Sviluppo di modelli QSAR predittivi mediante tecniche di Machine Learning: applicazione ad inibitori di HDAC6

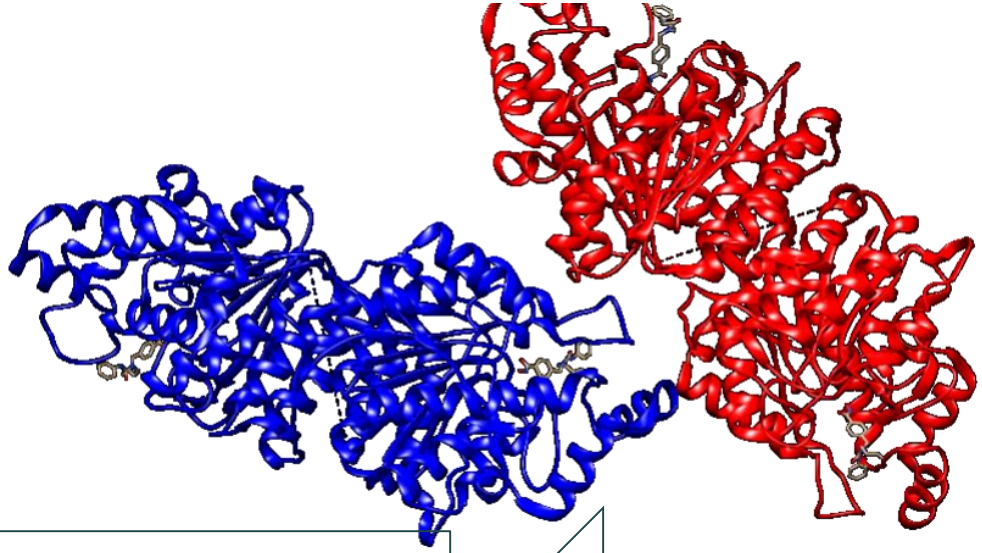
SAPIENZA
UNIVERSITÀ DI ROMA

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

Laureanda: Beatrice Foti
Matricola: 1754003

Relatore: prof. Rino Ragno

Scopo del Lavoro



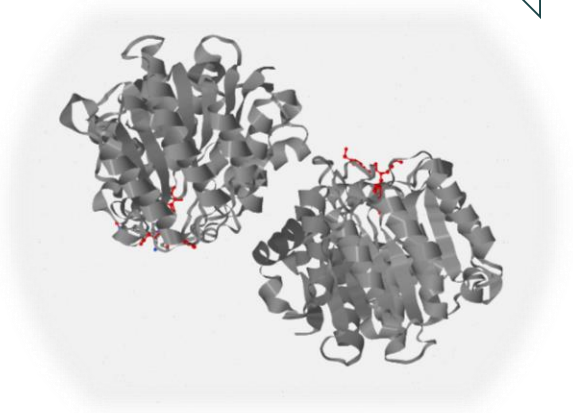
Apprendimento Automatico Relazione Struttura-Attività di Tipo Quantitativa (QSAR)

