Relazioni Quantitative Struttura-Attività (QSAR) Mediante Algoritmi di Machine Learning. Sviluppo di Modelli per la Predizione di Attività di inibitori di Janus Chinasi 2 (JAK2)

Tesi sperimentale
in
Chimica Farmaceutica

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Farmacia e Medicina
Corso di Laurea in Chimica e Tecnologia Farmaceutiche
Tesi Sperimentale in Chimica Farmaceutica
a.a. 2020/2021

Laureanda: Giorgia Canini
Matricola: 1583394

Relatore: prof. Rino Ragno

Obiettivo della ricerca

Dataset inibitori JAK 2

QSAR

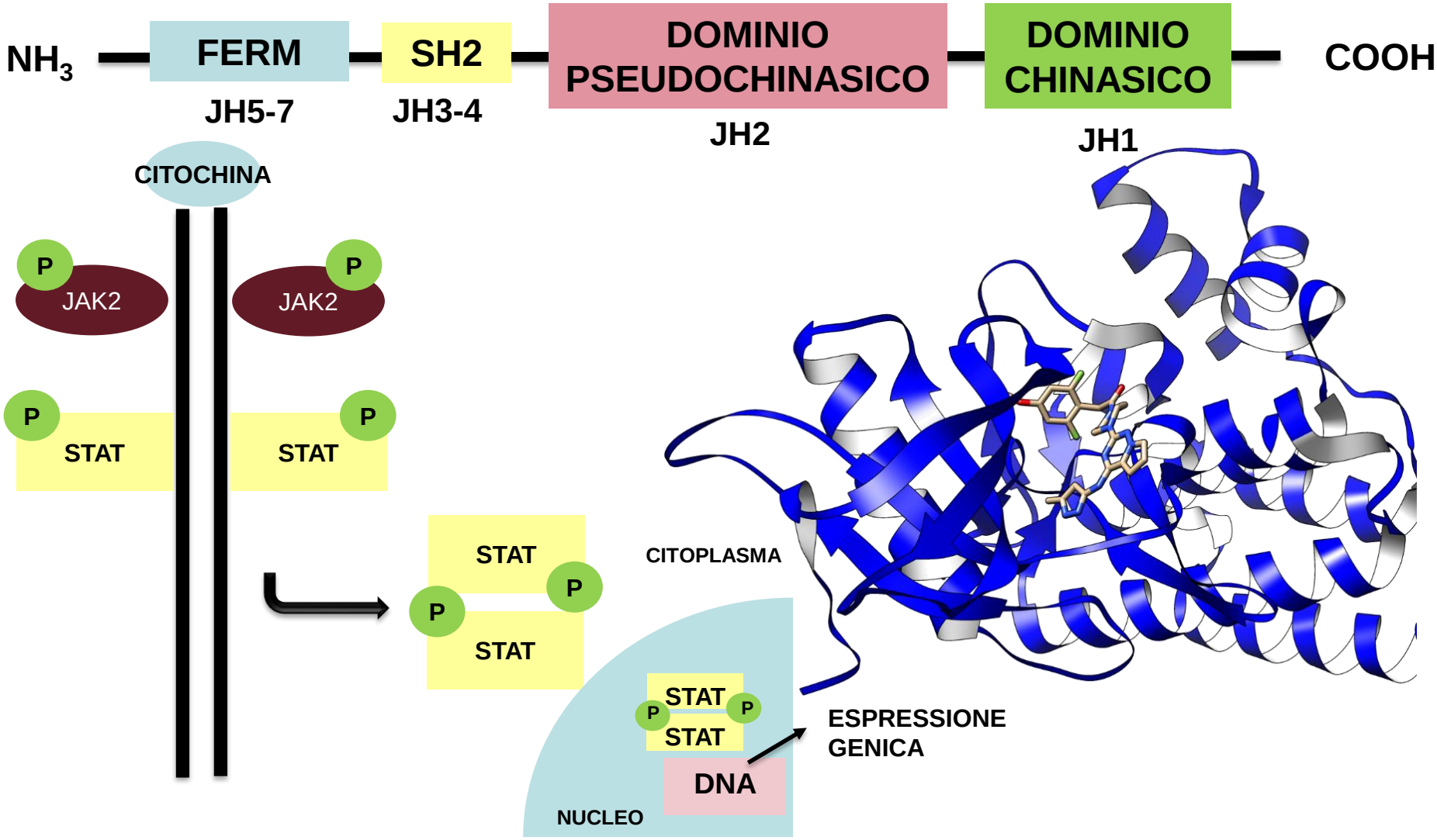
Machine Learning

Virtual Screening

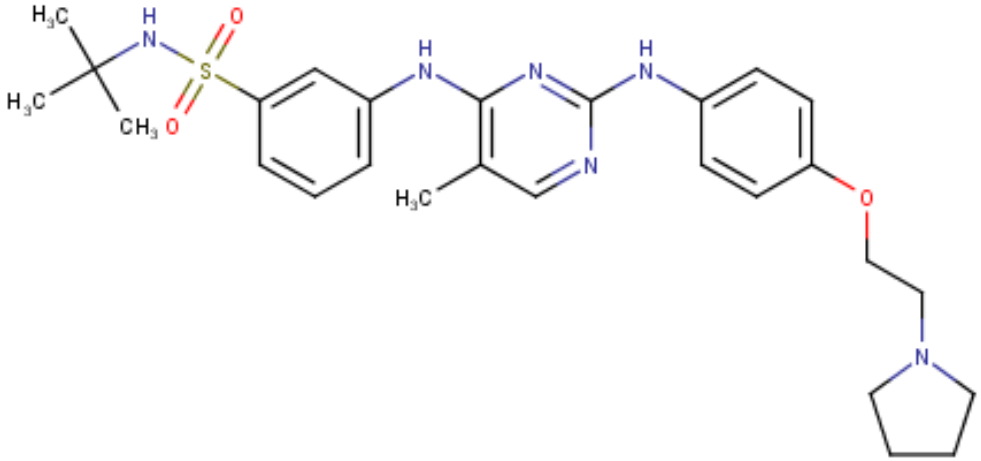
Nuovi potenziali inibitori JAK2

Scopo del lavoro

Target: JAK2


NH3
FERM
JH5-7
SH2
JH3-4
DOMINIO PSEUDOCHINASICO
JH2
DOMINIO CHINASICO
JH1
COOH
CITOCHINA
P
JAK2
P
JAK2
P
STAT
P
STAT
STAT
P
P
STAT
CITOPLASMA
STAT
P
P
STAT
DNA
NUCLEO
ESPRESSIONE GENICA

Inibitori JAK2

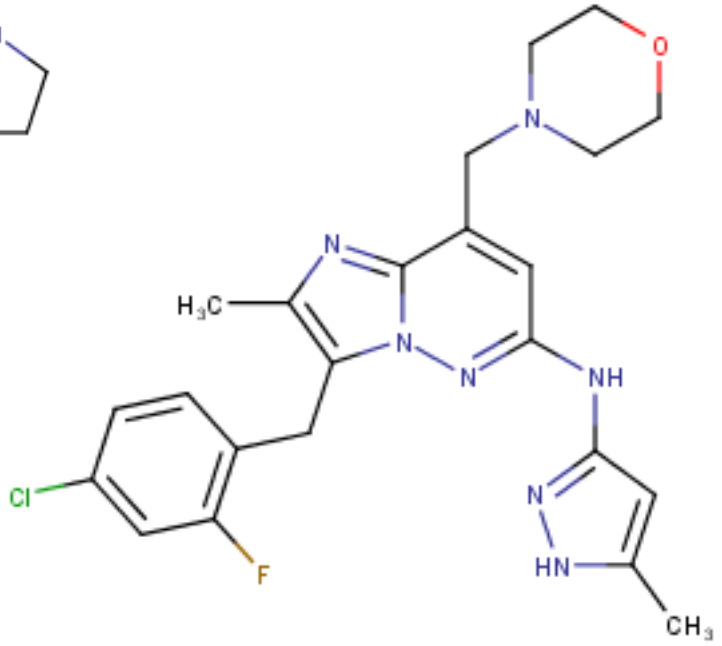


Fedratinib

- Inibitore selettivo di JAK2
- Approvato nel 2019 negli Stati Uniti per mielofibrosi primaria o secondaria di livello intermedio o ad alto rischio

