Analisi Dataset Assessing Mathematics Learning in Higher Education (AMLHE)

Simone Manunza 2156783 manunza.2156783@studenti.uniroma1.it
Allegra Berardi 2116628 berardi.2116628@studenti.uniroma1.it
Sofia Bruno 2146543 bruno.2146543@studenti.uniroma1.it

1. Abstract

Partendo da un dataset che analizza report di studi matematici di studenti delle superiori di paesi differenti, si analizzano i dati tramite Exploratory Data Analysis (EDA) per ottenere grafici di valutazione che studiano le domande riportate nel dataset: i loro livelli, come gli studenti rispondono, chi sono gli studenti e i loro paesi. L'EDA mette in luce alcuni sbilanciamenti e bias presenti nel dataset, come la differenza numerica tra risposte errate e risposte corrette, o la differenza numerica tra domande "Basic" e "Advanced".
In secondo luogo si addestra un modello di Machine Learning (ML) con metodi di Random Forest (RF) e Logistic Regression (LR) per ottenere un modello che riesca a prevedere quale sia la risposta a una domanda matematica di uno studente, dati fattori come la nazione di provenienza dello studente e l'argomento della domanda.
Il metodo di addestramento di LR risulta avere un'accuratezza maggiore del metodo RF. Entrambi, tuttavia, risultano avere una maggiore frequenza nel predire che una risposta sia errata piuttosto che corretta, anche quando non dovrebbero, risultato di bias portati alla luce durante l'EDA.

2. Introduzione e Definizione del Problema

Il dataset AMLHE analizza un contesto educativo, precisamente lo studio della matematica in contesti di educazione superiore di più paesi. Lo studio del dataset si pone i seguenti quesiti:

1. Quante sono le domande poste agli studenti?
2. A quante di queste gli studenti rispondono correttamente o erroneamente?
3. Qual è la distribuzione del livello di difficoltà della domanda (Basic o Advanced)?
4. Qual è la percentuale di successo a seconda del livello della domanda (Basic o Advanced)?
5. Quanti studenti per nazione prende in analisi il dataset?
6. Quali paesi hanno la percentuale di performance media più alta?
7. Quali sono i topic più frequenti?
8. Quali sono i topic più difficili?

3. Descrizione del Dataset

Il dataset è stato scelto tra le proposte della Prof.ssa Monti su Classroom e riporta il titolo Assessing Mathematics Learning in Higher Education (AMLHE).
Il numero di feature è otto e sono le seguenti:

1. Student ID
2. Student Country
3. Question ID
4. Type of Answer
5. Question Level
6. Topic
7. Subtopic
8. Keywords

Invece, le osservazioni sono 9.546, e sono di tipo numerico, categorico, binario e testuale.

Nel dataset originale è presente un carattere errato che avrebbe dovuto rappresentare un carattere di apostrofo, tuttavia questo non riusciva ad essere letto da Python se il file era aperto con encoding UTF-8. Per risolvere il problema Gemini ha generato un codice che ha trovato la riga del carattere errato, dopodiché attraverso Visual Studio Code il carattere è stato cambiato in tutte le sue occorrenze nel file "dataset_IA_corretto.csv" con l'effettivo carattere corretto indicante l'apostrofo.

```
## CODICE UTILIZZATO PER TROVARE CARATTERE ERRATO
byte_position = 1017828 # The problematic byte position from the error
message
line_number = 0
bytes_read = 0

with open('dataset_IA.csv', 'rb') as f:
    for line in f:
        bytes_read += len(line)
        line_number += 1
        if bytes_read > byte_position:
            print(f"The error likely occurred on line: {line_number}")
            # Now read that specific line with the correct encoding to
see its content
            f.seek(bytes_read - len(line)) # Go back to the start of
the current line
            problematic_line_bytes = line # Store the bytes of the
problematic line
            try:
                problematic_line_str =
problematic_line_bytes.decode('latin1')
                print(f"Content of the problematic line (decoded with
latin1):\n{problematic_line_str.strip()}")
            except UnicodeDecodeError as e:
                print(f"Could not decode the problematic line even with
latin1: {e}")
            break
    else:
        print(f"Byte position {byte_position} not found in the file.")
```

