UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingeniería

Modelos computacionales tipo ensamble para predecir la actividad biológica de fármacos contra la malaria: Un nuevo enfoque para mejorar los métodos tradicionales

.

Martín Alejandro Moreno Armas

Ingeniería Química

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero Químico

Quito, 20 de diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingeniería

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Modelos computacionales tipo ensamble para predecir la actividad biológica de fármacos contra la malaria: Un nuevo enfoque para mejorar los métodos tradicionales

Martín Alejandro Moreno Armas

Nombre del profesor, Título académico **José Ramón Mora, PhD**

Quito, 20 de diciembre de 2022

© DERECHOS DE AUTOR

Nombres y apellidos: Martín Alejandro Moreno Armas

Código: 00206923

Cédula de identidad: 1724392145

Lugar y fecha: Quito, 20 de diciembre de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El *P.falciparum,* virus causante de la mortalidad de la malaria, es cada vez más resistente a los fármacos utilizados para el tratamiento. En consecuencia, los científicos se han respaldado de nuevas herramientas computacionales para la búsqueda de nuevos fármacos. En el presente estudio se realizó un modelado tipo ensamble para predecir el pEC_{50}. Como punto de partida, se utilizó la Malaria Box, una base de datos con mucha variabilidad estructural para asegurar una cobertura amplia del dominio de aplicación. Se combinaron descriptores topográficos y mecano-cuánticos, junto con una variedad de técnicas de aprendizaje automático para el desarrollo del modelo. Inicialmente se intentó la construcción de un modelo tipo QSAR utilizando toda la base de datos, pero no se encontraron resultados satisfactorios. Por esta razón, se realizó una partición de la base de datos, y se construyó un modelo tipo ensamble para clasificar moléculas de acuerdo con su actividad biológica en activa o muy activa. El mejor modelo presentó los siguientes valores de exactitud: $Acc_{10\text{-}fold} = 0.738$ y $Acc_{Ext} = 0.675$, así como de sensibilidad y especificidad de 0.585 y 0.769, respectivamente para la validación externa, demostrando que es un modelo confiable para la predicción de las dos clases. La siguiente fase fue un modelado de regresión por separado para cada clase. De la misma manera, se combinaron descriptores topográficos y mecano-cuánticos junto con técnicas de regresión, y se construyeron dos modelos de tipo ensamble. Después de una extensa validación, el ensamble de la clase activa alcanzó un $Q^2_{10\text{-}fold} = 0.793$ y $Q^2_{Ext} = 0.765$, mientras que el ensamble de la clase muy activa alcanzó un $Q^2_{10\text{-}fold} = 0.810$ y $Q^2_{Ext} = 0.749$. Estos parámetros demostraron que la predictibilidad de ambos ensambles es muy buena y superan a los resultados encontrados con los modelos individuales, confirmando la necesidad de trabajar con modelos de tipo ensamble para obtener un mejor desempeño estadístico.

Palabras clave: diseño de fármacos, malaria, QSAR, modelos predictivos, ensamble, clasificación, regresión, descriptores moleculares, aprendizaje automático

ABSTRACT

P.falciparum, the virus that causes malaria mortality, is increasingly resistant to the drugs used for treatment. Consequently, scientists have relied on new computational tools to search for new drugs. In the present study, ensemble modeling was performed to predict pEC50. As a starting point, the Malaria Box, a database with high structural variability, was used to ensure broad coverage of the applicability domain. Topographical and quantum mechanical descriptors were combined, along with a variety of machine learning techniques for the model development. Initially, the construction of a QSAR-type model using the entire database was attempted, but no satisfactory results were found. For this reason, the database was partitioned, and an ensemble type model was built to classify molecules according to their biological activity in two classes: active or more active. After an extensive validation, the model presented $Acc_{10\text{-}fold} = 0.738$ and $Acc_{Ext} = 0.675$ as well as sensitivity and specificity of 0.585 and 0.769, respectively for external validation, demonstrating that it is a reliable model for predicting the two classes. The next phase was a separate regression modeling for each class. In the same way, topographic and quantum mechanical descriptors were combined with regression techniques, and two ensemble-type models were built. After an extensive validation, the active class ensemble model achieved $Q^2_{10\text{-}fold} = 0.793$ and $Q^2_{Ext} = 0.765$, while the more active class ensemble model achieved $Q^2_{10\text{-}fold} = 0.810$ and $Q^2_{Ext} = 0.749$. These parameters showed that the predictability of both ensembles is very good and exceeds the results found with the individual models, confirming the need to work with ensemble-type models to obtain better statistical performance.

Key words: drug design, malaria, QSAR, predictive models, ensemble, classification, regression, molecular descriptors, machine learning

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En la actualidad, se reportan aproximadamente 250 millones de casos anuales de malaria, de los cuales alrededor de 650 000 se derivan en la muerte de los pacientes [1,2]. Esta enfermedad se ha esparcido a lo largo de algunos países de África y Asia, y se ha convertido en un asunto de salud pública muy controversial debido a su impacto social y económico [3,4]. La malaria es causada por parásitos protozoarios de la familia *Plasmodium* que se encuentra en especies de mosquitos de la familia *Anopheles* [4]. De todas las especies de parásitos, el *Plasmodium falciparum* es el principal responsable de los casos mortales [5]. En algunos casos, la enfermedad se puede controlar apropiadamente, sin embargo, en los últimos años surgieron algunos obstáculos para tratarla de manera efectiva[4]. La barrera más importante para la farmacoterapia es la resistencia que han desarrollado algunas variantes del *P.falciparum* a los medicamentos anti malaria [5,6]. El surgimiento gradual de esta barrera ha provocado que los fármacos para tratar la enfermedad sean cada vez más limitados, en consecuencia, el manejo de los contagios es deficiente y el número de muertos aumenta de manera significativa [4,5].

A raíz de esto, el desarrollo de nuevos medicamentos contra la malaria se ha convertido en un problema urgente para la industria farmacéutica. Sin embargo, este es un proceso muy complejo y requiere una cantidad considerable de tiempo. Una de las etapas más importantes del descubrimiento de nuevos medicamentos es la selección de los compuestos que pasarán a estudios *in vitro* e *in vivo.* En las últimas décadas, las herramientas computacionales han sido de gran ayuda para escoger los candidatos más apropiados. El diseño de fármacos asistido por computadora (CADD de sus siglas en inglés) consta de una serie de enfoques que se utilizan para descubrir y analizar compuestos químicos que presentan actividad biológica[7]. En la actualidad, la relación cuantitativa estructura actividad (QSAR de sus siglas en inglés) es una de las técnicas de CADD con mayor uso.

QSAR es un método en el que se utiliza la relación entre las estructuras moleculares de un set de moléculas y su actividad biológica, para construir modelos que se utilicen en la búsqueda de compuestos con reactividades similares [7]. Los modelos se construyen a partir de descriptores moleculares como variables independientes. Los descriptores son representaciones matemáticas de las propiedades físicas y químicas de una molécula, cuyos valores numéricos son generados por algoritmos computacionales a partir de sus estructuras. En un estudio QSAR se busca la correlación entre los descriptores y la actividad biológica de una serie de compuestos químicos, utilizando diferentes técnicas de regresión o métodos más complejos de aprendizaje automático (ML) [8]. La ventaja que presenta un estudio QSAR frente a otro tipo de enfoques es la versatilidad en términos de tiempo y dinero. Si un modelo se construye con el set más apropiado de descriptores moleculares, existe la posibilidad de cribar bases de datos externas y encontrar fármacos potenciales con estructuras diferentes que puedan superar los problemas de resistencia. Además, se pueden identificar patrones estructurales que influyen en la actividad biológica de un fármaco. Sin embargo, la complejidad de estos estudios radica en tres puntos importantes: el cálculo de los descriptores moleculares, la elección de la técnica estadística más apropiada, y la selección de la base de datos.

En los últimos 6 años se han reportado un gran número de estudios QSAR para el descubrimiento de fármacos anti malaria [3]. En la literatura, los enfoques más utilizados para la selección de descriptores moleculares son: 2D-QSAR (uso de descriptores 2D), 3D-QSAR (uso de descriptores 3D) y 2D-3D-QSAR (uso mixto de descriptores 2D y 3D) [3,9–11]. Con respecto a la elección de la técnica estadística, la gran mayoría de estudios utiliza técnicas de regresión lineal o técnicas no lineales de ML [3,12]. Para el presente estudio, se emplearon descriptores moleculares 2D y 3D combinando una gran variedad de técnicas de ML. De esta manera, se aseguró un nivel de exploración muy profundo para la fase de modelado.

El problema de la gran mayoría de estudios de la literatura radica en la selección de la base de datos, donde el común denominador es el limitado dominio de aplicación [3]. El dominio de aplicación es una región teórica en el espacio que está delimitada por los descriptores moleculares que componen el modelo. Si la base de datos utilizada para el modelado contiene moléculas estructuralmente similares, el dominio de aplicación se reduce considerablemente ya que los descriptores moleculares abarcan muy poco espacio del dominio. Esto se traduce en que las predicciones son confiables únicamente para moléculas con estructuras parecidas. Esto implica que, si el propósito de los estudios QSAR es la búsqueda de compuestos novedosos que sean candidatos como fármacos anti malaria, los resultados de los últimos 6 años son desalentadores. Por esta razón, se planteó un estudio de modelado desde una perspectiva diferente.

El uso de modelos tipo ensamble es uno de los caminos que se ha utilizado con menor frecuencia en estudios QSAR contra la malaria. Un modelo tipo ensamble es una técnica de ML en la que se combinan las predicciones de múltiples modelos individuales con el objetivo de mejorar el desempeño estadístico [13]. En los últimos 6 años, se realizó un único estudio de clasificación con modelos tipo ensamble [14]. Se utilizaron 5697 compuestos activos biológicamente contra la malaria y el desempeño estadístico fue bueno [14]. Adicionalmente, estudios pertenecientes a otras ramas demostraron que los modelos tipo ensamble son bastante prometedores, motivo por el cual se escogió este enfoque para el presente trabajo de titulación [13,15].

La base de datos que se seleccionó para el modelado lleva el nombre de Malaria Box, la cual consta de 400 compuestos químicos que presentan actividad biológica contra el *Plasmodium falciparum* [2]. Malaria Box recopila la actividad anti malaria utilizando la mitad de la concentración efectiva máxima (EC_{50}) [2]. En el área de la farmacología, el EC_{50} se define como la concentración de fármaco necesaria para causar el 50% del máximo del posible efecto [16].

Los valores de EC_{50} para las 400 moléculas se encuentran entre 30 nM y 4 μM, lo que implica que el rango de la actividad biológica es bastante amplio [2]. Malaria Box tiene como ventaja su variabilidad molecular, es decir, las moléculas que la componen no presentan estructuras similares. Esto beneficia al modelado ya que el propósito es mejorar el dominio de aplicación. Por lo tanto, el objetivo de este trabajo fue construir modelos tipo ensamble, utilizando descriptores moleculares 2D-3D y técnicas de ML, para predecir la actividad biológica de fármacos contra la malaria.

METODOLOGÍA

Construcción de la base de datos

La base de datos para el modelado se construyó con la Malaria Box. De los 400 compuestos, se consideraron únicamente aquellos cuyo EC_{50} se encontraba reportado contra el *P.falciparum* 3D7 en la base de datos ChEMB, dejando un total de 317 moléculas. La variabilidad molecular de este set indica que es viable construir un modelo con un dominio de aplicación bastante amplio.