Gandotinib

- LY2784544, inibitore selettivo JAK2
- Studi fase II nel 2018 per neoplasie mieloproliferative

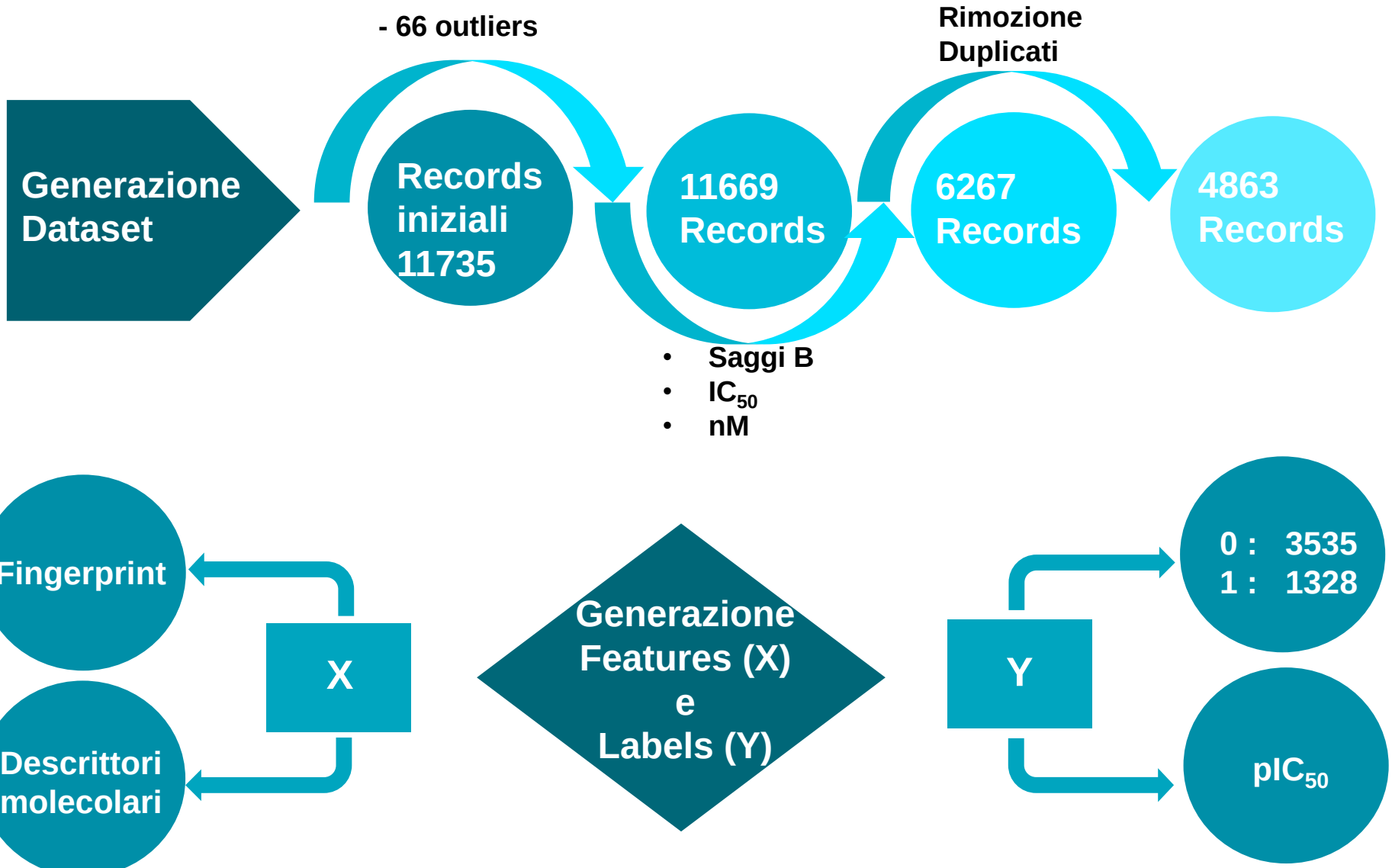


Procedura sperimentale

METODOLOGIA SPERIMENTALE

1. Generazione training set senza attività binarizzata
2. Modelli di classificazione e regressione
3. Validazione esterna modelli
4. Analisi similarità molecolare
5. Studio di selettività JAK2 / JAK1
6. Virtual Screening

Generazione train set



Modelli di classificazione

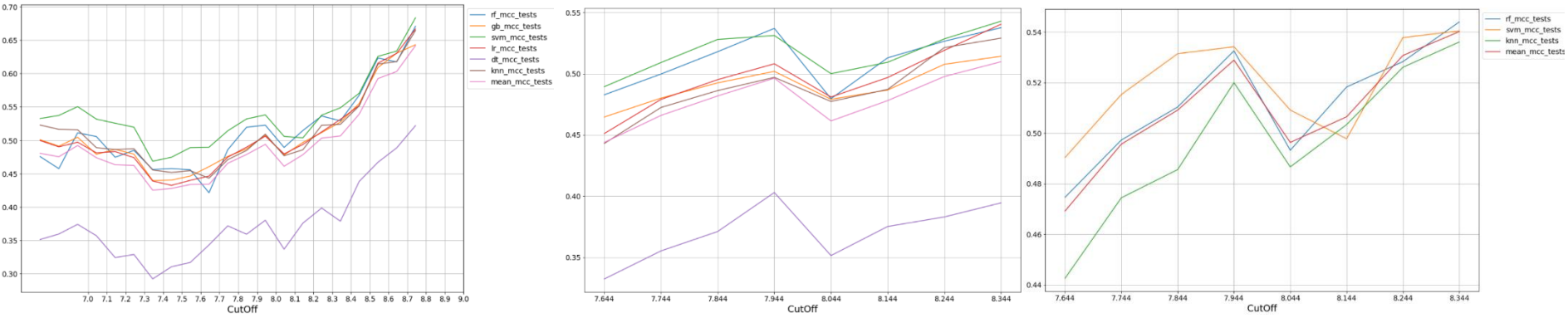
Stratified K-Fold
Train set 10%
Test set 90%

10 iterazioni
Stabilire l'intorno del cutoff

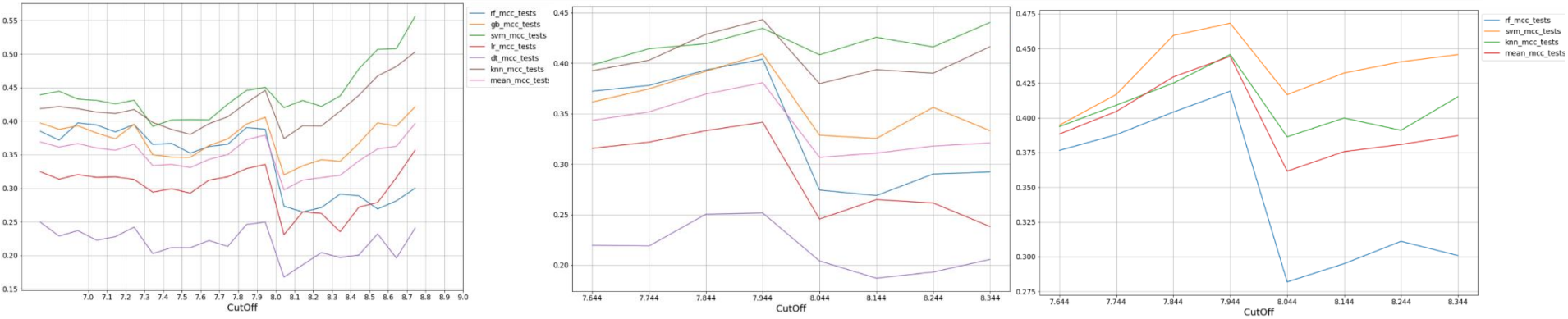
100 iterazioni
Selezione algoritmi migliori

