*** NOTE SULLA STRUTTURA DEL REPORT ***

1. **Abstract** (10-15 righe)

2. Scelta del dataset

Motivare la scelta del dataset e fornire breve sommario delle sue caratteristiche.

-- Dimensioni del dataset, con numero e tipo di feature e numero di osservazioni.
-- Eventuale presenza di valori mancanti (che richiede quindi una fase di data cleansing).
-- Verifica di bilanciamento o sbilanciamento dei dati.
-- Breve descrizione delle feature (con eventuali statistiche)

NB Interessa poco la correlazione tra feature. Può interessare invece la correlazione tra feature e target (se il problema è di classificazione o regressione, ovviamente). Va però detto che se c'è un gruppo di feature (al limite anche una sola) altamente correlato con il target allora il problema è facile. Ma questo tipo di problemi non è interessante per il ML. Interessano invece problemi in cui non ci sia questo indicatore forte di "facilità"; nel senso che il problema potrebbe essere facile, ma per effetto di una combinazione di feature che lo rendono tale...

3. Tipologia del problema e architettura della soluzione

-- Indicare la tipologia di problema (classificazione, regressione, clustering, predizione).
-- Indicare la scelta del modello (o dei modelli), motivandola in 2-3 righe.
-- Indicare la strategia scelta per training e test.
-- Indicare se si intende utilizzare una strategia per l'ottimizzazione degli iperparametri dei modelli.
-- Indicare molto brevemente (es. 5-6 righe) come si intendono gestire i dati (da sviluppare nella sezione successiva)

4. Gestione dei dati

Indicare in particolare:

-- Data cleansing (opzionale; dipende dal dataset).
-- Data preprocessing (feature reduction e/o selection, scaling, etc.).
-- Eventuale presenza di sbilanciamento dei dati.

NB Si può fare feature selection anche utilizzando preliminarmente una RF; infatti le RF restituiscono oltre al modello anche la cosiddetta "feature importance".

NB L'eventuale sbilanciamento dei dati ha tipicamente un grande impatto sulla scelta del modello, e bisogna COMUNQUE gestirlo. Infatti, in presenza di grande sbilanciamento occorre tipicamente utilizzare a) un ensemble con strategia di bagging oppure b) data augmentation. Per gli ensemble: supponiamo per esempio che la percentuale della classe

minoritaria sia 10%. In tal caso sceglieremmo random il 7-8% degli esempi da quella classe e circa l'1% degli esempi della classe maggioritaria. Per la data augmentation: consultare wikipedia o un chatBot.

NB Ricordare che le RF sono il modello più robusto rispetto allo sbilanciamento.

5. Training e test del modello (o dei modelli) + valutazione delle prestazioni

-- Indicare con qualche dettaglio in più il motivo della scelta del modello (se si sperimentano più modelli motivarne comunque la scelta).
-- Descrivere le eventuali tecniche utilizzate per l'ottimizzazione degli iperparametri dei modelli.
-- Descrivere la soluzione indicando come viene stratificato il software (es. DAG delle procedure o delle classi).
-- Per training e test usare training e test classico oppure k-fold CV.
-- Scegliere le misure di valutazione delle prestazioni (ricordare che sono misure e non metriche).
-- Valutare i risultati (riportare obbligatoriamente delle tabelle in caso di più misure).

NB Non soffermarsi mai sul "come" (es. non parafrasare il codice con pseudo-codice). Ci si deve concentrare soltanto sul "cosa" e sulla stratificazione del software.

NB Se il numero di osservazioni è almeno dell'ordine delle decine di migliaia non occorre usare k-fold CV, anche perché potrebbe essere computazionalmente pesante...

NOT LEAST OF ALL: Non scrivere la relazione come un "racconto", tipo ho fatto questo e quello...