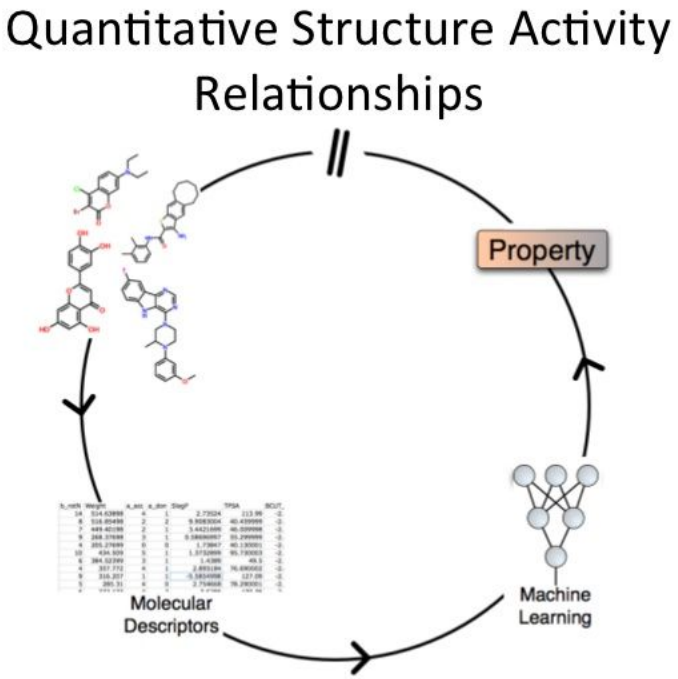
Modelo QSAR para drogas con objetivos tipo citocina en humanos usando IC_{50} y descriptores moleculares de SMILES

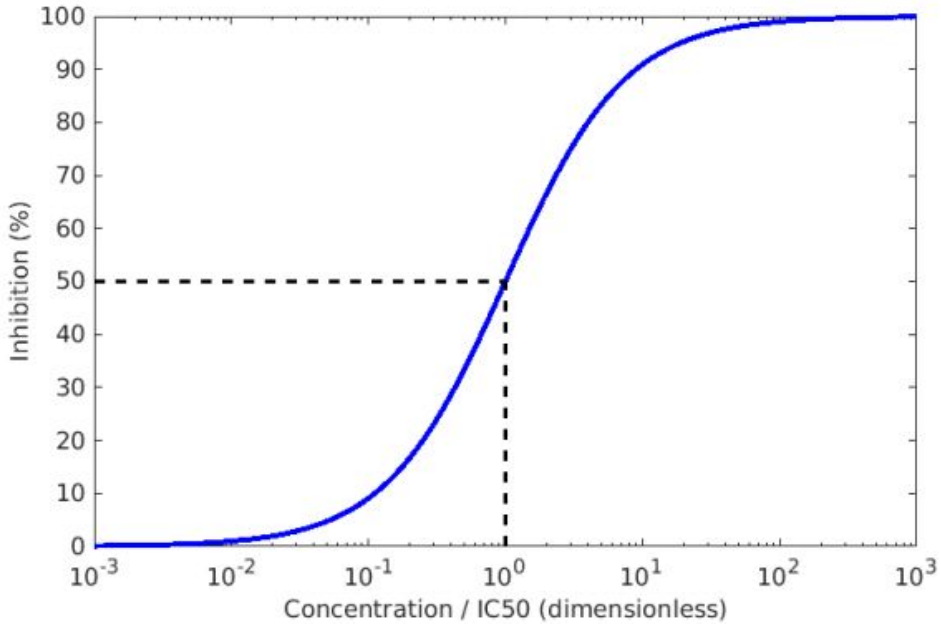
Por: *Aja Macaya, Pablo*
Rodriguez Arias, Alejandro
Romera de los Santos, Juan
Serantes Raposo, Santiago

Introducción



Objetivo

Predicción del índice IC50:

