Unione europea
Fondo sociale europeo

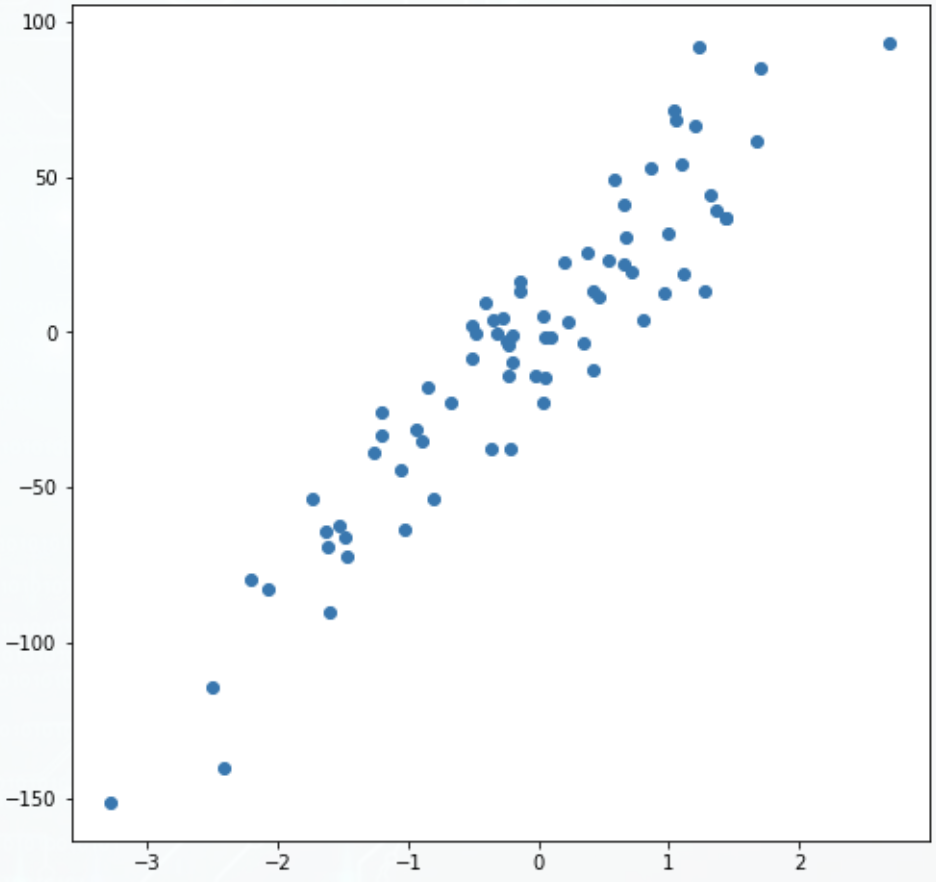
RegioneEmilia-Romagna

RETE POLITECNICA
La tecnica per crescere

Materiale didattico per partecipante al corso "**TECNICO ESPERTO NELL'ANALISI E NELLA VISUALIZZAZIONE DEI DATI**" – Rif.P.A. 2021-15998/RER – approvata con DGR n. 1263 del 02/08/2021 di IFOA – Istituto Formazione Operatori Aziendali

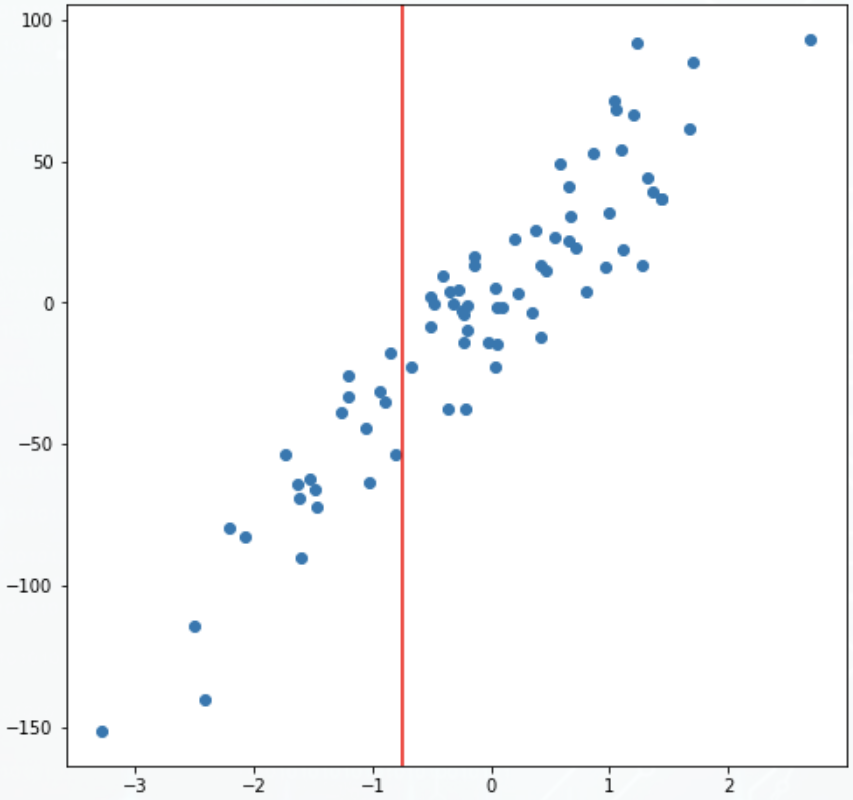
1.
K-Nearest Neighbors (KNN)

