

Machine Learning

Progetto d'esame

Progetto d'esame (1)

- Il progetto d'esame prevede di essere svolto in gruppo - **max 3 persone**
- Ogni gruppo deve identificare un dominio di suo interesse per il quale intende indurre almeno **2 modelli** di classificazione supervisionata e/o non supervisionata, individuando il relativo dataset.
- I **dataset** possono essere acquisiti da siti di benchmark quali:
 - UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>)
 - KAGGLE (www.kaggle.com)
 - AWS (https://registry.opendata.aws/?source=post_page-----bb6d0dc3378b-----)
 - GITHUB (<https://github.com/awesomedata/awesome-public-datasets>)
- In alternativa, i dataset possono essere creati ottenendoli da siti individuati dagli studenti sotto forma di open data, dati aziendali, dati ottenuti da scraping di siti web, dati in streaming prodotti da sensori o canali social.

Progetto d'esame (2)

- **Materiale da consegnare:**

1. Codice sorgente (auto-consistente)
2. Dataset
3. Relazione
4. Presentazione

- **Data di consegna:**

- **2 giorni** prima dell'appello orale (vedi Date Appelli d'Esame).

- **Modalità di consegna:**

- la consegna di tutto il materiale (codice sorgente, dataset, relazione e presentazione) deve avvenire mediante condivisione di un folder **Google Drive** con i seguenti indirizzi: elisabetta.fersini@unimib.it e claudio.ferretti@unimib.it.
- il folder deve contenere un Readme.txt che indichi Nome, Cognome e Matricola di ogni partecipante al gruppo

Progetto d'esame (3)

- **Relazione e presentazione** devono contenere le seguenti parti:
 - Descrizione del dominio di riferimento e obiettivi dell'elaborato
 - Scelte di design per la creazione del data set, eventuali ipotesi o assunzioni
 - Descrizione del training set: analisi esplorativa del training set (analisi delle covariate e/o PCA)
 - Descrizione e motivazione dei modelli di machine learning scelti (almeno due modelli)
 - Esperimenti:
 - Esecuzione di una 10-fold cross validation e stima delle seguenti misure di performance:
 - Matrice di confusione complessiva
 - Precision, recall, f-measure, ROC e AUC
 - Analisi dei risultati ottenuti
 - Conclusioni

Progetto d'esame (4)

- Linguaggio di programmazione:
 - R, salvo casi eccezionali (es.: studenti non vedenti) da comunicare via email direttamente a elisabetta.fersini@unimib.it
- Presentazione:
 - La presentazione avrà una durata massima di 15 minuti, durante i quali ciascun membro del gruppo dovrà esporre parte del lavoro realizzato