# Big Data Analytics
## Comunicazione, marketing e innovazione digitale
## Università degli Studi di Teramo

# Progetto di esame

## Obiettivo del progetto

L'obiettivo del progetto è dimostrare la capacità di:

1. comprendere un dataset reale
2. preparare correttamente i dati (pre-processing)
3. applicare tecniche di machine learning **supervisionato** e **non supervisionato**
4. valutare criticamente i risultati ottenuti
5. comunicare in modo chiaro e motivato le scelte effettuate

## Dataset

Gli studenti possono:

- utilizzare **un dataset assegnato dal docente**, **oppure**
- scegliere **un dataset pubblico** (Kaggle, UCI, open data), previa breve descrizione e approvazione

Il dataset deve:

- contenere **almeno 1.000 righe**
- includere **variabili numeriche e/o categoriche**
- permettere sia analisi supervisionate sia non supervisionate

## Parte 1 — Comprensione del dataset (obbligatoria)

Gli studenti devono fornire:

- descrizione del contesto del dataset
- significato delle variabili principali
- identificazione di:
    - variabile/e target (se presenti)
    - variabili predittive

- prime osservazioni esplorative (distribuzioni, correlazioni, sbilanciamenti)

*Questa sezione valuta la capacità di "leggere" i dati prima di modellare.*

# Parte 2 — Pre-processing dei dati (obbligatoria)

Gli studenti devono **documentare e motivare** tutte le operazioni di preparazione dei dati, ad esempio:

## 2.1 Pulizia dei dati

- gestione dei valori mancanti (rimozione, imputazione, motivazione)
- gestione di outlier o valori anomali
- rimozione di colonne non informative o ridondanti

## 2.2 Trasformazioni

- codifica di variabili categoriche
- normalizzazione o standardizzazione (se necessaria)
- creazione di nuove feature (feature engineering), se rilevante

*Non conta "fare tutto", ma sapere **perché** si fa una certa operazione.*

# Parte 3 — Machine Learning supervisionato (obbligatoria)

Applicare **almeno un modello supervisionato**, coerente con il dataset:

- **classificazione** (target categorico)

  oppure

- **regressione** (target numerico)

Gli studenti devono:

## 4.1 Costruzione del modello

- suddivisione training/test (o cross-validation)
- scelta del modello (motivata)
- configurazione degli iperparametri principali

## 4.2 Valutazione

- utilizzo di metriche appropriate:
    - classificazione: accuracy, precision, recall, F1, confusion matrix
    - regressione: RMSE, MAE, $R^2$
- confronto tra almeno **due configurazioni o modelli**, se possibile

*Si valuta la correttezza metodologica, non la performance assoluta.*

# Parte 4 — Machine Learning non supervisionato (obbligatoria)

Applicare **almeno una tecnica non supervisionata**, ad esempio:

- clustering (es. K-means, hierarchical clustering)
- riduzione della dimensionalità (es. PCA)
- analisi esplorativa basata su similarità

Gli studenti devono:

- motivare la scelta dell'algoritmo
- descrivere i parametri principali
- interpretare i risultati (cluster, componenti, pattern)
- discutere **il senso** dei risultati rispetto al dominio del dataset

*L'obiettivo non è "indovinare" cluster giusti, ma interpretare strutture nei dati.*

# Parte 5 — Confronto e riflessione critica (obbligatoria)

Gli studenti devono discutere:

- differenze tra analisi supervisionata e non supervisionata sul dataset
- punti di forza e limiti dei modelli utilizzati
- eventuali problemi riscontrati (overfitting, sbilanciamento, rumore)
- possibili miglioramenti futuri del lavoro svolto

# Output richiesti

## Obbligatori

- **Relazione scritta** (PDF, ~10–15 pagine)
- **Workflow / codice** (KNIME workflow, notebook, script)
- **Breve presentazione** (slide) che riassuma:
    - dataset
    - scelte principali
    - risultati

### Facoltativi (bonus)

- visualizzazioni avanzate
- confronto con modelli alternativi
- interpretabilità del modello

# Criteri di valutazione

| Aspetto | Peso |
|---|---|
| Comprensione del dataset | 15% |
| Pre-processing | 25% |
| Analisi non supervisionata | 20% |
| Analisi supervisionata | 25% |
| Chiarezza e riflessione critica | 15% |