Machine Learning **Anno accademico 2024-25**

Prof. Giuseppe Polese
Prof.ssa Loredana Caruccio

Specifiche per lo svolgimento del progetto

Il progetto ha l'obiettivo di realizzare una (piccola) applicazione/soluzione di Machine Learning (ML), la quale potrà essere svolta in gruppo (di massimo due persone) ed esposta ai docenti del corso. La soluzione proposta dovrà applicare, nei vari step della pipeline di ML, gli approcci e le tecniche studiate nell'ambito del corso. Ad esempio, verranno valutate positivamente una corretta analisi e risoluzione delle problematiche legate alla presenza di missing value, di outlier, di dati sbilanciati, feature ridondanti, ecc. Inoltre, risulta importante giustificare la selezione del/dei modello/i, valutare le performance (almeno con le metriche di valutazione tradizionali) e i tipi di errore effettuati dal modello. Nota: non è strettamente necessario realizzare soluzioni che impieghino i soli modelli studiati nell'ambito del corso; tuttavia, è necessario dimostrare padronanza nell'esposizione di tutte le tecniche impiegate per lo sviluppo del progetto.
Il progetto dovrà essere consegnato tramite il link messo a disposizione sulla piattaforma e-learning allegando un file .zip che contenga il codice (o eventualmente un link ad un repository GitHub accessibile) e un report di progetto nel quale venga dettagliato:

1. Lo scenario/problema/task analizzato.
2. Il dataset utilizzato e le sue caratteristiche.
3. Le issue analizzate, le soluzioni proposte e le scelte di progettazione effettuate.
4. L'analisi delle prestazioni, che comprenda dati numerici e grafici.
5. Considerazioni finali e possibili sviluppi futuri.

N.B.: I risultati del progetto dovranno essere presentati nelle sessioni di discussione attraverso una presentazione (Power Point, Google Slides, ecc.).

Di seguito sono riportati alcuni scenari e dataset di **esempio**. Saranno ben accolti progetti su problematiche/task alternativi a quelli presentati di seguito.

- **U.S.-Airbnb-Open-Data:** Una raccolta di più dataset trovati su Inside Airbnb con informazioni riguardanti diversi attributi degli immobili (contiene dati testuali come il nome dell'annuncio che si possono ignorare).
- **Cirrhosis Patient Survival Prediction:** Un dataset con 17 feature di parametri clinici per predire lo stato di sopravvivenza dei pazienti con cirrosi epatica. Le classi includono 0 = D (death), 1 = C (censored), 2 = CL (censored due to liver transplantation).
- **Spaceship Titanic:** Un dataset realizzato nell'ambito di una competition di Kaggle nel quale sono presenti dati mancanti e nuove feature.

- **Regression with an Insurance Dataset:** Un dataset costruito con l'obiettivo di facilitare lo sviluppo e il test di modelli di regressione per la previsione dei premi assicurativi in base a varie caratteristiche del cliente e dettagli della polizza.
- **Bank Marketing:** Una raccolta di dati relativi alle campagne di marketing diretto di un istituto bancario portoghese. Le campagne di marketing si basavano su telefonate, al fine di verificare se il prodotto (deposito bancario a termine) sarebbe stato sottoscritto (sì) o non sottoscritto (no).
- **Predict Students' Dropout and Academic Success:** Un dataset che mira alla riduzione dell'abbandono accademico e dell'insuccesso nell'istruzione superiore, utilizzando tecniche di apprendimento automatico per identificare gli studenti a rischio in una fase precoce del loro percorso accademico. Il problema è formulato come un compito di classificazione a tre categorie (dropout, enrolled, e graduate) alla fine della normale durata del corso.