QSAR

MACHINE LEARNING

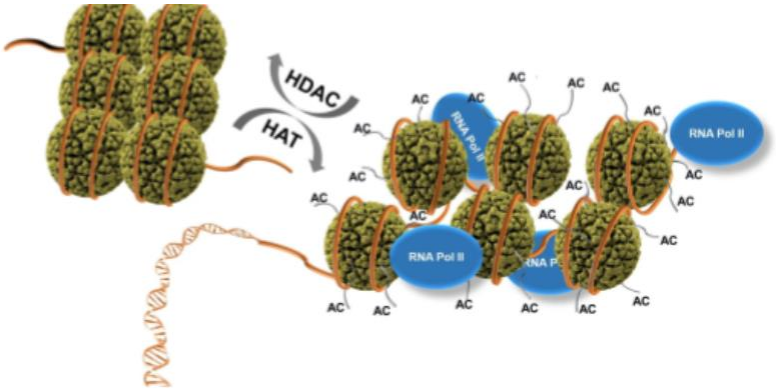
VIRTUAL SCREENING

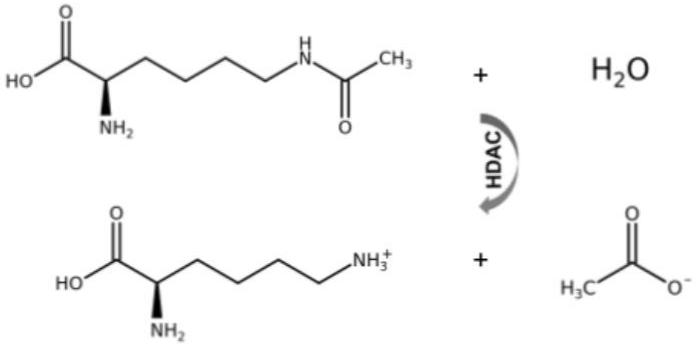
IDENTIFICAZIONE POTENZIALI HIT COMPOUNDS



Target: HDAC6

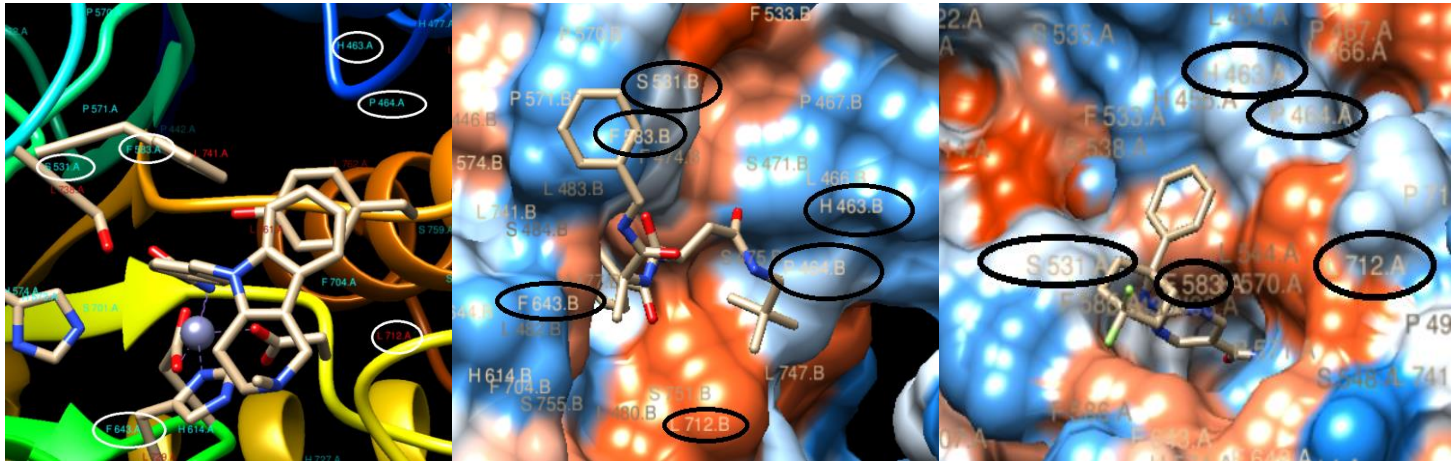
Group	Class	Name	Location in cell	Location in body
Classical (Zn dependent)	Class I (Rpd3)	HDAC1 HDAC2 HDAC3 HDAC8	Nucleus	Ubiquitou
	Class IIa (Hda1)	HDAC4 HDAC5 HDAC7 HDAC9	Nucleus/ cytoplasm	Tissue specific
	Class IIb (Hda1)	HDAC6 HDAC10	Cytoplasm	Tissue specific
	Class IV (Rpd3/Hda1)	HDAC11	Nucleus/ cytoplasm	Tissue specific
NAD dependent	Class III	SIRT (1-7)	Nucleus/ cytoplasm	

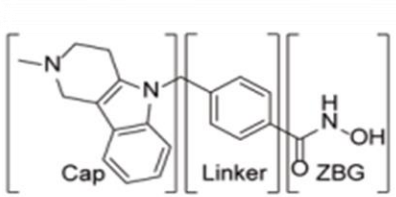




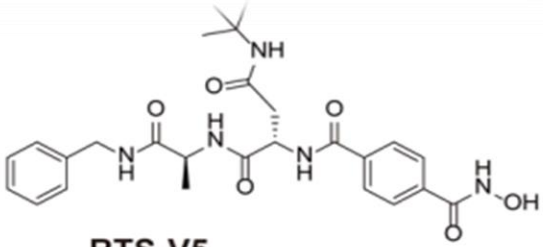
Inibitori HDAC6

CAP GROUP — LINKER — GRUPPO LEGANTE LO Zn^{2+}

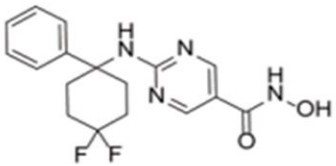




Tubastatin A
HDAC1, IC_{50} = 8,100 nM
HDAC6, IC_{50} = 4.4 nM



RTS-V5
HDAC1, IC_{50} = 6,900 nM
HDAC6, IC_{50} = 270 nM

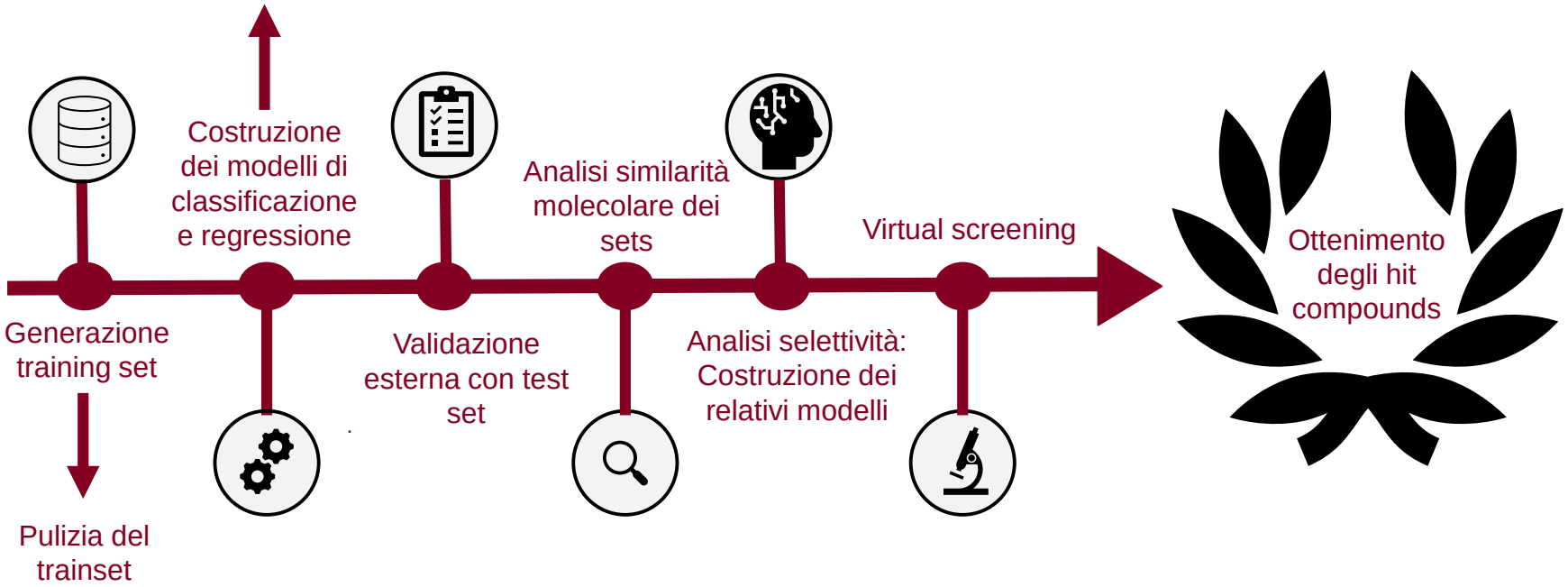


ACY-1083
HDAC1, IC_{50} = 780 nM
HDAC6, IC_{50} = 3 nM

Rational Design of Suprastat: A Novel Selective Histone Deacetylase6 Inhibitor with the Ability to Potentiate Immunotherapy in Melanoma Models

Procedura Sperimentale

- RandomForest(rf)
- SupporterVectorMachine (SVM)
- GradientBoosting(gb)
- DecisionTree(dt)
- LogisticRegression(lr)
- K-nearest neighbors(knn)



Classificazione

	Predetto 0	Predetto 1
Reale 0	**TN**	**FP**
Reale 1	**FN**	**TP**

Matrice di confusione

Matthews correlation coefficient (MCC)

