Progetti finali ASML 2025/2026

Leonardo Egidi, Nicola Torelli

4 dicembre 2025

Presentazione progetti finali

La presentazione dovrebbe includere quanto segue:

- una descrizione dell'obiettivo del progetto;
- un'analisi esplorativa dei dati, inclusa una possibile fase di pulizia dei dati;
- selezione, descrizione ed eventualmente confronto dei modelli di apprendimento statistico adottati più adatti;
- commenti sui risultati.

Istruzioni:

- Libera aggregazione in gruppi, al massimo 4 membri per gruppo.
- Iscrivetevi su esse3 all'appello desiderato: tutti i membri di un gruppo devono iscriversi allo stesso appello.
- Ogni dataset ha una sua descrizione e un link. Prima di iniziare a fare analisi, leggere attentamente la descrizione di dataset e variabili.

- Per alcuni file di dati probabilmente troverete alcune analisi in rete eseguite da altri. Potete guardarle e trarre un aiuto o un'ispirazione, ma cercate di trovare una chiave originale per l'analisi, la presentazione e la discussione dei risultati.
- *Cum grano salis*, ma potete aiutarvi anche con ChatGPT/Copilot.
- Ogni presentazione deve durare 25 minuti. Tutti i membri del gruppo devono essere a conoscenza di tutte le parti del progetto e prendere parte alla presentazione: ogni membro di un gruppo da 3 parlerà, dunque per 8/9 minuti.
- Gli iscritti all'appello sono caldamente invitati a sentire le relazioni degli altri gruppi iscritti al medesimo appello.
- Siete liberi di scegliere il tipo di file per organizzare la vostra presentazione, ad esempio solo le slides, oppure un report (con Rmarkdown o con altri strumenti a scelta, anche powerpoint se preferite).
- Siete liberi di scegliere un altro dataset di vostro gradimento, previa consultazione con i docenti.
- Spedite ai docenti via mail la presentazione, se possibile, il giorno/la sera prima dell'appello.
- *Non esitate a contattarci per qualsiasi problema* - mail e/o ricevimento.

Elenco progetti selezionati

Regressione

- **'Cars'** dataset: scaricabile qui: https://www.kaggle.com/datasets/CooperUnion/cardataset, 11914 righe e 16 colonne. Ogni riga rappresenta una singola macchina venduta, con certe caratteristiche. Utilizzare la variabile risposta prezzo di listino `MSRP` e usare modelli/algoritmi di apprendimento statistico per modellarla e prevederla in base alle covariate presenti. Quali sono le variabili che meglio spiegano il prezzo di listino? Qual è il metodo migliore per prevederlo/modellarlo? *Suggerimenti*: condurre una ricca analisi esplorativa, considerando correlazioni tra le covariate continue e sondando la

potenziale presenza di effetti non lineari sulla variabile risposta. Eventualmente, trasformare la variabile risposta. Dividere in training/test set e calcolare MSE per diversi algoritmi.

- **'College'** dataset: scaricabile qui: https://www.statlearning.com/resources-first-edition e disponibile sul pacchetto `ISLR2`, 777 righe e 18 colonne. Contiene statistiche per un grande numero di college americani dal 1995. Costruire modelli di regressione e machine learning per la variabile risposta `Apps`, il numero di domande ricevute. Quali sono le variabili influenti? Qual è il metodo migliore per prevedere il numero di applicazioni ricevute? *Suggerimenti*: condurre una ricca analisi esplorativa, considerando correlazioni tra le covariate continue e sondando la potenziale presenza di effetti non lineari sulla variabile risposta. Eventualmente, trasformare la variabile risposta. Dividere in training/test set e calcolare MSE per diversi algoritmi.
- **'Flight'** dataset: scaricabile qui: https://www.kaggle.com/code/varunsaikanuri/flight-fare-prediction-10-ml-models., 300153 righe, 11 colonne. Il prezzo di un biglietto aereo è influenzato da diversi fattori, come la durata del volo, i giorni rimanenti alla partenza, l'orario di arrivo e di partenza, ecc. Le compagnie aeree possono diminuire i costi nel momento in cui hanno bisogno di costruire il mercato e quando i biglietti sono meno accessibili. Costruire modelli di regressione/machine learning per la variabile `price` e spiegare quali variabili influenzino di più il prezzo del biglietto. *Suggerimenti*: cercare di sondare l'impatto 'causale' delle covariate presenti sul prezzo del biglietto aereo. Usare vari algoritmi di regressione e compararli usando una partizione tra training e test set.
- **'Housing'** dataset: scaricabile qui https://www.kaggle.com/datasets/yasserh/housing-prices-dataset, 545 righe e 13 colonne, ogni riga esprime il prezzo di una casa. Usare la variabile risposta `price` e costruire modelli/algoritmi di apprendimento statistico per modellare e prevedere il prezzo di una casa in base alle variabili esplicative disponibili. Quali sono le variabili che meglio spiegano il prezzo di una casa?
- **'Videogames'** dataset: scaricabile qui https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings, 16719 righe e 16 colonne. Riporta le vendite a dicembre 2016 dei maggiori videogiochi presenti sul mercato. Ogni riga esprime le caratteristiche di un dato videogioco. Usare come variabile risposta la variabile `Global_Sales` e usare modelli/algoritmi di apprendimento statistico per prevedere e modellare questa grandezza in funzione delle variabili disponibili. *Suggerimenti*: starer attenti a come valutare e usare le covariate categoriali. Eventualmente, trasformare la variabile risposta.