En primer lugar, se obtuvieron los archivos del SMILE canónico de las 317 de moléculas. A partir de ellos, se construyeron las estructuras 2D y 3D de los compuestos utilizando Open Babel. Todas las estructuras 3D se optimizaron con el software RDKit, en el nivel de teoría de mecánica molecular, utilizando el campo de fuerza universal (UFF). Se calcularon tres tipos de descriptores moleculares: topológicos de dos dimensiones (2Dt), topológicos de tres dimensiones (3Dt) y mecano-cuánticos (MC).

Los descriptores 2Dt se calcularon con el software ToMoCoMD QuBiLs-MAS, mientras que los descriptores 3Dt se calcularon con ToMoCoMD QuBiLs-MIDAS. Se optimizaron nuevamente las estructuras a nivel semi empírico con el método PM6, utilizando Gaussian 16, para calcular los descriptores MC. A partir de los archivos de salida de Gaussian, se extrajeron los siguientes descriptores: energía interna (U), entalpía (H), energía libre de Gibbs (G), energía del orbital molecular ocupado más alto (E_HOMO), energía del orbital molecular no ocupado más bajo (E_LUMO), polarizabilidad, momento dipolar (E_dipole_m), Log10 (Q), energía térmica (E), capacidad calorífica (CV) y entropía (S). Adicionalmente, se agregó el peso molecular, el número de donantes de hidrógeno y el coeficiente de partición (AlogP), cuyos valores se reportaron en la Malaria Box. Finalmente, se eliminaron los descriptores colineales

utilizando un coeficiente de correlación de Spearman y una entropía de Shannon de 0.7, obteniendo un total de 317 descriptores 2Dt-MC y 499 descriptores 3Dt-MC.

Construcción de modelos de regresión global

En primera instancia, se planteó un modelado de regresión utilizando dos enfoques 2Dt-MC y 3Dt-MC. Se realizó una selección de atributos con el método ¨wrapper¨ de Weka 3.8. A pesar del alto costo computacional del método, este genera subconjuntos de descriptores que producen mejores resultados estadísticos que otros métodos [15]. Las técnicas de regresión utilizadas fueron las siguientes: procesos gaussianos (GP), aprendizaje basado en instancias con parámetro k (IBK), regresión lineal múltiple (LR), regresión vectorial de soporte (SMOR) y bosque con árboles aleatorios (RF). Cada una de ellas se combinó con tres métodos de búsqueda: el mejor primero (BF), búsqueda codiciosa (GS) y búsqueda por algoritmo genético (GEN).

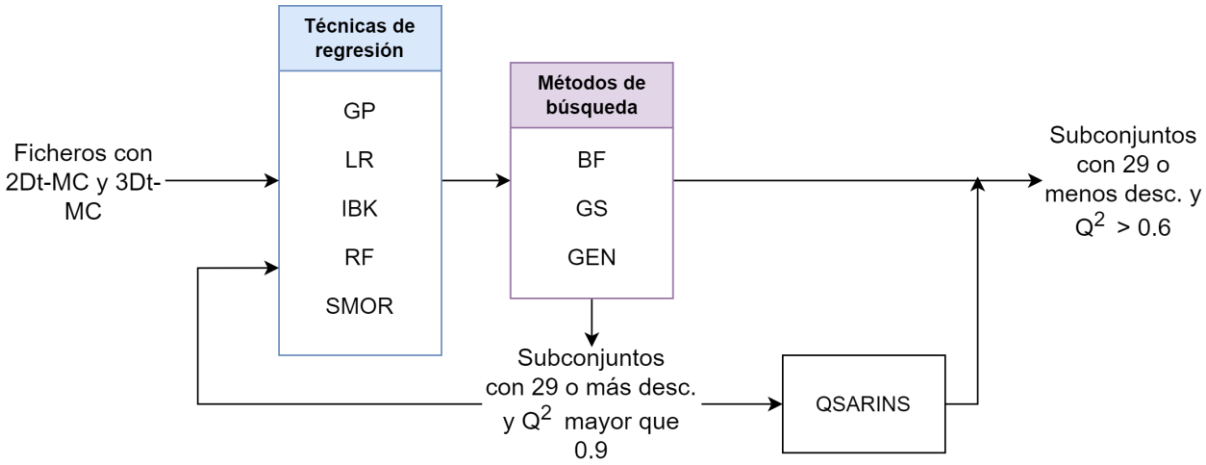


Figura 1. Diagrama del bloque de la selección de atributos para el modelado de regresión.

Finalmente, se analizó el coeficiente de la validación cruzada de 10 folds ($Q^2_{10\text{-fold}}$) como la métrica principal para el desempeño de los modelos. Aquellos subconjuntos con un número de descriptores mayor que 29 y un valor de $Q^2_{10\text{-fold}}$ cercano a 1, pasaron por una segunda búsqueda. Esto asegura que la relación entre el número de instancias y el número de atributos

sea de aproximadamente 11:1, tal como sugiere el marco teórico [15]. Para el modelado se utilizó como variable dependiente el pEC_{50}, definido en la ecuación 1.

$$pEC_{50} = -log\ (EC_{50}, M) \quad (1)$$

Construcción y validación de modelos de clasificación global

Se consideró un modelado de clasificación global debido al mal desempeño estadístico de los modelos de regresión. La base de datos se dividió en dos clases: muy activas (A) y activas (B), ya que todos los compuestos presentan actividad biológica contra el *P. falciparum*. Si el EC_{50} era igual o menor al corte se etiquetó como A, mientras que si era mayor al corte se etiquetó como B. Se exploraron cortes en el EC_{50} entre 0.6 y 1.2, con un intervalo de 0.1. De esta manera, se generaron 7 ficheros 2D-MC y 7 ficheros 3D-MC que se utilizaron para encontrar la partición de las moléculas y el subconjunto de descriptores con el mejor desempeño estadístico.

A continuación, se realizó una selección de atributos supervisada para cada uno de los 14 ficheros, utilizando el método ¨wrapper¨. Los subconjuntos con un número de descriptores mayor que 29 se descartaron. Las técnicas de clasificación utilizadas fueron las siguientes: aprendizaje de red de bayes (BN), función discriminante lineal de Fisher (FLDA), aprendizaje basado en instancias con parámetro k (IBK), árboles de decisión basados en la teoría de la información (J48), modelo de regresión logística multinomial con un estimador de cresta (LOG), bosque con árboles aleatorios (RF) y el algoritmo de optimización mínima secuencial de John Platt para entrenar un clasificador de vectores de soporte (SMO). En cuanto a los métodos de búsqueda, se utilizó el algoritmo de optimización de enjambre de partículas (PSO) y la selección de subconjuntos incrementales (IWSS), adicional a los métodos de búsqueda que se aplicaron en la etapa de regresión global. Este procedimiento se resume de manera gráfica en la Figura 2.

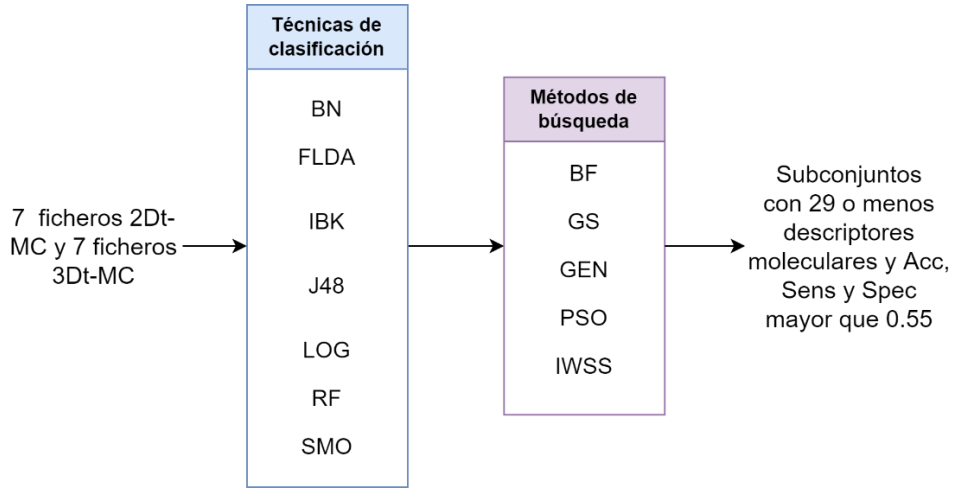


Figura 2. Diagrama del bloque de selección de atributos para el modelado de clasificación

Se efectuó una prueba de validación cruzada de 10 folds con los subconjuntos que pasaron los criterios de selección. Se construyeron los modelos con Weka 3.8, y se analizó su exactitud ($Acc_{10\text{-}fold}$), sensibilidad ($Sens_{10.fold}$) y especificidad ($Spec_{10\text{-}fold}$). Se seleccionó el mejor corte y sus mejores modelos pasaron al modelado individual. Este proceso se implementó en un script en Python aprovechando las funciones de la librería python-weka-wrapper3.

Como primer paso para la validación, se dividió la base de datos en un set de entrenamiento y en un set de prueba, utilizando un algoritmo de agrupación por k-medias. Este algoritmo ha sido ampliamente utilizado en la literatura para la evaluación de modelos predictivos. La agrupación en clústeres se realizó con el software para estadística Minitab, y la métrica que se utilizó para analizar la similitud entre las instancias fue la distancia euclidiana. Finalmente, se seleccionó 75% de las moléculas al azar dentro de cada clúster como parte del set de entrenamiento, y el 25% restante como set de prueba.

A continuación, se efectuó un análisis de dominio de aplicación (DA) para evaluar la confiabilidad de las predicciones. Se utilizaron cuatro métodos implementados en AMBIT Discovery: distancia de una manzana, distancia euclidiana, análisis de rango y densidad de

probabilidad. El criterio que se empleó para determinar que una molécula está fuera del DA del modelo es que más de dos de los métodos la consideren fuera.

El desempeño de los modelos se evaluó utilizando tres pruebas estadísticas: la validación cruzada de 10 folds para el set de entrenamiento, la predicción del set de prueba (Ext), y la validación cruzada dejando uno fuera (LOO). Para cada prueba se calcularon los siguientes estadísticos: exactitud (Acc), sensibilidad (Sens), especificidad (Spec), puntaje F, área bajo la curva ROC y el coeficiente de correlación de Mathews (MCC).

Como primer paso del modelado tipo ensamble se calculó la probabilidad de las predicciones (p) para cada uno de los modelos utilizando los ficheros de salida de Weka 3.8. A continuación, se definió el ΔP con la ecuación 2, si el modelo individual la predice como A, y con la ecuación 3, si la predice como B. Se utilizaron los ΔP de cada modelo individual para construir los ficheros, y se realizó una selección de atributos para escoger los mejores modelos individuales para construir el ensamble. Finalmente, se efectuaron las pruebas de validación descritas anteriormente y se comparó el desempeño del mejor modelo individual frente al modelo tipo ensamble.

$$\Delta P_A = (1 - 2p) \quad (2)$$

$$\Delta P_B = (2p - 1) \quad (3)$$

Construcción y validación de modelos de regresión por cada clase

Una vez construido el modelo de clasificación, se separaron ambas clases y se realizó un modelado de regresión para A y otro para B. Si bien las instancias son diferentes, los pasos del modelado fueron los mismos. En primer lugar, se eliminaron los descriptores colineales utilizando un coeficiente de correlación de Spearman de 0.6 y una entropía de Shannon de 0.7,

obteniendo un total de 1185 descriptores 3Dt-MC. Los descriptores 2Dt-MC se descartaron debido al bajo desempeño estadístico en las etapas de regresión y clasificación global.