K Nearest Neighbors (k-NN)



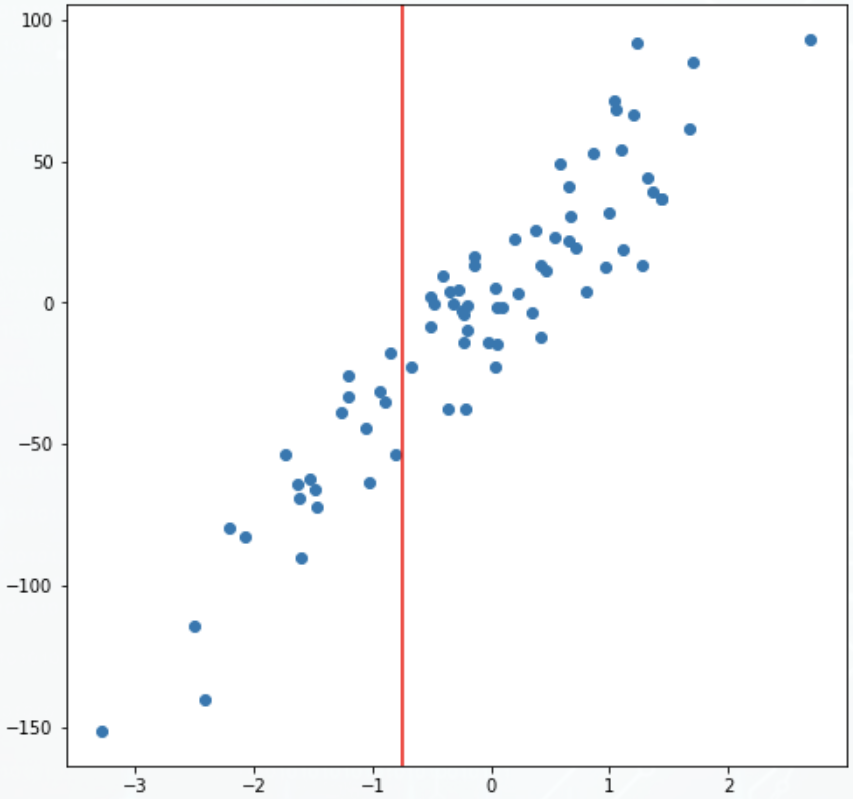
K Nearest Neighbors (k-NN)

Prendiamo un valore di X

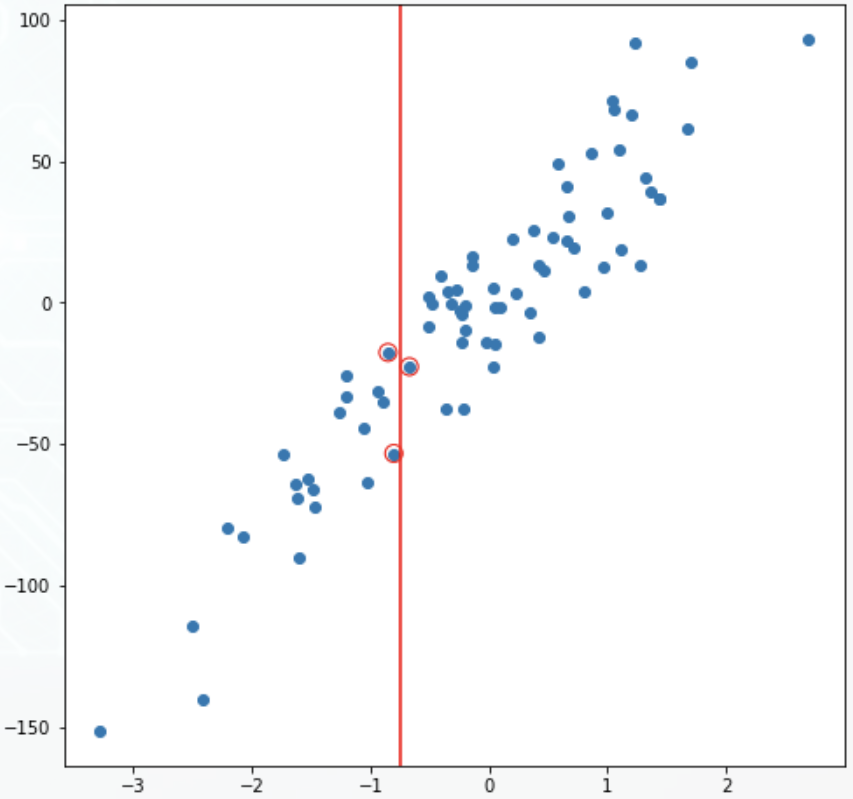


K Nearest Neighbors (k-NN)

