Heart Failure Prediction Dataset

Il **dataset** si trova al link https://bit.ly/3CLdwU7 (preso da kaggle - https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction) e contiene dati relativi ad alcuni pazienti:

- **Age**: età del paziente [anni]
- **Sex**: sesso del paziente [M: maschio, F: femmina]
- **ChestPainType**: tipo di dolore toracico [TA: Angina tipica, ATA: Angina atipica, NAP: Dolore non anginoso, ASY: Asintomatico]
- **RestingBP**: pressione sanguigna a riposo [mm Hg]
- **Cholesterol**: colesterolo sierico [mm/dl]
- **FastingBS**: zucchero nel sangue a digiuno [1: se FastingBS > 120 mg/dl, 0: altrimenti]
- **RestingECG**: elettrocardiogramma a riposo [Normale, ST: anomalia dell'onda T-ST, LVH: ipertrofia ventricolare sinistra]
- **MaxHR**: frequenza cardiaca massima raggiunta [valore numerico tra 60 e 202]
- **ExerciseAngina**: angina indotta da esercizio [SÌ, NO]
- **Oldpeak**: valore numerico misurato in depressione
- **ST_Slope**: pendenza del picco [Up: in salita, Flat: piatto, Down: in discesa]
- **HeartDisease**: classe da predire [1: cardiopatia, 0: normale]

Trasformazione del dataset e predizione con le pipeline

1. A partire dal dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare eventuali istanze che contengono valori nulli, trasformare opportunamente valori categorici e dividere il dataset in modo che 3/4 degli elementi siano contenuti in un nuovo dataset “**train**” e 1/4 nel dataset “**test**” preservando le proporzioni delle classi nel target.
2. Creare una pipeline in cui i valori degli attributi Age, RestingBP, Cholesterol e MaxHR sono discretizzati in 5 intervalli (**KBinsDiscretizer**) e gli altri attributi sono lasciati invariati. Valutare le performance di predizione del modello **DecisionTree**.
3. Aggiungere alla pipeline precedente la funzione **SelectKBest**. Utilizzare la funzione di **gridSearchCV** per selezionare: il valore migliore di K di SelectKBest, il numero migliore di intervalli in cui discretizzare i valori di Age, RestingBP, Cholesterol e MaxHR e i valori migliori di criterion e max_depth del modello DecisionTree.
4. Creare una nuova pipeline che applica la decomposizione **TruncatedSVD** al dataset del punto 1 e aggiunge le componenti ottenute alle componenti della pipeline del punto 2. Valutare il valore migliore per il numero di componenti di TruncatedSVD tra 2, 4 e 6.
5. Creare una pipeline in cui prima si applica la normalizzazione (**Normalizer**) a tutto il dataset del punto 1 e poi i valori di Oldpeak sono scalati nell’intervallo 0-1 (**MinMaxScaler**). Valutare le performance del modello **DecisionTree**.