Predicción de propiedades farmacológicas de moléculas para el virus Chikungunya mediante el uso de machine learning

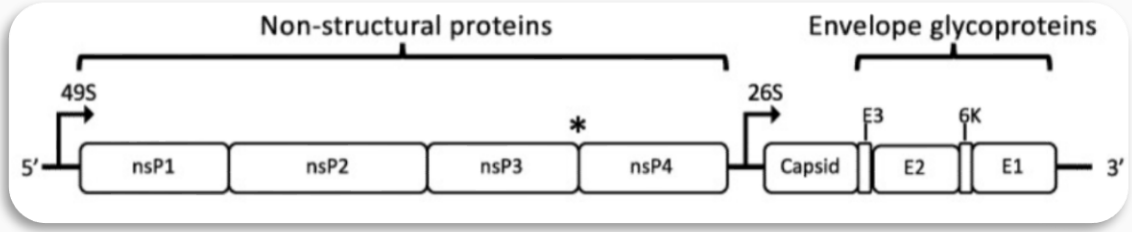
Miguel Jiménez Morcuende

Dr. Jorge Valencia Delgadillo

15/01/2023

UOC Universitat Oberta de Catalunya

Contexto y Justificación



Representación esquemática del genoma de un alfavirus (Bakar & Ng, 2018).

Aislado a mediados del siglo XX en Tanzania

- Alfavirus de ARN monocatenario positivo
- Aproximadamente 11.8 kb
- 2 ORF:
  - 5′ → proteínas no estructurales
  - 3′ → estructuras proteicas

- Síntomas (7-10 días):
  - Fiebre alta
  - Dolor punzante
  - Irrupción de sarpullidos
  - Fatiga intensa
  - Vómitos
  - Diarrea

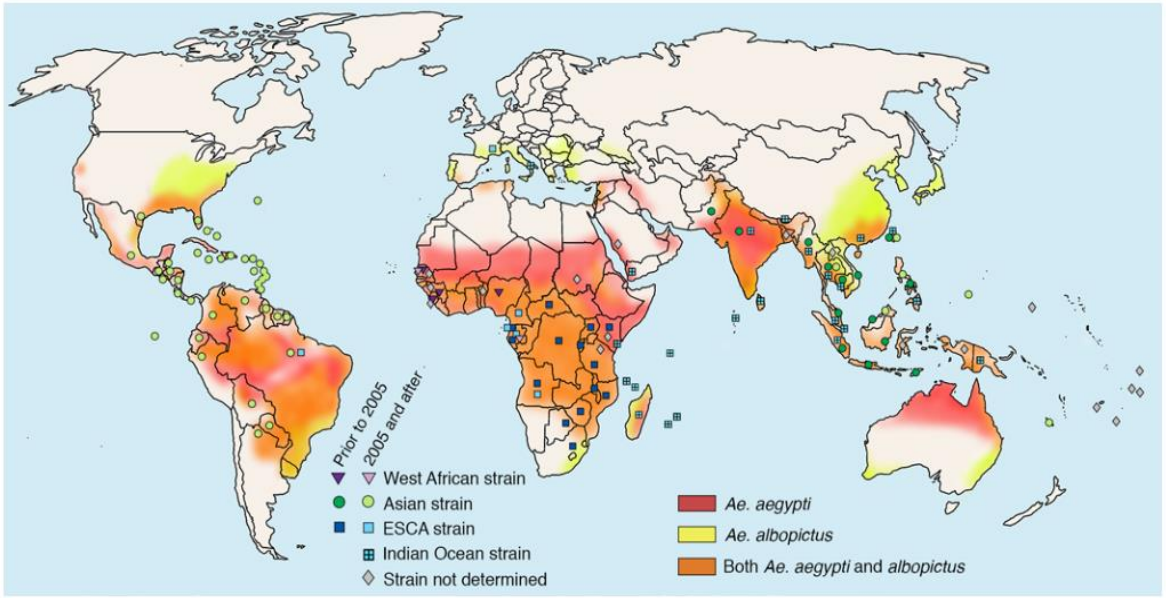
- Síntomas crónicos
  - Artritis
  - Poliartralgias
  - Depresión
  - Fatiga

Ross R.W., 1956 / Robinson, 1955 / Matusali et al., 2019 / Burt et al., 2017 / Galán-Huerta et al., 2015 / da Silva-Júnior et al., 2017 / Chen et al., 2017

