ANALISI SUL NUMERO DI DOTTORI VISITATI DA UN CAMPIONE DI INDIVIDUI

Utilizzo della classificazione, in particolare degli alberi di decisione e delle random forest, per fare previsioni sul numero di dottori visitati da un individuo a partire da altri dati

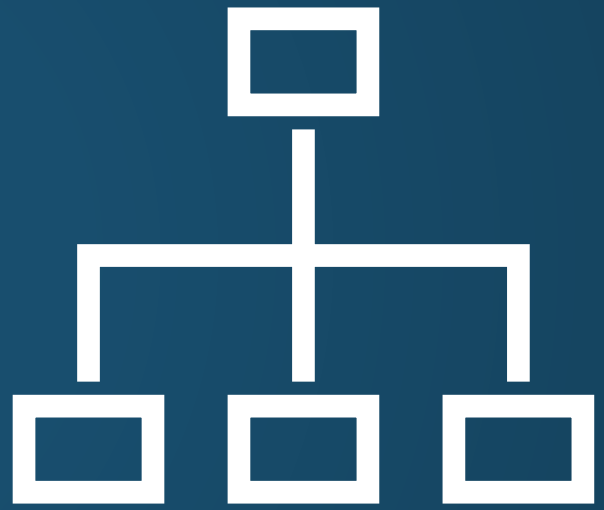
IL DATASET

- Il dataset utilizzato per l'analisi è «National Poll on Healthy Aging», creato dall'Università del Michigan;
- Il dataset ha 15 classi e gli individui intervistati sono 714, quindi non si tratta di un dataset particolarmente grande;
- Le classi del dataset sono nello specifico:
  - **Number_of_Doctors_Visited** (valori possibili: 1 = 0-1 dottori, 2 = 2-3 dottori, 3 = 4 o più dottori);
  - **Age** (valori possibili: 1 = 50-64, 2 = 65-80);
  - **Physical_Health** (valori possibili: -1 = rifiutato, 1 = eccellente, 2 = molto buona, 3 = buona, 4 = discreta, 5 = scarsa);
  - **Mental_Health** (valori possibili : -1 = rifiutato, 1 = eccellente, 2 = molto buona, 3 = buona, 4 = discreta, 5 = scarsa);
  - **Dental_Health** (valori possibili : -1 = rifiutato, 1 = eccellente, 2 = molto buona, 3 = buona, 4 = discreta, 5 = scarsa);
  - **Employment** (valori possibili : -1 = rifiutato, 1 = lavoro full-time, 2 = lavoro part-time, 3 = in pensione, 4 = disoccupato al momento);
  - **Stress_Keeps_Patient_from_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Medication_Keeps_Patient_from_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Pain_Keeps_Patient_from_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Bathroom_Needs_Keeps_Patient_from_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Unknown_Keeps_Patient_from_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Trouble_Sleeping** (valori possibili : 0 = no, 1 = si);
  - **Prescription_Sleep_Medication** (valori possibili : -1 = rifiutato, 1 = uso regolare, 2 = uso occasionale, 3 = non usa);
  - **Race** (valori possibili : -2 = non è stato chiesto, -1 = rifiutato, 1 = bianco non-ispanico, 2 = nero non ispanico, 3 = altro non ispanico, 4 = ispanico, 5 = 2+ races non ispaniche);
  - **Gender** (valori possibili : -2 = non è stato chiesto, -1 = rifiutato, 1 = uomo, 2 = donna).

  Il dataset non presenta valori mancanti.

SCOPO DELL'ANALISI

Dato il dataset, lo scopo dell'analisi era quello di prevedere il numero di dottori visitati da un individuo a partire da altri dati.