Prendiamo un valore di X

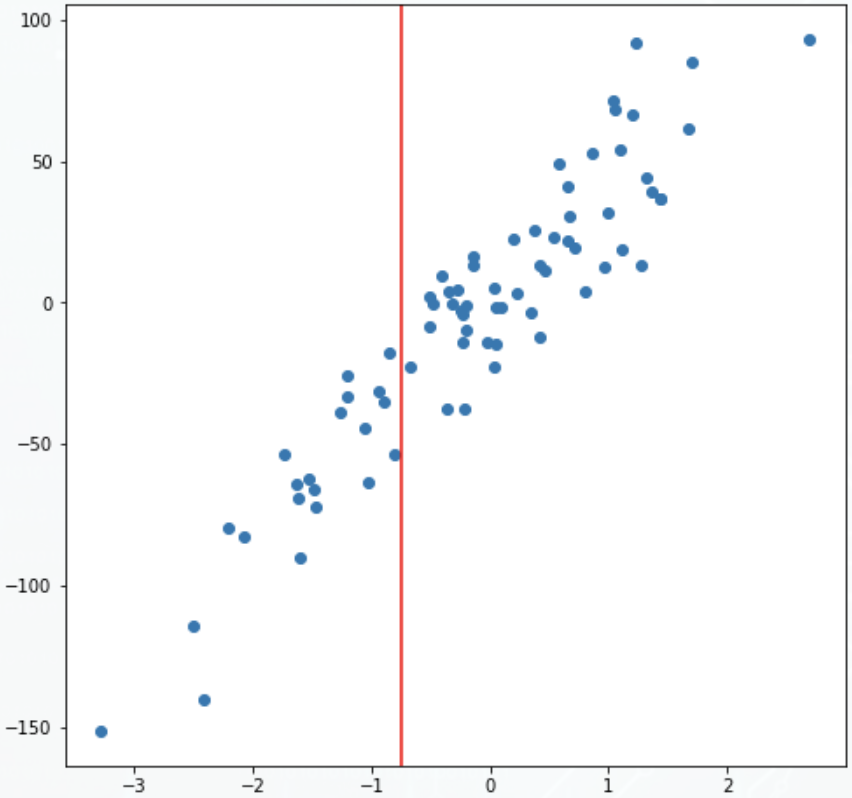


Individuo i 3 (k) punti più ‘vicini’

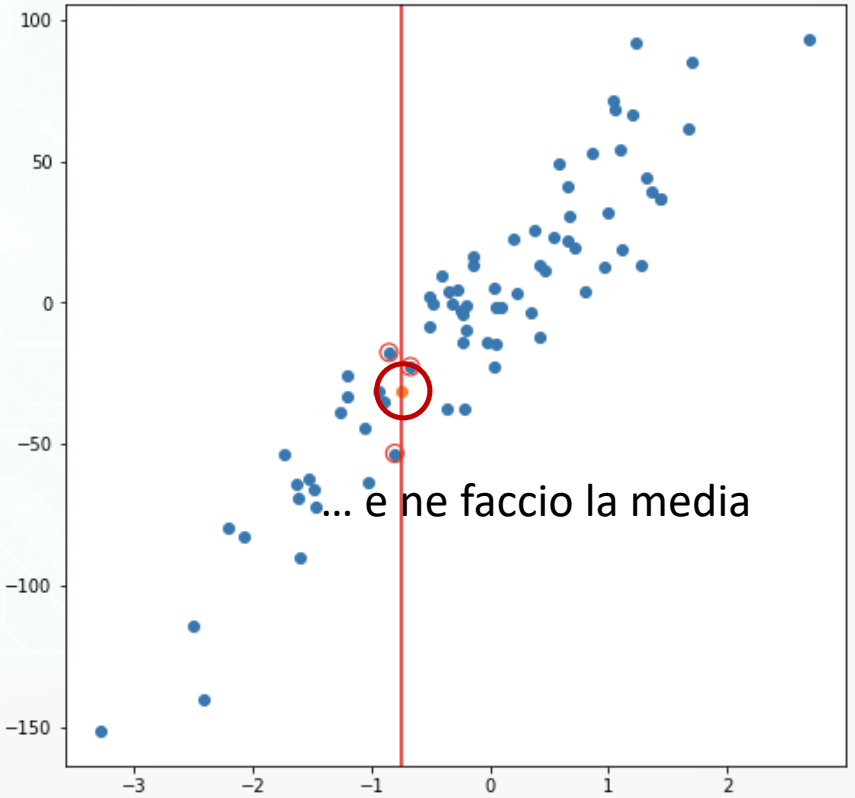


K Nearest Neighbors (k-NN)