Universitat Oberta de Catalunya

Vector de transmisión

- *Aedes aegypti*
- *Aedes albopictus*



Linajes enzoóticos:

- África Occidental
- África Este, Centro y Sur
- Asiática
- Océano Índico

- Transmisión a través de la saliva del mosquito hembra
- Aumento del riesgo:
  - Cambio climático
  - Globalización

- Entre 2.004 y 2.019:
  - Más de 100 países
  - Más de 10 millones de casos
  - Riesgo de infección a 1,3 billones de personas

Silva et al., 2018 / Lo Presti et al., 2012 / Trolle, 2019 / Ensenrink, 2007 / Rezza et al., 2007 / LaBeaud et al., 2011 / Suhrbier, 2019

Objetivos del trabajo

Objetivos generales:

1. Creación de una serie de modelos que ayuden a la predicción de nuevas moléculas con actividad inhibitoria frente al virus chikungunya.
2. Comparativa de los modelos creados y selección del más fiable.

Objetivos secundarios:

1.1. Obtención del set de datos y cálculo de descriptores

1.2. Identificación de las variables más descriptivas

1.3. Entrenamiento de los diferentes modelos

2.1. Análisis de rendimiento y realización curvas ROC

2.2. Optimización del modelo seleccionado

Impactos

- No presenta ningún impacto a nivel de sostenibilidad.
- No presenta impacto del tipo ético-social.
- Si que supone un impacto positivo en la dimensión de diversidad y derechos humanos

↓

Países en vías de desarrollo = Gran riesgo social y económico

↙ Hábitat de los mosquitos del género *Aedes*

↓ Peores condiciones de vida e higiene

↓ Falta de material de prevención

↘ Peor sanidad pública

Enserink, 2007

Enfoque y Método

- Desarrollo novel de un medicamento
- Técnicas de investigación de estructuras 3D
- Evita los ensayos con animales
- Gran reducción de costes

→ Alto coste y tiempo

- Selección de moléculas activas frente a Chikungunya (ChEMBL ID 4296563) y creación de descriptores
- Realización de los modelos y curvas ROC mediante Rstudio (2022.07.2 +576)

DiMasi et al., 2016 / Dara et al., 2022 / Elbadawi et al., 2021

Planificación del trabajo

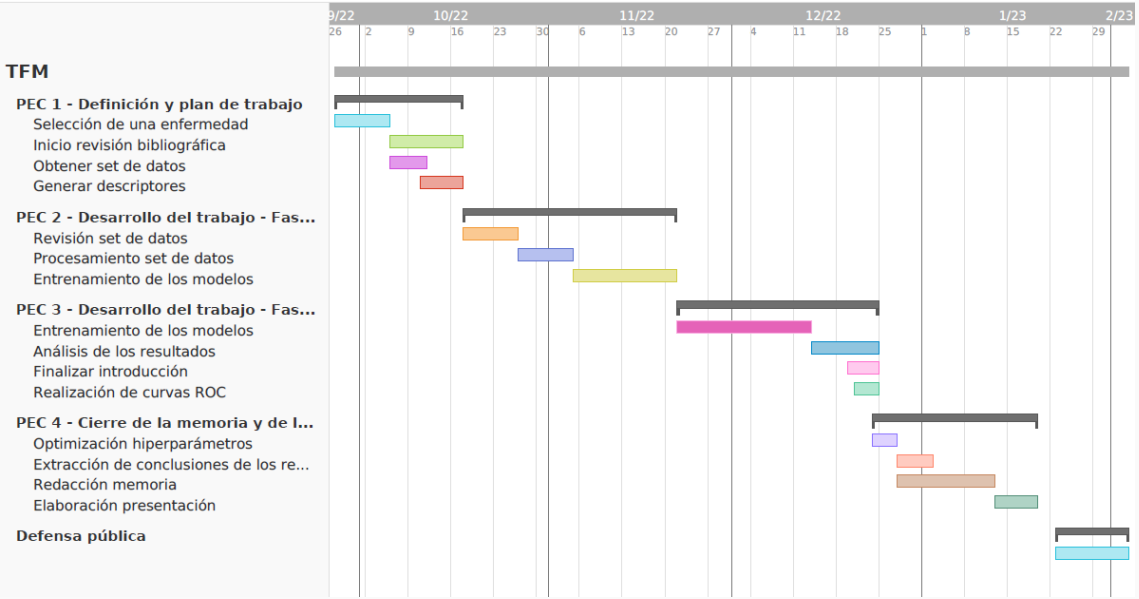


Diagrama de Grantt

Planificación del TFM.

