Progetto di Big Data Analytics & Machine Learning

Cuicchi Manila,

De Bartolomeo Gabriele,

Giannelli Edoardo

(LM Ingegneria Informatica e dell'Automazione, UNIVPM, A.A 2020-2021)

Obiettivo – Analisi di Regressione

- Nel nostro progetto ci occupiamo di sviluppare dei modelli di regressione per la predizione su un dataset relativo a misure svolte su campioni di biomassa prelevati da centrali elettriche che usano la biomassa come carburante.

Il Dataset

Il dataset utilizzato è un **foglio in formato Excel (.xlsx)** composto da **5217 record** (righe) relativi alle singole misurazioni effettuate in laboratorio e dai seguenti **18 attributi** (colonne), in elenco, corrispondenti ai vari parametri misurati durante ciascuna combustione:

- ID CAMPIONE
- CENTRALE
- INFO: campo descrittivo
- MATERIALE: descrive la tipologia di biomassa
- GRUPPO ISO: raggruppa diversi materiali
- DATA RICEVIMENTO: campo descrittivo
- UMIDITA’
- PCN
- CENERI
- PCS
- PCI
- CARBONIO
- IDROGENO
- AZOTO
- OSSIGENO
- CLORO
- ZOLFO
- UMIDITA’ DI CORREZIONE

Goal dell'Analisi

- "Quesito 1" - dato il valore della feature ceneri predire per ogni materiale i valori di PCS, PCI, Azoto, Cloro e Zolfo [quindi un modello di regressione per ogni materiale e per ognuna delle 5 feature indicate];

- "Quesito 2" - dati i valori delle feature ceneri, carbonio, idrogeno e azoto, per ogni materiale predire i valori di PCS,PCI, cloro e zolfo [anche in questo caso diversi modelli di regressione];

- "Quesito 3" - date tutte le feature relative a misure, per ogni materiale predire i valori di PCS,PCI, cloro e zolfo.

Con l'obiettivo finale di individuare i modelli di predizione che restituiscano, per ognuno dei Goal, i risultati migliori, secondo le misure di

- "**Root Mean Squared Error**" **("RMSE")**
- e **relativa varianza(standard deviation)**.

Algoritmi

Si è optato per l'utilizzo degli algoritmi:

- **Decision Tree**;
- **Random Forest**;
- **SVMdot**;
- **SVMradial**;
- **SVMpolynomial**;

