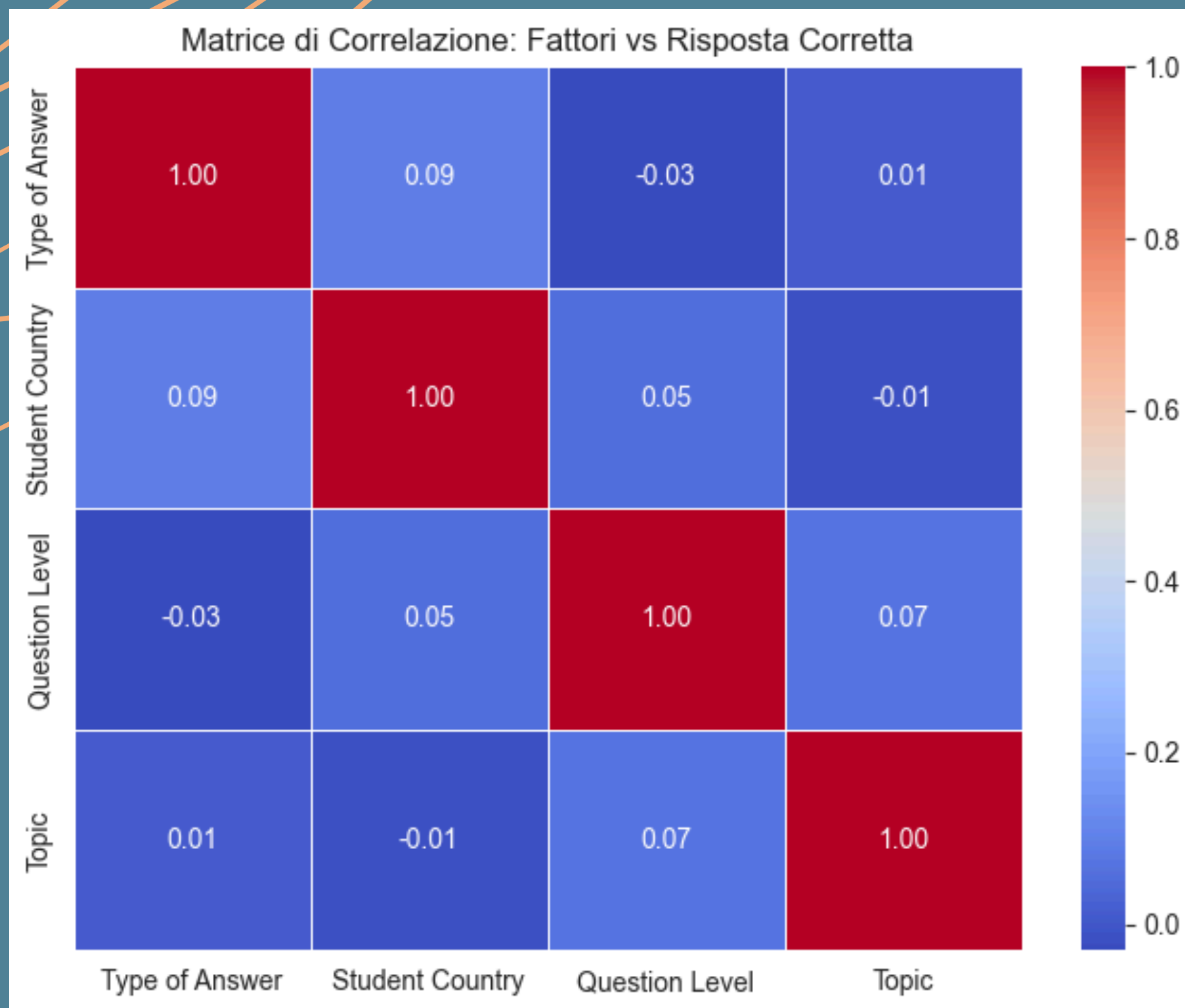


Risultati EDA

L'analisi esplorativa dei dati mette in luce diversi sbilanciamenti che troviamo nel dataset:

- Numero di Domande Corrette ed Errate
 - Numero di Domande Basic e Advanced
 - Numero di Studenti per Nazione
 - Numero di Domande per Topic
- 

Preprocessing & Feature Engineering



Abbiamo usato Label Encoder per generare la matrice di correlazione, NLP e One-Hot Encoding per l'addestramento dei modelli, trasformando i valori testuali in valori numerici.

Metodi e Modelli

Abbiamo addestrato il modello su 80% dei dati e testato la sua capacità di previsione sul restante 20%, usando la Stratificazione, con "stratify=y", per mantenere la giusta proporzione tra risposte corrette ed errate in entrambi i set.

- Logistic Regression: modello lineare come baseline.
- Random Forest: modello basato su Ensemble Learning per catturare relazioni non lineari tra le variabili.