Tareas
PEC1- Definición y plan de trabajo
T.1. Selección de una enfermedad
T.2. Revisión bibliográfica
T.3. Set de datos
T.4. Generación de descriptores
PEC2- Desarrollo del trabajo - fase I
T.4. Revisión del set de descriptores
T.5. Procesamiento de los datos
T.6. Entrenamiento de los modelos
PEC3- Desarrollo del trabajo – fase II
T.6. Entrenamiento de los modelos
T.7. Análisis de los resultados y curvas ROC
T.8. Optimización de hiperparámetros
T.9. Conclusiones
PEC4-Cierre de la memoria y de la presentación
Redacción memoria
Elaboración presentación
Defensa pública

Universitat Oberta de Catalunya

Análisis de Riesgos

- Valores de confianza inferiores a lo esperado

- No se observan diferencias aparentes entre modelos

Producto obtenido

- Modelo de predicción de actividad frente al Chikungunya

Estado del arte

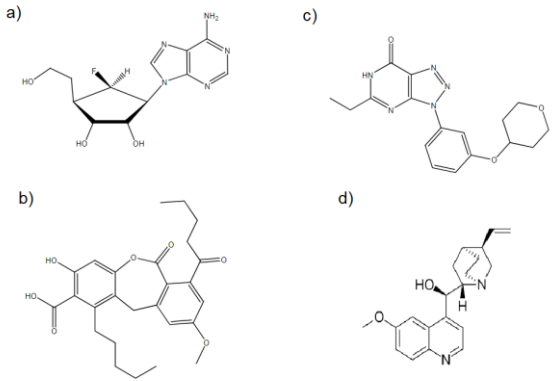
Actualmente no existe ningún tipo de medicamento o vacuna, solo el tratamiento de los síntomas

- La vacuna en estudio VLA1553 obtuvo niveles de seroprotección en el 98.9% de los participantes con un 0.45% de efectos adversos

En cuanto al diseño de fármacos

- Tipos de enfoque:
  1. Reutilización de medicamentos
  2. Detección basada en células fenotípicas
  3. Detección basada en ensayos
  4. Diseño basado en estructuras

- Función:
  1. Inhibidores de entrada
  2. Inhibidores de proteínas no estructurales
  3. Inhibidores de la proteasa de la cápside
  4. Inhibidores de la proteína 6K

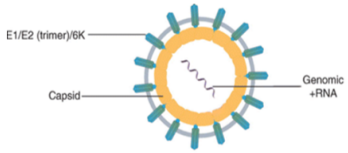


Inhibidores de la proteína nsP1: a) 6'-β-fluoro-homoaristeromicina, b) Ácido lobárico, c) MADTP, d) Quininas.

Mourad et al., 2022 / Abdelnabi et al., 2017 / Pérez-Pérez et al., 2019 / Powers, 2018 / Tsao et al., 2018 / Coles et al., 2003 / Malet et al., 2009 / Battisti et al., 2021 / da Silva-Júnior et al., 2017

Universitat Oberta de Catalunya

Metodología



Chikungunya virus (CHIKV)

ChEMBL

Moléculas con actividad frente a CHIKV

272 resultados

Generación de 129 descriptores utilizando RDkit

Selección de descriptores en función de la correlación y la varianza
Se emplean un total de 63 descriptores para el entrenamiento de los modelos

Training/Testing (67%)

Desarrollo de modelos machine learning

Support Vector Machine | Random Forest | K-Nearest Neighbor | Decision Tree | Naïve Bayes | Extreme Gradient Boosting

Evaluación de la actuación y robustez de los modelos predictivos desarrollados Anti-CHIKV

Esquema general de la metodología anti-CHIKV para el desarrollo de algoritmos predictivos para identificar inhibidores

Resultados y Discusión

K-Nearest Neighbor

Resultados obtenidos del modelo kNN para valores de k de 1,5 y 11. K (número de vecinos a considerar), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

K	VN	VP	FN	FP	TE	P	Kappa	SE	SP
1	31	29	14	16	0.3333	0.6667	0.3333	0.6744	0.6596
5	29	27	16	18	0.3778	0.6222	0.2444	0.6279	0.6170
11	29	26	17	18	0.3889	0.6111	0.2215	0.6047	0.6170