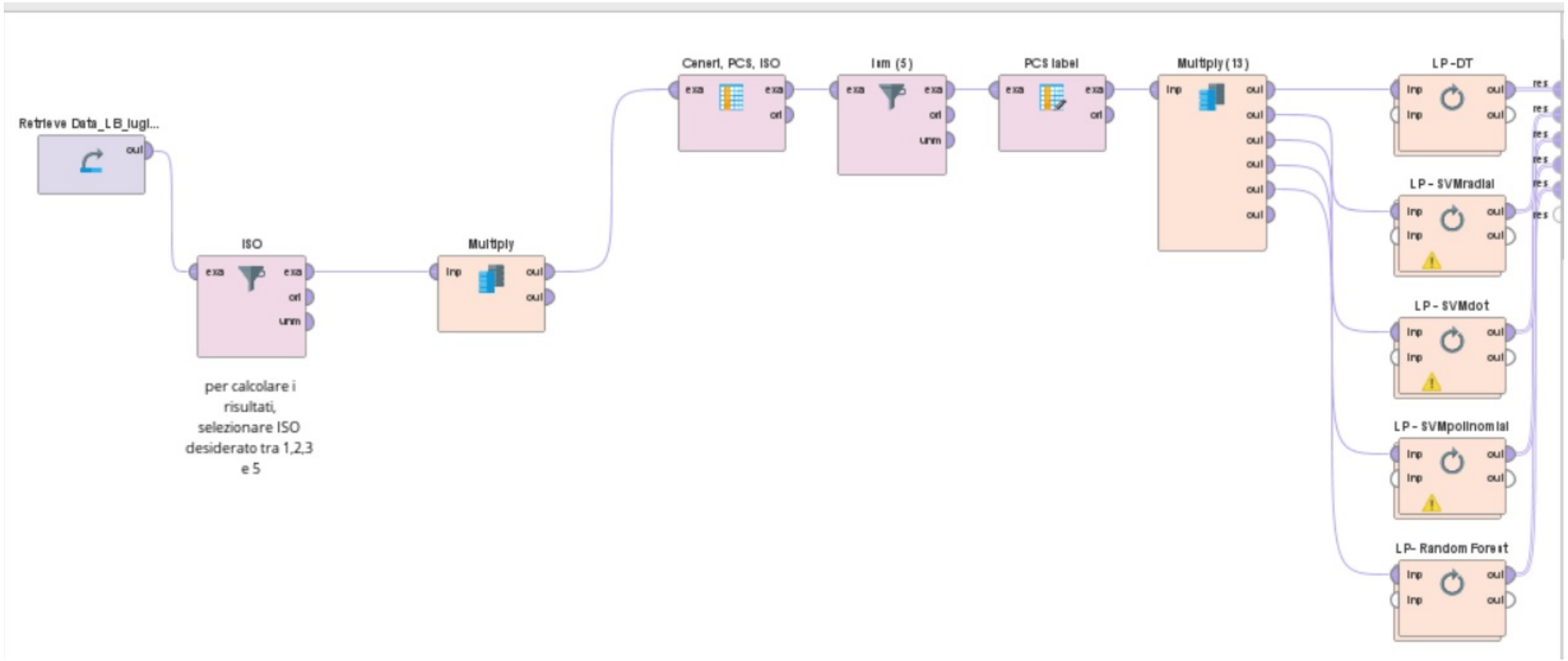
per la generazione di un modello di predizione tramite addestramento basato su **X-Fold Cross Validation**, su cui sono stati utilizzati **10 Folds**.

Implementazione

Per raggiungere gli obiettivi richiesti, a partire dal dataset a disposizione, è stato utilizzato il software “Rapid Miner Studio”, abbiamo implementato 3 processi, ciascuno relativo rispettivamente ai Goal descritti precedentemente.