Se realizó una primera selección de atributos utilizando Weka 3.8.0. Esta última con las técnicas de regresión y métodos de búsqueda que se emplearon para la fase de regresión global (Figura 1). Finalmente, se evaluó el $Q^2_{10\text{-fold}}$ para seleccionar los mejores modelos y se dividieron las instancias en set de entrenamiento y set de prueba, conservando los subconjuntos obtenidos en el modelado de clasificación global.

Como primera parte de la validación, se analizó el DA con el enfoque explicado para los modelos de clasificación global. Si una molécula está fuera del DA en más de dos de las cuatro técnicas, esta molécula está fuera del DA. El desempeño de los modelos se evaluó con cinco pruebas estadísticas: análisis de correlación, validación cruzada de 10 folds para el set de entrenamiento, la predicción del set de prueba (Ext), la validación cruzada dejando uno fuera (LOO), y aleatorización de la variable dependiente (y-scrambling). Se calcularon los parámetros estadísticos más relevantes tales como: el coeficiente de determinación (R^2), los valores de Q^2 , el error medio absoluto (MAE) y el error cuadrático medio (RMSE).

Los ficheros para el ensamble se construyeron con las predicciones para la prueba de validación cruzada de 10 folds de cada modelo. A continuación, se realizó una selección de atributos para escoger los mejores modelos individuales para el ensamble. Finalmente, se evaluó el desempeño estadístico del ensamble con las pruebas de validación, y se compararon sus resultados con el mejor modelo individual.

RESULTADOS Y DISCUSIÓN

Modelado de regresión global

Como primera aproximación para construir un modelo predictivo, se realizó una selección de atributos para regresión utilizando las 317 moléculas de la base de datos. La búsqueda de los mejores descriptores se realizó por separado para 2Dt-MC y 3Dt-MC. Cada subconjunto de descriptores se etiquetó con su respectiva técnica de regresión, método de búsqueda, y su número de descriptores moleculares. El primer parámetro estadístico que se utilizó para la selección de los mejores modelos fue el $Q^2_{10\text{-fold}}$, seguido del $MAE_{10\text{-fold}}$. Dentro del marco teórico, los mejores modelos presentan valores de $Q^2_{10\text{-fold}}$ cercanos a 1, y valores de $MAE_{10\text{-fold}}$ cercanos a 0 [17].

Tabla 1. Los cinco mejores modelos de regresión global para una primera selección de atributos, construidos con 2Dt-MC y 3Dt-MC, junto con sus parámetros estadísticos para la validación cruzada de 10 folds sin la partición entrenamiento/prueba ($Q^2_{10\text{-fold}}$ y MAE)

Nombre del modelo	$Q^2_{10\text{-fold}}$	$MAE_{10\text{-fold}}$	Tipo de descriptores
GP_BF_67	0.363	0.266	2Dt-MC
GP_GS_77	0.360	0.268	
SMOR_BF_27	0.259	0.278	
SMOR_GS_22	0.259	0.275	
LR_BF_14	0.219	0.292	
GP_BF_144	0.792	0.159	3Dt-MC
GP_GS_144	0.770	0.168	
LR_BF_38	0.545	0.232	
LR_GS_38	0.545	0.232	
SMOR_BF_36	0.495	0.237	

En total, se construyeron 29 modelos utilizando Weka 3.8 cuyos estadísticos se encuentran en el Anexo A. En la Tabla 1, se resumen los cinco mejores modelos para 2Dt-MC y los cinco mejores modelos para 3Dt-MC. Cabe recalcar que estos subconjuntos se obtuvieron después

de una primera búsqueda. Como se puede observar, los modelos construidos con 2Dt-MC no superan el 0.4 para su $Q^2_{10\text{-fold}}$, mientras que para los 3Dt-MC, hay dos modelos interesantes con un $Q^2_{10\text{-fold}}$ menor que 0.8. Sin embargo, si se toma en cuenta que apenas se alcanzó 0.792 de $Q^2_{10\text{-fold}}$ con 144 descriptores moleculares, una reducción en el número de descriptores no es un camino viable ya que automáticamente el $Q^2_{10\text{-fold}}$ cae por debajo de 0.5 . En un 100% de los casos, una segunda selección de atributos con una reducción muy significativa (10 o más descriptores) perturba en gran medida los parámetros estadísticos del modelo. Por lo tanto, se planteó el estudio desde otra perspectiva debido al mal desempeño de los modelos globales en términos de sus parámetros estadísticos

Modelado de clasificación global

El camino alterno que se escogió fue un modelado de clasificación global. En primera instancia, se dividió la base de datos en dos clases: muy activas (A) y activas (B). Se exploraron siete diferentes particiones de la base de datos y se crearon ficheros con una variable dependiente nominal. Para cada corte, se realizó una selección de atributos por separado para 2Dt-MC y 3Dt-MC, obteniendo en total 490 subconjuntos. Se descartaron aquellos modelos con más de 30 descriptores y se obtuvieron un total de 243 modelos para analizar. Los estadísticos de estos modelos se encuentran en el Anexo B. El parámetro principal que se consideró para escoger los mejores modelos fue la $Acc_{10\text{-fold}}$, seguido de $Sens_{10\text{-fold}}$ y $Spec_{10\text{-fold}}$. De acuerdo con la literatura, los mejores modelos presentan valores más grandes de $Acc_{10\text{-fold}}$, y valores de $Sens_{10\text{-fold}}$ y $Sens_{10\text{-fold}}$ cercanos a 1. Estos dos últimos parámetros son indicadores de la fiabilidad con la que el modelo clasifica A, y de la fiabilidad con la que el modelo clasifica B, respectivamente.

Tabla 2. Los mejores modelos de clasificación global, construidos con 2Dt-MC y 3Dt-MC para una primera selección de atributos, para cada corte en el valor del EC50 junto con sus parámetros estadísticos para la validación cruzada de 10 folds sin la partición entrenamiento/prueba ($Acc_{10\text{-}fold}$, $Sens_{10\text{-}fold}$ y $Sens_{10\text{-}fold}$)

Nombre del modelo	**Corte**	**$Acc_{10\text{-}fold}$**	**$Sens_{10\text{-}fold}$**	**$Spec_{10\text{-}fold}$**	**Tipo de descriptores**
J48_GS_17	0.6	0.710	0.450	0.888	2Dt-MC
IBK_IWSS_17	0.7	0.672	0.644	0.701	
IBK_BF_8	0.8	0.672	0.726	0.606	
RF_BF_8	0.9	0.722	0.855	0.516	
IBK_IWSS_8	1.0	0.729	0.892	0.394	
IBK_BF_8	1.1	0.804	0.988	0.213	
BN_GEN_6	1.2	0.845	1.00	0.00	
SMO_BF_12	0.6	0.732	0.481	0.904	3Dt-MC
RF_BF_12	0.7	0.707	0.688	0.726	
SMO_BF_9	0.8	0.697	0.766	0.613	
IBK_BF_11	0.9	0.729	0.928	0.419	
IBK_IWSS_10	1.0	0.757	0.925	0.414	
LOG_BF_19	1.1	0.814	0.975	0.293	
IBK_BF_12	1.2	0.880	0.993	0.265	

En la Tabla 2, se resume el mejor modelo de cada corte tanto para 2Dt-MC como 3Dt-MC . Cada modelo se rotuló con su respectiva técnica de clasificación, método de búsqueda y número de descriptores. En cuanto a sus parámetros estadísticos, se puede apreciar que la gran mayoría de los modelos presentan una exactitud ($Acc_{10\text{-}fold}$) mayor a 0.7 siendo los cortes mayores a 1.0 aquellos con una exactitud superior a 0.8. Con respecto a la sensibilidad ($Sens_{10\text{-}fold}$) y especificidad ($Spec_{10\text{-}fold}$), se observa que no hay un equilibrio entre ambos parámetros para los cortes mayores que 0.9 y menores que 0.7. Por un lado, la sensibilidad es muy baja para el corte de 0.6, mientras que la especificidad es muy baja para los cortes de 0.9 a 1.2. Esto se debe a que las clases están muy desbalanceadas para estos casos. Por lo tanto, la construcción

de modelos de clasificación con estos cortes no es un camino factible y se descartaron, dejando como candidatos principales los cortes de 0.7 y 0.8.

Al realizar la comparación entre los mejores modelos de ambos cortes, se observa que los dos presentan estadísticos similares para 2Dt-MC, con diferencias poco perceptibles en sus valores de sensibilidad y especificidad. Sin embargo, de manera general, los modelos 3Dt-MC para el corte de 0.7 presentan mejor $Acc_{10\text{-}fold}$ que los modelos con el corte de 0.8. Esta tendencia se evidencia de mejor manera en el Anexo B. Por lo tanto, se escogió a 0.7 como el mejor corte para el modelado de clasificación, y 3Dt-MC como el mejor tipo de descriptores para el modelado. En total, nueve modelos de clasificación 3Dt-MC pasaron a la siguiente etapa, en donde el mejor fue RF_BF_12 (M1_CLASS).

El siguiente paso fue la partición en los sets entrenamiento/prueba. Se realizó un agrupamiento en clústeres con la técnica de k-medias y, como resultado, las moléculas se agruparon en 13 clústeres por similitud. Para cada clúster, se escogió un 75% de moléculas para conformar el set de entrenamiento, y un 25% de moléculas para el set de prueba. Una vez creados los ficheros para los nueve modelos, se realizó un análisis del dominio de aplicación utilizando AMBIT Discovery. Los resultados fueron satisfactorios ya que los sets de entrenamiento tuvieron una cobertura del 100% para los sets de prueba. Finalmente, se evaluó el desempeño y robustez de los modelos de clasificación utilizando las siguientes pruebas: validación cruzada de 10-folds para el set de entrenamiento, predicción del set de prueba (Ext) y validación cruzada de dejar uno fuera (LOO). En el Anexo C, se recopilan los estadísticos para la validación de los nueve mejores modelos de clasificación global.

Para continuar con el modelado de clasificación, se construyó un modelo tipo ensamble a partir de cuatro de los modelos del Anexo C (J48_GS_12, IBK_IWSS_8, RF_IWSS_12, RF_BF_12), y se evaluó su desempeño con las pruebas de validación descritas anteriormente. La Tabla 3

resume los parámetros estadísticos para el mejor modelo de clasificación individual y del modelo tipo ensamble. Se observa como E_CLASS supera en la gran mayoría de parámetros estadísticos a M1_CLASS. Esto sucede ya que E_CLASS fue construido con cuatro modelos individuales, lo que disminuye considerablemente el error en la clasificación de un subconjunto de moléculas. Por último, se realizó un análisis de ANOVA para $Acc_{10\text{-fold}}$ y $MCC_{10\text{-fold}}$ utilizando un 95% de significancia. Se calcularon diez observaciones de estos parámetros para cada modelo utilizando una semilla al azar en la validación cruzada, y se encontró que M1_CLASS y E_CLASS son diferentes estadísticamente. Los resultados del análisis de ANOVA se encuentran en el Anexo D.

Tabla 3. Parámetros estadísticos para la validación del mejor modelo individual (M1_CLASS) y del modelo tipo ensamble (E_CLASS) con la partición entrenamiento/prueba.