METODI UTILIZZATI

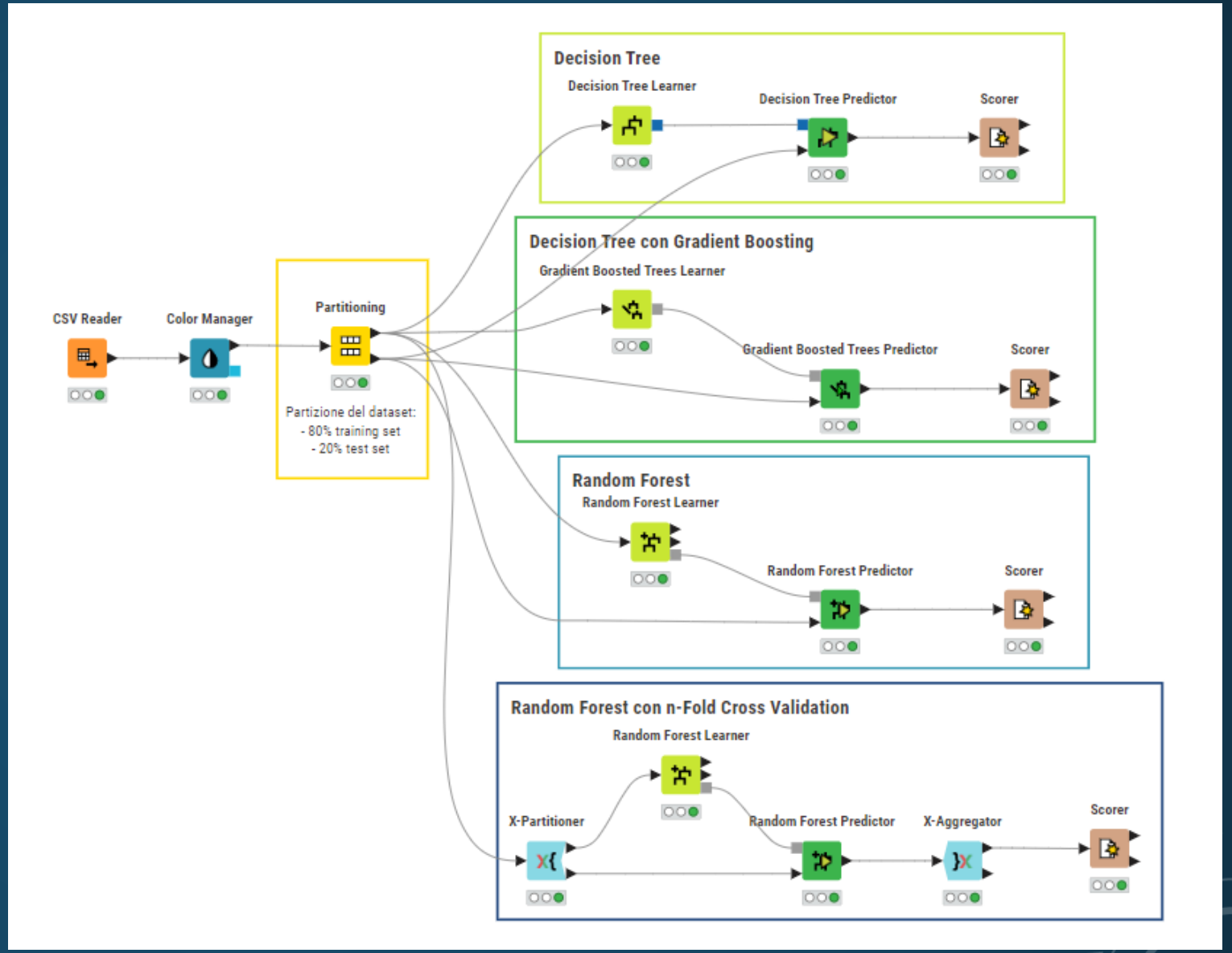
Per l'analisi sono stati utilizzati:

- Albero di decisione (semplice)
- Albero di decisione con la tecnica del Gradient Boosting
- Random Forest
- Random Forest con n-Fold Cross Validation