Resultados y Discusión

Naive Bayes

Resultados obtenidos del modelo NB para valores de laplace de 0,1 y 5. Laplace (valor añadido a las tablas de frecuencia), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

Laplace	VN	VP	FN	FP	TE	P	Kappa	SE	SP
0	36	22	21	11	0.3556	0.6444	0.2804	0.5116	0.7660
1	36	22	21	11	0.3556	0.6444	0.2804	0.5116	0.7660
5	36	22	21	11	0.3556	0.6444	0.2804	0.5116	0.7660

Resultados y Discusión

Support Vector Machine

Resultados obtenidos del modelo SVM para kernel (transformación y combinación de vectores) lineal o Gaussiano, valores de C de 1,3 y 7. C (coste de violación de las restricciones), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

Kernel	C	VN	VP	FN	FP	TE	P	Kappa	SE	SP
Lineal	1	31	30	13	16	0.3222	0.6778	0.3562	0.6977	0.6596
	3	30	26	17	17	0.3778	0.6222	0.2429	0.6047	0.6383
	7	29	28	15	18	0.3667	0.6333	0.2674	0.6512	0.6170
Gaussiano	1	29	27	16	18	0.3778	0.6222	0.2444	0.6279	0.6170
	3	32	25	18	15	0.3667	0.6333	0.263	0.5814	0.6809
	7	32	28	15	15	0.3333	0.6667	0.332	0.6512	0.6809

Universitat Oberta de Catalunya

Resultados y Discusión

Decision Trees

Resultados obtenidos del modelo DT para el parámetro trial valores de 1, 15, 40 y 50 respectivamente. Trials (número de iteraciones de refuerzo), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

Trials	VN	VP	FN	FP	TE	P	Kappa	SE	SP
1	29	29	14	18	0.3556	0.6444	0.2903	0.6744	0.6710
15	33	30	13	14	0.3	0. 7	0.3994	0.6977	0.7021
40	33	31	12	14	0.2889	0.7111	0.4222	0.7209	0.7021
50	31	30	13	16	0.3222	0.6778	0.3562	0.6977	0.6596

Resultados y Discusión

Random Forest

Resultados obtenidos del modelo RF para el parámetro ntree valores de 50, 100, 150 y 200. Ntree (número de árboles), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

ntree	VN	VP	FN	FP	TE	P	Kappa	SE	SP
50	34	31	13	12	0.2778	0.7222	0.4439	0.7045	0.7391
100	33	32	14	11	0.2778	0.7222	0.445	0.6957	0.7500
150	33	31	14	12	0.2889	0.7111	0.4222	0.6889	0.7333
200	33	29	14	14	0.3111	0.6889	0.3765	0.6744	0.7021