Parámetro	**M1_CLASS**	**E_CLASS**
$Acc_{10\text{-fold}}$	0.696	0.738
$Sens_{10\text{-fold}}$	0.689	0.748
$Spec_{10\text{-fold}}$	0.703	0.729
Puntaje $F_{10\text{-fold}}$	0.695	0.742
$ROC_{10\text{-fold}}$	0.765	0.792
$MCC_{10\text{-fold}}$	0.392	0.477
Acc_{Ext}	0.675	0.675
$Sens_{Ext}$	0.585	0.537
$Spec_{Ext}$	0.769	0.821
Puntaje F_{Ext}	0.649	0.629
ROC_{Ext}	0.690	0.708
MCC_{Ext}	0.360	0.371
Acc_{LOO}	0.717	0.734
$Sens_{LOO}$	0.739	0.731
$Spec_{LOO}$	0.695	0.737
Puntaje F_{LOO}	0.724	0.734
ROC_{LOO}	0.779	0.789
MCC_{LOO}	0.435	0.468

Adicionalmente, se realizó una comparación detallada de cuatro de los parámetros estadísticos más utilizados en la literatura para analizar la robustez de modelos de clasificación binarios

(exactitud, puntaje F, área bajo la curva ROC y coeficiente de correlación de Mathews). Estos cuatro parámetros en conjunto brindan una estimación buena del desempeño del modelo, el balance de las clases, y la confiabilidad de sus predicciones[18,19]. Un modelo estadístico fiable presenta valores para estos parámetros cercanos a 1. Tal como se observa en la Figura 3, E_CLASS supera o iguala a M1_CLASS en el desempeño y fiabilidad en las tres pruebas de validación. Esto demuestra que el uso de modelos tipo ensamble mejora de manera significativa el desempeño de un modelo predictivo de clasificación.

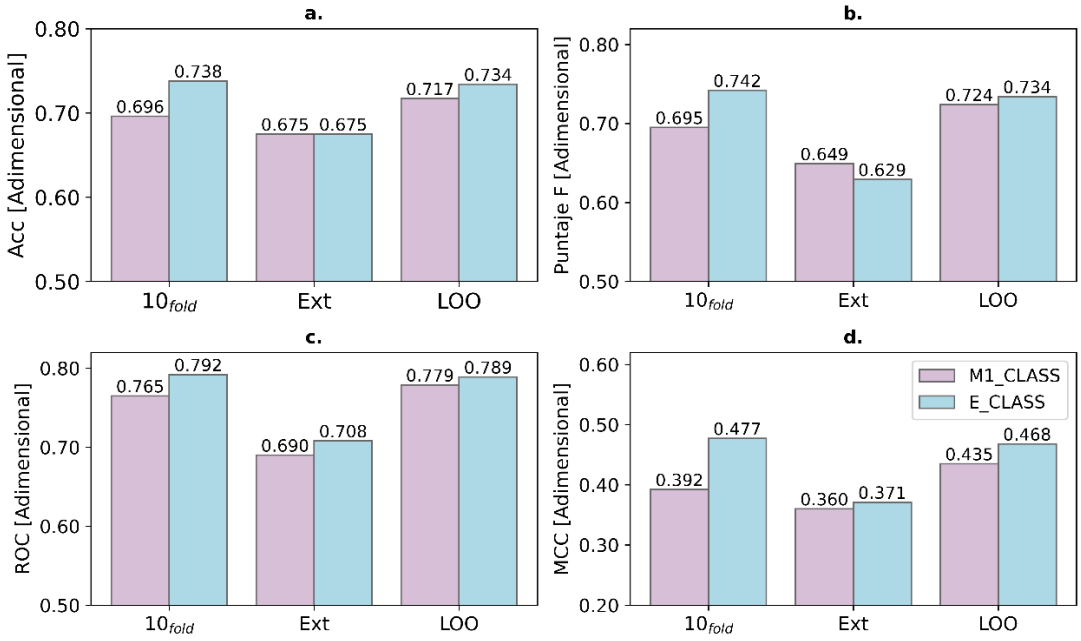


Figura 3. Comparación entre el mejor modelo individual de clasificación (M1_CLASS) vs el modeo tipo ensamble (E_CLASS) para su exactitud (a), puntaje F (b), área bajo la curva ROC (c) y coeficiente de correlación de Mathews (d).

Por último, se analizaron los descriptores moleculares de los modelos con los que se construyó E_CLASS. Dentro de los cuatro modelos, donde también está incluido M1_CLASS, al menos el 60% de los descriptores topográficos fueron calculados en función de tres propiedades químicas: polarizabilidad, carga y el coeficiente de partición octanol/agua. Esto sugiere que la distribución de la nube electrónica, las interacciones electrostáticas y el comportamiento en

medio acuoso influyen de manera significativa en la distinción de una molécula con mayor actividad biológica frente a una con menor actividad.

Modelado de regresión por separado para cada clase

Una vez que se realizó el modelado global de clasificación, la siguiente fase fue un modelado de regresión por separado. El procedimiento para ambas clases fue exactamente el mismo. En principio, se realizó una selección de atributos de regresión para la clase muy activa (A), como para la clase activa (B). Se utilizaron únicamente descriptores 3Dt-MC debido al bajo desempeño de los 2Dt-MC. El primer parámetro estadístico que se utilizó para la selección de los mejores modelos fue el $Q^2_{10\text{-fold}}$, seguido del $MAE_{10\text{-fold}}$. Inicialmente, se construyeron un total de 15 modelos por clase, y la Tabla 4 resume los parámetros estadísticos de los mejores cinco.

Tabla 4. Los 5 mejores modelos de regresión para las clases: Muy Activa (A) y Activa (B), construidos con 3Dt-MC, junto con sus parámetros estadísticos ($Q^2_{10\text{-fold}}$ y $MAE_{10\text{-fold}}$) para la validación cruzada de 10 folds sin la partición entrenamiento/prueba

Nombre del modelo	$Q^2_{10\text{-fold}}$	$MAE_{10\text{-fold}}$	Clase
LR_BF_70	0.944	0.079	A
LR_GS_71	0.944	0.079	
GP_BF_101	0.932	0.112	
GP_GS_131	0.931	0.114	
SMOR_BF_31	0.661	0.181	
GP_BF_112	0.930	0.046	B
GP_GS_117	0.922	0.048	
LR_BF_44	0.840	0.054	
LR_GS_44	0.840	0.054	
SMOR_BF_30	0.663	0.072	

En la Tabla 4 se observa que los modelos alcanzaron valores de $Q^2_{10\text{-fold}}$ muy cercanos a 1. Sin embargo, el número de descriptores de estos modelos es mayor que 30, tanto para la clase A

como para la clase B. Por lo tanto, se realizó una reducción en el número de descriptores que conforman los cuatro mejores modelos por cada clase utilizando Weka 3.8 y QSARINS. El riesgo que se asumió para este procedimiento es la perturbación en el desempeño del $Q^2_{10\text{-fold}}$. Los parámetros estadísticos para todos los modelos, incluyendo los de la primera búsqueda, se encuentran en el Anexo E. Se escogieron cinco modelos para la clase A, de los cuales el mejor fue GP_BF_101_QSARINS_20 (M5_REG_A), y cuatro modelos para la clase B, de los cuales el mejor fue GP_BF_112_LR_BF_28 (M1_REG_B). Estos nueve modelos fueron los candidatos óptimos para pasar a la siguiente fase del modelado.

Con respecto a la primera parte de la validación, se mantuvo la partición de entrenamiento-prueba realizada en el modelado de clasificación global. Una vez creados los ficheros, se analizó el dominio de aplicación (DA) de los modelos con AMBIT Discovery, obteniendo un 100% de cobertura del set de entrenamiento para el set de prueba. Por último, se evaluó el desempeño y la robustez de los modelos de regresión utilizando cinco pruebas estadísticas (prueba de correlación, validación cruzada de 10 folds para el set de entrenamiento, validación cruzada de dejar uno fuera, predicción del set de prueba, y aleatorización de la variable dependiente) . Los mejores modelos presentaron valores de R^2 y Q^2 cercanos a 1, lo que sugiere un buena correlación y predictibilidad. Por otro lado, los valores del error medio absoluto (MAE), error cuadrático medio (RMSE) y $Q^2_{\text{y-Scrambling}}$ fueron cercanos a 0. Esto implica que los modelo no realizan predicciones del pEC_{50} al azar, y lo hacen con buena exactitud. Los parámetros estadísticos para los modelos individuales de la clase A se encuentran en el Anexo F, y de la clase B en el Anexo G.

El siguiente paso fue la construcción de un modelo tipo ensamble para cada clase, a partir de los modelos de los Anexos F y G. En el caso de la clase A, el ensamble (E_REG_A) se construyó a partir de las predicciones de los modelos: GP_BF_101_QSARINS_20,

LR_GS_71_QSARINS_20 y LR_BF_70_SMOR_BF_22. Por otro lado, el ensamble para la clase B (E_REG_B) utilizando las predicciones de: LR_BF_44_QSARINS_20, GP_BF_112_QSARINS_20 y GP_BF_112_LR_BF_28. Se evaluaron ambos ensambles con las pruebas de validación descritas anteriormente, y se comparó su desempeño con el mejor modelo individual para cada clase. Los parámetros estadísticos para la validación se recopilaron en la Tabla 5, y se puede apreciar como los modelos tipo ensamble superan a los modelos individuales en todos sus estadísticos. Se realizó un análisis de ANOVA para $Q^2_{10\text{-fold}}$ y $RMSE_{10\text{-fold}}$ utilizando un 95% de significancia. Se calcularon diez observaciones de estos parámetros para cada modelo utilizando una semilla al azar en la validación cruzada, y se encontró que los modelos de tipo ensamble (E_REG_A y E_REG_B) son estadísticamente diferentes a sus modelos individuales (M5_REG_A y M1_REG_B). Los resultados del análisis de ANOVA para cada clase se encuentran en los Anexos H-I.

Tabla 5. Parámetros estadísticos para la validación de los mejores modelos individuales y de los modelos tipo ensamble con la partición entrenamiento/prueba para las clases: Muy Activa (A) y Activa (B)

Parámetro	**M5_REG_A**	**E_REG_A**	**M1_REG_B**	**E_REG_B**
R^2	0.815	0.831	0.843	0.855
MAE	0.127	0.130	0.056	0.053
RMSE	0.184	0.177	0.069	0.067
$Q^2_{10\text{-fold}}$	0.729	0.810	0.708	0.793
$MAE_{10\text{-fold}}$	0.174	0.139	0.075	0.065
$RMSE_{10\text{-fold}}$	0.221	0.188	0.096	0.080
Q^2_{LOO}	0.775	0.802	0.725	0.804
MAE_{LOO}	0.151	0.139	0.075	0.061
$RMSE_{LOO}$	0.202	0.192	0.093	0.077
Q^2_{Ext}	0.621	0.749	0.681	0.765
MAE_{Ext}	0.175	0.138	0.075	0.060
$RMSE_{Ext}$	0.230	0.193	0.093	0.076
$Q^2_{yScrambling}$	0.025	0.025	0.026	0.026

Adicionalmente, se realizó una comparación detallada del Q^2, MAE y RMSE para tres de las pruebas de validación. En las Figuras 4-5, se puede observar que los modelos de tipo ensamble superan con creces a los mejores modelos individuales en su desempeño. Las diferencias más significativas se encuentran en la validación cruzada de 10 folds y en la predicción del set de prueba. Esto quiere decir que las predicciones obtenidas con el modelo tipo ensamble son mucho más confiables. Por esta razón, uno de los proyectos con mayor impacto a largo plazo es la implementación de los modelos del presente estudio para el cribado de bases de datos externas, con el objetivo de encontrar nuevos candidatos contra la malaria.