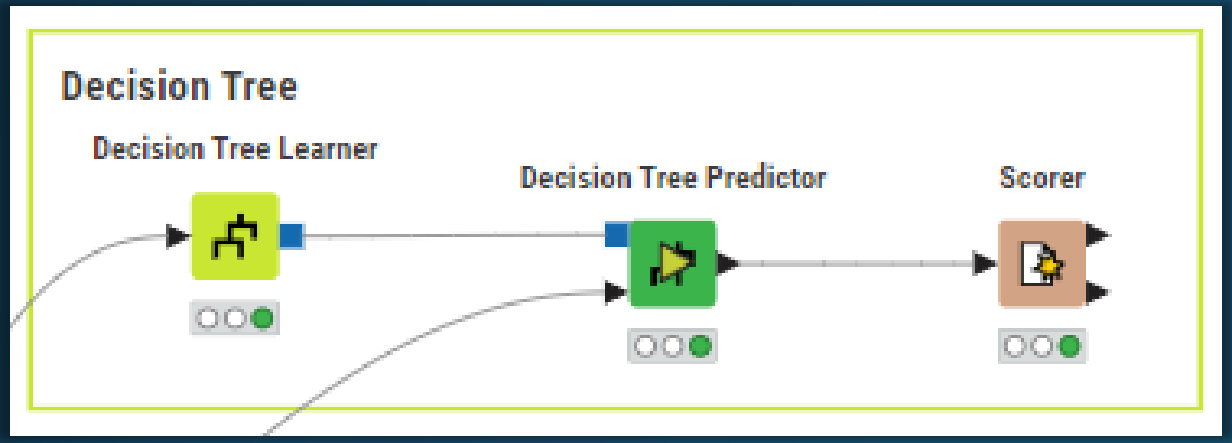
OVERVIEW

Dopo l'importazione del dataset, quest'ultimo è stato diviso in due subset:

- 80% utilizzato come set di addestramento per gli algoritmi
- 20% usato come set di test per verificare le previsioni