Universitat Oberta de Catalunya

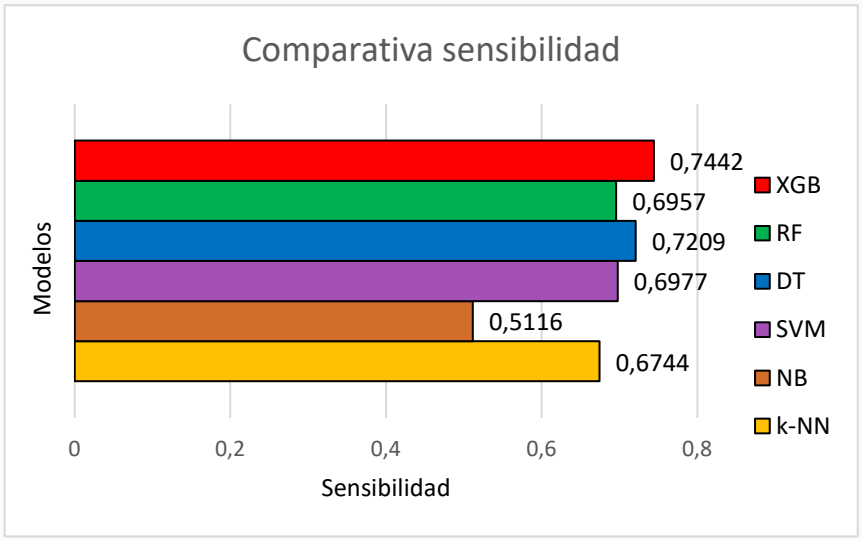
Resultados y Discusión

Extreme Gradient Boosting

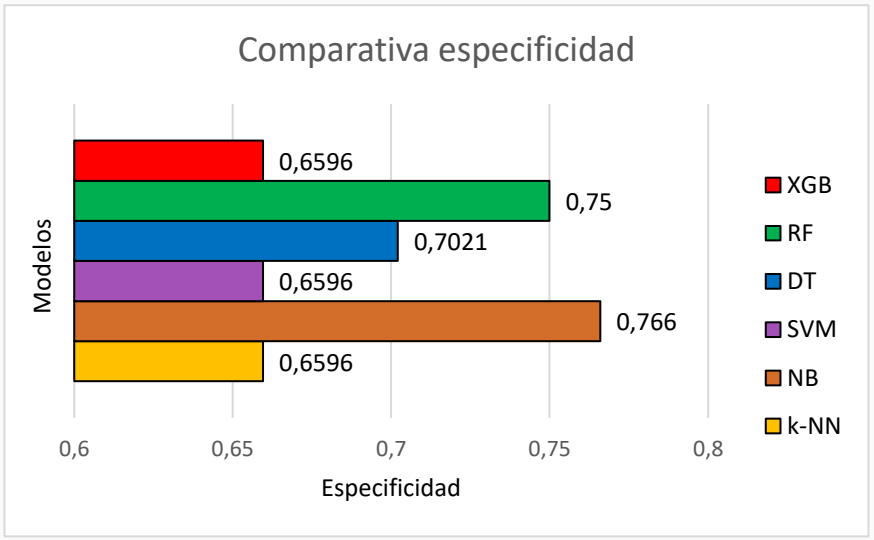
Resultados obtenidos del modelo XGB para los parámetros max.depth valores de 6 y 12; y nrounds 1, 6 y 12. Max.depth (profundidad máxima), nrounds (número de iteraciones de refuerzo), VN (verdaderos negativos), VP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), TE (tasa de error), P (precisión), Kappa, SE (sensibilidad) y SP (especificidad).

Max.depth	nrounds	VN	VP	FN	FP	TE	P	Kappa	SE	SP
6	1	33	24	19	14	0.3667	0.6333	0.2616	0.5581	0.7021
	6	31	32	11	16	0.3	0.7	0.4018	0.7442	0.6596
	12	32	30	13	15	0.311	0.6889	0.3778	0.6977	0.6809
12	1	33	24	19	14	0.3667	0.6333	0.2616	0.5581	0.7021
	6	33	24	19	14	0.3667	0.6333	0.2616	0.5581	0.7021
	12	29	33	10	18	0.311	0.6889	0.3778	0.7674	0.6170