Prendiamo un valore di X

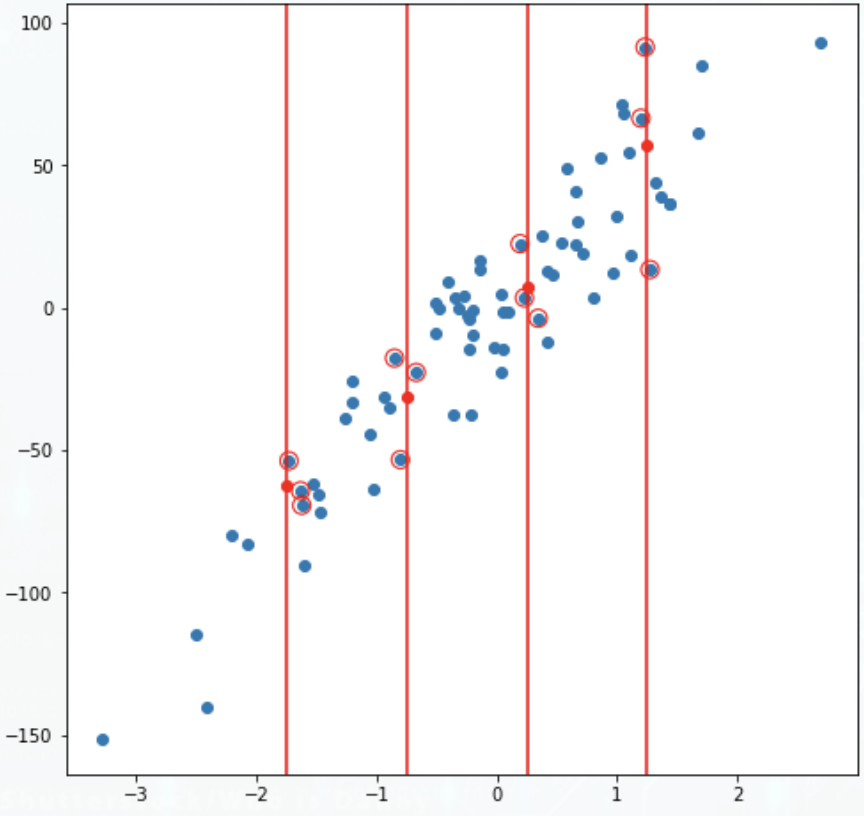


Individuo i 3 (k) punti più ‘vicini’



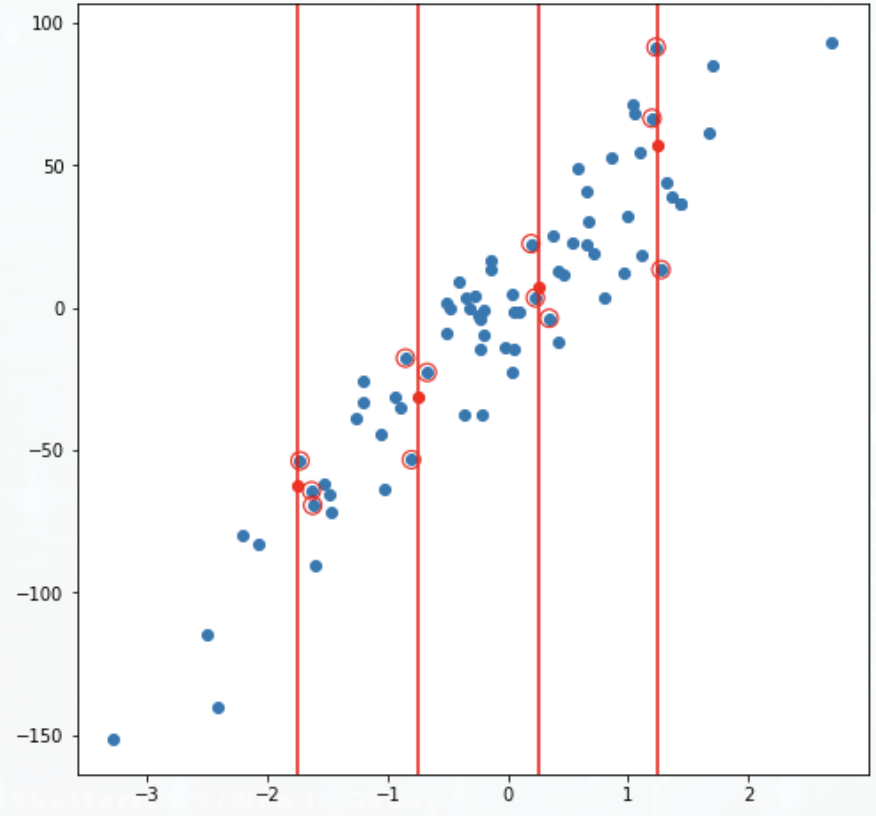
K Nearest Neighbors (k-NN)