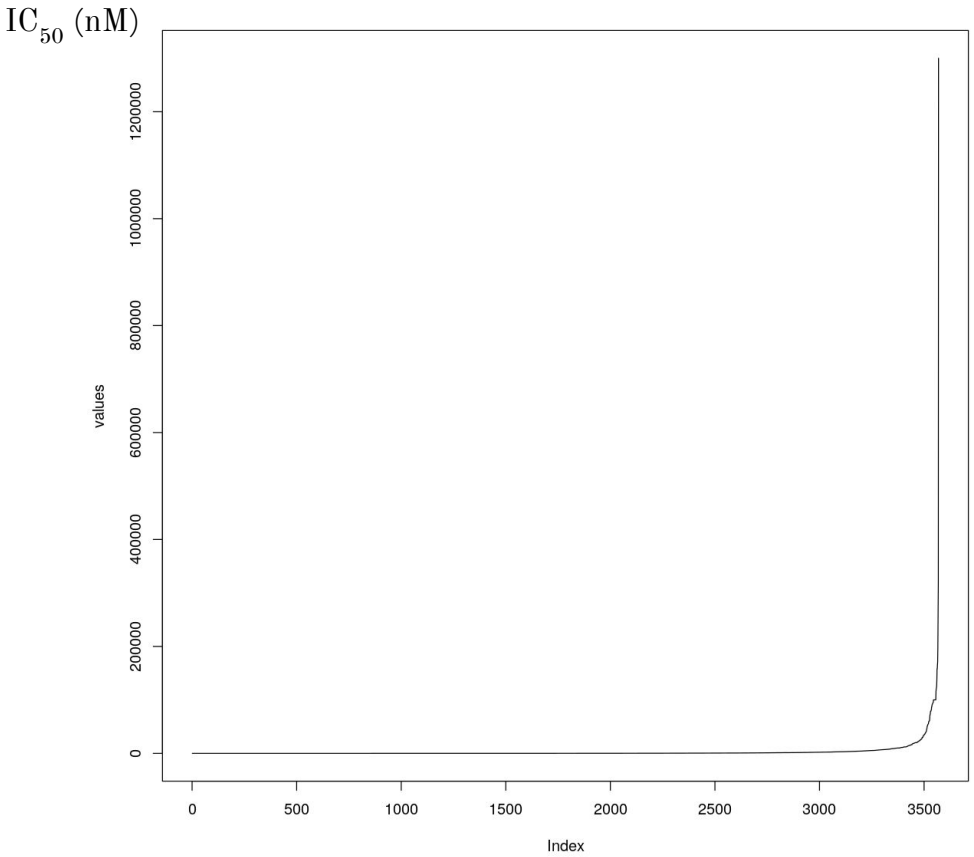


Parámetros de entrada (SMILES)

chemblid	smiles	value
CHEMBL357732	Cc1ccccc1C(=O)c2ccc(Nc3ccccc3N)cc2	47
CHEMBL325597	CCCn1c(SC)nc(c2ccc(F)cc2)c1c3ccncc3	1300
CHEMBL425494	COc1ccc(N2C(=O)Nc3c2ncnc3c4ccccc4C)c(OC)c1	5200
CHEMBL103667	Cc1ccc(cc1)n2nc(cc2NC(=O)Nc3ccc(OCCN4CCOCC4)c5ccccc35)C(C)(C)C	9

Etiquetado de datos

Es necesario clasificar los valores de IC50 en alto(High) y bajo(Low)



