Esercizio

1. Analizzare i dati a vostra disposizione e correggere eventuali errori presenti nell'insieme di dati

(dati mancanti, dati palesemente errati, etc...)

2. Individuare quali sono gli attributi che sembrano maggiormente correlati alla variabile target

3. Costruire almeno un paio di modelli previsivi, appartenenti a tipologie diverse (ad esempio alberi

di classificazione e classificatori a regole)

4. Studiare l'accuratezza dei modelli ottenuti

1. Preprocessing dei dati

* Importazione delle librerie: Importa le librerie necessarie, come Pandas, NumPy e Scikit-learn, che forniscono funzionalità per la manipolazione dei dati e l'analisi.
* Esplorazione dei dati: Esplora il dataset per ottenere familiarità con i dati. Controlla le prime righe del dataframe, ottieni informazioni sulla struttura dei dati, verifica la presenza di valori mancanti o outlier e analizza le statistiche di base.
* Rimuovere valori identificativi doppi
* Gestione dei valori mancanti: Tratta i valori mancanti nel dataset. Puoi eliminare le righe o colonne con valori mancanti, oppure imputare i valori mancanti con metodi come media, mediana o interpolazione.
* Gestione degli outlier: Identifica e gestisci gli outlier presenti nei dati. Gli outlier possono essere trattati eliminando le righe che li contengono o tramite tecniche di trasformazione dei dati come la winsorizzazione o la normalizzazione.
* Trasformazione dei dati: Applica trasformazioni ai dati, se necessario. Ad esempio, puoi standardizzare le variabili numeriche per far sì che abbiano media zero e deviazione standard unitaria.
* Codifica delle variabili categoriche: Se il dataset contiene variabili categoriche, codificale in modo appropriato. Ad esempio, puoi utilizzare la codifica one-hot per trasformare variabili categoriche in variabili binarie.
* Come migliorare risultato classificatori
* Addestrare e usare i classificatori anche sul new\_df
  + Fare il new df considerando anche solo le feature con correlazione positiva
* **NON** usare l’SVM
* Applicare lo SMOTE al dataset ?