Università Magna Graecia di Catanzaro

Corso di Laurea Magistrale in Ingegneria Biomedica

Infrastrutture di calcolo e algoritmi efficienti per la biologia e medicina

Progetti didattici Data Science A.A. 2021/2022

Prof. Mario Cannataro, Dr.ssa Chiara Zucco

Obiettivo del progetto è effettuare un tipico processo di Data Science, utilizzando tecniche statistiche congiunte ad algoritmi di Machine Learning per compiere un task di classificazione del dataset allegato relativo ai dati clinici contenuto nel dataset denominato **"Fetal health"** costruito manipolando i dati descritti in [1].

In particolare, ogni gruppo di candidati deve:

- Effettuare una descrizione introduttiva del dataset, dimostrando di:
    - avere familiarità con il dataset stesso,
    - conoscere a cosa si riferiscono i vari attributi presenti, indicarne il tipo ed esplicitare quale sia l'attributo target,
    - indicare le problematiche preliminari che sono state individuate in termini di numerosità del dataset, presenza di dati mancanti, presenza di duplicati, presenza di dati non consistenti o ridondanti ecc.,
    - considerato il dataset e lo stesso attributo target, è possibile effettuare task di analisi diversi dalla classificazione? Motivare la risposta.
    - considerato il dataset e il problema di predire il valore della variabile Fetal_health, che tipo di task di analisi è possibile svolgere?

- Effettuare un'analisi esplorativa del dataset e discutere i risultati, evidenziando in particolar modo eventuali problematiche riscontrate;

- Effettuare una opportuna fase di Data Cleaning e Data Integration ed eventualmente effettuare nuovamente l'analisi statistica al punto precedente.

- Effettuare, se opportuno, una fase di Data Transformation e/o Reduction.

- Definire un nuovo dataset formato da tutti i soggetti con fetal health= normal e tutti i soggetti con fetal health= pathological

- Definire una baseline di algoritmi di classificazione dicotomica e una o più modalità di convalida incrociata. Comparare i risultati ottenuti in modo consistente e commentare i risultati. Visualizzare matrice di confusione, la curva ROC e le curve di apprendimento;

- Provare a migliorare i risultati ottenuti al punto precedente attraverso una opportuna strategia che comprenda ad esempio un diverso approccio di Data Transformation e/o Data Reduction o una classificazione attraverso metodi di apprendimento d'insieme o un tuning

dei parametri dei classificatori. Riportare tale strategia (preferibilmente in una unica pipeline) e salvare il modello ottenuto in formato pikle o joblib per la fase di valutazione.

- Compiti Opzionali da eseguire in aggiunta alla classificazione dicotomica:

    - Classificazione multiclasse sul dataset di partenza. Lo svolgimento e le performance di questo punto avranno maggior peso ai fini della valutazione.

- Il/Gli script ottenuti vanno presentati in formato .py o ipynb (anche condivisi tramite Google Colab).

- Vanno inoltre presentati:
    - un breve Report (uno per ciascun gruppo) che contenga la descrizione argomentata delle attività eseguite (max 30 pagine A4), dei risultati ottenuti per ogni step del processo, e delle opportune conclusioni
    - una breve presentazione power point (una per ciascun gruppo) che riassuma le varie strategie.
- Si ricorda che la valutazione terrà conto dei seguenti tre aspetti:
    - Performance del modello
    - Originalità dell'approccio
    - Presentazione del lavoro svolto (Qualità del materiale presentato)

Riferimenti

1. Hoodbhoy, Z., Noman, M., Shafique, A., Nasim, A., Chowdhury, D., & Hasan, B. (2019). Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. *International Journal of Applied and Basic Medical Research*, *9*(4), 226.

Allegato 2. Informazioni per la consegna

Tutto il materiale utile ai fini della valutazione deve essere inviato con un'email al Professor Cannataro (cannataro@unicz.i) con in CC chiara.zucco@.unicz.it **ENTRO E NON OLTRE** una settimana prima della data di esame.

Alcune piccole note:

1) Il materiale dovrà essere consegnato in un unico file .zip dal nome "Prog_<NumeroGruppo>.zip" (dove <NumeroGruppo> è il numero del gruppo di progetto (come riportato nel file condiviso); ad esempio, se <NumeroGruppo>=3, allora il file da consegnare si chiamerà "Prog_3.zip") e dovrà comprendere i seguenti file:

   a) Cartella “Prog_<NumeroGruppo>” comprensiva dei seguenti file:

      i) “Codice_ Prog_<NumeroGruppo>”, contenente il codice relativo al progetto (.py o ipynb per Progetto Data Mining) o eventualmente il link di condivisione del file Colab;

      ii) “Relazione_<NumeroGruppo>.pdf”, contenente la relazione dell'attività di progetto.

2) **IMPORTANTE**: sarà cura del/dei candidato/i accertarsi della presenza e della corretta collocazione dei file/directory necessari per l'esecuzione dei codici presentati. A questo fine, il file .zip deve contenere anche **tutti i file necessari** (quindi anche i file relativi ai dataset forniti), opportunamente richiamati nel codice. Allo stesso modo se si condivide un file Colab questo deve contenere una cella per l’opportuno caricamento del dato in modo da renderlo eseguibile a tutti. Se i candidati usano librerie non viste durante le esercitazioni, devono allegare un file “Requirements.txt” contenente la lista delle suddette librerie. **L'impossibilità di eseguire il progetto avrà come conseguenza la valutazione negativa dell'elaborato.**

3) Si ricorda che la Presentazione Stilistica del Progetto di Gruppo è uno degli elementi di valutazione. Sarà cura del/dei candidato/i che gli eventuali grafici e tabelle risultino visibili.