Resultados y Discusión

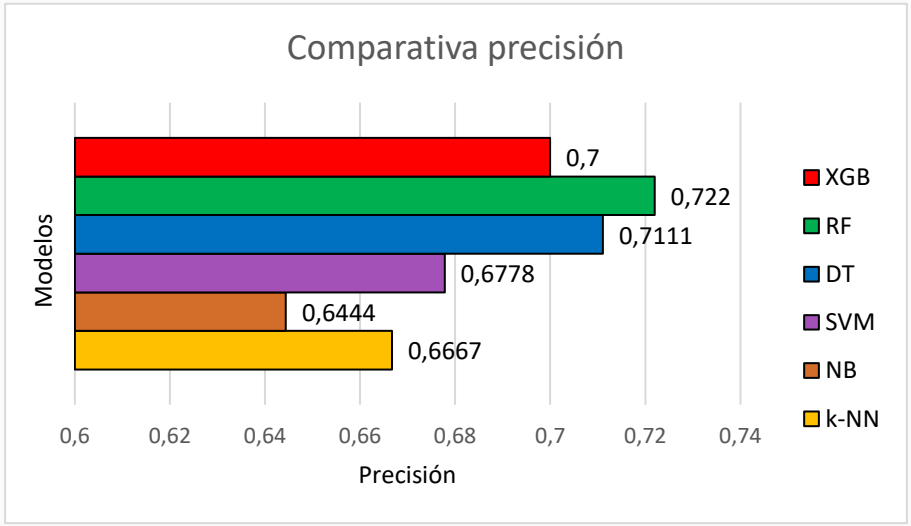


Gráfica comparativa de los resultados de sensibilidad obtenidos en cada modelo

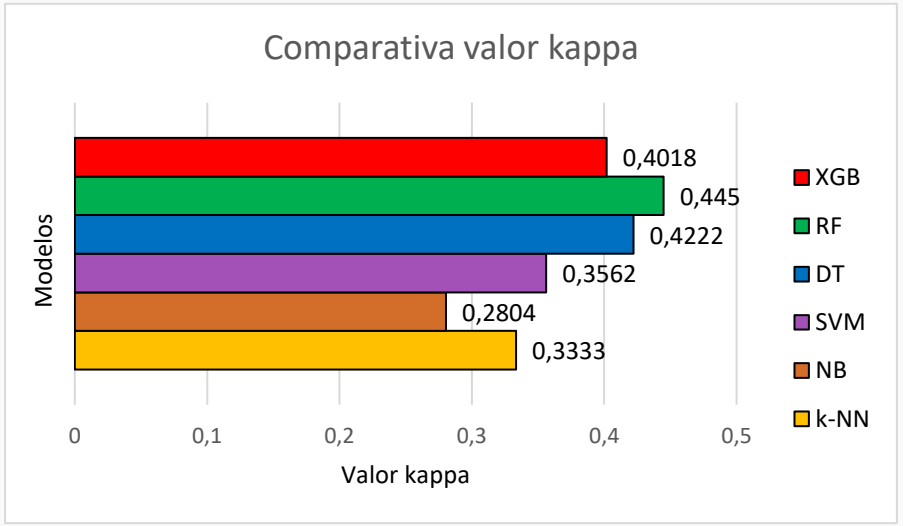


Gráfica comparativa de los resultados de especificidad obtenidos en cada modelo

Resultados y Discusión

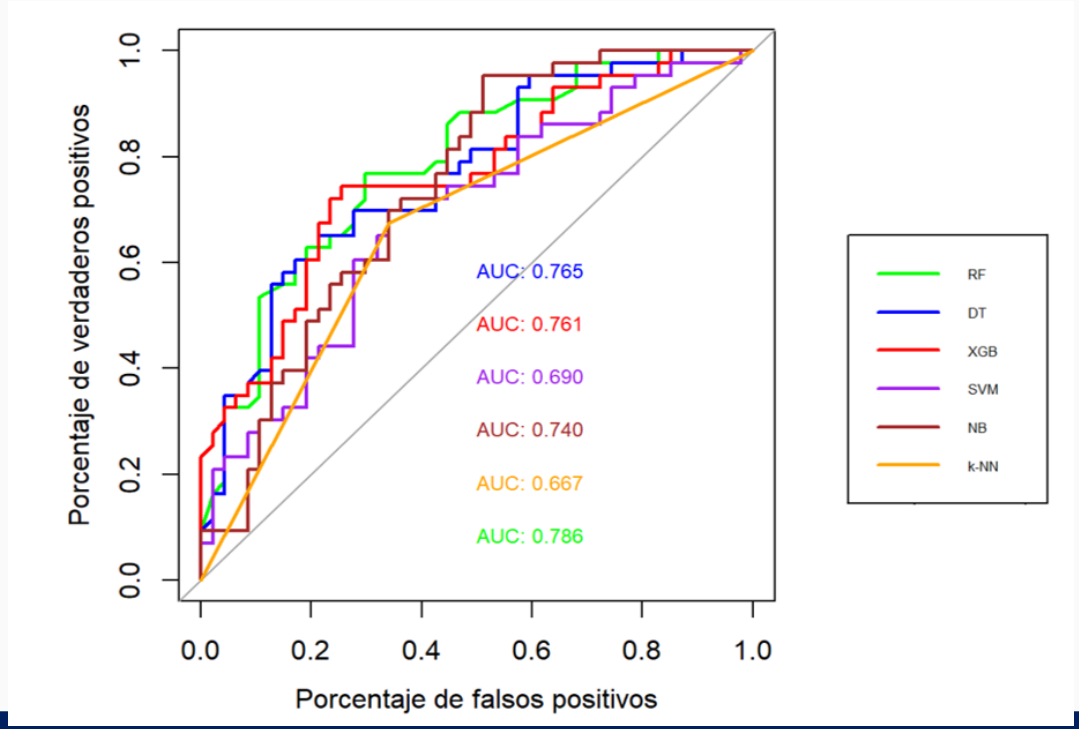


Gráfica comparativa de los resultados de precisión obtenidos en cada modelo.