Classificazione

- **'Breast'** dataset: scaricabile qui https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download, 569 righe e 33 colonne. Le 30 variabili disponibili descrivono misure derivate da immagini digitalizzate di una massa mammaria — in particolare misurano caratteristiche del nucleo cellulare (forma, dimensione, "texture", area, perimetro, concavità, simmetria, ecc.). Il target è binario, usare la variabile risposta binaria `diagnosis`: "M" (maligno – malignant) oppure "B" (benigno – benign). Nel dataset ci sono 212 casi maligni e 357 benigni. Usare modelli di classificazione/machine learning per prevedere l'esito e capire quali variabili spieghino meglio la probabilità di avere un tumore maligno. *Suggerimenti*: usare più algoritmi di classificazione e costruire metriche predittive (ad es. AUC, curve ROC).
- **'Bank_marketing'** dataset: https://www.kaggle.com/code/pkdarabi/marketing-campaign-patterns?select=bank_customers_train.csv., 39188 righe, 21 colonne. Per migliorare l'efficacia delle future campagne di marketing per un istituto finanziario, è necessario analizzare i modelli della precedente campagna di marketing. In questo modo, possiamo identificare le migliori strategie da implementare per ottenere un maggiore successo nelle campagne future. Usare la variabile `y` (risposta positiva/negativa del cliente alla campagna) come variabile risposta e usare modelli di classificazione/machine learning per prevedere l'esito e capire quali variabili spieghino meglio la probabilità di rispondere positivamente alla campagna. *Suggerimenti*: usare più algoritmi di classificazione e costruire metriche predittive (ad es. AUC, curve ROC).
- **'Churn'** dataset: fornito dai docenti su moodle, 9969 righe, > 20 colonne. Il dataset contiene

informazioni sull'abbandono dei clienti di una data compagnia assicurativa. La variabile risposta è `churn`, variabile binaria che è uguale a 1 se il cliente ha abbandonato la compagnia, e 0 se invece non l'ha abbandonata. Costruire algoritmi di classificazione per la previsione dell'abbandono comparandoli in base alle performance predittive studiate in classe e studiare quali covariate influenzino di più e come la scelta dei clienti. *Suggerimenti*: valutare in via preliminare quante e quali sono le variabili rilevanti, potenzialmente utilizzando tecniche quali LASSO. Usare diversi algoritmi di classificazione binaria e costruire metriche predittive (AUC, curve ROC).

- **'Marketing_campaign'** dataset: scaricabile qui https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign, 2240 righe, 25 colonne. Un modello di apprendimento statistico può dare una spinta significativa all'efficienza di una campagna di marketing sia aumentando l'adesione alla campagna che riducendo le spese. L'obiettivo è prevedere chi risponderà a un'offerta per un prodotto o servizio. L'obiettivo principale è quello di addestrare un modello predittivo che permetta all'azienda di massimizzare il profitto della successiva campagna di marketing. La variabile risposta è `Response`, uguale a 1 se si è accettata un' offerta promozionale nella precedente campagna, e 0 altrimenti. Vi sono 5 variabili che rguardano l'aver accettato l'offerta in 5 diverse occasioni. Potrebbe essere opportuno creare una singola variabile che assuma il valore 1 se il cliente ha accettato l'oferta in almeno una delle campagne. *Suggerimenti*: tentare di sondare l'impatto 'causale' delle covariate presenti sull'accettazione dell'offerta promozionale.
- **'Travel_ticket_cancellation'** dataset: scaricabile qui https://www.kaggle.com/datasets/pkdarabi/classification-of-travel-purpose, 101017 righe e 22 colonne. L'obiettivo è quello di sviluppare un modello che preveda se gli utenti annulleranno i loro biglietti, dove la variabile riposta è `Cancel`, 0 se non si cancella e 1 se si cancella. Ogni cancellazione comporta una multa per il sito di registrazione dei biglietti da parte della compagnia aerea. Pertanto, è fondamentale identificare i biglietti che hanno probabilità di essere annullati, consentendo una gestione efficace del rischio di cancellazione all'interno dell'azienda. Vi sono oltre 20 possibili variabili che possono esser utilizzate per costuire l'algoritmo di classificazione (alcune si noti che sono inutili come il numero identificativo del biglietto). *Suggerimenti*: usare diversi algoritmi di classificazione binaria e costruire metriche predittive (AUC, curve ROC).