Ripeto per tutti i valori di X

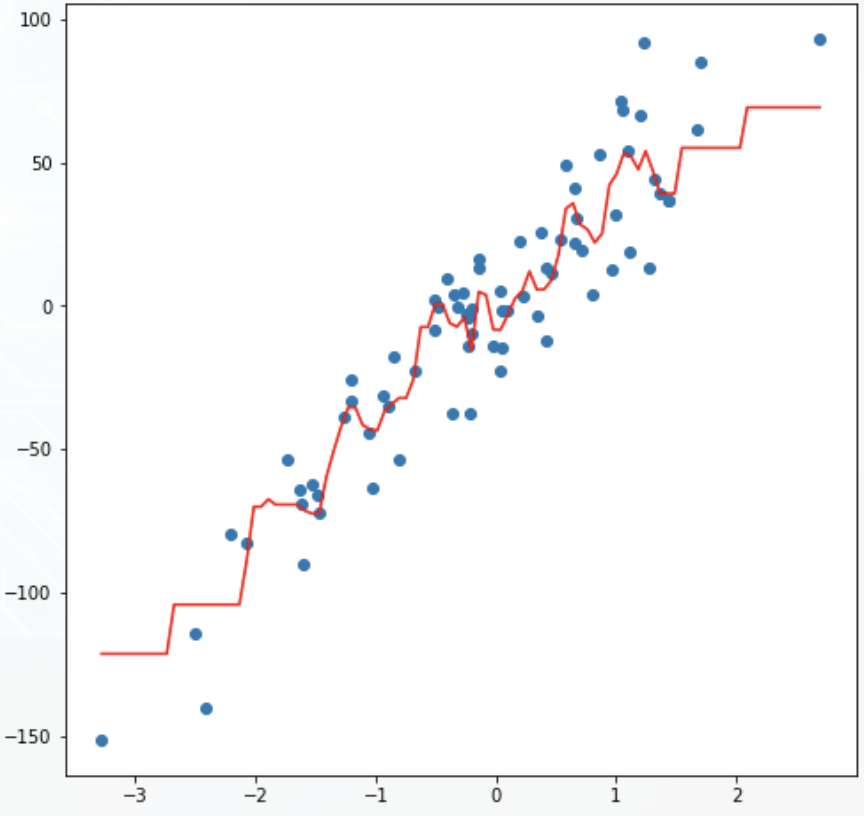


K Nearest Neighbors (k-NN)