1000 iterazioni
Scelta cutoff e algoritmi ottimali

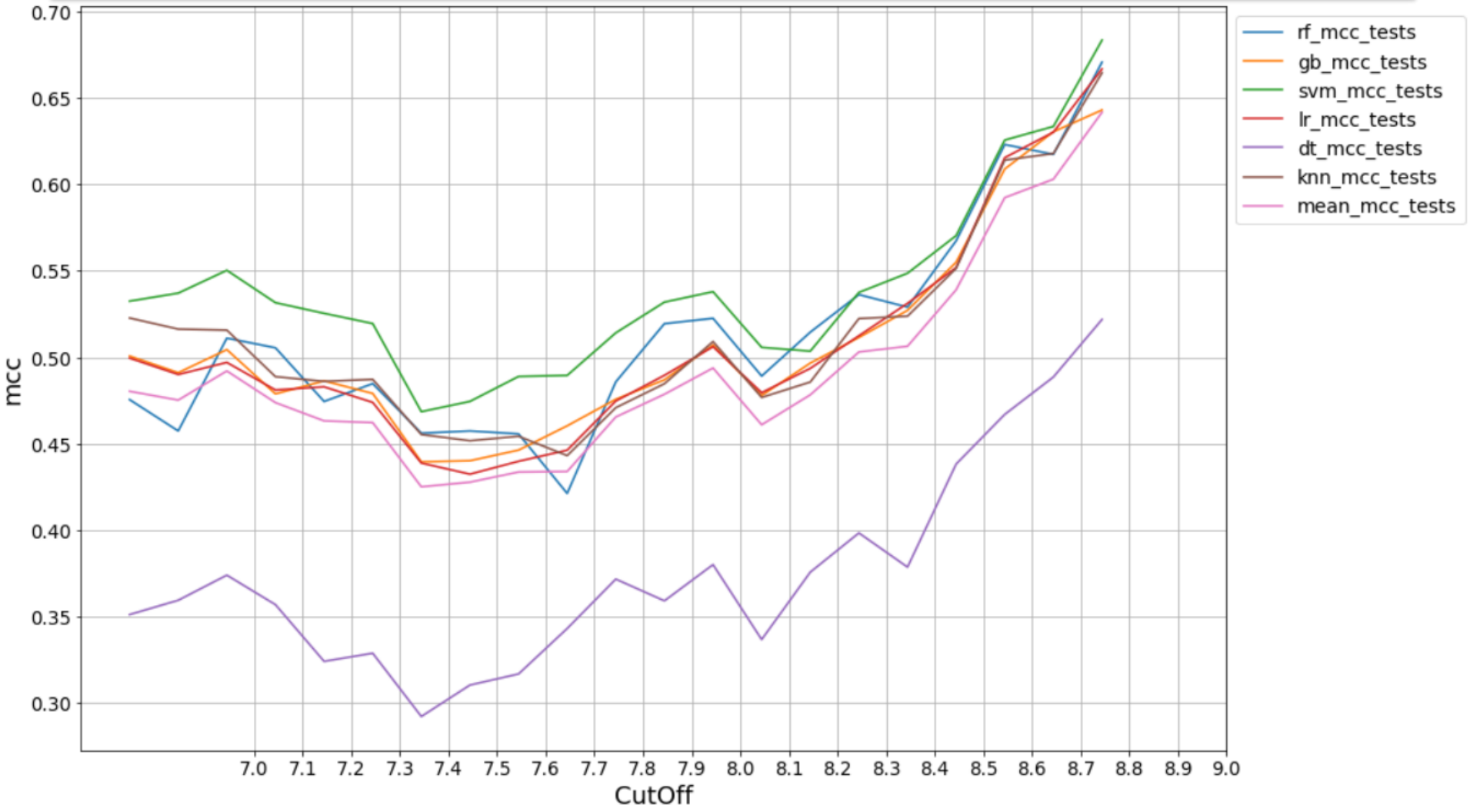
Fingerprint



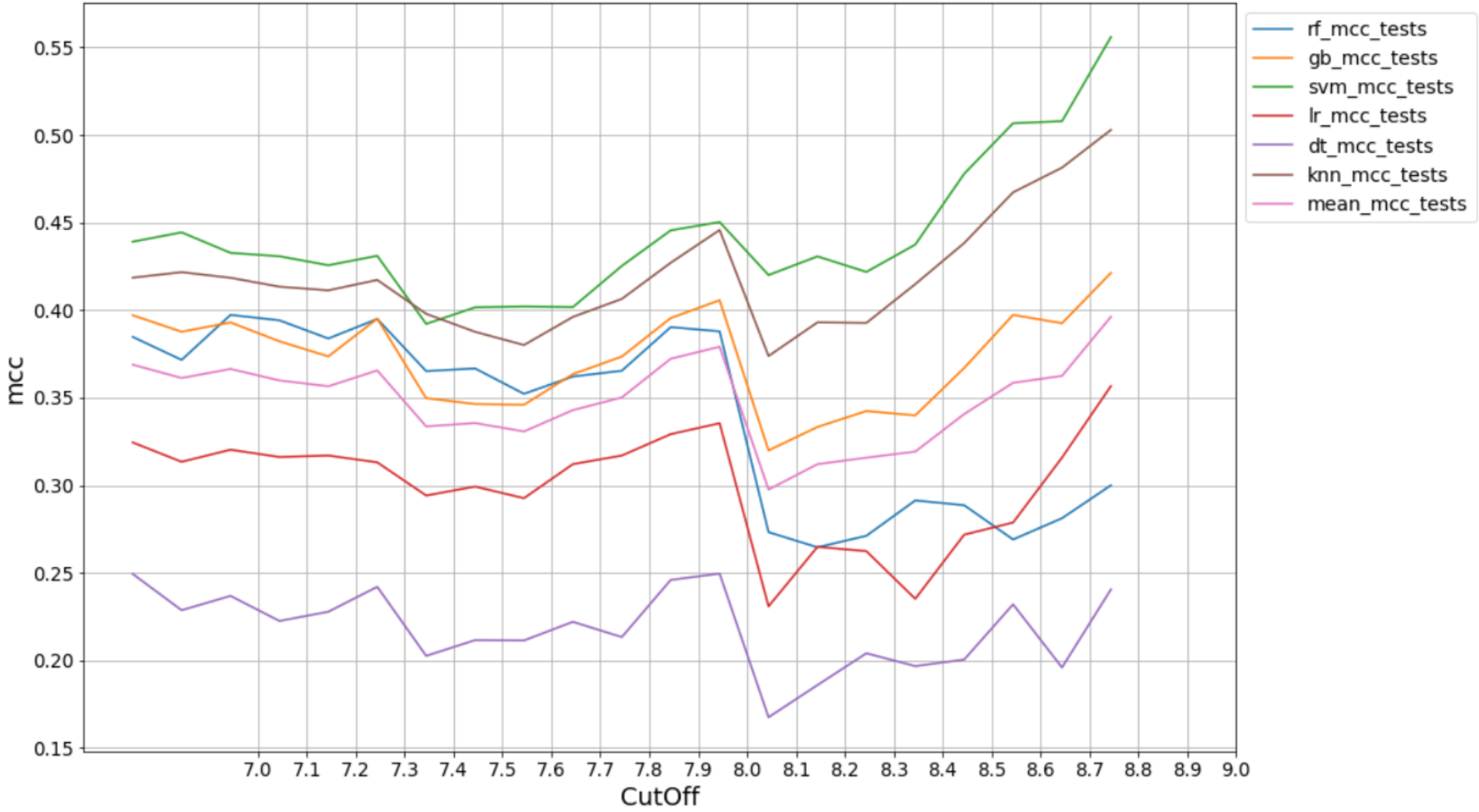
Descrittori



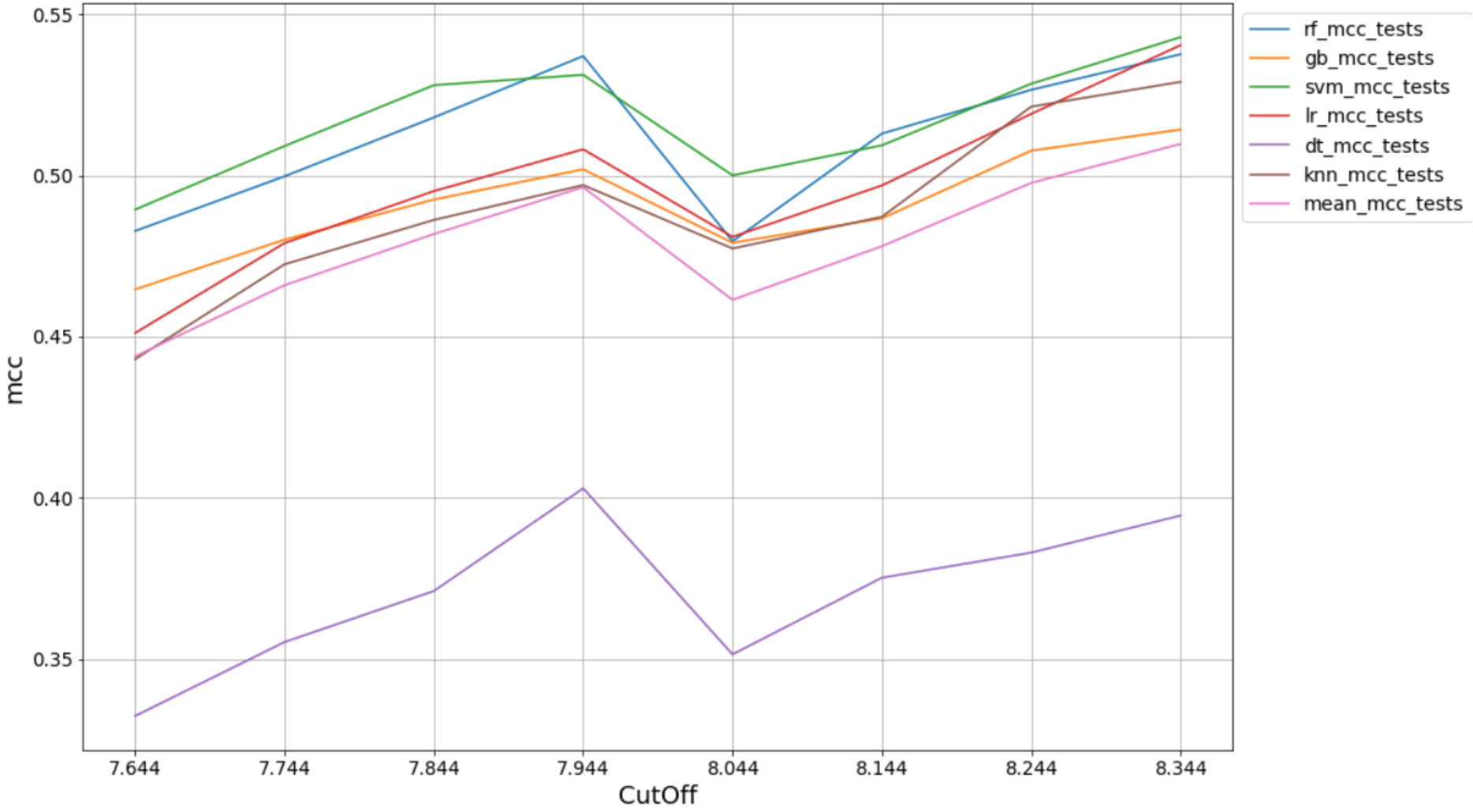
Fingerprint



Descrittori

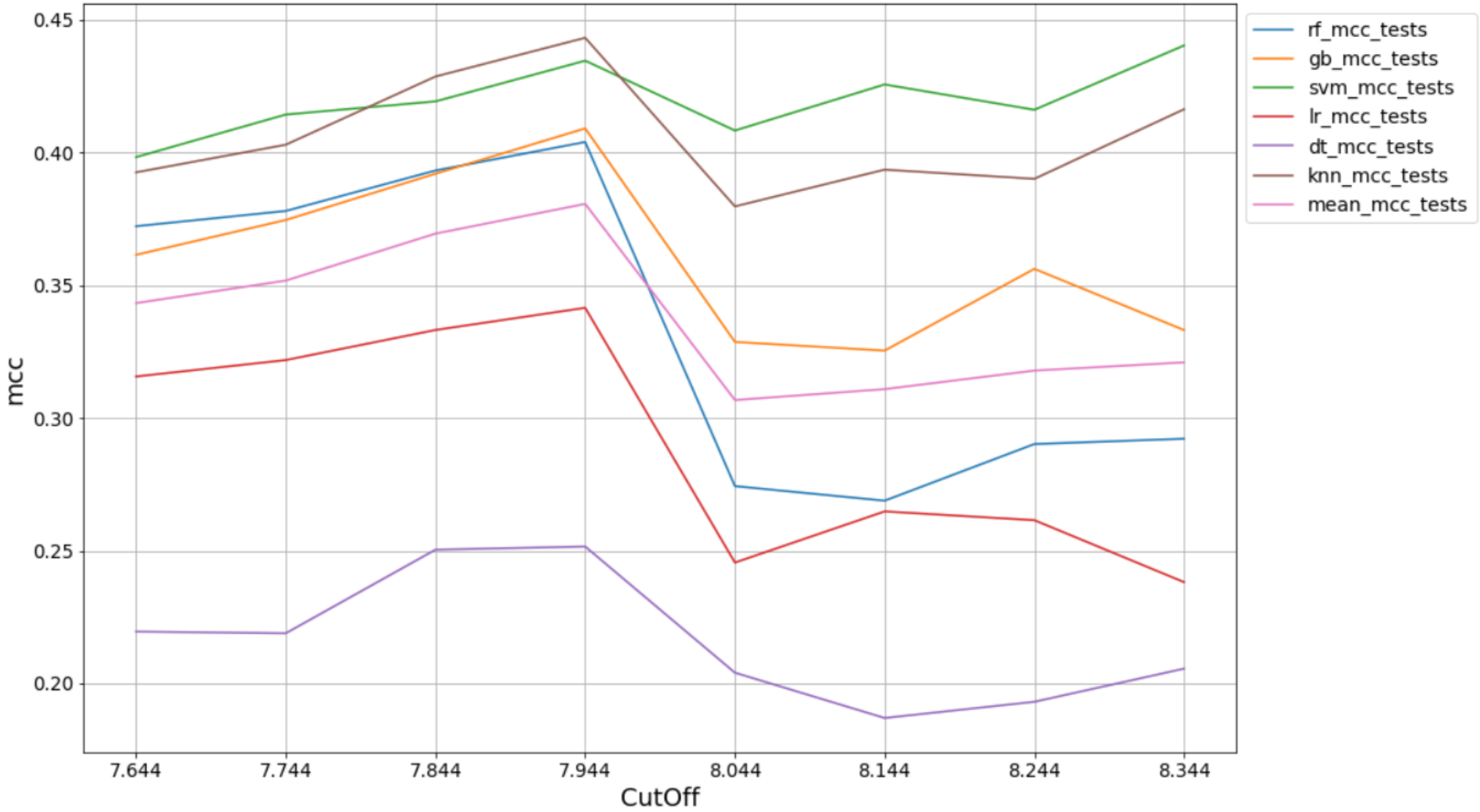


Fingerprint