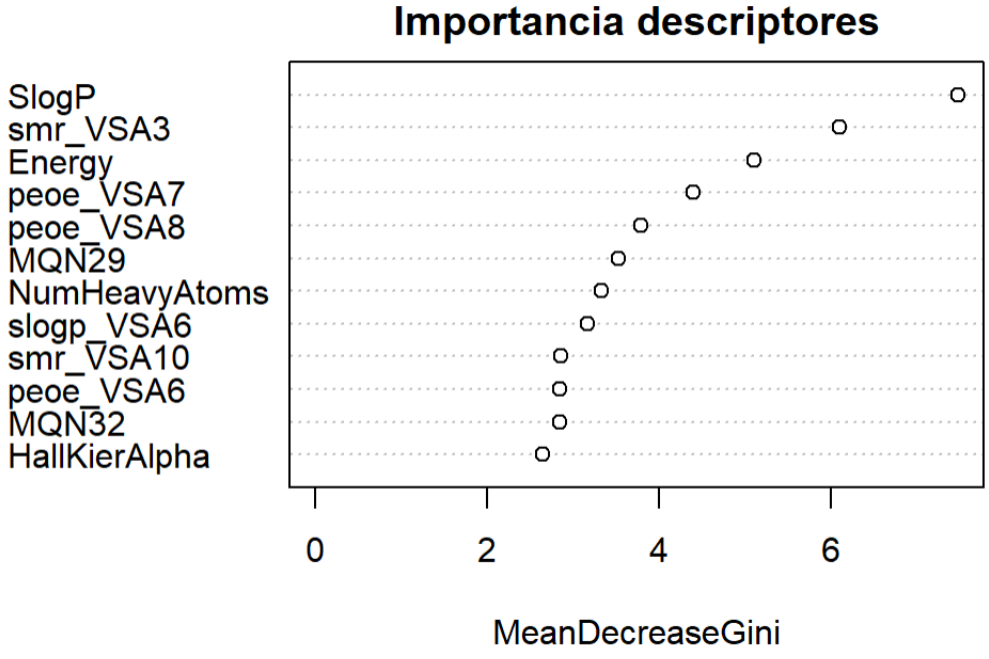


Gráfica comparativa de los resultados de kappa obtenidos en cada modelo

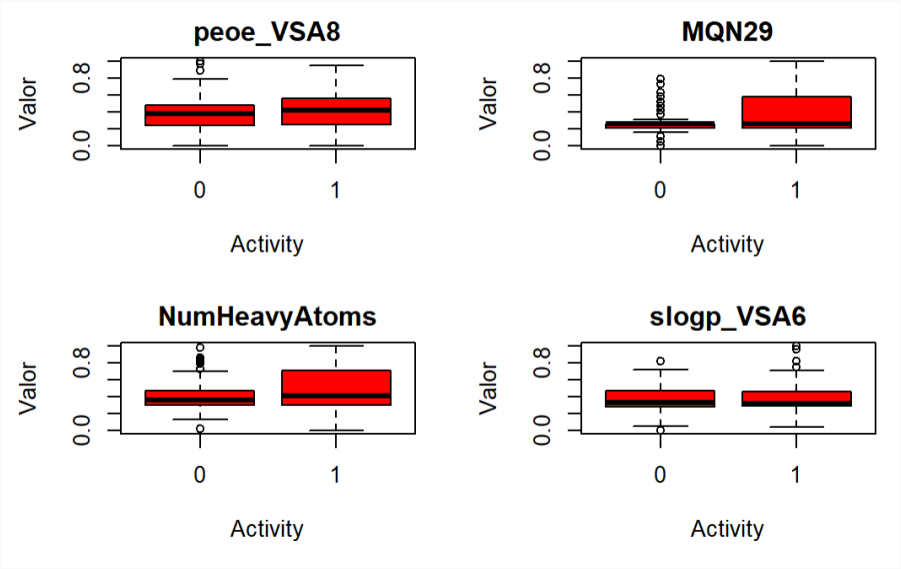
Resultados y Discusión



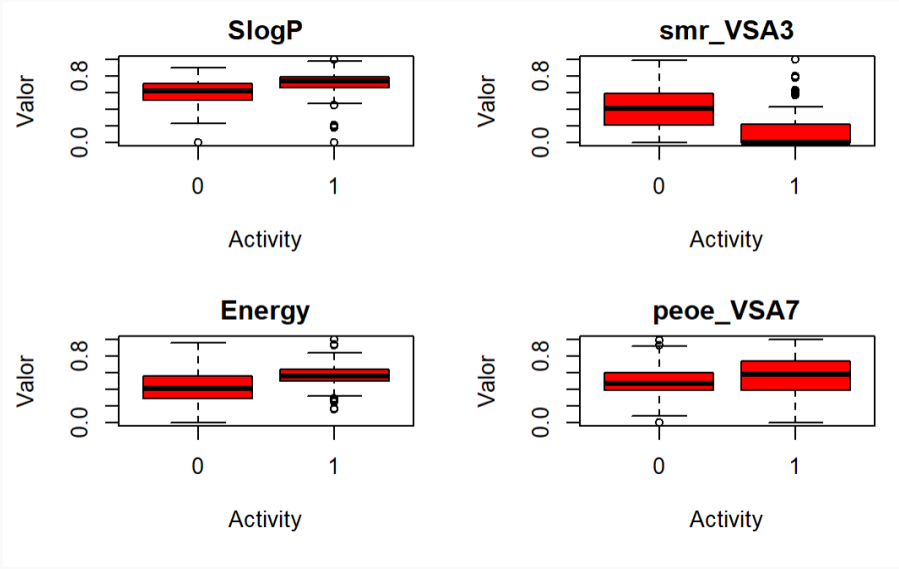
- Se selecciona a RF como el mejor modelo de predicción de moléculas con actividad biológica anti-CHIKV



Gráfica comparativa de la importancia de los descriptores para el modelo RF, usando el Gini de descenso medio que mide qué tan puros son los nodos al final de árbol sin cada variable.



Boxplot de la distribución de valores de los descriptores peoe_VSA8, MQN29, NumHeavyAtoms, slogp_VSA6 para el modelo RF teniendo en cuenta la actividad o no actividad de las moléculas de estudio.



Boxplot de la distribución de valores de los descriptores SlogP, smr_VSA3, Energy y peoe_VSA7 para el modelo RF teniendo en cuenta la actividad o no actividad de las moléculas de estudio.

Resultados y Discusión

La selección no significa que los modelos de RF sean lo modelos de referencia para la búsqueda de nuevos fármacos:

- Gawriljuk y su equipo trabajaron obuvieron mejores resultados en modelos de k-NN con una precisión del 0.85 y un valor AUC de 0.77, o de SVM con 0.82 y 0.81 respectivamente, , que los obtenidos con RF con una precisión de 0.83 y un AUC de 0.77.

- Kamboj et al, emplearon modelos de SVM, RF y k-NN, para descubrir moléculas capaces de actuar sobre proteínas no estructurales del virus de la hepatitis C. Obtuvieron unos valores del coeficiente de regresión (R2) de 0.72 para SVM, 0.62 para k-NN y RF.

