Esercizio: Predizione del prezzo delle auto

Preso da Kaggle (https://www.kaggle.com/datasets/hellbuoy/car-price-prediction), il dataset è scaricabile attraverso il link **http://bit.ly/3lHknB2** e contiene dati relativi a diversi modelli di auto. L'obiettivo è predire il prezzo dell'auto. Sono presenti le seguenti feature:

- **Car_ID**: identificativo dell'auto
- **Symboling**: valutazione del rischio assicurativo (da +3 rischiosa a -3 abbastanza sicura)
- **carCompany**: nome della società automobilistica
- **fueltype**: tipo di carburante, ad es. gas o diesel
- **aspiration**: aspirazione usata nell'auto
- …
- **citympg**: consumo in città (miles per gallon)
- **highwaympg**: consumo in autostrada (miles per gallon)
- **price**: prezzo dell'auto

Trasformazione e Predizione

1. Caricare il dataset, eliminare eventuali attributi inutili (giustificare la scelta), eliminare eventuali istanze duplicate, eliminare le istanze che contengono valori nulli, trasformare opportunamente i valori categorici (consiglio: usare Label Encoder) e dividere il dataset in train (3/4 del dataset) e test (1/4). Calcolare e valutare le predizioni di un RandomForestRegressor. Effettuare alcune considerazioni sui risultati ottenuti, calcolando la metrica RMSE.

2. Creare una pipeline in cui, a partire dal dataset utilizzato al punto precedente, i valori degli attributi `carlength`, `carwidth` e `carheight` sono discretizzati in 5 intervalli, `citympg` e `highwaympg` sono trasformati con uno StandardScaler e tutti gli altri attributi sono lasciati invariati.

3. Creare una nuova pipeline che applica la SelectKBest al dataset utilizzato al punto 1 e aggiunge le componenti ottenute alle componenti della pipeline del punto precedente. Valutare i valori migliori di `k` di SelectKBest, del numero di intervalli in cui discretizzare `carlength`, `carwidth` e `carheight` e dei parametri `criterion` e `max_depth` del RandomForestRegressor. Ignorare eventuali warning. Confrontare i risultati con quelli ottenuti precedentemente.

4. Creare una pipeline che, a partire dal dataset originale, trasforma le colonne testuali in valori numerici e le feature numeriche attraverso lo StandardScaler e applica il RandomForestRegressor. Come variano le performance?