4. Analisi Esplorativa dei Dati

La prima analisi del dataset è quella delle statistiche descrittive, che ci mostra la media, deviazione standard, minimo, massimo, ecc… Tuttavia, in questo caso di analisi questi risultati non hanno una

grande quantità informativa, in quanto feature come Student ID e Question ID sono numeri di identificazione, pertanto non è importante trovarne la media, il valore minimo e il valore massimo. Altre feature come Type of Answer sono binarie e quindi vale lo stesso caso delle prime feature menzionate.

Successivamente, per quanto riguarda le visualizzazioni essenziali i grafici rispondono alle domande riportate nel paragrafo introduttivo. I grafici hanno dimostrato che:

1. Le domande poste agli studenti sono 9546.
2. Gli studenti hanno risposto correttamente a 4470 domande e erroneamente a 5076 domande. Quindi, la percentuale di risposte corrette è del 46,8%, mentre quella di risposte errate è del 53,2%.
3. Il numero di domande di livello base ("Basic") è 7844, mentre il numero di domande di livello avanzato ("Advanced") è 1702.
4. Per le domande di livello base la percentuale di successo è del 46,1%, mentre la percentuale di errore è del 53,9%. Per le domande di livello avanzato la percentuale di successo è del 50,1%, mentre la percentuale di errore è del 49,9%.
5. Nell'analisi del numero di studenti per nazione il grafico riporta dei risultati che visivamente possono essere suddivisi in tre gruppi:
    a. Numero alto ($n > 5000$): Portogallo
    b. Numero medio ($755 \leq n \leq 1443$): Lituania, Italia, Slovenia
    c. Numero basso ($n \leq 300$): Irlanda, Russia, Romania, Spagna
6. Nell'analisi della performance media per studente di una nazione il grafico riporta la Slovenia nel primo posto, con una percentuale di risposta corretta di 68,7%, successivamente l'Irlanda, il Portogallo, l'Italia, la Lituania e la Spagna hanno una percentuale simile ($46\% \leq p \leq 42,9\%$), infine la Russia si trova all'ultimo posto con una percentuale di 34,6%.
7. Nei topic più frequenti risulta esserci un grande gap tra l'argomento più frequente, "Linear Algebra" con 5726 domande, e tutto il resto di argomenti che hanno un numero di domande estremamente inferiore.
8. I topic più difficili risultano essere: Differentiation (34,2% di successo) e Real Functions of a single variable (35,4%).

I grafici, infine, mettono in risalto e dimostrano uno sbilanciamento delle feature, per esempio nella differenza tra: il numero di domande basic e il numero di domande advanced, il numero di studenti per paese e il numero di domande per argomento.

5. Preprocessing e Feature Engineering

Nel dataset non sono presenti dei valori mancanti, l'unico errore era presente con il carattere di apostrofo, e la soluzione è individuata nel paragrafo Descrizione del Dataset § 3.

Per la creazione della matrice di confusione è stato usato il "`LabelEncoder`", un encoding numerico delle variabili categoriche, che, però, introduce un ordine numerico arbitrario tra i risultati sulle variabili categoriali.

Prima di eseguire il One-Hot Encoding eseguiamo un passaggio di Natural Language Processing della Classe "Keywords", usando un fattore di "`max_features`" di 30 nel "`CountVectorizer`", si semplifica il calcolo ignorando parole poco frequenti (che compaiono meno di 30 volte), riducendo la dimensionalità ed evitando overfitting. Tuttavia, questo potrebbe aver rimosso termini poco utilizzati ma utili per definire la difficoltà di alcune domande più specifiche.
Per processare il resto del dataset si usa un processo di One-Hot Encoding, in cui prendiamo in considerazione le classi "Student Country", "Question Level" e "Topic", ignorande le restanti classi, non importanti ai fini di addestramento del modello, cioè "Student ID", "Question ID", "Keywords" e "Subtopic".