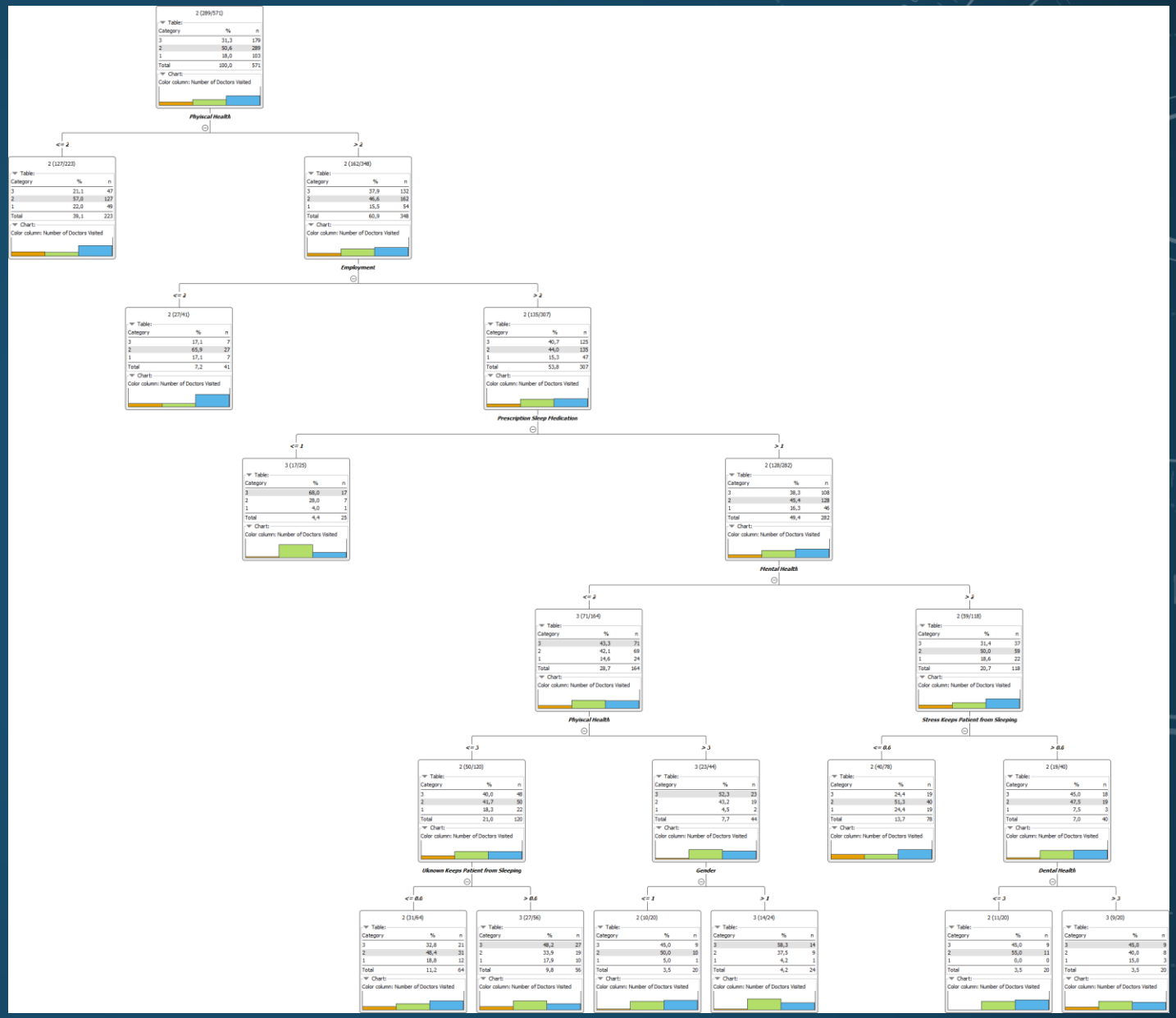
DECISION TREE



Il primo tipo di algoritmo di classificazione utilizzato è il Decision Tree (in italiano Albero di Decisione).
Il nodo ***Decision Tree Learner*** allena l'algoritmo con il training set; il risultato diventa poi input per il nodo ***Decision Tree Predictor***, insieme al test set.
Le previsioni fatte da quest'ultimo nodo sono poi date in input al nodo ***Scorer***, che le compara con i valori effettivi della classe di cui si vuole avere una previsione (nel caso di quest'analisi, la classe «Number of Doctors visited).
Lo ***Scorer*** produce la matrice di confusione e statistiche riguardo all'accuratezza dell'analisi, utilizzate per valutarne la qualità.

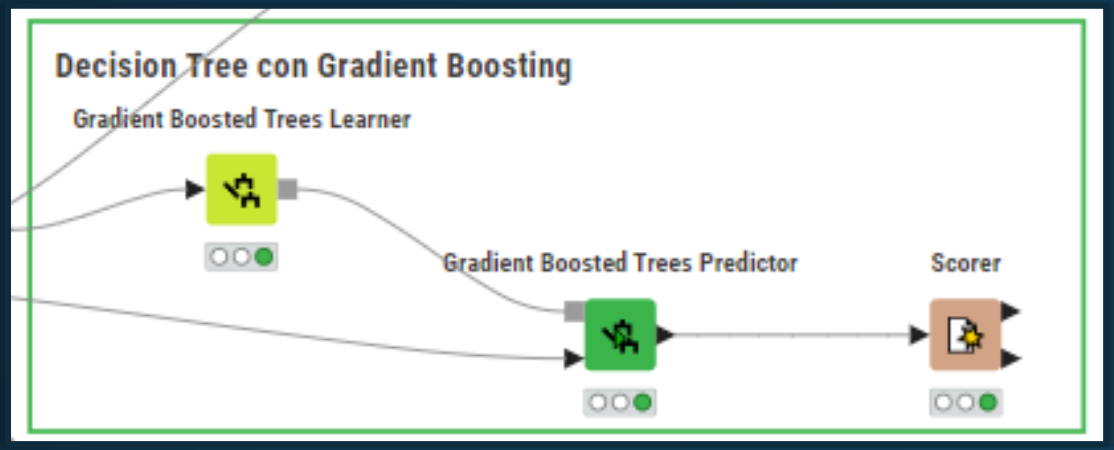
DECISION TREE

A lato si può osservare un esempio di decision tree prodotto dal *nodo Decision Tree Learner.*