Metodi e Modelli: Risultati a confronto

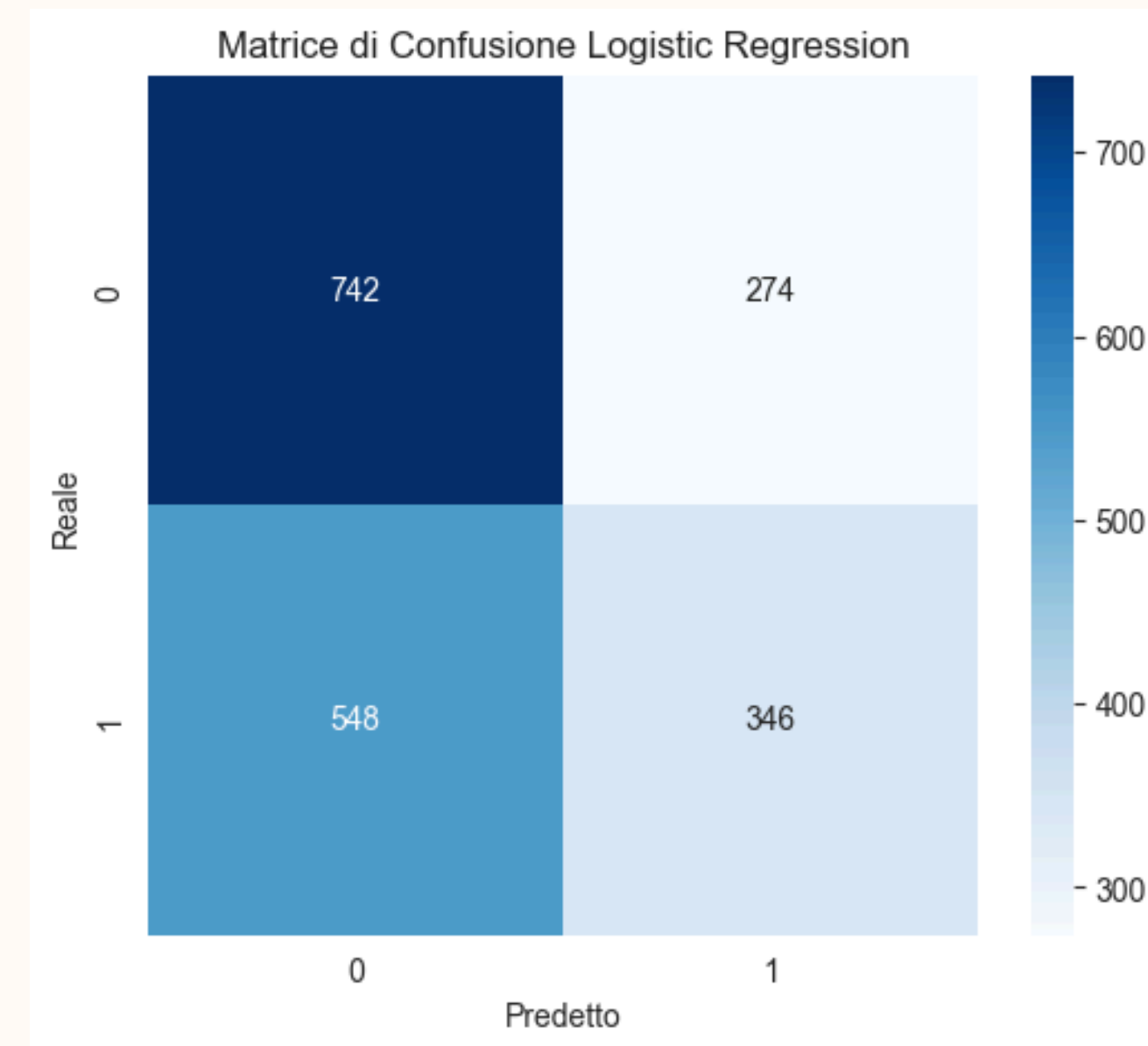
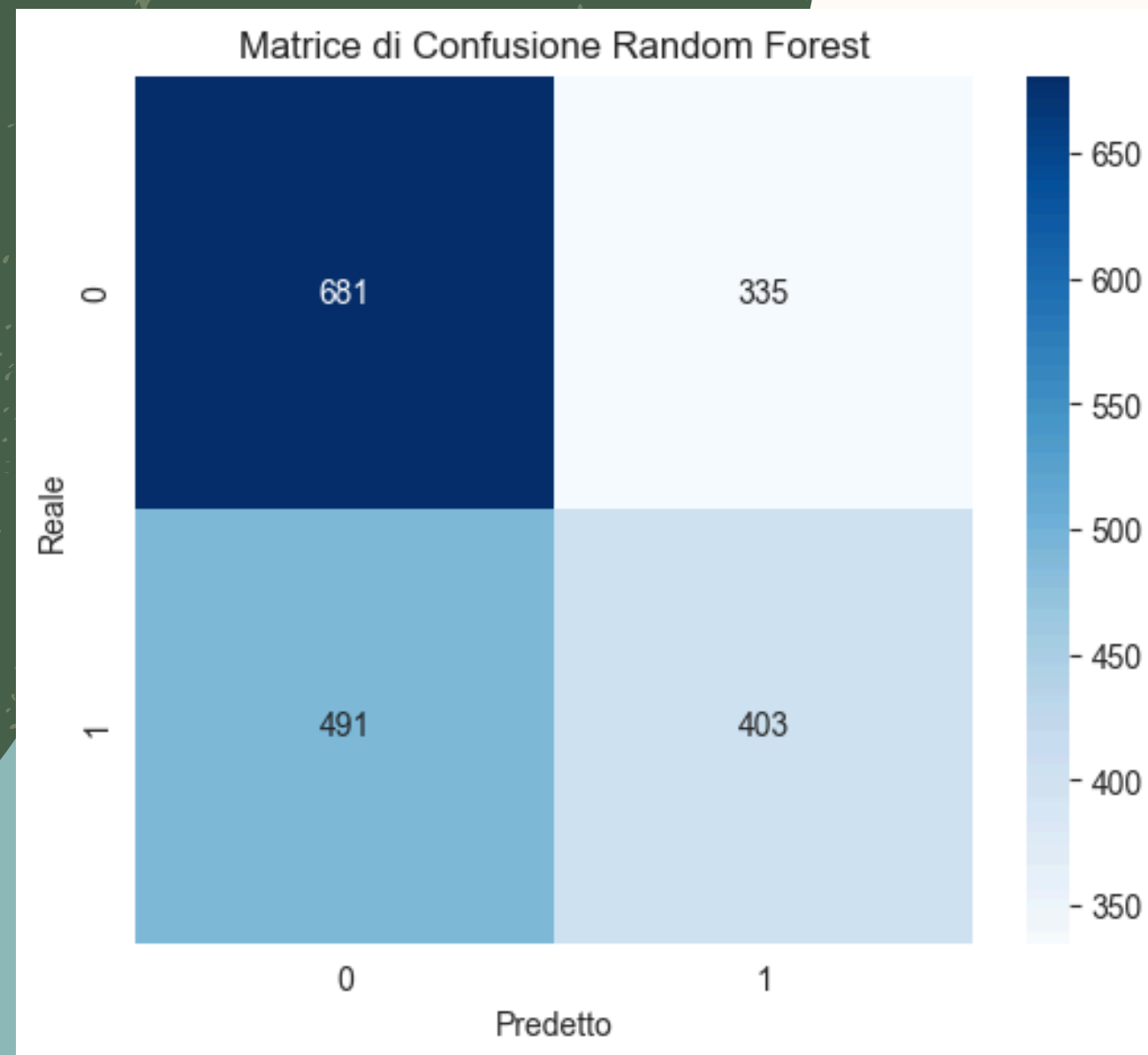
Valutazione Random Forest