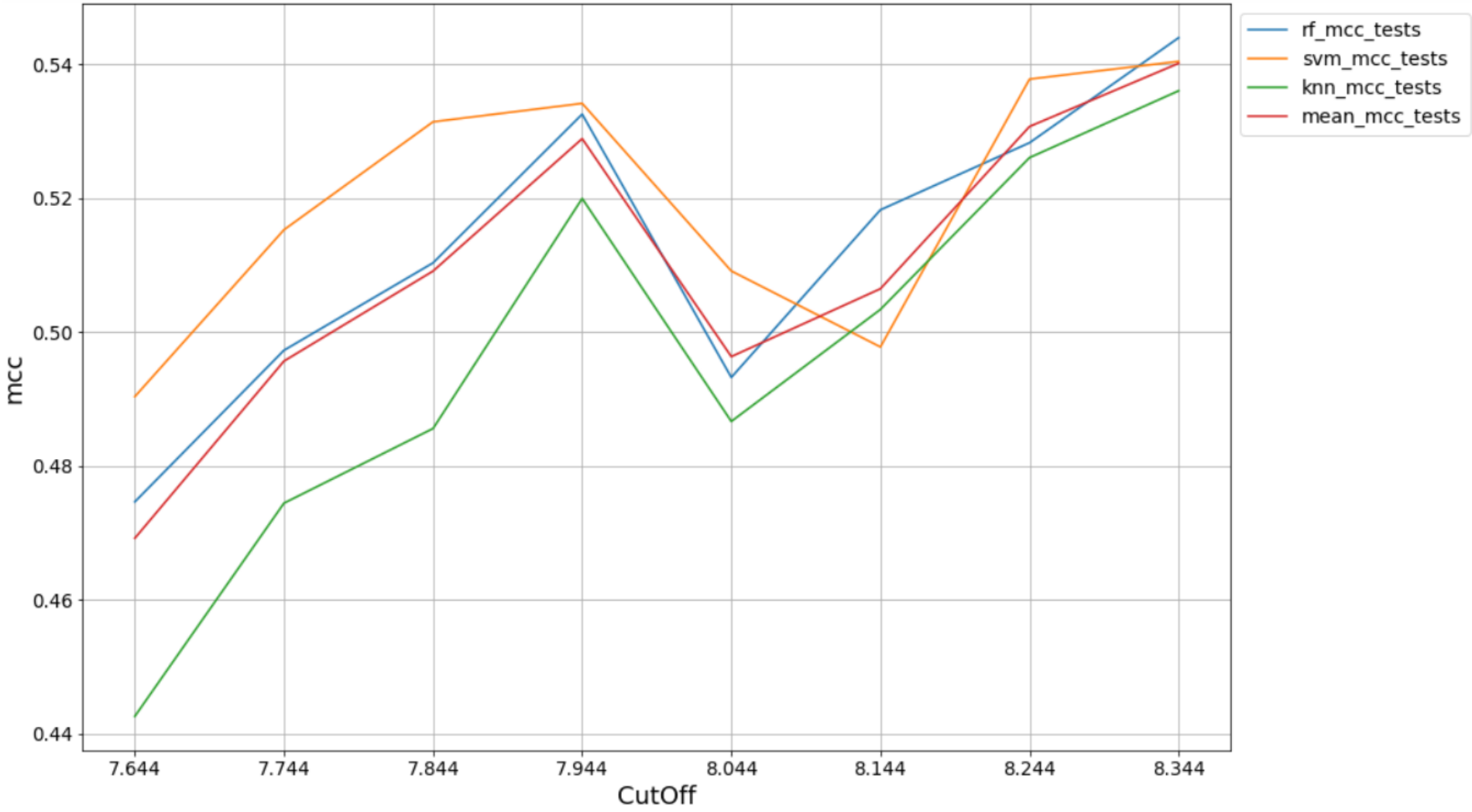


Modelli di classificazione

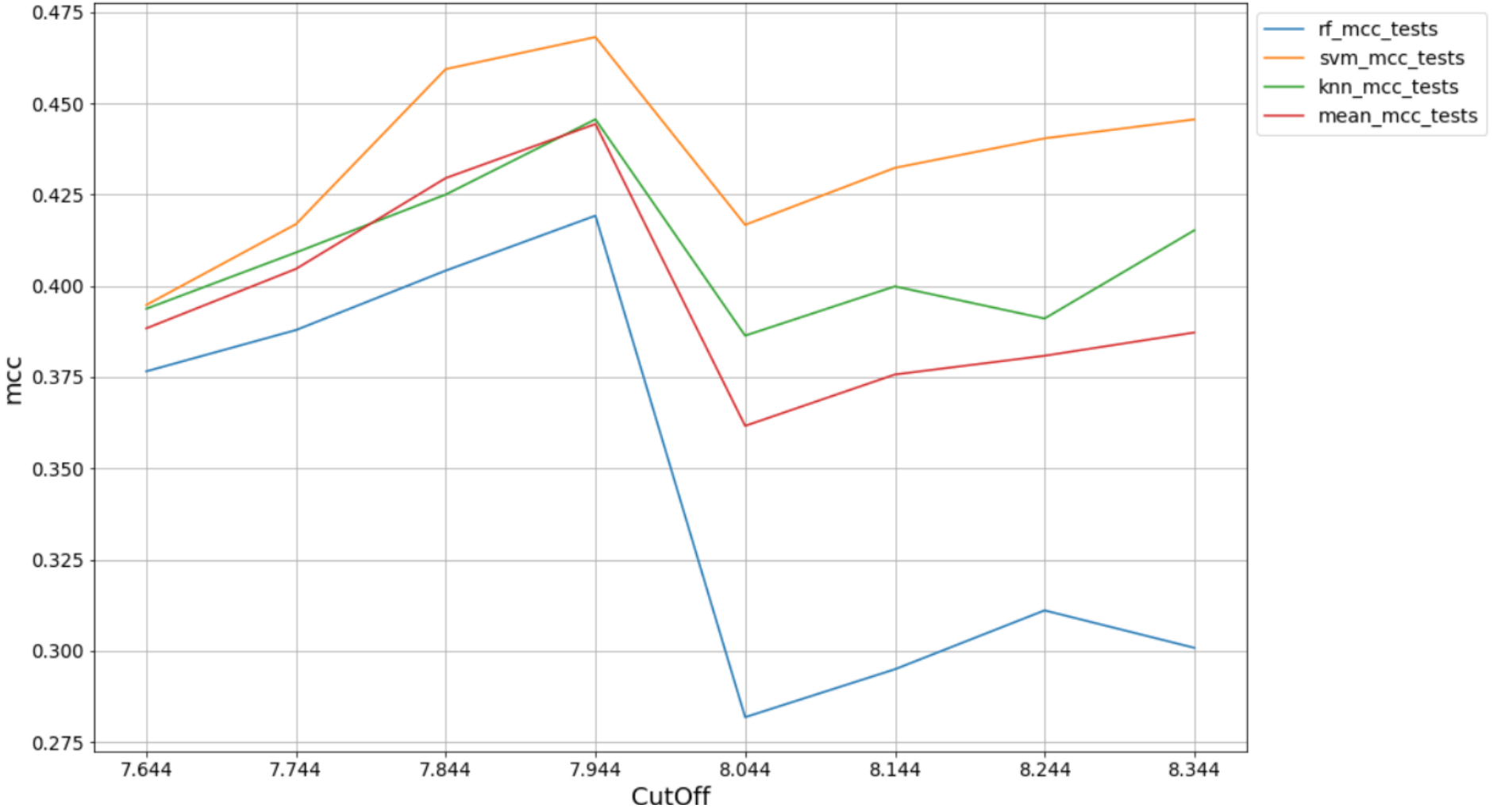
Descrittori



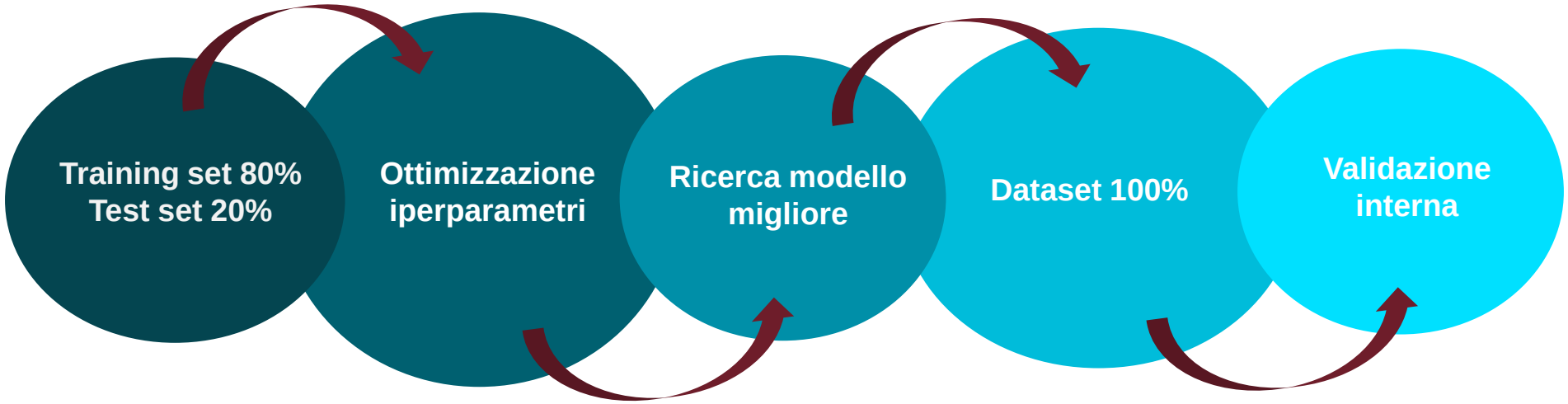
Fingerprint



Descrittori



Sviluppo modelli di classificazione e regressione



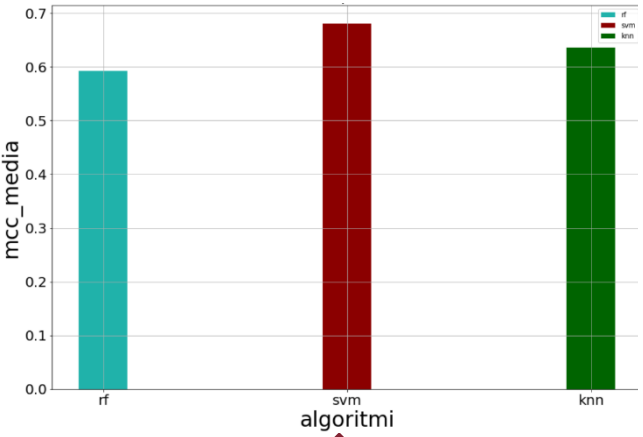