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

Regressione

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(Yexp,i - Ycalc,i)^2}{\sum_{i=1}^{n}(Yexp,i - \overline{Y})^2}$$

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(Yexp,i - Ypred,i)^2}{\sum_{i=1}^{n}(Yexp,i - \overline{Y})^2}$$

$$SDEP = \sqrt{\frac{\sum_{i=1}^{N}(Yexp,i - Ypred,i)^2}{N}}$$

Generazione Training Set

Generazione Dataset

RECORD INIZIALI 5476

- 51 OUTLIERS 5425

SAGGI B(95%) IC_{50}(84%) 3551

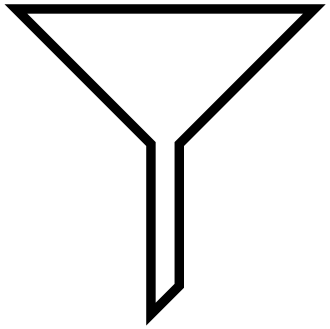
- 682 DUPLICATI 2869

Generazione Features (x)

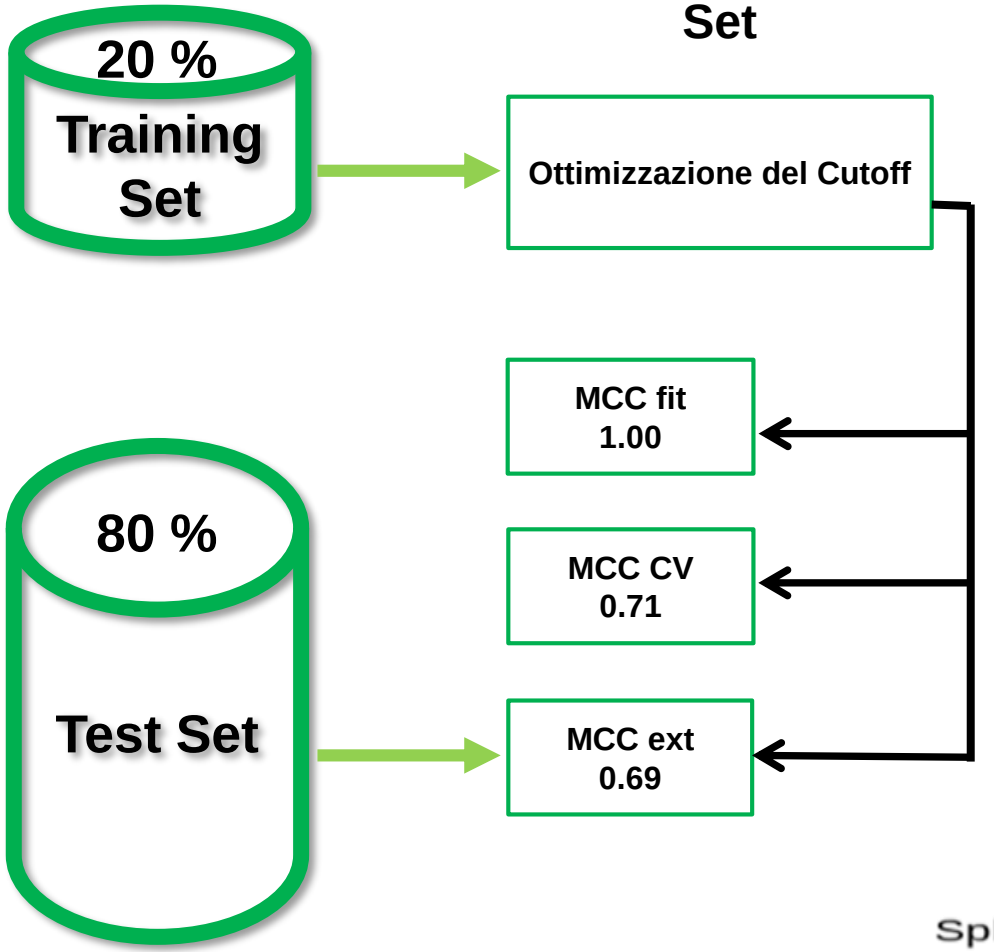
X

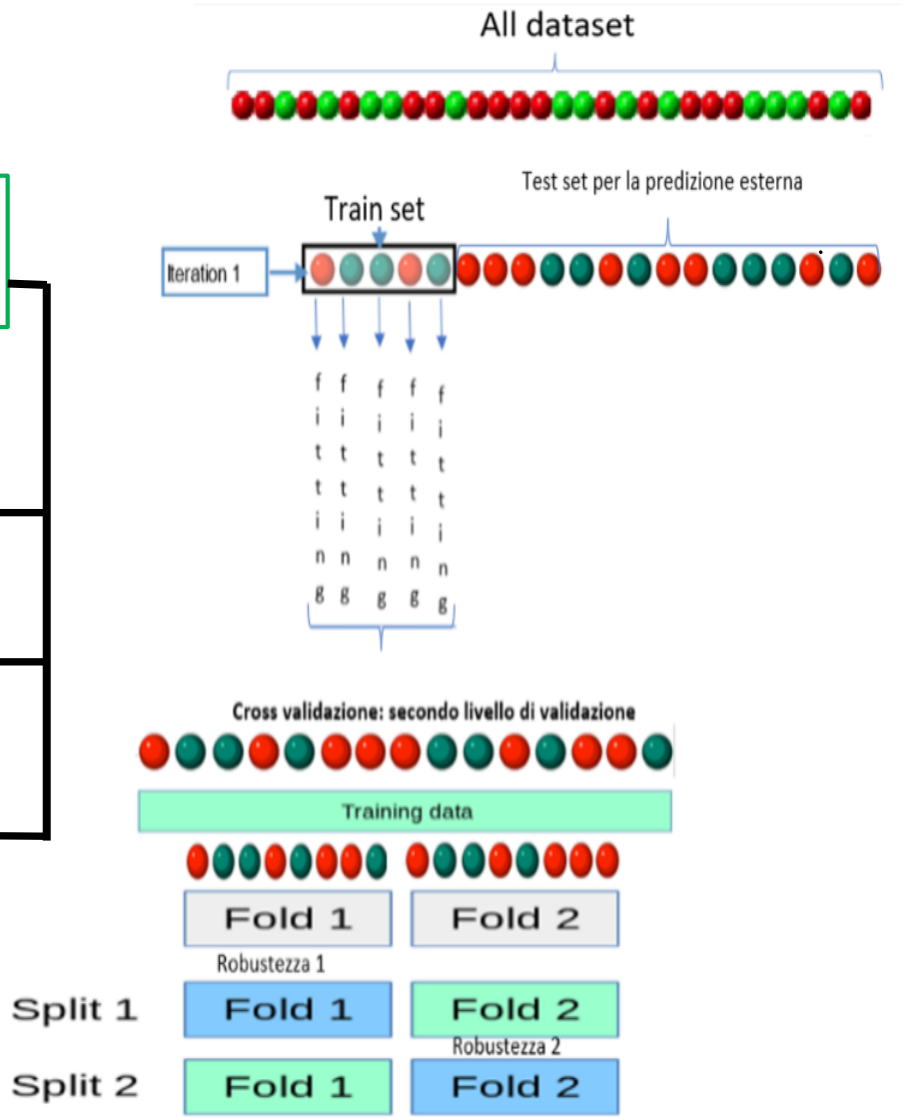
DESCRITTORI MOLECOLARI

FINGERPRINT

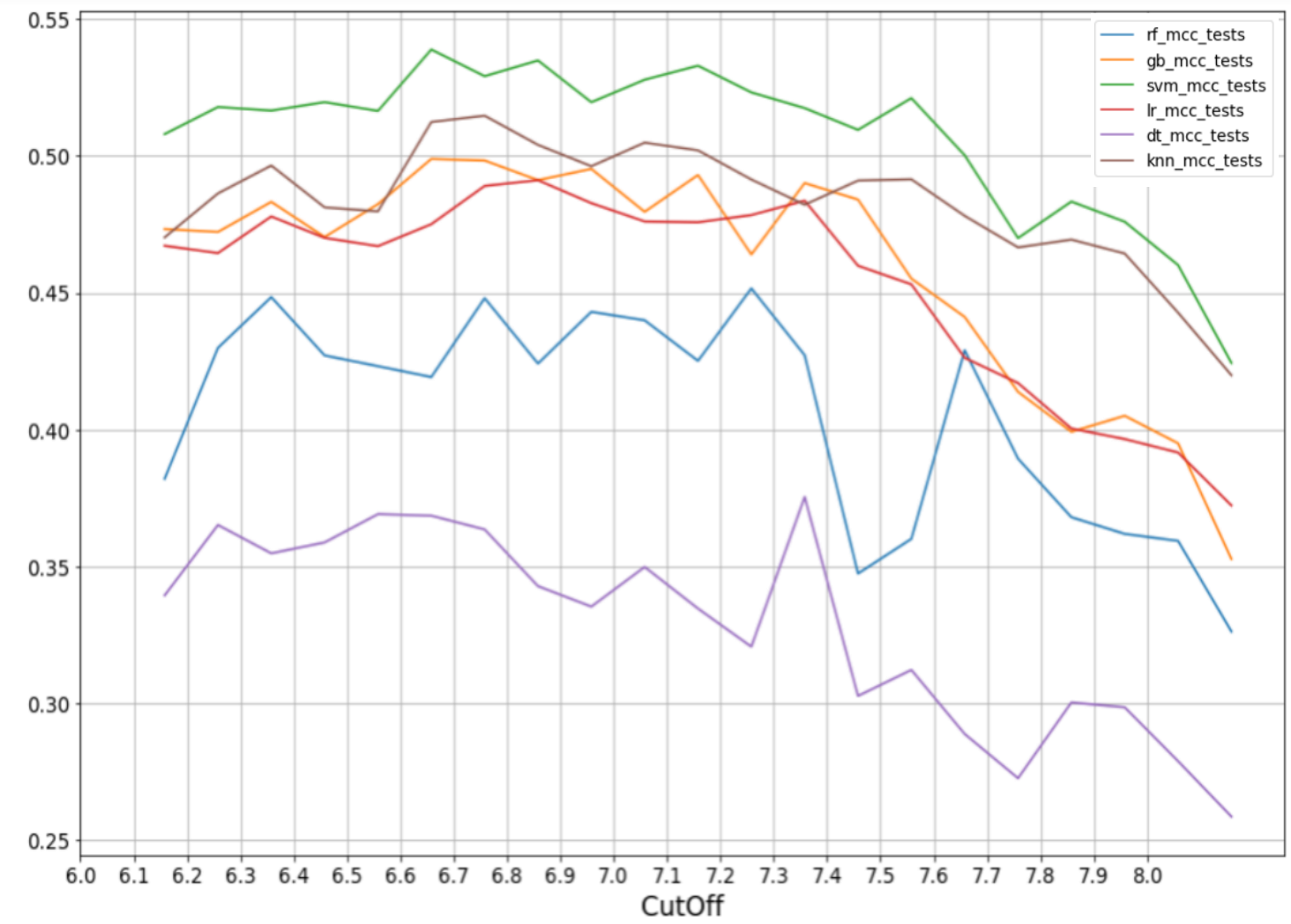


Sviluppo Modelli di Classificazione



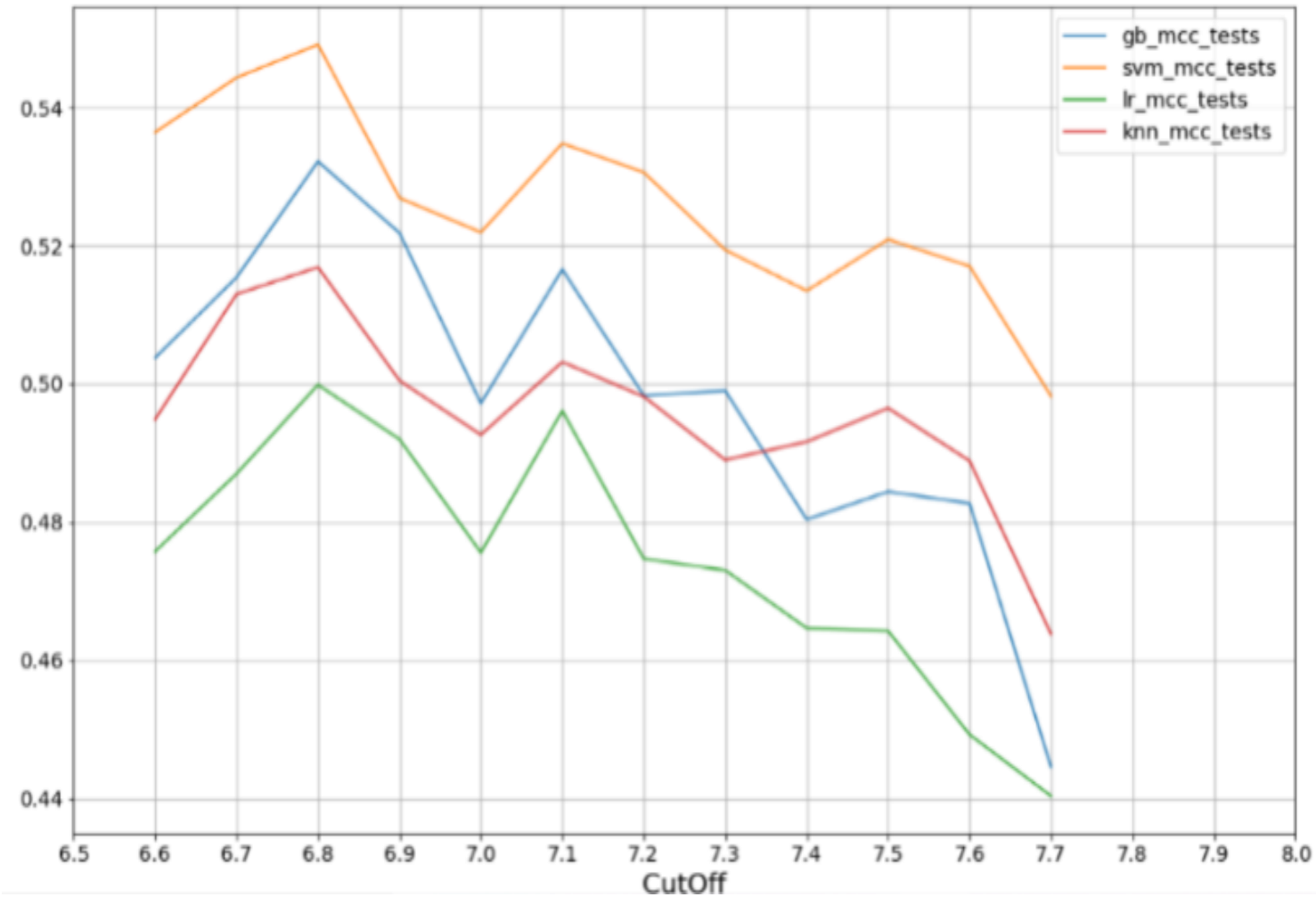


Ottimizzazione del Cutoff



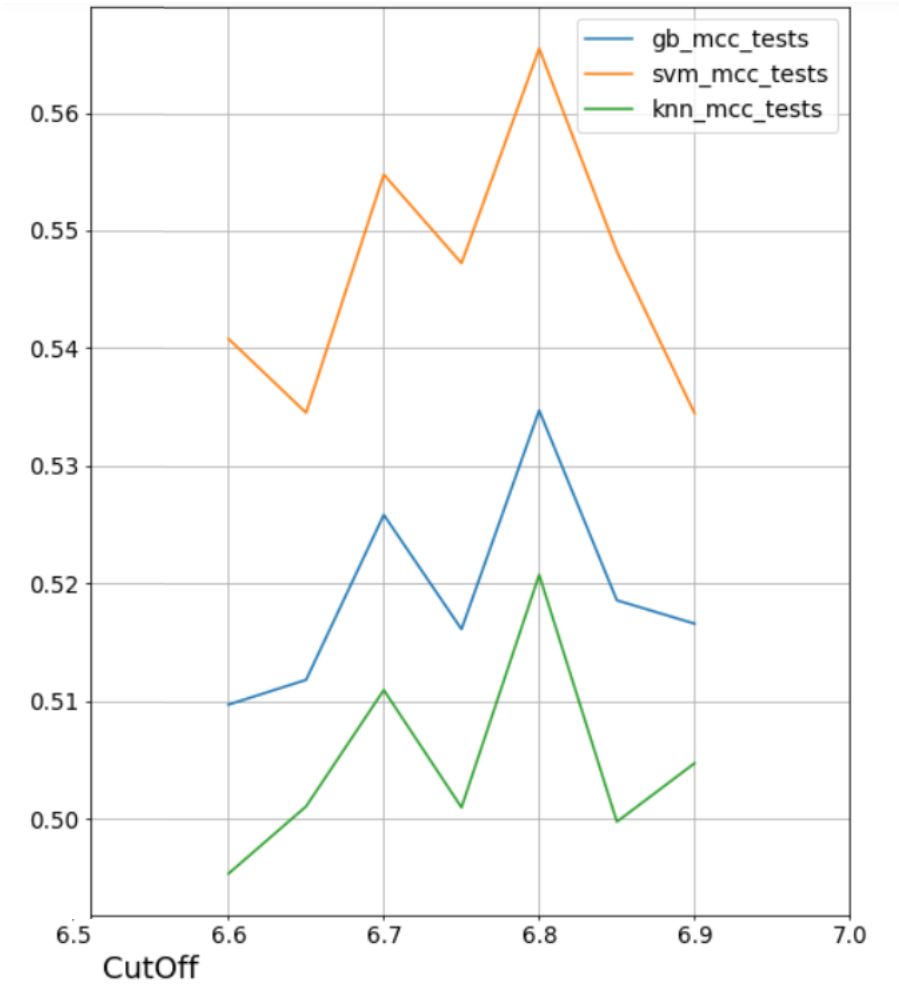