- Para la predicción de inhibición de proteínas resistentes al cáncer de mama, Jiang y su grupo emplearon múltiples modelos. Los mejores resultados de precisión los obtuvo el modelo SVM con una precisión de 0.911 y un AUC de 0.958. Por debajo se encontraba XGB con una precisión de 0.891 y un AUC de 0.957. Descendían notablemente para k-NN (0.857) y NB (0.78).

Gawriljuk et al., 2021 / Jiang et al., 2020 / Kamboj et al., 2022

Universitat Oberta de Catalunya

Resultados y Discusión

Por tanto,

- No se puede tomar un único modelo como referencia
- Los descriptores y la variable a predecir, afectan a la capacidad de predicción

Optimización de Hiperparámetros

- Se tuvo en cuenta los parámetros mtry y ntree
- Mediante 10-fold cross validation se obtuvo un máximo de precisión de 0.701 para mtry = 22 y ntree= 75 en comparación a 0.687

Conclusión

- Se ha creado un sistema de predicción de nuevas moléculas anti-CHIKV

- Valores de precisión entre 0.6444 a 0.722 // Valores AUC de 0.667 a 0.786

- Optimización modelo RF, precisión pasa de 0.68479 → 0.701177

- Se han cumplido los objetivos previstos

⬇

- Construcción de un método de predicción

Líneas futuras

- Optimización de los métodos → uso de diferentes softwares y mayor número de descriptores

- Predicciones con medicamentos en uso, molecular docking