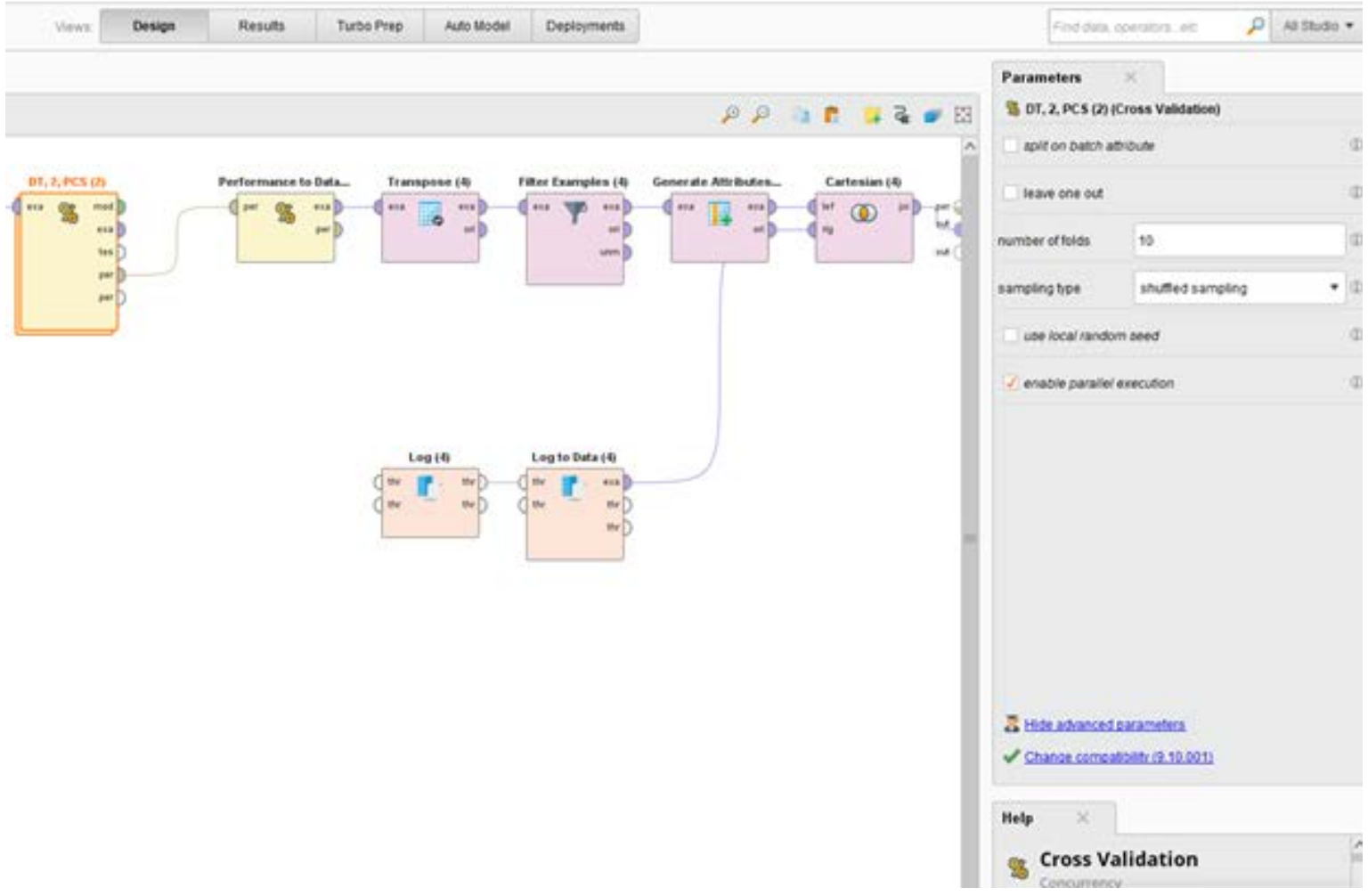
Implementazione (1)

Ognuno dei 3 processi è strutturato come segue:



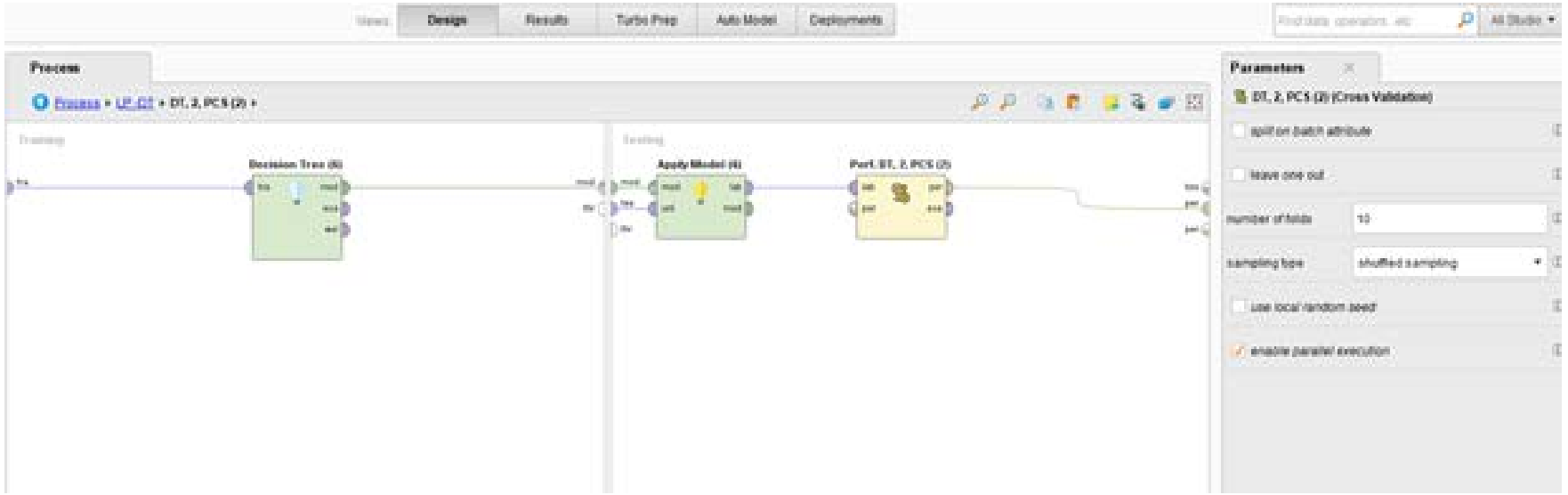
Implementazione (2)