Ripeto per tutti i valori di X

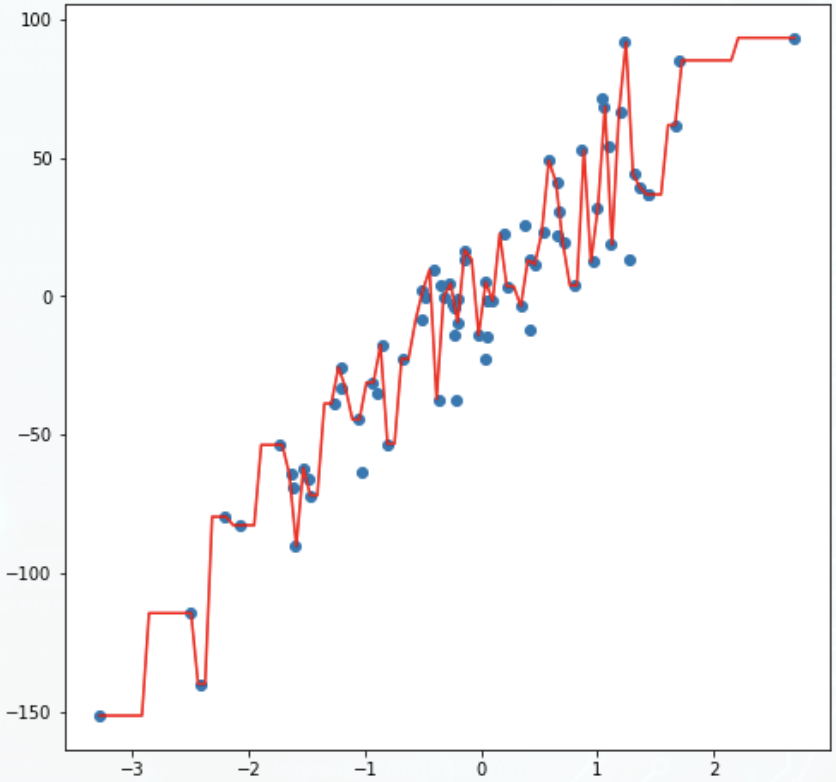


Connetto le medie trovate

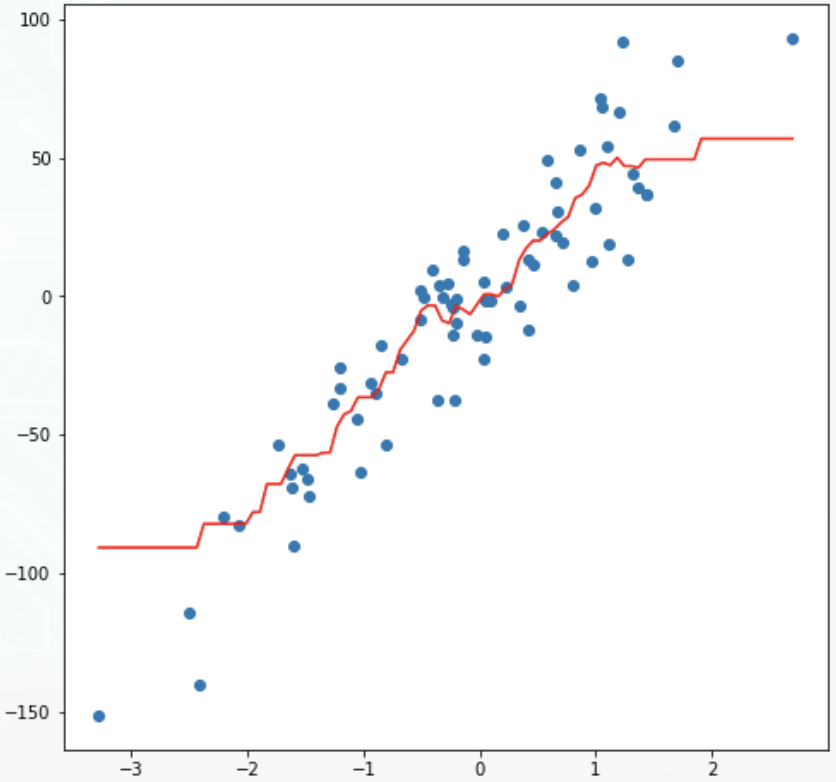


K Nearest Neighbors (k-NN)

K = 1 → probabile overfitting



K = 10 → curva più smooth



K Nearest Neighbors (k-NN)

IPER-PARAMETRI DEL MODELLO

- **K:** numero dei vicini. Più è alto più la curva di decisione sarà smooth, evitando l'overfitting ma anche perdendo un po' di accuratezza

- **Distance metric:** tipicamente si usa la distanza Euclidea, ma ci sono anche altre scelte (Manhattan, Minkowski…)

- **Weights**: volendo posso pesare la media sulla distanza dei singoli punti (i più vicini peseranno di più), ma occhio, si rischia di tornare a fare overfitting

K Nearest Neighbors (k-NN)

PRO

1. semplici da implementare (pochi parametri da impostare)
2. vedono bene relazioni non lineari

CONTRO

1. con l’aumentare dei dati la previsione rallenta molto
2. Non è interpretabile come può essere una regressione lineare, o un albero decisionale

2.
Decision Tree / Random Ferst

Decision Tree

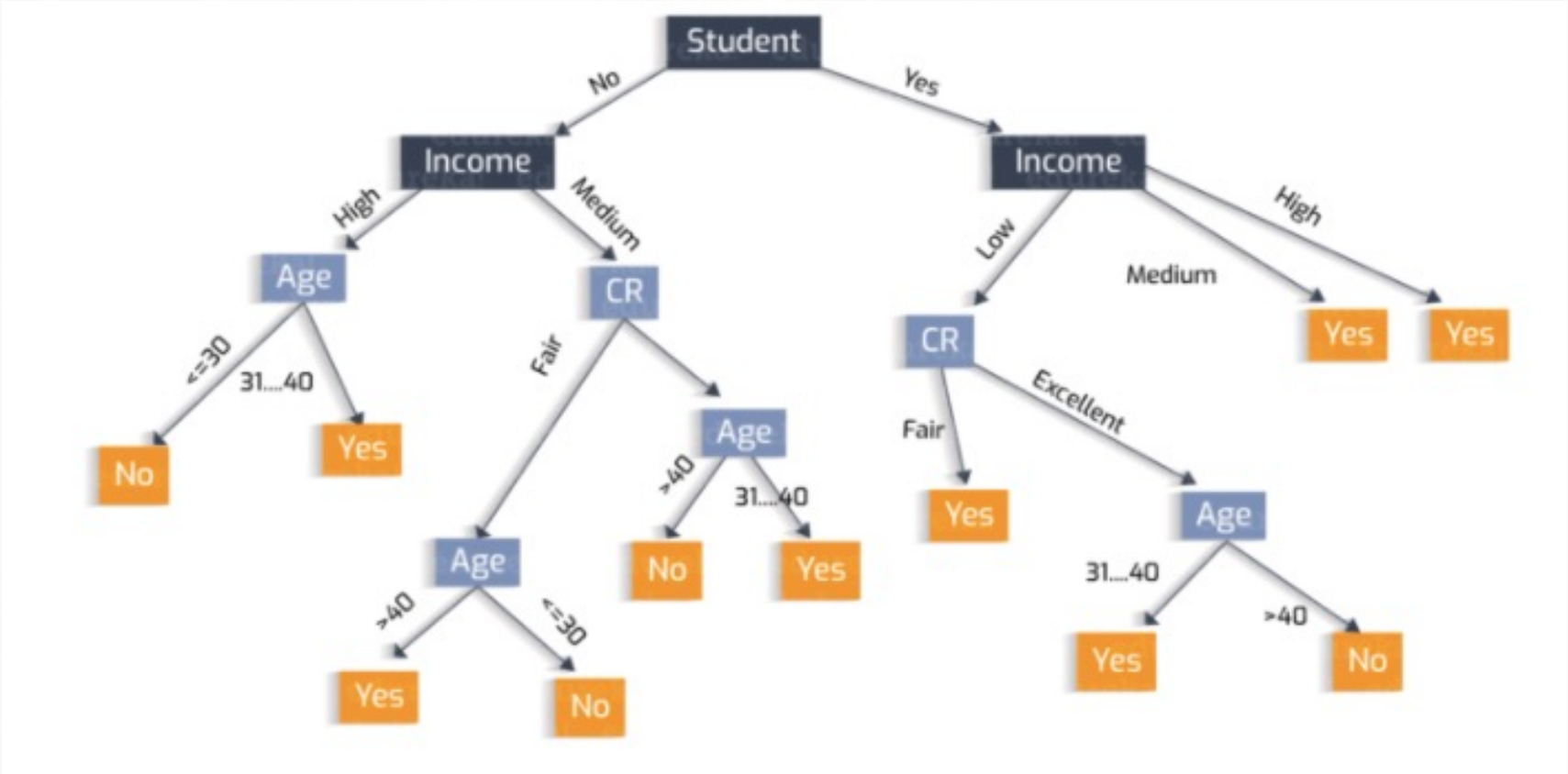
Age	Income	Student	Credit _rating	Buys_computer
<=30	Hight	No	Fair	No
<=30	Hight	No	Excellent	No
31...40	Hight	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
.=30	Low	Yes	Fair	Yes
>30	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

