ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Progetto del Corso

Data Analytics

Master in Computer Science (Second cycle degree)
University of Bologna

Prof. Marco Di Felice Prof. Giuseppe Lisanti

Department of Computer Science and Engineering, University of Bologna

marco.difelice3@unibo.it, giuseppe.lisanti@unibo.it

Progetto del Corso

- L'esame del corso di Data Analytics, a.a. 2022/2023, consiste di due prove: **prova orale** e **progetto**, entrambe **OBBLIGATORIE**.
- Le prove maturano voti distinti, il voto finale si calcola come media delle due prove:

$$\mathbf{Voto_{FINALE} = 0.5 * Voto_{PROGETTO} + 0.5\ Voto_{ORALE}}$$

- Le due prove avvengono contestualmente (nella stessa data). A valle della discussione del progetto, è previsto l'esame orale con domande di teoria sugli argomenti trattati durante il corso.

Progetto del Corso

- Il progetto può essere svolto **singolarmente** o in gruppi di massimo **DUE unità**.
- Sono previste le seguenti deadline di consegna del progetto:

1 Febbraio 2023, 1 Marzo 2023, 1 Aprile 2023, 1 Maggio 2023, 1 Giugno 2023, 1 Luglio 2023, 15 Settembre 2023

- Occorre consegnare i sorgenti ed una relazione (con le seguenti sezioni di massima: Introduzione, Metodologia, Implementazione, Risultati), lingua a scelta (inglese preferibile)

Progetto del Corso

- A valle della consegna, occorre preparare una **presentazione con slide** per la discussione del progetto: durante la presentazione, può essere chiesta una demo.
- Tutti i membri del gruppo devono essere presenti durante la discussione (**IN PRESENZA**), e devono conoscere il 100% del progetto svolto. La ripartizione del lavoro deve essere equa.
- La prova orale è indipendente dal progetto e verte su **TUTTI** gli argomenti del corso (anche quelli non trattati dal progetto).

Progetto del Corso

- Il progetto può essere svolto su **dataset/task** proposti dallo studente
(vedi punto sotto) oppure sviluppando il dataset/task proposto dai docenti. E' possibile estendere/modificare/customizzare quest'ultimo a piacimento, a patto di non ridurne la complessità.

- Nel caso di dataset proposto dallo studente, va richiesta l'approvazione da parte dei docenti. Il **dataset** deve essere significativo (sia in termini di numero di dimensioni, sia in termini di numero di istanze) e giustificare l'utilizzo di tecniche di data analytics al fine di estrarre conoscenza.

Progetto del Corso

- Il progetto consiste nella realizzazione di uno studio di data analytics, implementando **TUTTE le fasi della pipeline** viste a lezione:
  - Data Acquisition
  - Data Visualization
  - Data Preprocessing (quelle necessarie per il task)
  - Modeling (con tuning degli iperparametri)
  - Performance Evaluation

Progetto del Corso

- Il progetto prevede lo sviluppo di **funzionalità indipendenti, e con complessità crescente.**
- Ogni funzionalità è associata ad un **punteggio massimo** che si può conseguire se tali funzionalità sono state sviluppate nel modo corretto (a discrezione dei docenti).
- Il punteggio massimo che si può conseguire è **30 e lode**.
- Il progetto può essere ritenuto insufficiente (in tale caso, si devono ripetere la consegna e discussione seguente).
- In tutti gli altri casi, **NON sono previste riconsegne**.

Progetto del Corso

- **FUNZIONALITA' 1**. Il progetto prevede l'utilizzo di tecniche di ML supervised tradizionali non deep (metodi Bayesiani, alberi, KNN, SVM , etc) → **PUNTEGGIO MASSIMO OTTENIBILE: 25**
- **FUNZIONALITA' 2**. In aggiunta alla funzionalità 1, il progetto include l'utilizzo di tecniche di ML supervised basate su reti neurali, con utilizzo OBBLIGATORIO del framework Pytorch → **PUNTEGGIO MASSIMO OTTENIBILE: 28**
- **FUNZIONALITA' 3.** In aggiunta alle funzionalità 1 e 2, il progetto prevede l'utilizzo di almeno una tecnica di ML supervised con modelli deep per *Tabular* Data di recente introduzione (next slide)
  → **PUNTEGGIO MASSIMO OTTENIBILE: 30 e lode**

Progetto del Corso

- Nel caso della FUNZIONALITA' 3, far riferimento ai **modelli deep per Tabular Data** disponibili qui:

  https://github.com/manujosephv/pytorch_tabular

- Tali modelli **NON sono stati presentati durante il corso**.

- In caso di sviluppo della FUNZIONALITA' 3, si chiede anche di **inserire 3-4 slide di teoria** durante la presentazione del progetto, finalizzate ad illustrare caratteristiche di tali modelli (chiaramente, dimonstrandone di aver compreso il funzionamento).

Techniques

- Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data
- Mixture Density Networks
- TabNet: Attentive Interpretable Tabular Learning
- AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks
- TabTransformer
- Revisiting Deep Learning Models for Tabular Data (FT Transformer)

Progetto del Corso

DATASET https://grouplens.org/datasets/movielens/

Il dataset contiene dati provenienti da **MovieLens**, un recommendation system per contenuti video (film). Il dataset contiene rating e tag per piu` di 60000 film. I dati sono stati raccolti da più di 150K utenti negli anni 1995-2019. Ogni file dispone di un genoma che ne identifica una caratteristiche (es. ”world war I”) e rilevanza della stessa (0.75). Si rimanda alla lettura del file README (contenuto nell’archivio) per la descrizione del contenuto dei file. **L’obiettivo dello studio è predire il voto medio di un film, a partire dalle sue caratteristiche.**

GPU

Ai fini del progetto potete utilizzare Google Colab per addestrare i vostri modelli (Colab Tutorial).

Potete anche fare richiesta di accesso al Cluster GPU messo a disposizione dal Dipartimento DISI (il quale utilizza SLURM) scrivendo a tecnici@cs.unibo.it e specificando il corso, il docente/i e il motivo per cui avete bisogno di tali risorse; ricordatevi di metterci in copia alla mail in cui fate richiesta.

Progetti speciali