	precision	recall	f1-score	support
0	0.58	0.73	0.64	1016
1	0.56	0.39	0.46	894
accuracy			0.57	1910
macro avg	0.57	0.56	0.55	1910
weighted avg	0.57	0.57	0.56	1910
Accuratezza Random Forest: 0.5675				

Valutazione Logistic Regression

	precision	recall	f1-score	support
0	0.58	0.67	0.62	1016
1	0.55	0.45	0.49	894
accuracy			0.57	1910
macro avg	0.56	0.56	0.56	1910
weighted avg	0.56	0.57	0.56	1910
Accuratezza Logistic Regression: 0.5696				

Analisi della Matrice di Confusione



Le matrici di confusione mostrano alti valori sia di Precision che di Recall nel caso di risposta errata. Rivelano un bias che porta il modello a prevedere con più frequenza una risposta errata che corretta, dimostrando una più alta difficoltà (o più bassa F1-Score) per le risposte corrette.

Conclusioni

Abbiamo imparato a trasformare i dati prima di farli analizzare dal modello.

Attraverso i due modelli scelti, abbiamo avviato un confronto che mette in luce, da una parte una maggiore trasparenza e interpretazione, dall'altra un'analisi dei pattern complessi che riguardano i paesi di provenienza, gli argomenti e le difficoltà delle domande.

L'analisi delle matrici di confusione ha rilevato uno sbilanciamento del dataset verso le risposte errate che abitua il modello a fare previsioni troppo approssimate su nuovi dati e a prevedere l'errore anche quando assente.