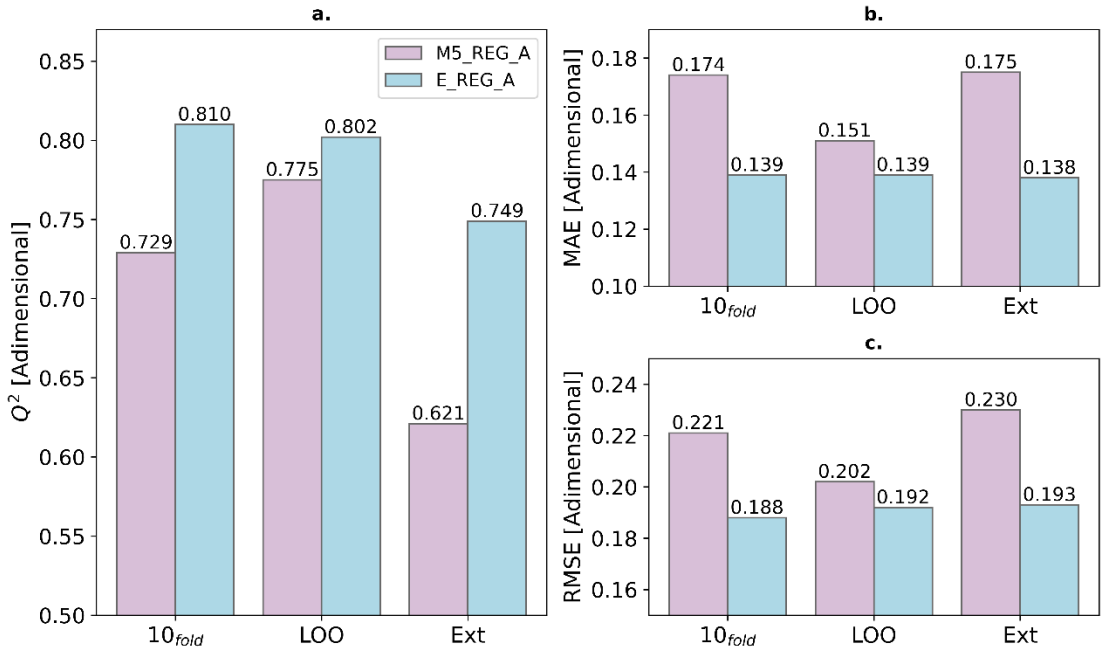


Figura 4. Comparación, para la clase Muy Activa (A), entre el mejor modelo individual de regresión (M5_REG_A) vs el modelo tipo ensamble (E_REG_A) para su Q^2 (a), error medio absoluto (b) y error cuadrático medio (d).

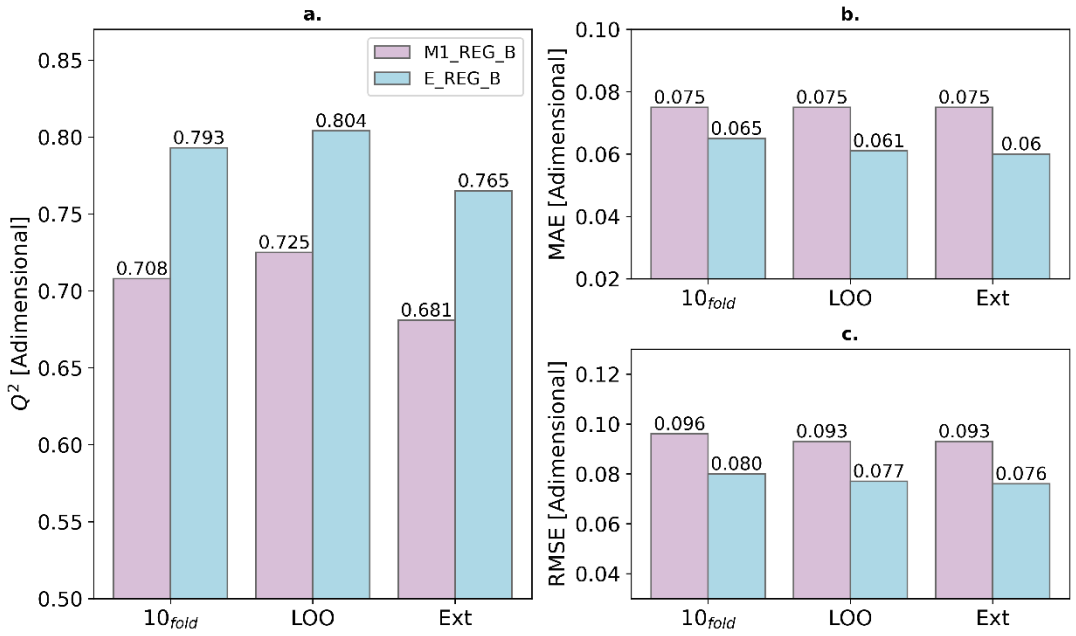


Figura 5. Comparación, para la clase Activa (B), entre el mejor modelo individual de regresión (M1_REG_B) vs el modelo tipo ensamble (E_REG_B) para su Q^2 (a), error medio absoluto (b) y error cuadrático medio (d).

Por último, se realizó un análisis mucho más detallado de los descriptores moleculares que participaron en los modelos de tipo ensamble. Tanto para la clase A como la clase B, se observa que al menos el 60% de los descriptores fueron calculados en función de cuatro propiedades químicas: superficie polar, electronegatividad, coeficiente de partición octanol/agua, y el volumen de van de van der Waals. Esto sugiere que la localización de los átomos más electronegativos, la redistribución de la carga, el tamaño de la molécula, y las interacciones en un medio acuoso juegan un papel muy importante en la predicción de la actividad biológica. Por lo tanto, estos son factores se deben considerar para la búsqueda y desarrollo de nuevos fármacos contra el *P. falciparum.*

CONCLUSIONES

En el presente estudio, se construyeron modelos tipo ensamble en función de descriptores moleculares topográficos 2D (2Dt), topográficos 3D (3Dt), y mecano-cuánticos (MC) calculados en el nivel semi empírico con el método PM6. En principio, se planteó un modelado de regresión para predecir el pEC_{50} utilizando las 317 moléculas de la base de datos Malaria Box. Sin embargo, los modelos construidos tanto con 2Dt-MC como 3Dt-MC, presentaron valores de $Q^2_{10\text{-fold}}$ inferiores a 0.7 y su número de descriptores fue muy superior a los límites establecidos en la literatura.

Por esta razón, se optó por un modelado de clasificación en donde las moléculas se etiquetaron como muy activas (A), si su EC_{50} era menor o igual que un corte establecido, y activas (B), si su EC_{50} era mayor o igual al corte. Se exploraron siete cortes diferentes y, se demostró a través del análisis estadístico de la exactitud, sensibilidad y especificidad para una validación cruzada de 10 folds, que los mejores modelos de clasificación se construyeron con un corte de 0.7 en el EC_{50}. Posteriormente, se realizó una partición en sets de entrenamiento/prueba con un algoritmo de agrupación en clústeres por k-medias para la validación estadística de los modelos. En primer lugar, se analizó el dominio de aplicación (DA) utilizando AMBIT Discovery, donde se obtuvo un 100% de cobertura para el set de entrenamiento. Esto implica que, a pesar de la variabilidad molecular de la Malaria Box, el modelo podría utilizarse para clasificar bases de datos externas con variedad de compuestos que no han sido explorados aún. Por último, se construyó un ensamble con los mejores modelos de clasificación, y se comparó su desempeño frente al mejor modelo individual. El ensamble (E_CLASS) superó al modelo individual (M1_CLASS) en casi todos sus parámetros. E_CLASS alcanzó una exactitud de 0.738, un puntaje F de 0.742, un área bajo la curva ROC de 0.792 y un coeficiente de correlación de Mathews de 0.477, para la validación cruzada de 10 folds. Este análisis demostró

que la predictibilidad del modelo tipo ensamble es muy buena, y supera a los parámetros del modelo individual.

La siguiente etapa fue un modelado de regresión por separado para predecir el pEC_{50}. Se siguió el mismo procedimiento tanto para la clase A como la clase B. Se realizó una selección de atributos con descriptores 3Dt-MC y se escogieron los mejores modelos de acuerdo con el $Q^2_{10\text{-}fold}$. Se conservó la partición entrenamiento/prueba realizada en el modelado de clasificación, y se analizó el DA con AMBIT Discovery. A pesar de que las moléculas no eran estructuralmente similares, se obtuvo el 100% de cobertura. Esto implica que los modelos podrían utilizarse para predecir actividad biológica de bases de datos con estructuras variadas. Como último paso, se construyó un ensamble con los mejores modelos de regresión de cada clase, y se comparó su desempeño estadístico con el mejor modelo individual. Los ensambles (E_REG_A y E_REG_B) superaron con creces a los modelos individuales (M5_REG_A y M1_REG_B) en todos sus parámetros. E_REG_A alcanzó un $Q^2_{10\text{-}fold}$ de 0.810 y un Q^2_{Ext} = 0.749, mientras que E_REG_B un $Q^2_{10\text{-}fold}$ de 0.793 y un Q^2_{Ext} de 0.765. Esto implica que la predictibilidad de los modelos tipo ensamble es muy buena, y supera con creces la de los modelos individuales.

Finalmente, se realizó un análisis a detalle de los descriptores que participaron en los modelos de clasificación y regresión. Se encontró que las propiedades relacionadas con la distribución de la nube electrónica, el volumen que ocupa la estructura, y la hidrofobicidad de la molécula juegan un rol importante en la distinción de clases y en la predicción del pEC_{50}. Por esta razón, estas propiedades deberían ser consideradas en estudios posteriores sobre el desarrollo de nuevos fármacos contra la malaria.

REFERENCIAS BIBLIOGRÁFICAS

1. Prevention, C.-C. for D. C. and. (2021, diciembre 16). *CDC - Malaria—Malaria Worldwide—Impact of Malaria*. https://www.cdc.gov/malaria/malaria_worldwide/impact.html
2. Spangenberg, T., Burrows, J. N., Kowalczyk, P., McDonald, S., Wells, T. N. C., & Willis, P. (2013). The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases. *PLOS ONE*, *8*(6), e62906. https://doi.org/10.1371/journal.pone.0062906
3. Ojha, P. K., Kumar, V., Roy, J., & Roy, K. (2021). Recent advances in quantitative structure–activity relationship models of antimalarial drugs. *Expert Opinion on Drug Discovery*, *16*(6), 659–695. Scopus. https://doi.org/10.1080/17460441.2021.1866535
4. Tibon, N. S., Ng, C. H., & Cheong, S. L. (2020). Current progress in antimalarial pharmacotherapy and multi-target drug discovery. *European Journal of Medicinal Chemistry*, *188*, 111983. https://doi.org/10.1016/j.ejmech.2019.111983
5. Hanboonkunupakarn, B., & White, N. J. (2022). Advances and roadblocks in the treatment of malaria. *British Journal of Clinical Pharmacology*, *88*(2), 374–382. Scopus. https://doi.org/10.1111/bcp.14474
6. Flores, M. C., Márquez, E. A., & Mora, J. R. (2018). Molecular modeling studies of bromopyrrole alkaloids as potential antimalarial compounds: A DFT approach. *Medicinal Chemistry Research*, *27*(3), 844–856. https://doi.org/10.1007/s00044-017-2107-3
7. Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, *224*, 113705. https://doi.org/10.1016/j.ejmech.2021.113705
8. Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, *11*(3), 225–239. https://doi.org/10.1517/17460441.2016.1146250
9. Sharma, K., Srivastava, A., Tiwari, P., Sharma, S., Shaquiquzzaman, M., Alam, M. M., & Akhter, M. (2019). 3D QSAR Based Virtual Screening of Pyrido[1,2-a] Benzimidazoles as Potent Antimalarial Agents. *Letters in Drug Design & Discovery*, *16*(3), 301–312.