Decision Tree

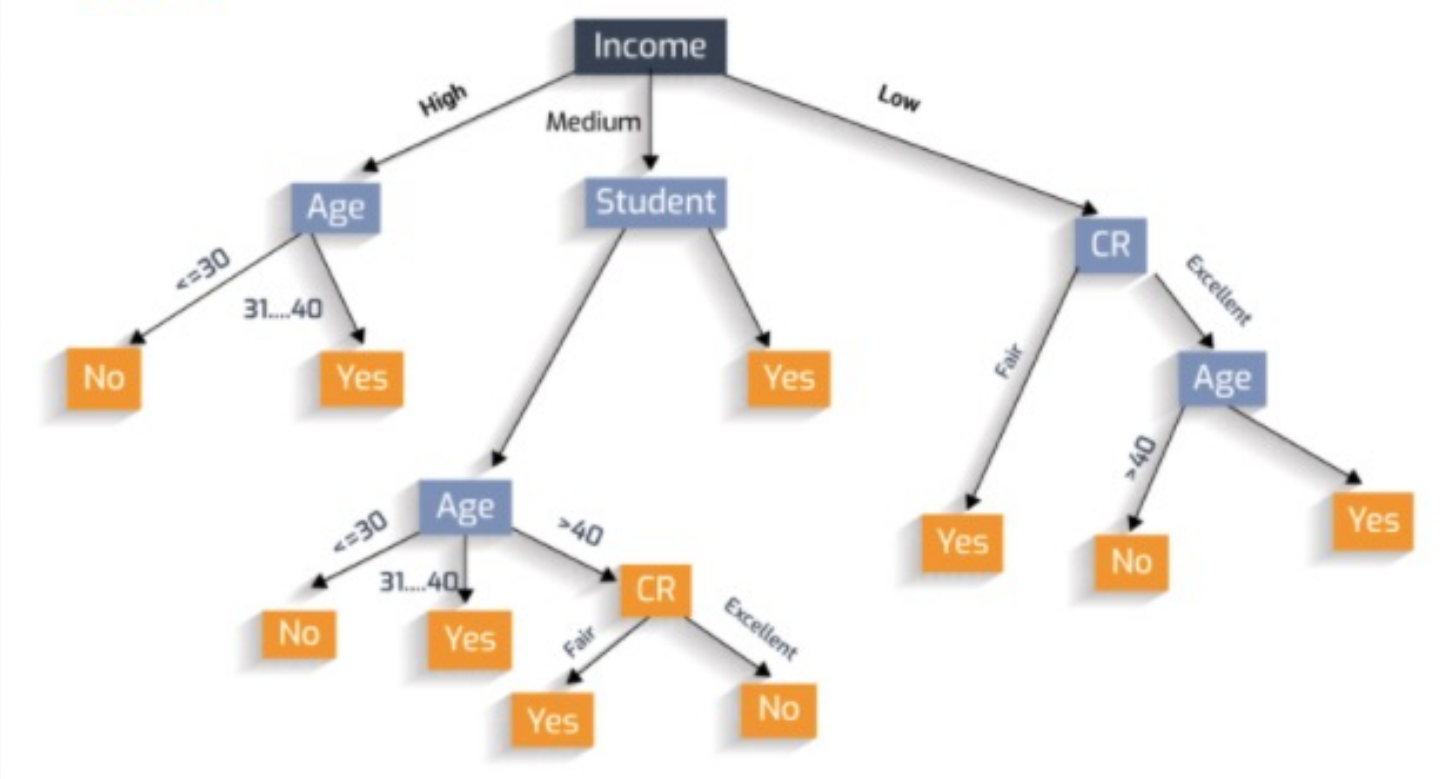
TARGET

Age	Income	Student	Credit _rating	Buys_computer
<=30	Hight	No	Fair	No
<=30	Hight	No	Excellent	No
31...40	Hight	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
.=30	Low	Yes	Fair	Yes
>30	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

