Progetto MOBD 2021-2022

1 Dataset

Il dataset che dovrete analizzare presenta 15 attributi: i primi 14 rappresentano le features mentre l'ultimo rappresenta la classe di appartenenza. Il dataset è misto in quanto contiene sia attributi continui che attributi categorici. Un attributo categorico prevede che il valore di ogni istanza appartenga ad un set di categorie predefinito (e.g. per la feature 'colore' potremmo avere come categorie possibili 'verde', 'giallo' e 'rosso').

2 Obiettivo

L'obiettivo del progetto è risolvere un problema di classificazione binaria. Il classificatore deve prevedere la classe corretta sulla base delle caratteristiche di ciascuna istanza. A seguito di un processo di analisi, dovrà essere prodotta come output **una** (e una sola) routine di addestramento, che preveda tutti gli step di preprocessing e classificazione che sono stati individuati durante l'analisi. È fortemente sconsigliato consegnare più di una routine di addestramento per la valutazione: in questo caso, verrà considerata in fase di correzione la routine che porta a risultati peggiori sul test set.

3 Modalità di svolgimento

Il progetto può essere svolto individualmente o in gruppi composti al massimo da 2 componenti. Il dataset fornito (*training_set.csv*) comprende una porzione delle istanze del dataset originale; la restante parte del dataset non verrà fornita ma sarà utilizzata in fase di correzione per la valutazione finale del progetto. La metrica di valutazione è l'accuratezza sul test set.

Per la correzione verranno considerati due parametri:

- Lo svolgimento del progetto, che tiene conto della fase di analisi svolta (che deve essere documentata chiaramente nella relazione e nel codice fornito) e le scelte implementative fatte;
- Il risultato ottenuto sul test set privato, che verrà valutato anche in relazione ai risultati raggiunti dagli altri gruppi.

I progetti potranno essere consegnati durante le prossime sessioni circa in concomitanza con le date degli esami: verranno stabilite delle deadline entro le quali consegnare il progetto. Non è necessario consegnare il progetto nella stessa sessione in cui si svolge l'esame teorico. Per ogni deadline, verrà comunicato un benchmark, ovvero il risultato migliore ottenuto dai gruppi che già

hanno consegnato sul test set privato. L'obiettivo dovrà essere quindi provare a migliorare, di volta in volta, il benchmark corrente.

Il benchmark iniziale è pari ad un'accuratezza sul test set privato dell'80%. **ATTENZIONE**: il raggiungimento del benchmark iniziale è una condizione necessaria per superare il progetto: se non viene raggiunto tale benchmark, il progetto verrà valutato come insufficiente e dovrete consegnare nuovamente ad un appello successivo.

4 Modalità di consegna

Il materiale da consegnare comprende:

- Una relazione breve ma chiara e completa in cui si giustificano le analisi e le scelte fatte;
- Il codice sorgente del progetto. Il codice deve essere chiaro, leggibile, e parzialmente commentato;
- Una routine di test che, specificato il file di test *test_set.csv*, esegua le operazioni di pre-processing, di classificazione e di valutazione in maniera coerente con le operazioni fatte sul training set;
- Un README in cui si specificano le informazioni necessarie per l'esecuzione della routine di test, tra cui:
  - la directory in cui inserire il file *test_set.csv*;
  - eventuali librerie aggiuntive da installare;

È fortemente consigliato (ma non obbligatorio) l'utilizzo di Google Colab in modo da semplificare la gestione delle librerie e evitare problemi in fase di correzione.
IMPORTANTE: la routine di test **non** deve prevedere l'intero processo di addestramento sul training set. Suggerimento: serializzate su file il modello già addestrato.