10. Flores-Sumoza, M., Alcázar, J. J., Márquez, E., Mora, J. R., Lezama, J., & Puello, E. (2018). Classical QSAR and Docking Simulation of 4-Pyridone Derivatives for Their Antimalarial Activity. *Molecules*, *23*(12), 3166. https://doi.org/10.3390/molecules23123166

11. Jarrahpour, A., Aye, M., Rad, J. A., Yousefinejad, S., Sinou, V., Latour, C., Brunel, J. M., & Turos, E. (2018). Design, synthesis, activity evaluation and QSAR studies of novel antimalarial 1,2,3-triazolo-β-lactam derivatives. *Journal of the Iranian Chemical Society*, *15*(6), 1311–1326. https://doi.org/10.1007/s13738-018-1330-2

12. Neves, B. J., Braga, R. C., Alves, V. M., Lima, M. N. N., Cassiano, G. C., Muratov, E. N., Costa, F. T. M., & Andrade, C. H. (2020). Deep Learning-driven research for drug discovery: Tackling Malaria. *PLOS Computational Biology*, *16*(2), e1007025. https://doi.org/10.1371/journal.pcbi.1007025

13. Wu, Z., Zhu, M., Kang, Y., Leung, E. L.-H., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D., & Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings in Bioinformatics*, *22*(4), bbaa321. https://doi.org/10.1093/bib/bbaa321

14. Caballero-Alfonso, A. Y., Cruz-Monteagudo, M., Tejera, E., Benfenati, E., Borges, F., Cordeiro, M. N. D. S., Armijos-Jaramillo, V., & Perez-Castillo, Y. (2019). Ensemble-Based Modeling of Chemical Compounds with Antimalarial Activity. *Current Topics in Medicinal Chemistry*, *19*(11), 957–969. https://doi.org/10.2174/1568026619666190510100313

15. Mora, J. R., Marrero-Ponce, Y., García-Jacas, C. R., & Suarez Causado, A. (2020). Ensemble Models Based on QuBiLS-MAS Features and Shallow Learning for the Prediction of Drug-Induced Liver Toxicity: Improving Deep Learning and Traditional Approaches. *Chemical Research in Toxicology*, *33*(7), 1855–1873. https://doi.org/10.1021/acs.chemrestox.0c00030

16. Waller, D. G., & Sampson, A. P. (2018). 1—Principles of pharmacology and mechanisms of drug action. En D. G. Waller & A. P. Sampson (Eds.), *Medical Pharmacology and Therapeutics (Fifth Edition)* (pp. 3–31). Elsevier. https://doi.org/10.1016/B978-0-7020-7167-6.00001-4

17. Cabrera, N., Mora, J. R., & Marquez, E. A. (2019). Computational Molecular Modeling of Pin1 Inhibition Activity of Quinazoline, Benzophenone, and Pyrimidine Derivatives. *Journal of Chemistry*, *2019*, 1–11. https://doi.org/10.1155/2019/2954250

18. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

19. Halimu, C., Kasem, A., & Newaz, S. H. S. (2019). Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification. *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, 1–6. https://doi.org/10.1145/3310986.3311023

ANEXO A: MODELOS DE REGRESIÓN GLOBAL PARA UNA PRIMERA SELECCIÓN DE ATRIBUTOS, CONSTRUIDOS CON 2DT-MC Y 3DT-MC, JUNTO CON SUS PARÁMETROS ESTADÍSTICOS PARA LA VALIDACIÓN CRUZADA DE 10 FOLDS SIN LA PARTICIÓN ENTRENAMIENTO/PRUEBA ($Q^2_{10\text{-FOLD}}$ Y MAE)

Nombre del modelo	$Q^2_{10\text{-fold}}$	$MAE_{10\text{-fold}}$	Tipo de descriptores
GP_BF_67	0.363	0.266	2Dt-MC
GP_GS_77	0.360	0.268	
GP_GEN_123	0.122	0.326	
IBK_BF_6	0.201	0.292	
IBK_GS_6	0.201	0.292	
IBK_GEN_126	0.156	0.293	
LR_BF_14	0.219	0.292	
LR_GS_14	0.219	0.292	
LR_GEN_107	0.145	0.345	
SMOR_BF_27	0.259	0.278	
SMOR_GS_22	0.259	0.275	
SMOR_GEN_119	0.152	0.323	
RF_BF_4	0.180	0.294	
RF_GS_4	0.180	0.294	
RF_GEN_107	0.127	0.303	
GP_BF_144	0.792	0.159	3Dt-MC
GP_GS_144	0.770	0.168	
GP_GEN_205	0.353	0.294	
IBK_BF_9	0.228	0.286	
IBK_GEN_269	0.172	0.283	
LR_BF_38	0.545	0.232	
LR_GS_38	0.545	0.232	
LR_GEN_185	0.144	0.380	
SMOR_BF_36	0.495	0.237	
SMOR_GS_36	0.495	0.237	
SMOR_GEN_149	0.319	0.311	
RF_BF_15	0.245	0.293	
RF_GS_8	0.188	0.301	
RF_GEN_217	0.183	0.298	

ANEXO B: MODELOS DE CLASIFICACIÓN GLOBAL, CONSTRUIDOS CON 2DT-MC Y 3DT-MC PARA UNA PRIMERA SELECCIÓN DE ATRIBUTOS, PARA CADA CORTE EN EL VALOR DEL EC50 JUNTO CON SUS PARÁMETROS ESTADÍSTICOS PARA LA VALIDACIÓN CRUZADA DE 10 FOLDS SIN LA PARTICIÓN ENTRENAMIENTO/PRUEBA ($ACC_{10\text{-FOLD}}$, $SENS_{10\text{-FOLD}}$ Y $SPEC_{10\text{-FOLD}}$)

Nombre del modelo	Corte	$Acc_{10\text{-fold}}$	$Sens_{10\text{-fold}}$	$Spec_{10\text{-fold}}$	Tipo de descriptores
BN_BF_1	0.6	0.662	0.217	0.968	2Dt-MC
BN_GS_1		0.662	0.217	0.968	
BN_IWSS_1		0.662	0.217	0.968	
SMO_BF_7		0.707	0.372	0.936	
SMO_GS_7		0.707	0.372	0.936	
SMO_IWSS_17		0.694	0.395	0.899	
LOG_BF_4		0.688	0.349	0.920	
LOG_GS_4		0.688	0.349	0.920	
LOG_IWSS_11		0.653	0.395	0.830	
FLDA_BF_5		0.653	0.558	0.718	
FLDA_GS_3		0.637	0.535	0.707	
FLDA_IWSS_14		0.634	0.496	0.729	
IBK_BF_6		0.621	0.419	0.761	
IBK_GS_6		0.621	0.419	0.761	
IBK_IWSS_5		0.647	0.512	0.739	
J48_BF_19		0.700	0.434	0.883	
J48_GS_17		0.710	0.450	0.888	
J48_IWSS_4		0.678	0.287	0.947	
RF_BF_10		0.688	0.473	0.835	
RF_GS_7		0.678	0.419	0.856	
RF_IWSS_12		0.669	0.465	0.809	
BN_BF_6		0.644	0.186	0.957	3Dt-MC
BN_GS_6		0.644	0.186	0.957	
BN_IWSS_2		0.644	0.140	0.989	
SMO_BF_12		0.732	0.481	0.904	
SMO_GS_9		0.719	0.504	0.867	
SMO_IWSS_17		0.726	0.504	0.878	
LOG_BF_8		0.697	0.465	0.856	
LOG_GS_4		0.681	0.450	0.840	
LOG_IWSS_12		0.669	0.465	0.809	
FLDA_BF_7		0.700	0.651	0.734	
FLDA_GS_7		0.700	0.651	0.734	
FLDA_IWSS_13		0.669	0.643	0.686	
IBK_BF_2		0.672	0.535	0.766	

IBK_GS_2		0.672	0.535	0.766	
IBK_IWSS_6		0.644	0.457	0.771	
J48_BF_22		0.656	0.403	0.830	
J48_GS_16		0.666	0.388	0.856	
J48_IWSS_12		0.707	0.504	0.846	
RF_BF_10		0.710	0.496	0.856	
RF_GS_6		0.713	0.496	0.862	
RF_IWSS_22		0.700	0.488	0.846	
BN_BF_4	0.7	0.524	0.475	0.573	2Dt-MC
BN_GS_4		0.524	0.475	0.573	
BN_IWSS_1		0.517	0.569	0.465	
SMO_BF_10		0.628	0.588	0.669	
SMO_GS_8		0.606	0.575	0.637	
SMO_IWSS_19		0.644	0.475	0.815	
LOG_BF_3		0.599	0.631	0.567	
LOG_GS_3		0.599	0.631	0.567	
LOG_IWSS_14		0.596	0.613	0.580	
FLDA_BF_4		0.603	0.600	0.605	
FLDA_GS_2		0.612	0.625	0.599	
FLDA_IWSS_19		0.634	0.644	0.624	
IBK_BF_5		0.666	0.644	0.688	
IBK_GS_3		0.618	0.631	0.605	
IBK_IWSS_17		0.672	0.644	0.701	
J48_BF_16		0.615	0.575	0.656	
J48_GS_15		0.625	0.613	0.637	
J48_IWSS_2		0.568	0.169	0.975	
RF_BF_5		0.637	0.669	0.605	
RF_GS_5		0.637	0.669	0.605	
RF_IWSS_7		0.628	0.613	0.643	
BN_BF_4		0.552	0.156	0.955	3Dt-MC
BN_GS_4		0.552	0.156	0.955	
BN_IWSS_3		0.539	0.369	0.713	
SMO_BF_10		0.694	0.681	0.707	
SMO_GS_10		0.694	0.681	0.707	
SMO_IWSS_15		0.650	0.588	0.713	
LOG_BF_10		0.669	0.719	0.618	
LOG_GS_5		0.662	0.706	0.618	
LOG_IWSS_16		0.666	0.638	0.694	
FLDA_BF_6		0.612	0.606	0.618	
FLDA_GS_6		0.618	0.619	0.618	
FLDA_IWSS_17		0.653	0.631	0.675	

IBK_BF_2		0.644	0.694	0.592	
IBK_GS_2		0.644	0.694	0.592	
IBK_IWSS_13		0.653	0.650	0.656	
J48_BF_18		0.681	0.706	0.656	
J48_GS_12		0.678	0.694	0.662	
J48_IWSS_16		0.678	0.531	0.828	
RF_BF_12		0.707	0.688	0.726	
RF_GS_1		0.631	0.638	0.624	
RF_IWSS_17		0.659	0.631	0.688	
BN_PSO_12	0.8	0.552	1.000	0.000	2Dt-MC
BN_GEN_6		0.552	1.000	0.000	
BN_IWSS_1		0.552	1.000	0.000	
SMO_BF_5		0.596	0.840	0.296	
SMO_GS_5		0.596	0.840	0.296	
SMO_IWSS_1		0.552	1.000	0.000	
LOG_BF_6		0.647	0.737	0.535	
LOG_GS_6		0.647	0.737	0.535	
LOG_IWSS_14		0.612	0.720	0.479	
FLDA_BF_11		0.644	0.646	0.641	
FLDA_GS_2		0.603	0.634	0.563	
FLDA_IWSS_20		0.558	0.549	0.570	
IBK_BF_8		0.672	0.726	0.606	
IBK_GS_2		0.659	0.863	0.409	
IBK_IWSS_10		0.647	0.737	0.535	
J48_BF_17		0.625	0.823	0.380	
J48_GS_8		0.640	0.903	0.317	
J48_IWSS_8		0.634	0.851	0.366	
RF_BF_6		0.666	0.749	0.563	
RF_GS_3		0.618	0.686	0.535	
RF_IWSS_10		0.647	0.726	0.549	
BN_IWSS_1		0.527	0.760	0.239	3Dt-MC
SMO_BF_9		0.697	0.766	0.613	
SMO_GS_9		0.697	0.766	0.613	
SMO_IWSS_18		0.625	0.726	0.500	
LOG_BF_11		0.653	0.794	0.479	
LOG_GS_5		0.625	0.806	0.401	
FLDA_BF_7		0.666	0.669	0.662	
FLDA_GS_7		0.666	0.669	0.662	
FLDA_IWSS_21		0.653	0.629	0.683	
IBK_BF_2		0.647	0.731	0.542	
IBK_GS_2		0.647	0.731	0.542	