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

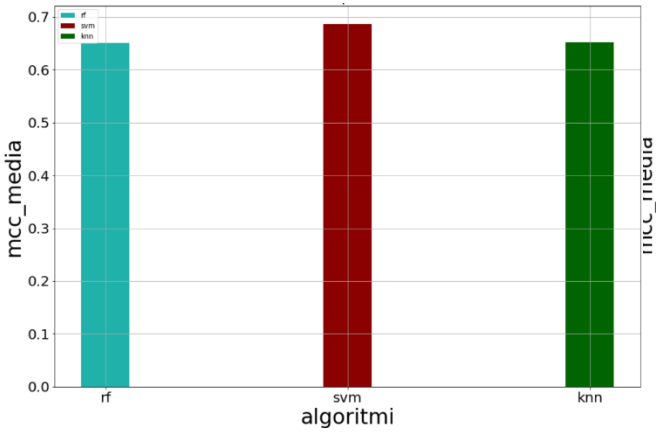
- -1 < MCC < 1;
- Misura della qualità delle classificazioni binarie;
- Metrica per validare modelli di classificazione.

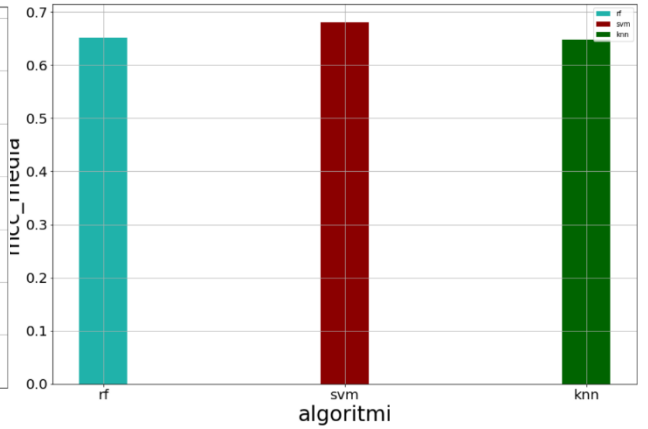
- veri positivi (TP): composti attivi correttamente predetti come attivi;
- falsi negativi (FN): composti attivi erroneamente classificati come inattivi;
- falsi positivi (FP): composti inattivi erroneamente classificati come attivi;
- veri negativi (TN): categoria che include i composti inattivi correttamente predetti inattivi