Infine, si divide il dataset con "`train_test_split`" per avere l'80% del dataset destinato al Training Set e il restante 20% per il Test Set. In quanto si è individuato uno sbilanciamento nell'EDA tra il numero di risposte corrette ed errate, in questo processo si utilizza una stratificazione, tale che la proporzione tra risposte corrette ed errate rimane uguale nel set di allenamento e in quello di test.

6. Metodi e Modelli

Per l'addestramento del modello si implementano due tipi di classificazione: la Logistic Regression e la Random Forest.

La Logistic Regression è un modello di Machine Learning che calcola la probabilità di trovare un dato in una classe specifica, restituendo 1 se il dato è presente, 0 altrimenti. È utile per i problemi di classificazione perché traccia una linea che separa le risposte che più si avvicinano a 0 da quelle che si più avvicinano a 1. Traccia un confine decisionale più solido rispetto ad altri modelli: restituisce un valore binario attraverso la funzione sigmoide, che schiaccia il valore forzatamente nell'intervallo $[0, 1]$. Si tratta di un modello semplice ed efficace per un dataset simile a quello utilizzato, ma può riscontrare difficoltà nell'analisi di dataset più complessi, composti da dati non separabili linearmente.

La Random Forest sostituisce, invece, le funzioni algebriche alla logica dei test, combinando molti alberi decisionali per ottenere una previsione più accurata e robusta. L'idea è quella del "potere ai molti", un singolo albero può sbagliare o essere troppo specifico, ma la media di molti alberi tende, invece, ad approssimare l'errore per catturare la reale struttura dei dati.

Nel dataset in analisi la Random Forest cattura le interazioni presenti tra le variabili di diverse colonne che il modello di LR potrebbe ignorare. Si immagini che nel dataset analizzato la Random Forest genera un singolo albero che pone le seguenti domande:

1. Il livello è "Advanced"? ⇒ SI/NO
2. L'argomento è "Algebra"? ⇒ SI/NO
3. Lo studente viene dalla nazione "Italy"? ⇒ SI/NO

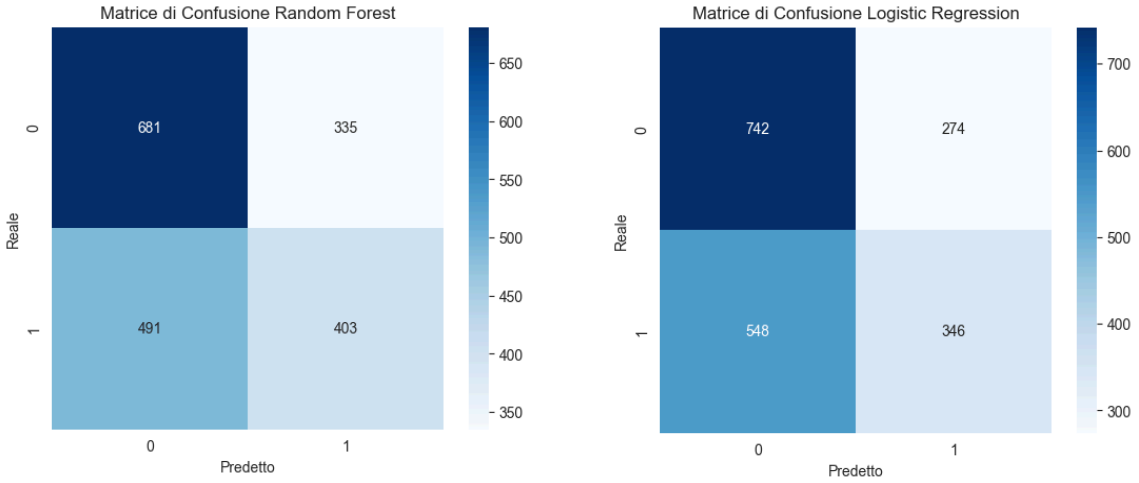
In base alla maggioranza delle risposte restituite si calcola il risultato. Chiaramente, però un singolo albero può essere troppo preciso, ponendo il modello in rischio di overfitting, portando ad errori sul test set.