IBK_IWSS_15		0.640	0.703	0.563	
J48_BF_23		0.647	0.749	0.521	
J48_GS_21		0.640	0.743	0.514	
J48_IWSS_17		0.546	0.594	0.486	
RF_BF_9		0.631	0.703	0.542	
RF_GS_2		0.621	0.674	0.556	
RF_IWSS_14		0.593	0.651	0.521	
BN_BF_1	0.9	0.609	0.995	0.008	2Dt-MC
BN_GS_1		0.609	0.995	0.008	
BN_PSO_6		0.609	0.995	0.008	
BN_GEN_6		0.609	0.995	0.008	
BN_IWSS_1		0.609	1.000	0.000	
SMO_IWSS_1		0.609	1.000	0.000	
LOG_BF_9		0.691	0.917	0.339	
LOG_GS_6		0.691	0.917	0.339	
LOG_IWSS_20		0.625	0.731	0.460	
FLDA_BF_5		0.631	0.643	0.613	
FLDA_GS_5		0.631	0.643	0.613	
FLDA_IWSS_12		0.662	0.637	0.702	
IBK_BF_3		0.707	0.725	0.677	
IBK_GS_3		0.707	0.725	0.677	
IBK_IWSS_11		0.669	0.865	0.363	
J48_BF_17		0.653	0.855	0.339	
J48_GS_15		0.647	0.860	0.315	
J48_IWSS_16		0.703	0.824	0.516	
RF_BF_8		0.722	0.855	0.516	
RF_GS_4		0.710	0.834	0.516	
BN_IWSS_1		0.609	1.000	0.000	3Dt-MC
SMO_IWSS_1		0.609	1.000	0.000	
LOG_BF_10		0.703	0.845	0.484	
LOG_GS_10		0.703	0.845	0.484	
LOG_IWSS_15		0.621	0.793	0.355	
FLDA_BF_7		0.659	0.658	0.661	
FLDA_IWSS_15		0.637	0.601	0.694	
IBK_BF_11		0.729	0.928	0.419	
IBK_GS_1		0.634	0.881	0.250	
IBK_IWSS_6		0.669	0.839	0.403	
J48_BF_21		0.615	0.824	0.290	
J48_GEN_139		0.539	0.565	0.500	
J48_IWSS_12		0.685	0.834	0.452	
RF_GS_6		0.716	0.845	0.516	

RF_IWSS_17		0.681	0.850	0.419	
BN_PSO_11	1	0.672	1.000	0.000	2Dt-MC
BN_GEN_6		0.672	1.000	0.000	
BN_IWSS_1		0.672	1.000	0.000	
SMO_IWSS_1		0.672	1.000	0.000	
LOG_BF_15		0.726	0.892	0.385	
LOG_GS_7		0.710	0.906	0.308	
LOG_IWSS_17		0.722	0.864	0.433	
FLDA_IWSS_10		0.662	0.676	0.635	
IBK_BF_10		0.726	0.944	0.279	
IBK_GS_5		0.719	0.920	0.308	
IBK_IWSS_8		0.729	0.892	0.394	
J48_BF_10		0.710	0.953	0.212	
J48_GS_10		0.710	0.953	0.212	
J48_PSO_27		0.656	0.967	0.019	
J48_IWSS_9		0.732	0.836	0.519	
RF_IWSS_15		0.691	0.878	0.308	
BN_IWSS_1		0.672	1.000	0.000	3Dt-MC
SMO_IWSS_1		0.672	1.000	0.000	
LOG_BF_12		0.726	0.948	0.269	
LOG_GS_6		0.713	0.958	0.212	
LOG_IWSS_13		0.707	0.892	0.327	
FLDA_IWSS_10		0.716	0.728	0.692	
IBK_BF_9		0.738	0.939	0.327	
IBK_GS_4		0.732	0.944	0.298	
IBK_IWSS_10		0.757	0.925	0.414	
J48_BF_13		0.681	0.977	0.077	
J48_GS_11		0.672	0.967	0.067	
J48_IWSS_10		0.675	0.897	0.221	
RF_IWSS_12		0.726	0.916	0.337	
BN_PSO_11	1.1	0.763	1.000	0.000	2Dt-MC
BN_GEN_6		0.763	1.000	0.000	
BN_IWSS_1		0.763	1.000	0.000	
SMO_IWSS_1		0.763	1.000	0.000	
LOG_BF_3		0.767	0.996	0.027	
LOG_GS_3		0.767	0.996	0.027	
LOG_PSO_6		0.760	0.996	0.000	
LOG_GEN_17		0.751	0.967	0.053	
LOG_IWSS_1		0.763	1.000	0.000	
FLDA_IWSS_20		0.653	0.694	0.520	
IBK_BF_8		0.804	0.988	0.213	

IBK_GS_1		0.767	0.971	0.107	
IBK_IWSS_8		0.754	0.963	0.080	
J48_BF_9		0.767	0.992	0.040	
J48_PSO_23		0.760	0.992	0.013	
J48_IWSS_1		0.763	1.000	0.000	
RF_BF_8		0.785	0.975	0.173	
RF_IWSS_13		0.751	0.950	0.107	
BN_IWSS_1		0.763	1.000	0.000	
SMO_IWSS_1		0.763	1.000	0.000	
LOG_BF_19		0.814	0.975	0.293	
LOG_GS_5		0.792	0.988	0.160	
LOG_IWSS_12		0.773	0.963	0.160	
FLDA_IWSS_13		0.650	0.657	0.627	
IBK_BF_1		0.785	0.967	0.200	3Dt-MC
IBK_GS_1		0.785	0.967	0.200	
IBK_IWSS_10		0.782	0.971	0.173	
J48_BF_14		0.763	0.992	0.027	
J48_GS_10		0.757	0.984	0.027	
J48_IWSS_1		0.763	1.000	0.000	
RF_IWSS_13		0.770	0.975	0.107	
BN_PSO_11		0.845	1.000	0.000	
BN_GEN_6		0.845	1.000	0.000	
BN_IWSS_1		0.845	1.000	0.000	
SMO_PSO_11		0.845	1.000	0.000	
SMO_GEN_6		0.845	1.000	0.000	
SMO_IWSS_1		0.845	1.000	0.000	
LOG_PSO_17		0.845	0.996	0.020	2Dt-MC
LOG_GEN_6		0.845	1.000	0.000	
LOG_IWSS_1		0.845	1.000	0.000	
FLDA_IWSS_16		0.691	0.709	0.592	
IBK_IWSS_10	1.2	0.861	0.985	0.184	
J48_PSO_29		0.845	1.000	0.000	
J48_IWSS_1		0.845	1.000	0.000	
RF_IWSS_7		0.836	0.978	0.061	
BN_IWSS_1		0.845	1.000	0.000	
SMO_IWSS_1		0.845	1.000	0.000	
LOG_BF_8		0.871	1.000	0.163	
LOG_GS_2		0.845	1.000	0.000	3Dt-MC
LOG_IWSS_1		0.845	1.000	0.000	
FDLA_IWSS_20		0.820	0.955	0.082	
IBK_BF_12		0.880	0.993	0.265	

IBK_GS_2		0.852	1.000	0.041	
IBK_IWSS_5		0.849	0.993	0.061	
J48_IWSS_1		0.845	1.000	0.000	
RF_IWSS_8		0.836	0.985	0.020	

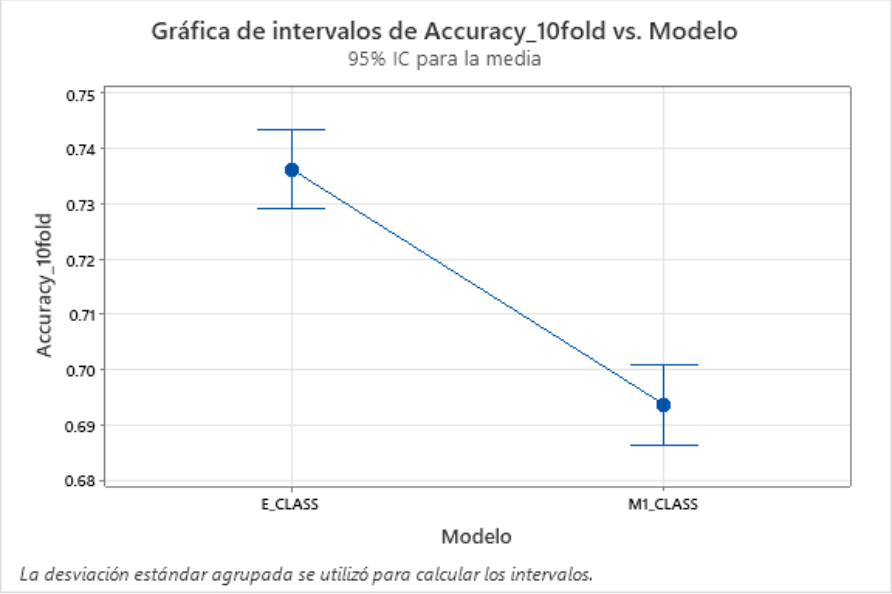
ANEXO C: PARÁMETROS ESTADÍSTICOS PARA LA VALIDACIÓN DE LOS MEJORES MODELOS INDIVIDUALES DE CLASIFICACIÓN CON LA PARTICIÓN ENTRENAMIENTO/PRUEBA.