Modelli di classificazione

Fingerprint

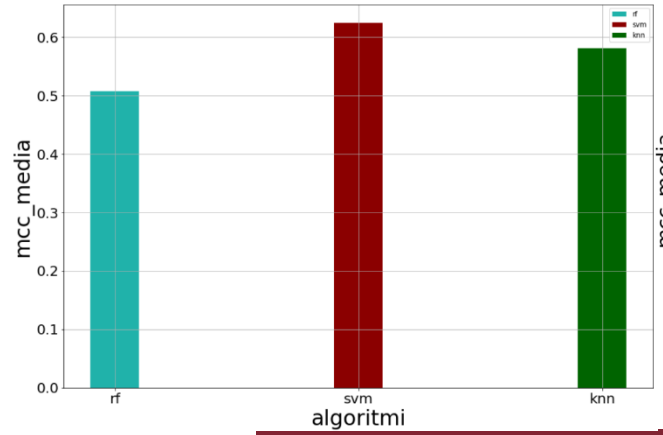


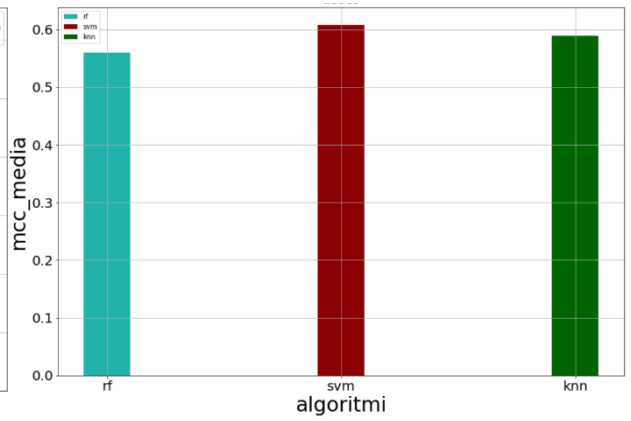


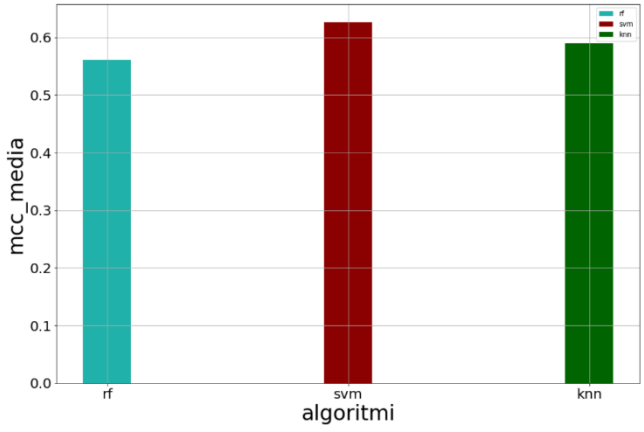


10 iterazioni | **100 iterazioni** | **1000 iterazioni**

Descrittori





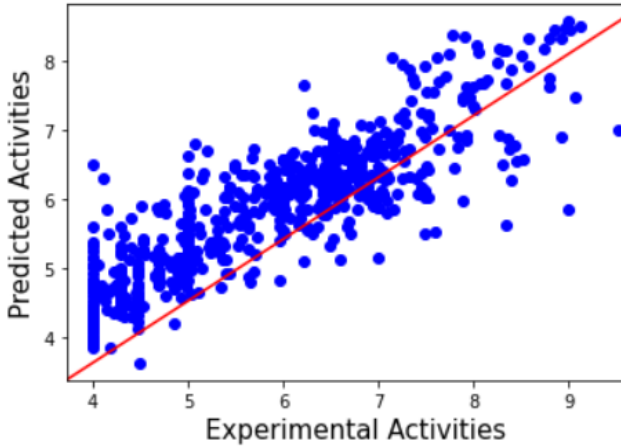


Training regressione

R^2	Fingerprint	Descrittori	Descrittori scalati
SupportVectorMachine	0.77	0.64	0.63
RandomForest	0.69	0.54	0.65
Ridge	0.67	0.23	0.34

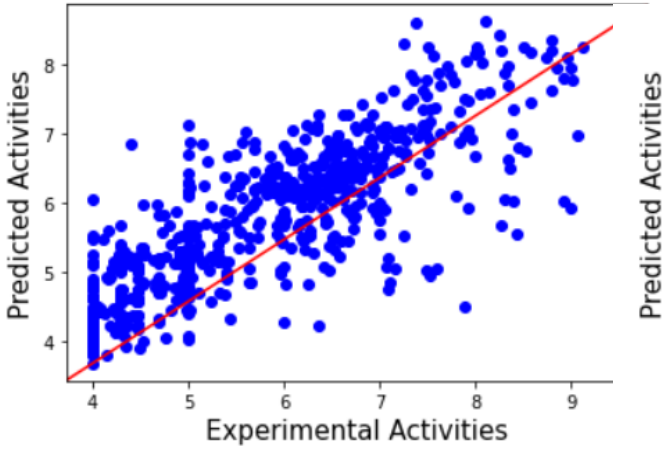
SVR Fingerprint

C=5, Kernel = rbf



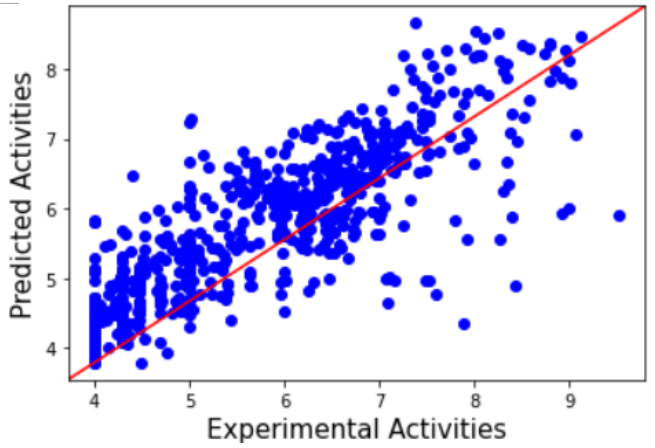
SVR Descrittori

C=9, Kernel = rbf



RF Descrittori scalati

Max_depth=20, n_estimators=104



Validazione interna classificazione e regressione

CLASSIFICAZIONE

	MCC_FIT	MCC_CV	MCC_EXT
Fingerprint	0.95	0.67	0.72
Descrittori molecolari	0.80	0.59	0.65

REGRESSIONE

	Q^2	SDEP
Fingerprint	0.77	0.58
Descrittori molecolari	0.66	0.71
Descrittori scalati	0.54	0.82

- Quanto un modello è predittivo

SDEP (Standard Deviation Error in Prediction)