Extracción de descriptores

rcdk

CanonicalSmiles	nSmallRings	nAromRings	nRingBlocks	nAromBlocks	nRings3	nRings4	nRings5	nRings6	nRings7	...
Cc1ccccc1C(=O)c2ccc(Nc3ccccc3N)cc2	3	3	3	3	0	0	0	3	0	...
CCCn1c(SC)nc(c2ccc(F)cc2)c1c3ccncc3	3	3	3	3	0	0	1	2	0	...
COc1ccc(N2C(=O)Nc3c2ncnc3c4ccccc4C)c(OC)c1	4	4	3	3	0	0	1	3	0	...
:2NC(=O)Nc3ccc(OCCN4CCOCC4)c5ccccc35)C(C)(C)C	5	4	4	3	0	0	1	4	0	...
:c(n1)C2=C(C(=O)N3CCCN23)c4ccc(F)cc4)C5CCCCC5	5	3	4	3	0	0	2	3	0	...
Cc1ccccc1C(=O)c2ccc(Nc3ccccc3N)cc2Cl	3	3	3	3	0	0	0	3	0	...
C@]23CCCN2CCc4cc5OCOc5cc4[C@@H]3[C@@H]1O	5	1	1	1	0	0	3	1	1	...
CN(C)c1ccc2c(c3ccncc3)c([nH]c2n1)c4ccc(F)cc4	4	4	3	3	0	0	1	3	0	...
CSc1ccc(CSc2ncc(c3ccc(F)cc3)c(n2)c4ccncc4)cc1	4	4	4	4	0	0	0	4	0	...
Fc1ccc(cc1)c2nc3SCCn3c2c4ccncc4	4	3	3	3	0	0	2	2	0	...
.ccc(cc1)C2=C(N3CCCN3C2=O)c4ccnc(NCc5ccccn5)n4	5	4	4	4	0	0	2	3	0	...
S(=O)(=O)c1ccc(cc1)c2cc(c3ccncc3)c([nH]2)c4ccc(F)cc4	4	4	4	4	0	0	1	3	0	...
:c(OCc2ccc(F)cc2)c3c(c4ccncc4)c([nH]c3n1)c5ccc(F)cc5	5	5	4	4	0	0	1	4	0	...
.cnc2cnc(NCCN3CCOCC3)cc2c1Nc4ccc(Sc5ccccc5)cc4	5	4	4	3	0	0	0	5	0	...
Cc1ccccc1C(=O)c2ccc(Nc3ccccc3N)c(C)c2	3	3	3	3	0	0	0	3	0	...
:1ccc(cc1)C(=O)Nc2cc(ccn2)c3c(nc(SC)n3C)c4ccc(F)cc4	4	4	4	4	0	0	1	3	0	...
][(C)Nc1nccc(n1)C2=C(C(=O)N3CCCCN23)c4ccc(F)cc4	4	3	3	3	0	0	1	3	0	...
CSc1nc(c2ccc(F)cc2)c([nH]1)c3ccnc(F)c3	3	3	3	3	0	0	1	2	0	...
+][([O-])c1ccc(CSc2nc(c3ccc(F)cc3)c([nH]2)c4ccncc4)cc1	4	4	4	4	0	0	1	3	0	...
Cc1ccccc1C(=O)c2ccc(Nc3ccc(Br)cc3N)cc2Cl	3	3	3	3	0	0	0	3	0	...
Cn1cc(cn1)c2conc2c3ccnc(Nc4ccc(cc4)N5CCOCC5)c3	5	4	5	4	0	0	2	3	0	...
O=C(c1ccccc1)c2ccc(Nc3ccccc3)cc2	3	3	3	3	0	0	0	3	0	...
nccc(n1)C2=C(C(=O)N3CCCN23)c4ccc(F)cc4)c5ccccc5	5	4	4	4	0	0	2	3	0	...