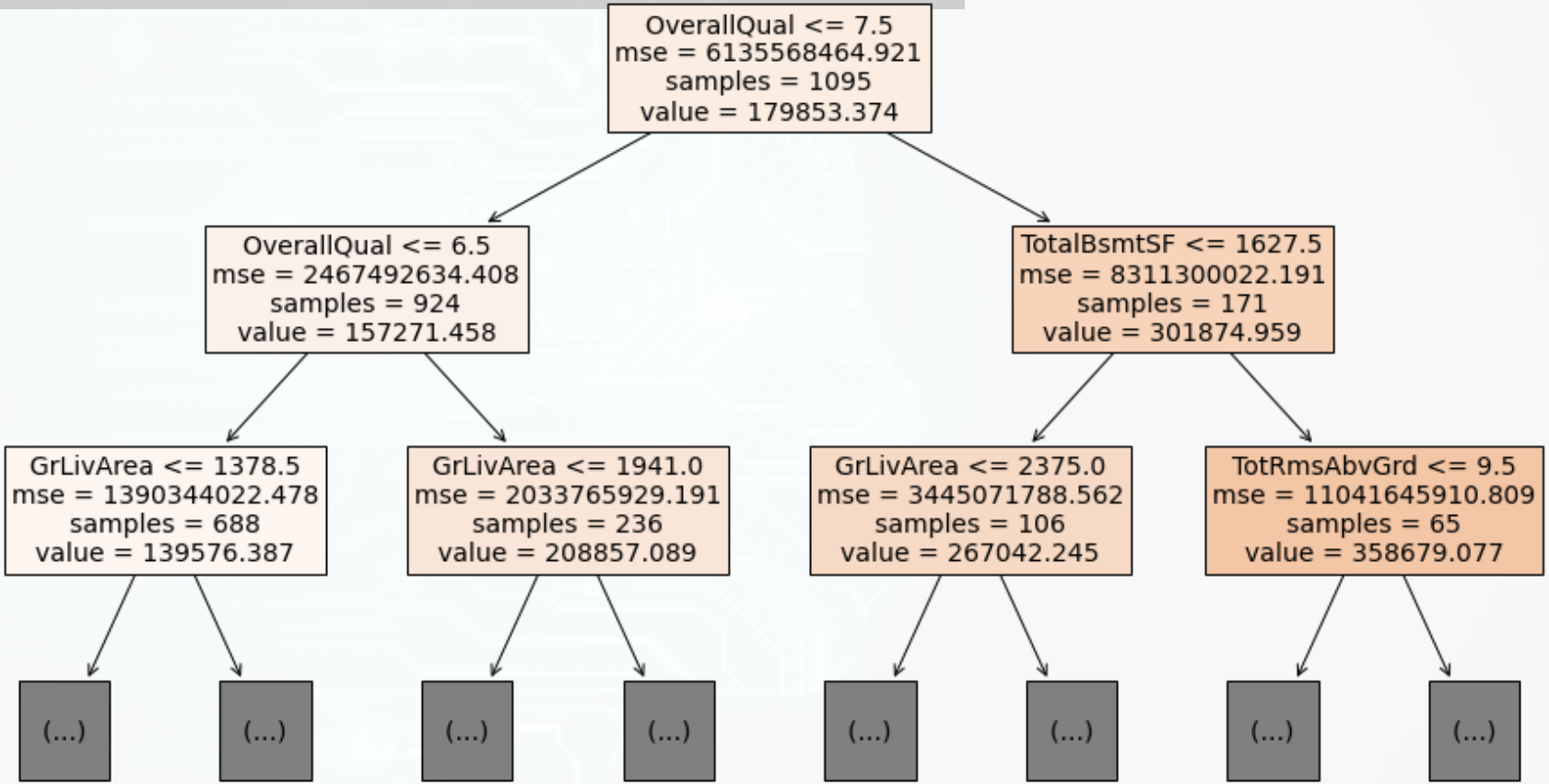
Decision Tree



Decision Tree



Decision Tree



Decision Tree

PRO

1. facili da **interpretare**
2. li si può usare per avere un'idea dell'**importanza relativa delle feature**
3. non hanno problemi con valori mancanti e non richiedono la normalizzazione dei dati
4. si possono usare sia per la regressione che per la classificazione

CONTRO

1. tendono a fare **overfitting**
2. sono **molto sensibili a piccoli cambiamenti nel training set** (per esempio cambiando di poco il numero di esempi nel training set si può arrivare ad alberi completamente differenti)
3. un singolo albero decisionale non è un buon predittore

Random Forest

Si creano N alberi decisionali variando in modo random le combinazioni di feature e il numero di samples usati per costruirli, e si media il risultato finale. Sono più stabili per quanto riguarda l'overfitting, ma computazionalmente più costose.