- Confrontabile con SDEP_CV

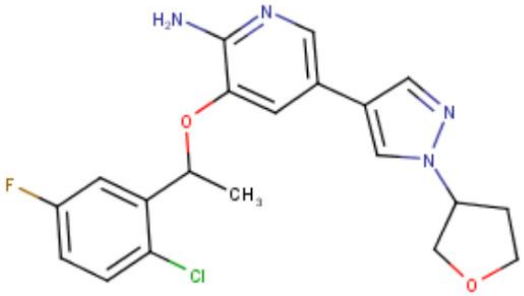
Analisi della similarità

Ricerca test set

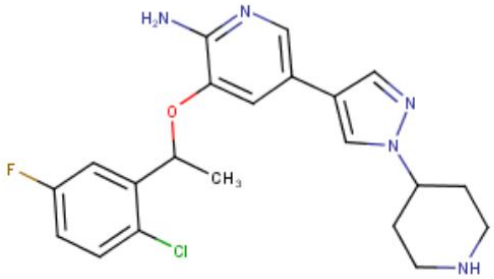
Confronto train e test set

Generazione features e labels

62 molecole:
17 attive
45 inattive



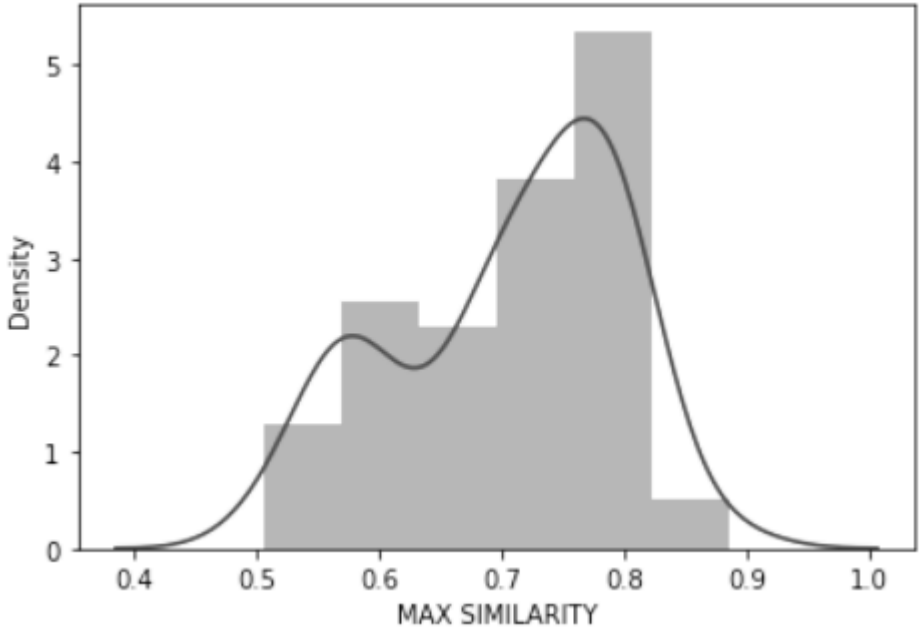
Test



Train

Max similarity: 0.73

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



Validazione esterna classificazione e regressione

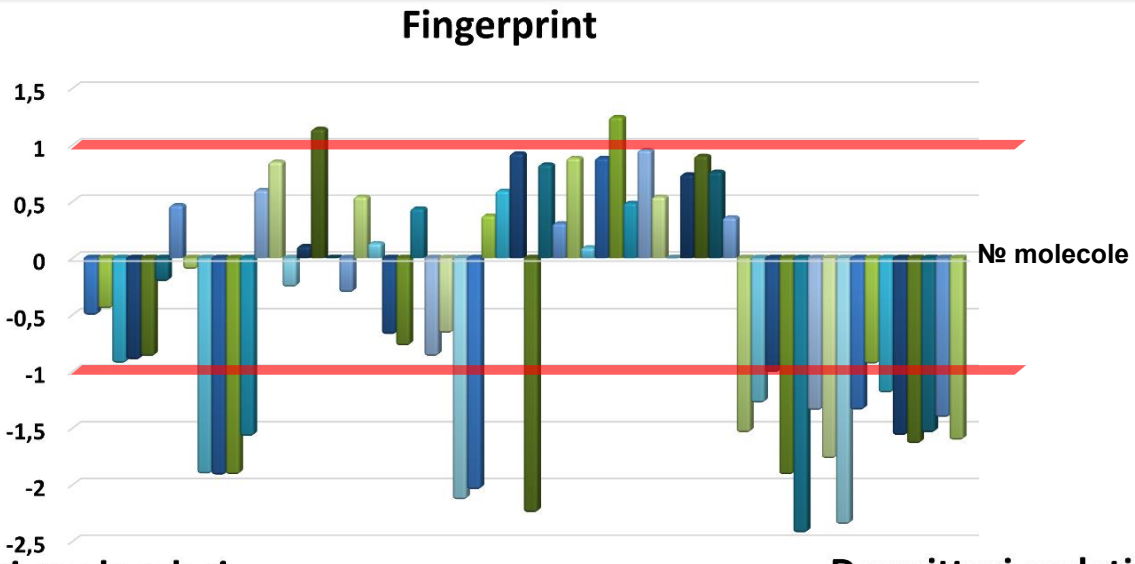
Classificazione

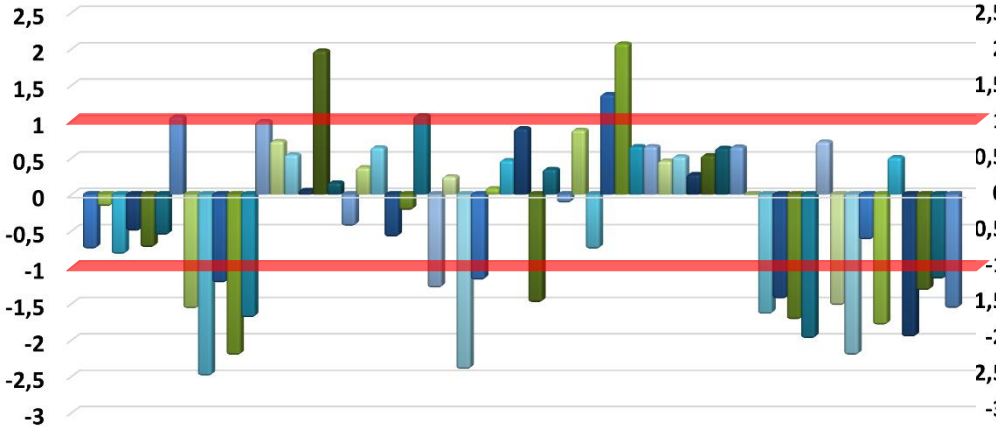
	Fingerprint	Descrittori molecolari
Accuracy	0.56	0.55
F1 Score	0.53	0.48
Confusion Matrix	[[20,25] [2,15]]	[[21,24] [4,13]]
Recall	88%	76%

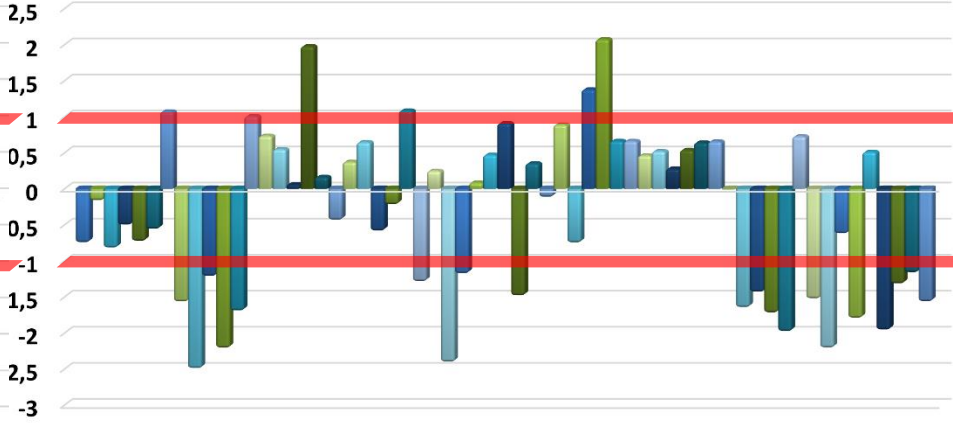
Regressione

	Fingerprint	Descrittori molecolari	Descrittori scalati
SDEP	1.18	1.20	1.06

Errori di Predizione







Selettività JAK2 – JAK1

Confronto dataset JAK2 – JAK1 → Selezione molecole con doppia attività → Definizione del rapporto tra i due pIC_{50} → Modelli di classificazione e regressione

182 Molecole selettive JAK2

1756 Molecole selettive JAK1

944 Molecole non selettive

Modelli di classificazione e regressione	MCC	Q^2
Fingerprint	0.90	0.80
Descrittori molecolari	0.90	0.79
Descrittori molecolari scalati	/	0.78

Virtual screening e sviluppi futuri

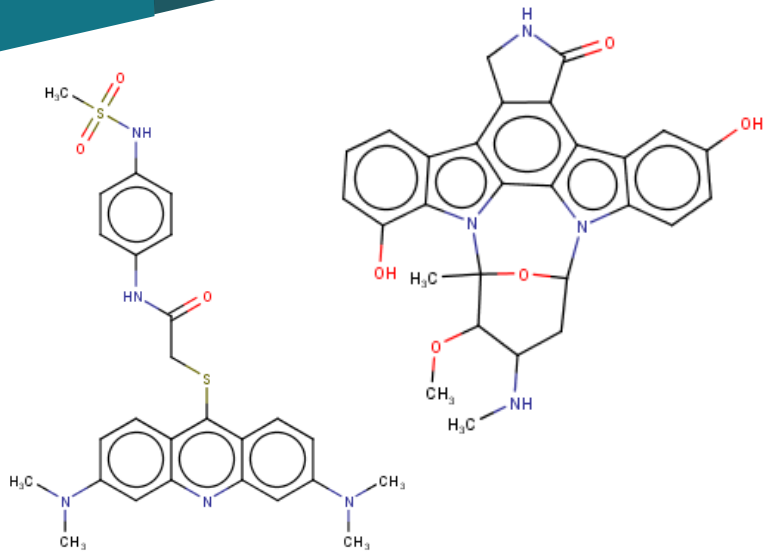
National Cancer Institute 252128 records

Features selection

Predizione attività

Consensus scoring

Selezione delle top 40 molecole più attive



Sviluppi futuri

- Saggi biologici sulle top 40 molecole
- Applicazione modelli elaborati a tutte le JAK
- Individuare molecole che indicano la selettività verso tutte le altre JAK

RINGRAZIAMENTI

- Professor Rino Ragno
- Dottoressa Manuela Sabatino
- Dottore Lorenzo Antonini
- Dottoressa Giulia Fantera
- Dottoressa Paola Caprioli
- Beatrice Foti
- Maria Stella Delle Chiaie