7. Valutazione e Analisi dei Risultati

Per la valutazione delle tecniche utilizzate per l'addestramento dei modelli si utilizzano le seguenti metriche: Matrice di Confusione, Precision, Recall e F1-Score.

1. Matrice di Confusione

Si calcola la matrice di confusione per entrambi i modelli.



Questa mostra nella diagonale principale il numero di predizioni corrette fatte dal modello addestrato con la precisa tecnica: nel caso della tecnica di RF il numero di predizioni corrette sono 1.084, mentre per la tecnica di LR il numero di predizioni corrette sono 1.088.
Per cui, la tecnica di addestramento del modello di Logistic Regression risulta ottenere dei risultati più accurati rispetto alla tecnica di Random Forest, anche solo per una differenza di 4 predizioni in più corrette.
È, però, da notare che in entrambi i modelli il numero di False Negatives è molto alto, risultato del fatto che i modelli tendono a predire più risposte false che vere in quanto addestrati su una percentuale maggiore di risposte errate che corrette.

2. Precision, Recall e F1-Score

La tabella di seguito è riferita al modello di 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

La tabella di seguito, invece, è riferita al modello di 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

Le tabelle, mostrando un valore di Recall alto per 0 in entrambi i modelli, confermano il bias verso una risposta errata osservato nelle matrici di confusione. Inoltre, dimostrano, di nuovo come già dimostrato dalla Matrice di Confusione, che i due modelli hanno un'accuratezza molto vicina, infatti, tutti i valori di Precision, Recall e F1-Score tra i due modelli sono molto vicini e l'accuratezza finale varia di 0,0021 in favore del modello di Logistic Regression.

8. Conclusioni e Limiti del Modello

Grazie all'analisi approfondita di questo dataset si può comprendere l'importanza del Preprocessing: i dati grezzi contenuti nel dataset scelto devono essere trasformati attraverso Natural Language Processing e One-Hot Encoding per poter essere analizzati computazionalmente dal nostro modello. Sbagliare tale codifica significherebbe dare al modello informazioni errate o distorte.

Il confronto tra Logistic Regression e Random Forest mostra che non esiste un modello assoluto: la Logistic Regression è ottima per la trasparenza, ma scarsa in relazioni lineari tra i dati, d'altra parte il Random Forest cattura pattern complessi come quelli che legano le Classi "Student Country" e "Topic" o "Topic" e "Question Level".

I limiti del modello sono stati individuati nello studio delle matrici di confusione e tabelle di valutazione dei modelli.
Il dataset presenta un maggior numero di risposte errate che corrette, il modello, quindi, tenderà a predire che uno studente sbagli più volte di quante predirà che lo studente non sbagli. I grafici di matrice di confusione e le tabelle di valutazione delle tecniche di addestramento del modello lo mostrano chiaramente:

- Entrambi i modelli hanno una Precision più alta su 0 che su 1, cioè sono più accurati nel predire risposte errate che risposte corrette.
- Entrambi i modelli, però, hanno anche una Recall più alta su 0 che su 1, cioè tendono a predire più volte del necessario che una risposta sia errata rispetto a quante volte lo fanno per una risposta corretta.

In conclusione, i modelli portano il peso di essere stati addestrati su un dataset con degli sbilanciamenti verso le risposte errate, che li porterà, quindi, a essere più corretti nel predire risposte errate, ma allo stesso tempo anche più sensibili in ciò, risultando nella generazione di un numero elevato di False Negatives, piuttosto che di False Positives.