10 iterazioni

Ottimizzazione del Cutoff



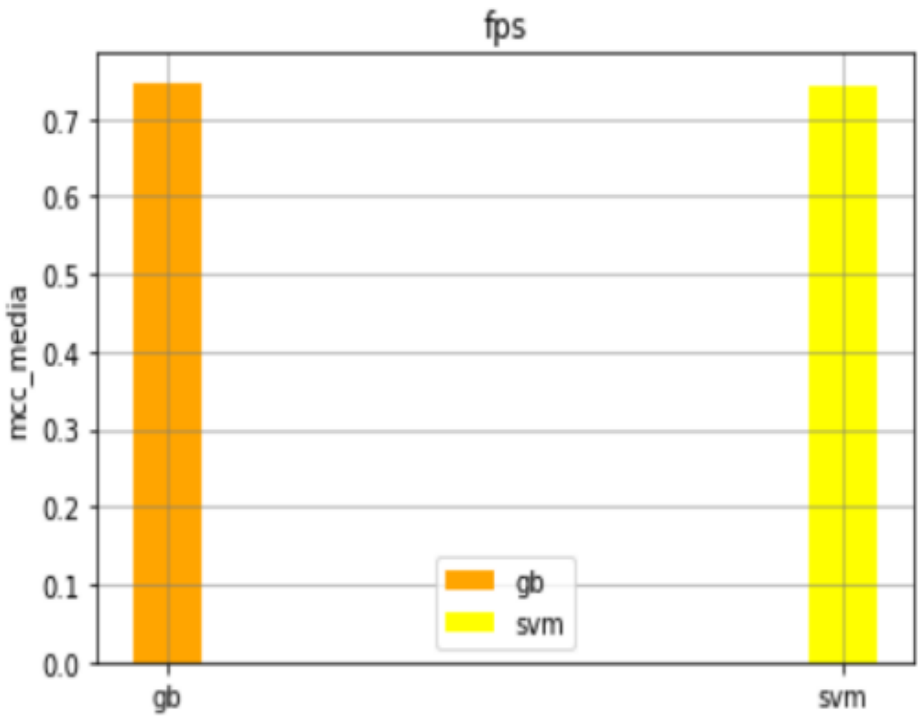
100 iterazioni

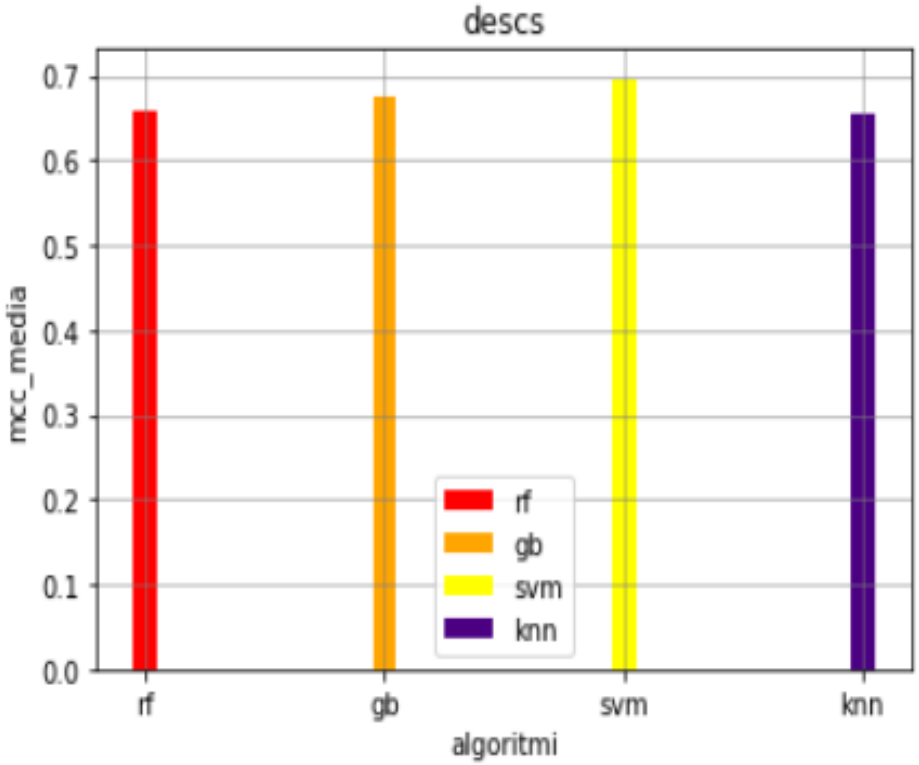
Ottimizzazione del Cutoff



1000 iterazioni

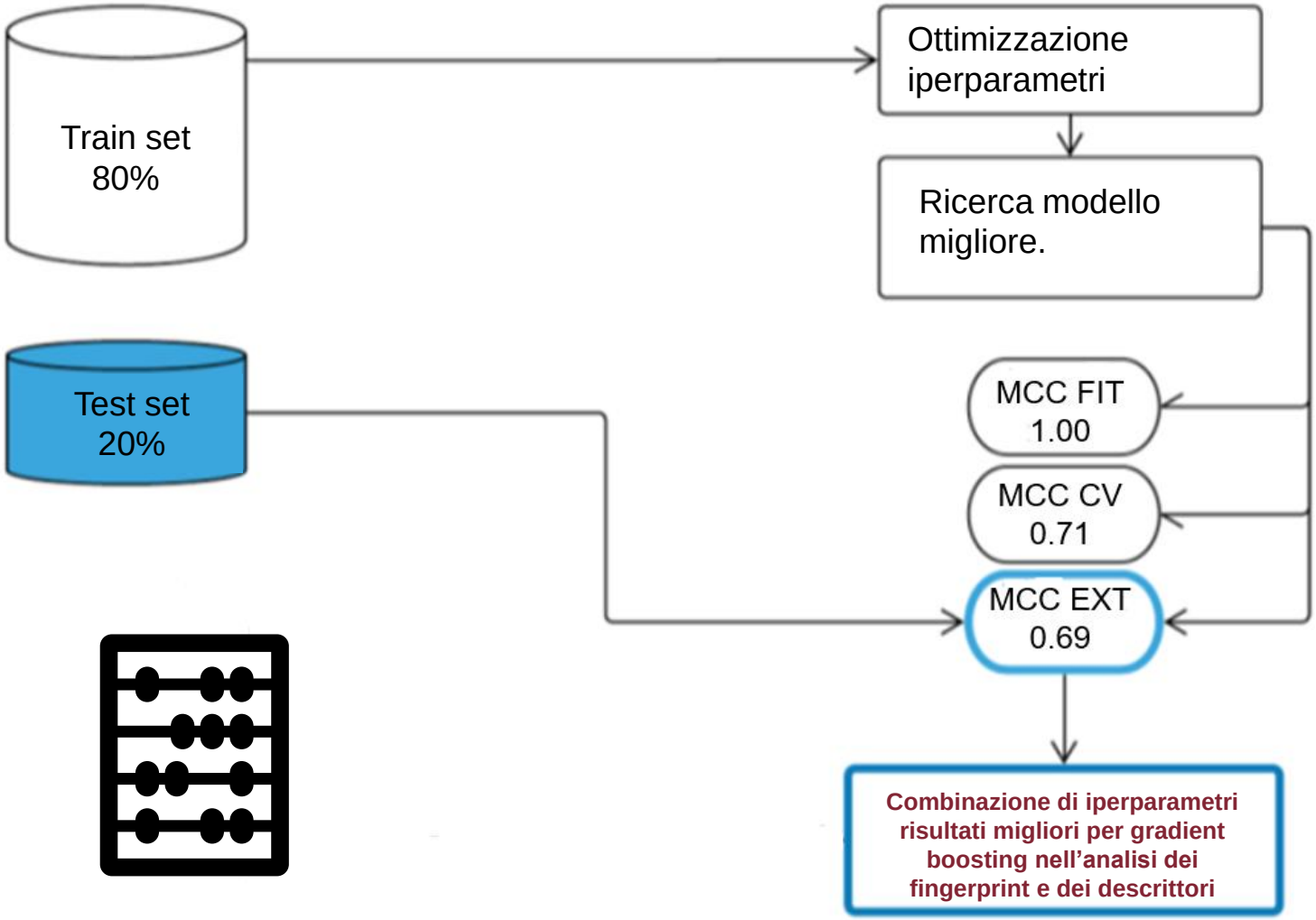
Selezione dell'Algoritmo Migliore



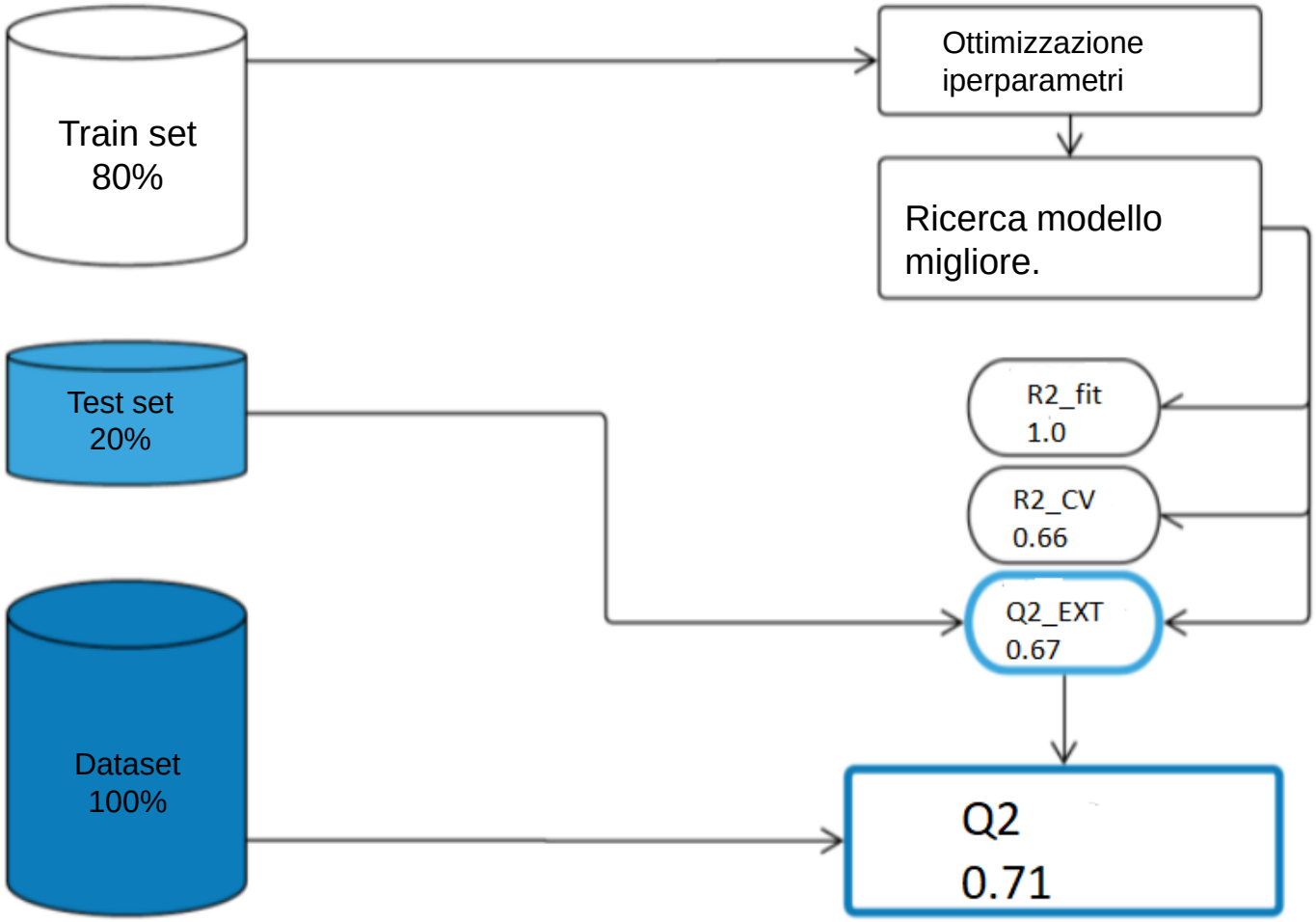