[i valori della classe «Numero di Dottori Visitati» sono colorati in base ai possibili valori 1, 2, 3 rispettivamente di arancione, azzurro e verde]

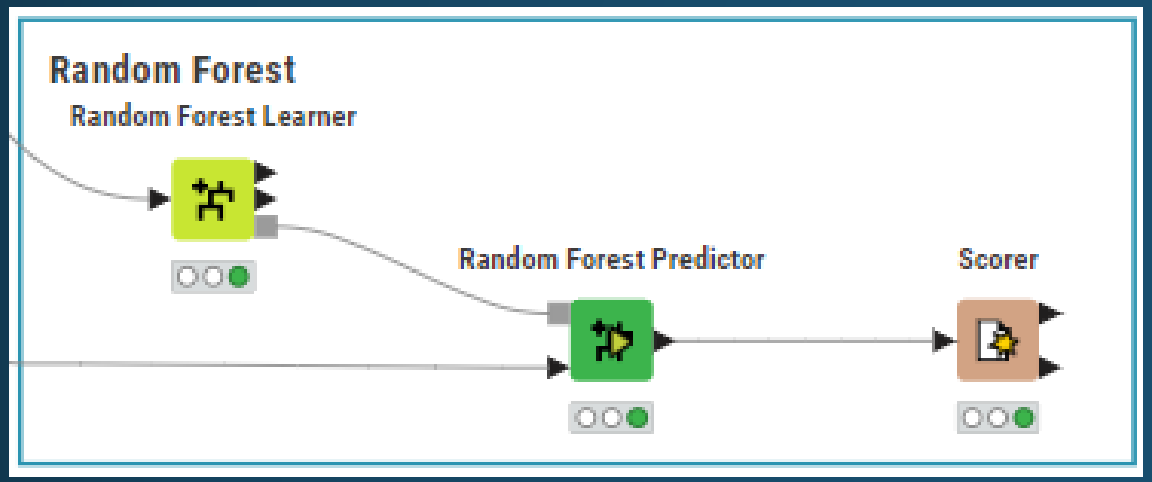
DECISION TREE CON GRADIENT BOOSTING



Il Gradient Boosted Decision Tree consiste in un insieme di alberi di decisione i cui risultati vengono aggregati mano a mano che il modello viene allenato.

L'idea di base è quella di unire tanti risultati «deboli» del learner di un decision tree in modo da poter ottenere poi una previsione più accurata.