Ognuno dei 3 processi è strutturato come segue:



Implementazione (3)

Ognuno dei 3 processi è strutturato come segue:



Strategia di Analisi

Per analizzare i risultati ottenuti, una volta esportate su file con estensione .collection da Rapid Miner le relative tabelle, si è scelto di **individuare i risultati migliori (ossia più bassi)** in termini di:

- **RMSE**
- **standard deviation.**

Strategia di Analisi - Approfondimento

L'analisi è stata approfondita in due modi, qualora necessario:

- **aggiungendo valori intermedi** nei loop parameter relativamente ai parametri caratteristici di ciascun algoritmo, in corrispondenza delle situazioni più interessanti, quali salti di valore rilevanti e/o andamenti irregolari;

- **effettuando un "controllo incrociato"**: per valori di RMSE medi analoghi, si è valutata anche la somiglianza tra le corrispondenti misure della standard deviation, in quanto un'eventuale standard deviation maggiore per uno stesso valore di RMSE indica una peggiore capacità di predizione del modello addestrato nel fold interessato da detto fenomeno.

Esempio di Tabella

Row No.	id	algoritmo	C	kernel_type	rmse	st_deviation
1	Value	SVMdot	0	dot	1513.593	545.893
2	Value	SVMdot	1	dot	1531.449	546.717
3	Value	SVMdot	10	dot	1551.804	590.829
4	Value	SVMdot	100	dot	1628.484	743.559
5	Value	SVMdot	1000	dot	1798.653	943.235
6	Value	SVMdot	10000	dot	1871.897	1011.442

Risultati Migliori

Risultati Migliori – Quesito 1

- "Quesito 1" - dato il valore della feature ceneri predire per ogni materiale i valori di PCS, PCI, Azoto, Cloro e Zolfo [quindi un modello di regressione per ogni materiale e per ognuna delle 5 feature indicate]

ISO1 – PCI,PCS -> SVMdot | Azoto -> DT+SVMradial | Cloro -> DT+RF | Zolfo -> RF+SVMRadial

ISO2 – PCI,PCS->SVMdot| Azoto, Cloro, Zolfo -> SVMdot

ISO3 – PCS -> SVMPolynomial | PCI, Azoto, Zolfo,Cloro -> RF

ISO5 – PCS,PCI- >SVMdot | Cloro,Zolfo,Azoto ->RF

Risultati Migliori – Quesito 2

- "Quesito 2" - dati i valori delle feature ceneri, carbonio, idrogeno e azoto, per ogni materiale predire i valori di PCS,PCI, cloro e zolfo [anche in questo caso diversi modelli di regressione]

ISO1 – PCS, PCI->SVMdot| Cloro,Zolfo->RF

ISO2 – PCS,PCI->SVMdot | Cloro->RF | Zolfo->DT

ISO3 – PCS->RF | PCI->SVMdot | Cloro,Zolfo->RF

ISO5 - PCS,PCI->SVMdot | Cloro,Zolfo->RF

Risultati Migliori – Quesito 3

- "Quesito 3" - date tutte le feature relative a misure, per ogni materiale predire i valori di PCS,PCI, cloro e zolfo

ISO1 - PCS,PCI->SVMdot | Cloro,Zolfo->RF

ISO2 – PCS,PCI->SVMdot | Cloro,Zolfo->RF

ISO3 - (dati non sufficienti per effettuare l'analisi)

ISO5 - PCS,PCI->SVMdot | Cloro,Zolfo->SVMdot