1000 iterazioni

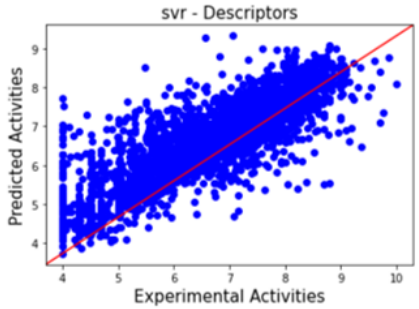
Ottimizzazione Iperparametri



Sviluppo modelli di regressione

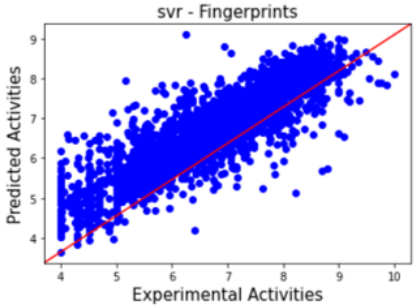


Risultati training regressione



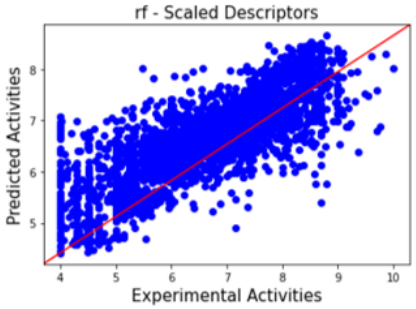
	R2_EXT
svr	0.704209
rf	0.603799
knr	0.672903