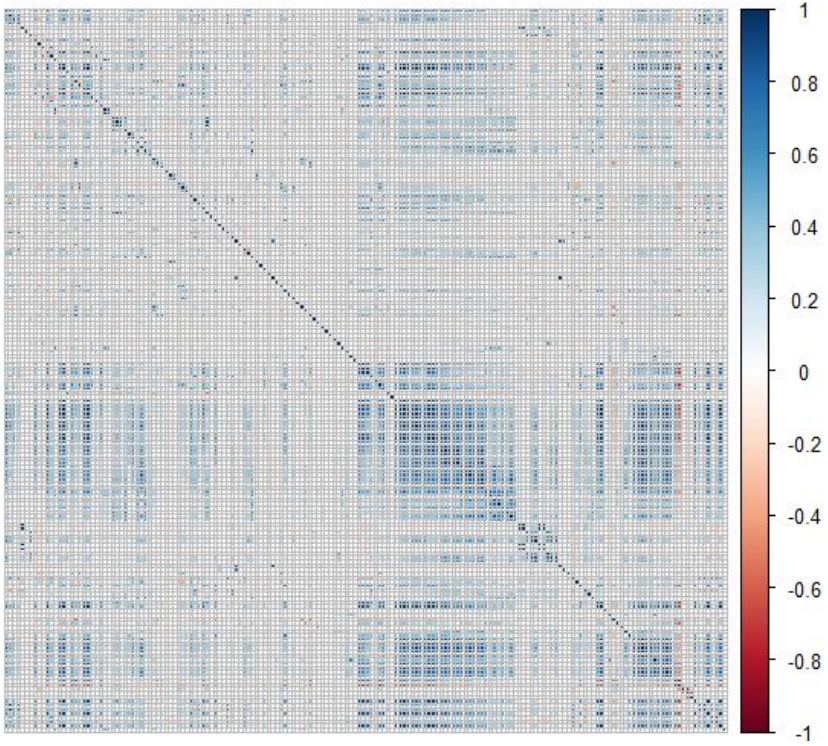
Limpieza de datos

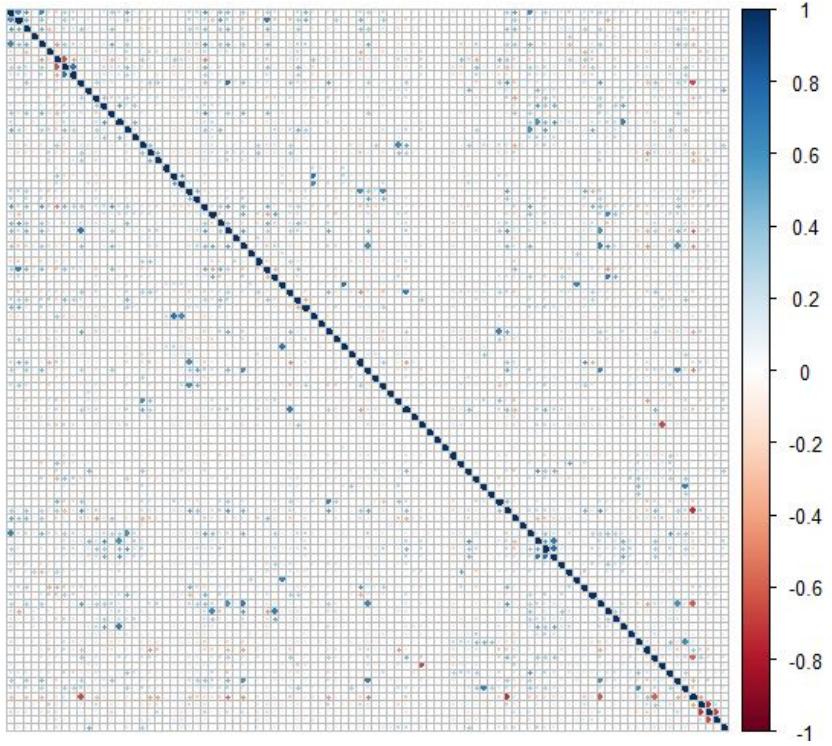
	Datos	Características
Inicial	3569	286
Eliminar col. valores cte.	3569	176
Eliminar filas NA	3559	176
Eliminar col. alta correlación	3559	93

Limpieza de datos

nSmallRings	nAromRings	nRingBlocks	nRings4	nRings6	nRings7	tpsaEfficiency	XLogP	LipinskiFailures
-0,50	-0,2	-0,2	-1	-0,3333333	-1	-0,4253837	-0,078859620	-1,0000000
-0,50	-0,2	-0,2	-1	-0,6666667	-1	-0,4710728	0,007335779	-1,0000000
-0,25	0,2	-0,2	-1	-0,3333333	-1	-0,3070373	-0,330443481	-1,0000000
0,00	0,2	0,2	-1	0,0000000	-1	-0,5495563	0,038846282	-0,3333333