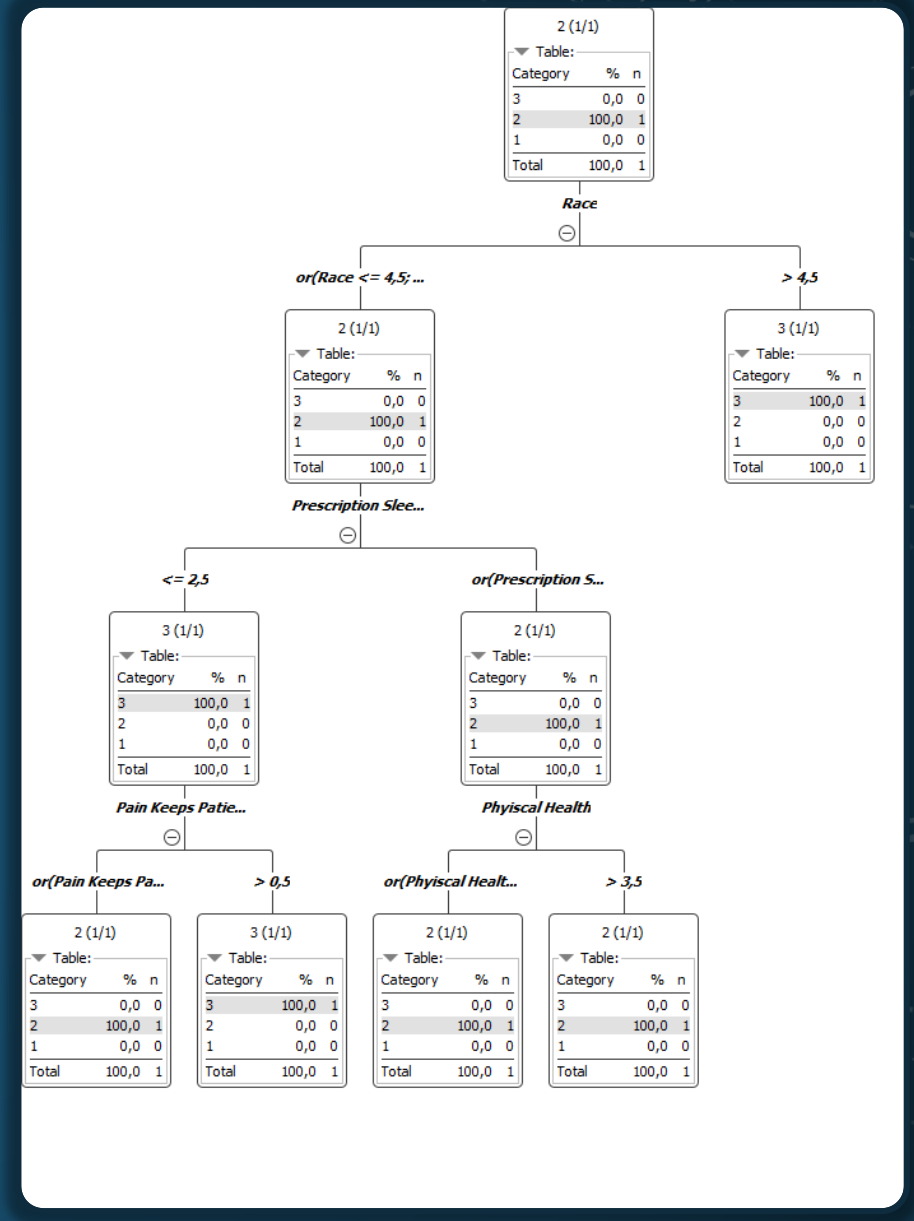
RANDOM FOREST



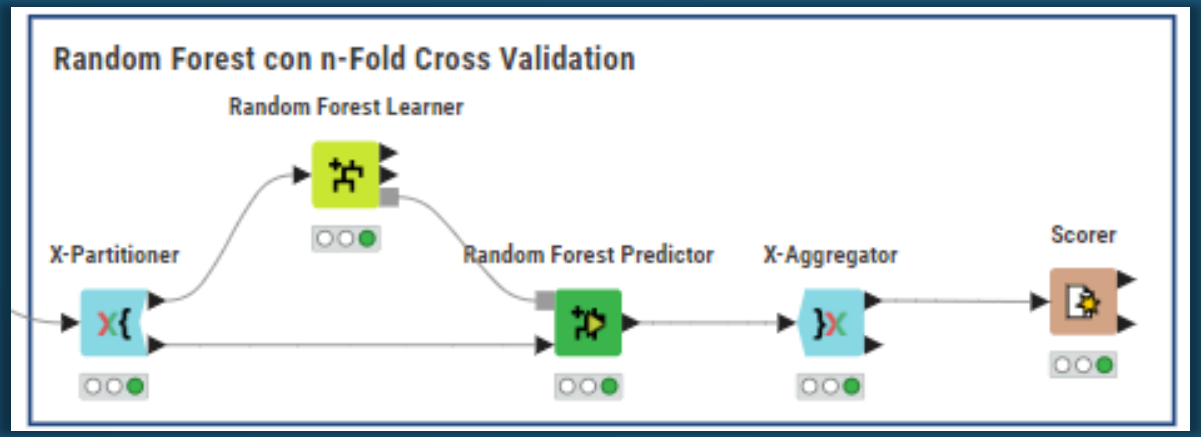
Una Random Forest consiste in un insieme di alberi di decisione, ciascuno costruito con un subset di istanze del dataset e un subset di colonne (classi) scelto randomicamente.