SVR (C=6,KERNEL='rbf')



	R2_EXT
rf	0.775902
knr	0.711704
svr	0.784629

SVR (C=5,KERNEL='rbf')



	R2_EXT
rf	0.743856
knr	0.719293
svr	0.716737

rf(max_depth=49,n_estimators=32)

Validazione esterna

TEST SET

37 MOLECOLE ATTIVE

4 MOLECOLE INATTIVE

	MCC	MATRICE DI CONFUSIONE	SDEP EXT	SDEP CV
DESC	0.33	[[2, 2], [4, 33]]	0.99	0.71
FPS	0.63	[[3, 1], [2, 35]]	0.86	0.60

Analisi selettività : costruzione modelli

CONFRONTO TRA IL DATASET DEGLI HDAC6i e QUELLO DEI HDAC4i

SELEZIONE MOLECOLE CON DOPPIA ATTIVITÀ

DEFINIZIONE DEL RAPPORTO TRA I DUE pIC_{50}

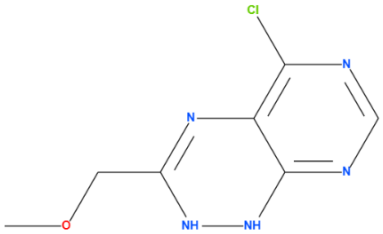
```
selectivity = []
for i in df_hdac6_4['RATIO_HDAC6_4']:
    if i >= 0.5 : selectivity.append(1)
    if i <= -0.5 : selectivity.append(0)
    if i > -0.5 and i < 0.5 : selectivity.append(-1)
```

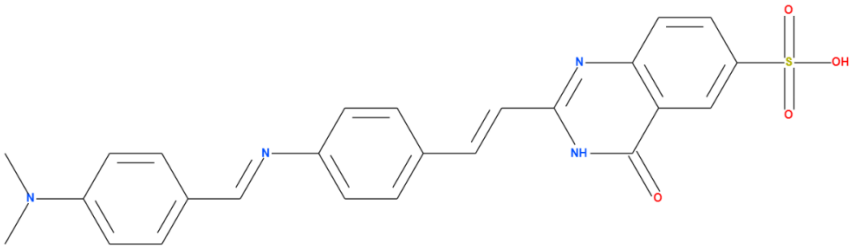
530 molecole
186 HDAC6 selettive
160 HDAC4 selettive
184 non selettive

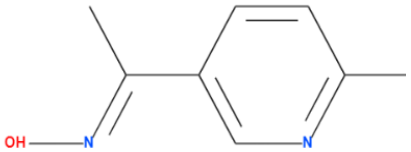
Risultati training regressione e classificazione selettività

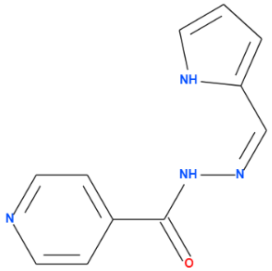
MCC	FINGERPRINT	DESCRITTORI MOLECOLARI
SupportVectorMachine	0.947	0.840
GradientBoosting	0.937	0.919
KneighborsClassifier	0.947	0.947

R^2	FINGERPRINT	DESCS.MOL.	DESCS.MOL.SCAL
SupportVectorMachine	0.783	0.739	0.777
KneighborsClassifier	0.784	0.716	0.734
RandomForest	0.759	0.773	0.742









SVILUPPI FUTURI

SAGGI BIOLOGICI SULLE 40 MOLECOLE

APPLICAZIONE DEI MODELLI ELABORATI A TUTTE LE HDAC

STUDIO DI SELETTIVITà PER TUTTE LE CLASSI DI HDAC

Ringraziamenti

Un grande ringraziamento al Professore Rino Ragno, per la sua disponibilità e il suo impegno verso i suoi studenti e il suo lavoro.

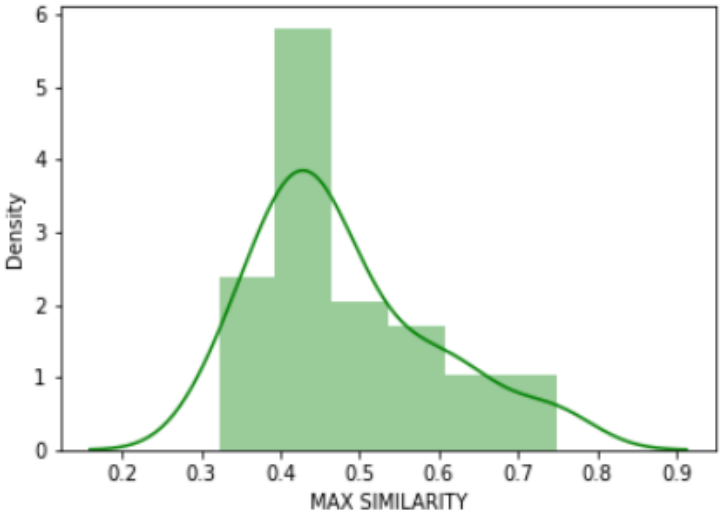
Le mie compagne di lavoro Giorgia Canini e Maria Stella Della Chiaie.

Le dottoresse Giulia Fantera e Paola Caprioli.

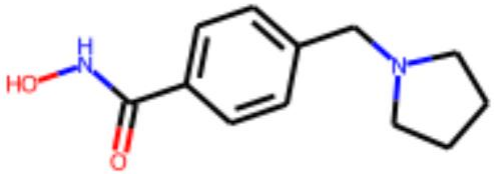
Validazione Esterna

Ricerca Test set → Confronto Test-Train set → Generazione features e labels

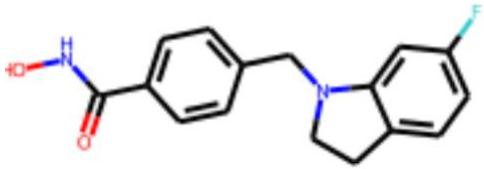
41 molecole
37 attive
4 inattive



Test



Train



Max similarità:
0.43 del coefficiente di tanimoto