Limpieza de datos





Conjuntos de entrenamiento y test

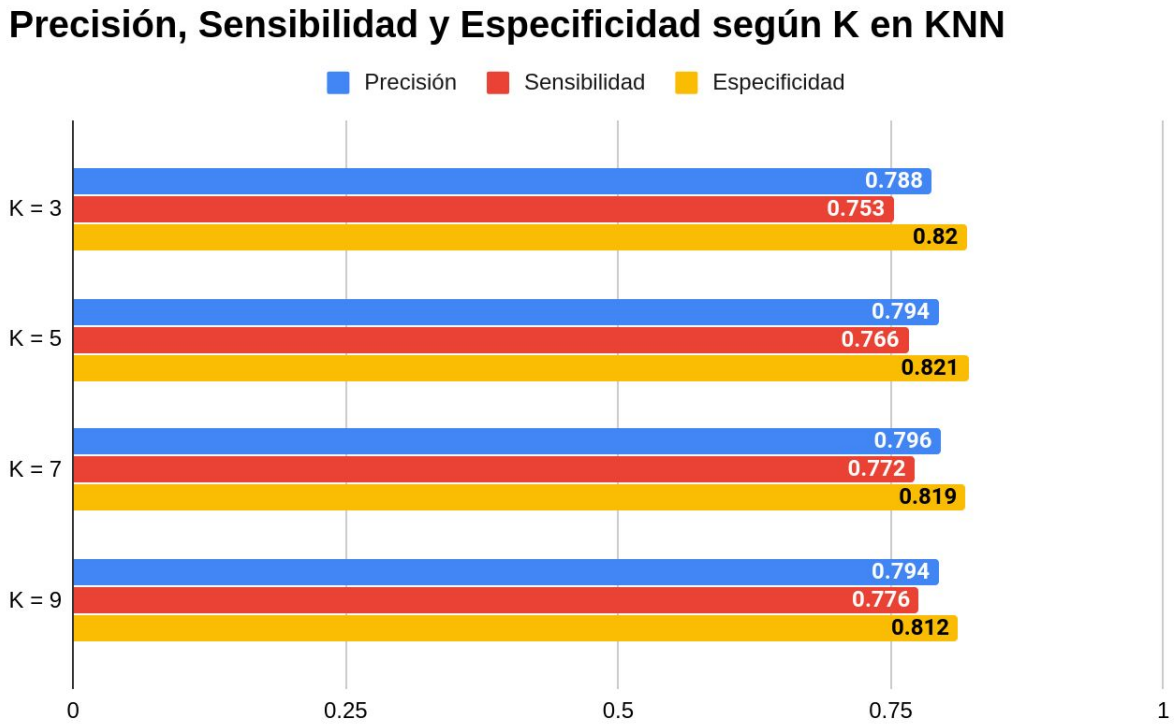
90% - 10%

Regresión logística

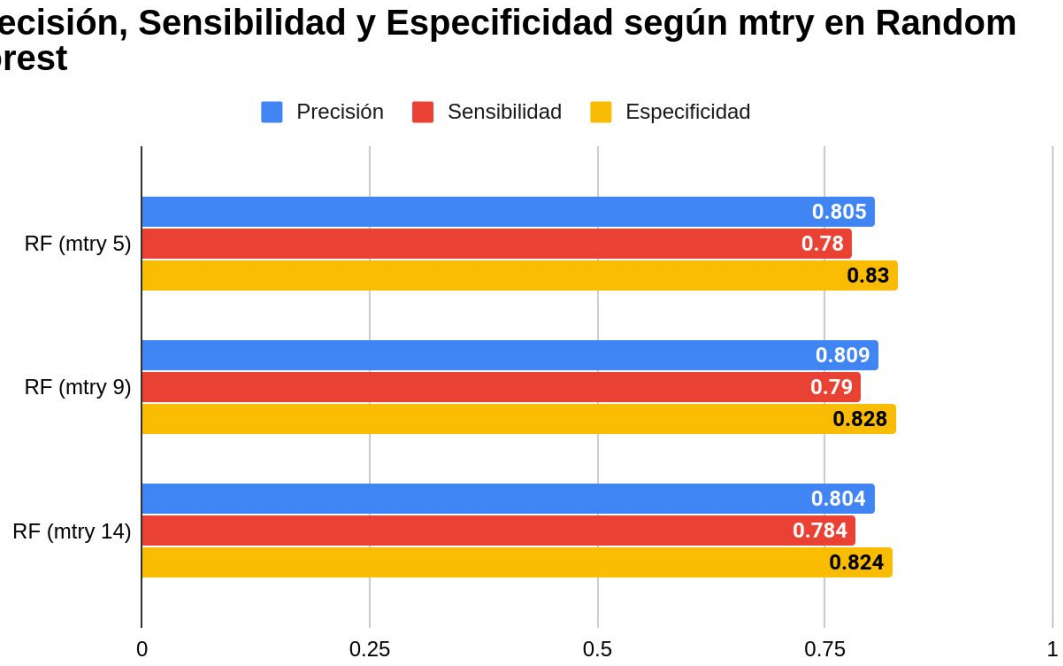
Valores reales		
Predicción del Modelo	**IC50 ALTO**	**IC50 BAJO**
IC50 ALTO	134	46
IC50 BAJO	40	126