Una volta calcolati i risultati di tutti gli alberi di decisione, viene fatta una media e viene restituita come risultato unico della random forest.

ESEMPIO DI DECISION TREE USATO NELLA RANDOM FOREST



RANDOM FOREST CON N-FOLD CROSS VALIDATION



Il metodo n-Fold Cross Validation (in questo caso applicato ad una random forest) viene utilizzato su dataset non molto grandi e consiste nella suddivisione del dataset in n subset disgiunti. Poi ciascun gruppo viene tenuto da parte come test set, mentre i restanti n-1 subset vengono utilizzati per l'addestramento.

Il risultato unico finale corrisponde alla media dei singoli risultati ottenuti.

RISULTATI DELL'ANALISI

In tutti e quattro i casi precedentemente descritti, i risultati ottenuti non sono particolarmente buoni: come si può infatti vedere dalle matrici di confusione sotto e a lato riportate, nessuno dei quattro metodi utilizzati permette di raggiungere un'accuratezza maggiore del 59%.

Number of Doctors Visited \ Prediction (Number of Doctors Visited)	3	2	1
3	10	22	0
2	15	68	0
1	5	23	0

Correct classified: 78 | Wrong classified: 65

Accuracy: 54,545% | Error: 45,455%

Cohen's kappa (κ): 0,081%

Matrice di confusione dell'albero di decisione

Number of Doctors Visited \ Prediction (Number of Doctors Visited)	3	2	1
3	4	28	0
2	6	77	0
1	4	24	0

Correct classified: 81 | Wrong classified: 62

Accuracy: 56,643% | Error: 43,357%

Cohen's kappa (κ): 0,046%

Matrice di confusione dell'albero di decisione con gradient boosting

Number of Doctors Visited \ Prediction (Number of Doctors Visited)	3	2	1
3	2	30	0
2	1	82	0
1	2	26	0

Correct classified: 84 | Wrong classified: 59

Accuracy: 58,741% | Error: 41,259%

Cohen's kappa (κ): 0,045%

Matrice di confusione della random forest

Number of Doctors Visited \ Prediction (Number of Doctors Visited)	3	2	1
3	22	157	0
2	13	276	0
1	2	101	0

Correct classified: 298 | Wrong classified: 273

Accuracy: 52,189% | Error: 47,811%

Cohen's kappa (κ): 0,056%

Matrice di confusione della random forest con n-fold cross validation

RISULTATI DELL'ANALISI

Il fatto che i risultati ottenuti siano così scarsi può essere dovuto a diversi fattori, come ad esempio:

- La grandezza del dataset: un dataset più grande permette un migliore addestramento degli algoritmi e dovrebbe portare quindi a previsioni più accurate;
- Overfitting, ovvero quando si ha una buona accuratezza sui dati di addestramento ma scarsa sui dati di test. In questa analisi sono stati usati metodi di pruning dove possibile per cercare di evitare questo fenomeno.
- Presenza di rumore o outliers nel dataset.

In particolare in questo caso è possibile ipotizzare che ci sia qualche problema con la gestione del valore 1. Ciò si può notare osservando per esempio le matrici di confusione (visibili alla pagina precedente), nelle quali il valore 1 risulta non essere mai classificato, né correttamente né erroneamente.
L'assente classificazione di tale valore può essere stata, a parer mio, la causa di tali risultati, ma non mi è stato possibile giungere ad una soluzione che portasse a risolvere il problema, che non si è invece verificato provando ad utilizzare altri dataset.