Precisión: 0,739
Sensibilidad: 0,736
Especificidad: 0,743

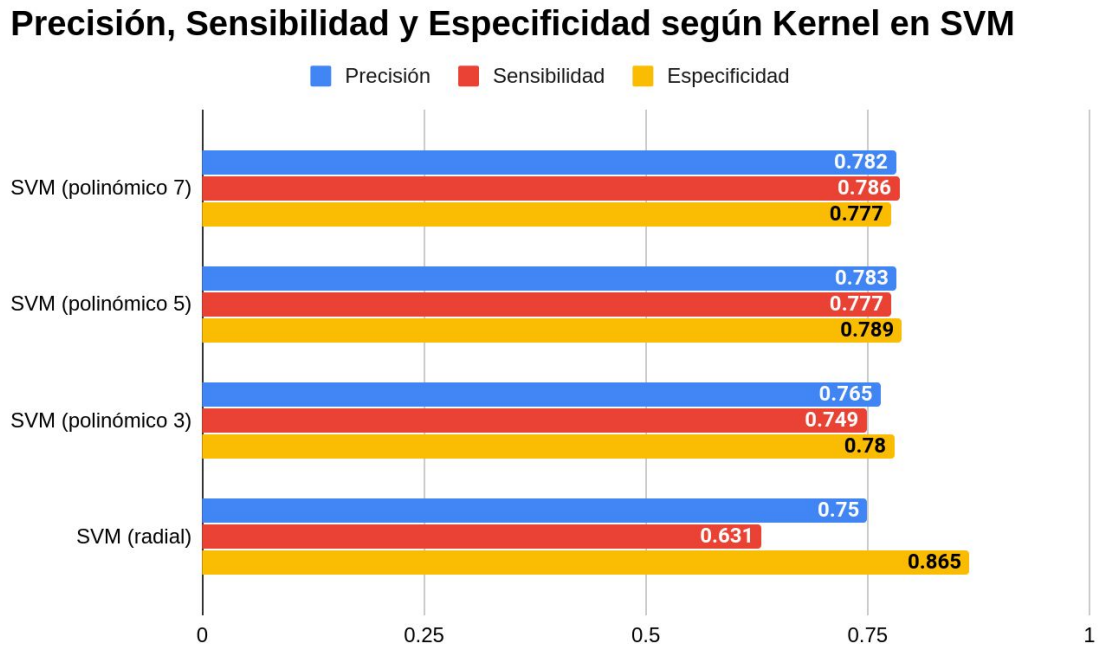
KNN


Precisión, Sensibilidad y Especificidad según K en KNN
Precisión
Sensibilidad
Especificidad
K = 3
0.788
0.753
0.82
K = 5
0.794
0.766
0.821
K = 7
0.796
0.772
0.819
K = 9
0.794
0.776
0.812
0
0.25
0.5
0.75
1

Random Forest



Support Vector Machine (SVM)



Precisión, Sensibilidad y Especificidad según modelos