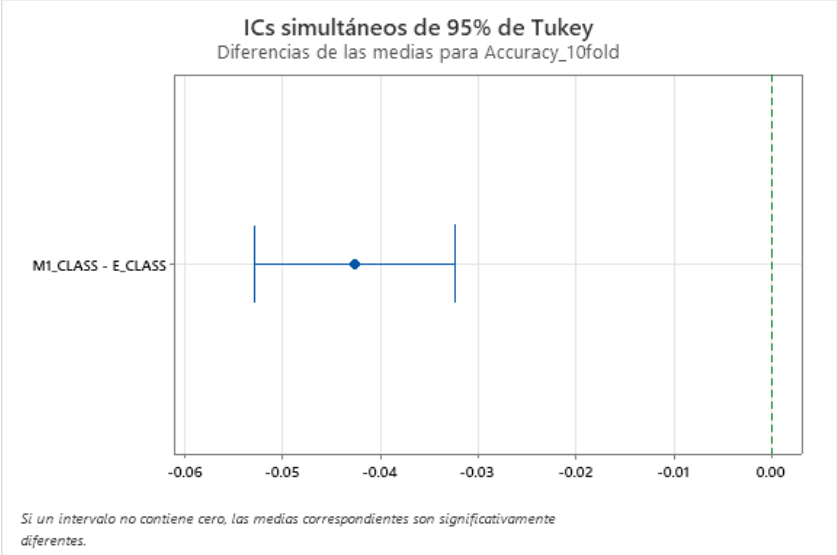
Parámetro	RF_BF_12	IBK_IWSS_10	J48_GS_12	RF_IWSS_12	J48_BF_18
$Acc_{10\text{-}fold}$	0.696	0.595	0.587	0.595	0.591
$Sens_{10\text{-}fold}$	0.689	0.756	0.513	0.605	0.487
$Spec_{10\text{-}fold}$	0.703	0.432	0.661	0.585	0.695
Puntaje $F_{10\text{-}fold}$	0.695	0.652	0.555	0.600	0.545
$ROC_{10\text{-}fold}$	0.765	0.599	0.590	0.633	0.590
$MCC_{10\text{-}fold}$	0.392	0.199	0.176	0.190	0.186
Acc_{Ext}	0.675	0.563	0.588	0.563	0.500
$Sens_{Ext}$	0.585	0.805	0.585	0.561	0.561
$Spec_{Ext}$	0.769	0.308	0.590	0.564	0.436
Puntaje F_{Ext}	0.649	0.653	0.593	0.568	0.535
ROC_{Ext}	0.690	0.559	0.612	0.558	0.461
MCC_{Ext}	0.360	0.130	0.175	0.125	-0.003
Acc_{LOO}	0.717	0.633	0.608	0.574	0.603
$Sens_{LOO}$	0.739	0.840	0.630	0.580	0.613
$Spec_{LOO}$	0.695	0.424	0.585	0.568	0.593
Puntaje F_{LOO}	0.724	0.697	0.617	0.577	0.608
ROC_{LOO}	0.779	0.621	0.631	0.625	0.604
MCC_{LOO}	0.435	0.291	0.215	0.148	0.207

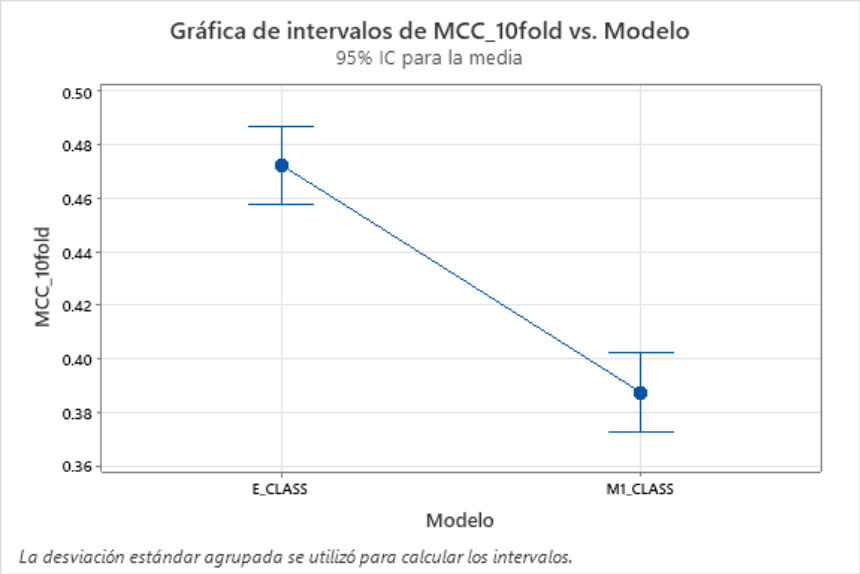
Parámetro	RF_IWSS_17	SMO_BF_10	J48_IWSS_16	IBK_IWSS_8
$Acc_{10\text{-}fold}$	0.650	0.679	0.608	0.624
$Sens_{10\text{-}fold}$	0.605	0.613	0.546	0.706
$Spec_{10\text{-}fold}$	0.695	0.746	0.669	0.542
Puntaje $F_{10\text{-}fold}$	0.634	0.658	0.583	0.654
$ROC_{10\text{-}fold}$	0.714	0.680	0.596	0.636
$MCC_{10\text{-}fold}$	0.301	0.362	0.217	0.252
Acc_{Ext}	0.575	0.650	0.575	0.563
$Sens_{Ext}$	0.415	0.659	0.634	0.537
$Spec_{Ext}$	0.744	0.641	0.513	0.590
Puntaje F_{Ext}	0.500	0.659	0.605	0.557
ROC_{Ext}	0.633	0.650	0.616	0.580
MCC_{Ext}	0.167	0.300	0.148	0.126
Acc_{LOO}	0.692	0.692	0.620	0.620
$Sens_{LOO}$	0.672	0.622	0.706	0.697
$Spec_{LOO}$	0.712	0.763	0.534	0.542
F_Measure_LOO	0.687	0.670	0.651	0.648
ROC_LOO	0.723	0.692	0.586	0.628
MCC_LOO	0.384	0.388	0.243	0.243

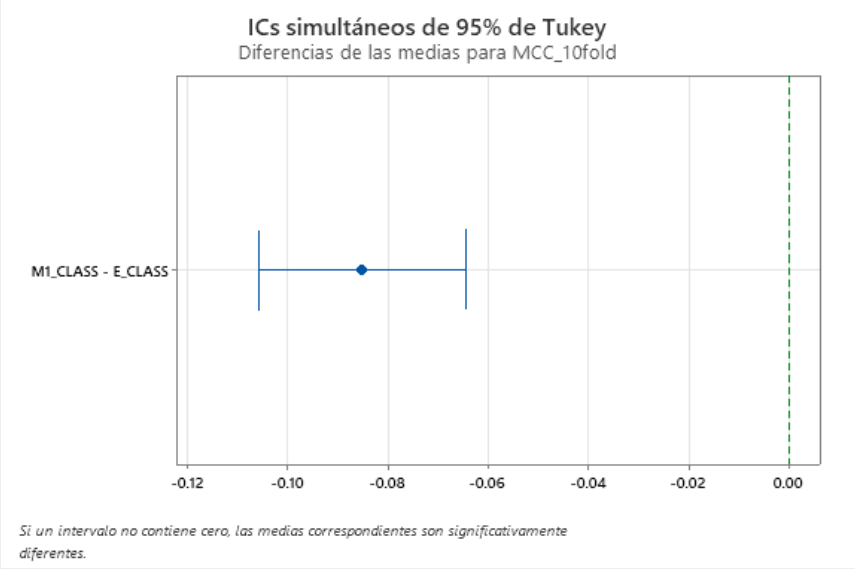
ANEXO D. ANÁLISIS DE ANOVA ENTRE EL MEJOR MODELO INDIVIDUAL DE CLASIFICACIÓN (M1_CLASS) Y EL MODELO TIPO ENSAMBLE (E_CLASS)

Nombre del modelo	$Acc_{10\text{-}fold}$	$MCC_{10\text{-}fold}$
E_CLASS	0.738	0.477
E_CLASS	0.743	0.485
E_CLASS	0.730	0.460
E_CLASS	0.738	0.477
E_CLASS	0.743	0.485
E_CLASS	0.730	0.460
E_CLASS	0.734	0.468
E_CLASS	0.738	0.477
E_CLASS	0.734	0.468
E_CLASS	0.734	0.468
M1_CLASS	0.696	0.392
M1_CLASS	0.675	0.350
M1_CLASS	0.709	0.418
M1_CLASS	0.679	0.359
M1_CLASS	0.705	0.409
M1_CLASS	0.675	0.350
M1_CLASS	0.700	0.401
M1_CLASS	0.696	0.393
M1_CLASS	0.684	0.367
M1_CLASS	0.717	0.435




ICs simultáneos de 95% de Tukey
Diferencias de las medias para Accuracy_10fold
M1_CLASS - E_CLASS
-0.06
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.


Gráfica de intervalos de MCC_10fold vs. Modelo
95% IC para la media
MCC_10fold
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0.36
E_CLASS
M1_CLASS
Modelo
La desviación estándar agrupada se utilizó para calcular los intervalos.


ICs simultáneos de 95% de Tukey
Diferencias de las medias para MCC_10fold
M1_CLASS - E_CLASS
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.

ANEXO E. MODELOS DE REGRESIÓN PARA LAS CLASES: MUY ACTIVA (A) Y ACTIVA (B), CONSTRUIDOS CON 3DT-MC, JUNTO CON SUS PARÁMETROS ESTADÍSTICOS ($Q^2_{10\text{-FOLD}}$ Y $MAE_{10\text{-FOLD}}$) PARA LA VALIDACIÓN CRUZADA DE 10 FOLDS SIN LA PARTICIÓN ENTRENAMIENTO/PRUEBA

Nombre del modelo	$Q^2_{10\text{-fold}}$	$MAE_{10\text{-fold}}$	Clase
GP_BF_101	0.932	0.112	Muy Activa (A)
GP_GS_131	0.931	0.114	
IBK_BF_10	0.439	0.224	
IBK_GS_8	0.389	0.234	
LR_BF_70	0.944	0.079	
LR_GS_71	0.944	0.079	
SMOR_BF_31	0.661	0.181	
SMOR_GS_31	0.661	0.181	
RF_BF_19	0.459	0.247	
RF_GS_13	0.450	0.244	
GP_GEN_612	0.350	0.254	
IBK_GEN_489	0.191	0.271	
LR_GEN_603	0.428	0.232	
SMOR_GEN_573	0.393	0.246	
RF_GEN_485	0.104	0.291	
Segunda búsqueda GP_BF_101			
GP_BF_88	0.938	0.110	
GP_GS_101	0.932	0.112	
GP_GEN_67	0.771	0.160	
IBK_BF_9	0.363	0.235	
IBK_GS_5	0.311	0.251	
IBK_GEN_47	0.317	0.254	
LR_BF_53	0.896	0.105	
LR_GS_60	0.898	0.106	
LR_GEN_61	0.765	0.152	
SMOR_BF_7	0.422	0.229	
SMOR_GS_7	0.422	0.229	
SMOR_GEN_58	0.714	0.176	
RF_BF_9	0.386	0.252	
RF_GS_6	0.369	0.249	
RF_GEN_54	0.193	0.285	
QSARINS_20	0.716	0.161	
Segunda búsqueda GP_GS_131			
GP_BF_84	0.913	0.114	

GP_GS_131	0.931	0.114	
GP_GEN_78	0.778	0.159	
IBK_BF_17	0.453	0.228	
IBK_GS_8	0.317	0.252	
IBK_GEN_74	0.350	0.254	
LR_BF_50	0.899	0.103	
LR_GS_51	0.887	0.112	
LR_GEN_73	0.695	0.183	
SMOR_BF_29	0.753	0.151	
SMOR_GS_29	0.753	0.151	
SMOR_GEN_75	0.692	0.184	
RF_BF_6	0.386	0.247	
RF_GS_5	0.371	0.253	
RF_GEN_76	0.192	0.282	
QSARINS_20	0.700	0.176	
Segunda búsqueda LR_BF_70			
GP_BF_49	0.854	0.137	
GP_GS_53	0.861	0.135	
GP_GEN_42	0.742	0.166	
IBK_BF_11	0.395	0.242	
IBK_GS_7	0.291	0.256	
IBK_GEN_37	0.302	0.256	
LR_BF_47	0.908	0.101	
LR_GS_48	0.908	0.101	
LR_GEN_46	0.823	0.14	
SMOR_BF_22	0.671	0.176	
SMOR_GS_20	0.654	0.177	
SMOR_GEN_34	0.727	0.164	
RF_BF_15	0.337	0.259	
RF_GS_6	0.332	0.264	
RF_GEN_41	0.227	0.277	
QSARINS_20	0.724	0.170	
Segunda búsqueda LR_GS_71			
GP_BF_49	0.854	0.137	
GP_GS_53	0.861	0.135	
GP_GEN_46	0.752	0.167	
IBK_BF_11	0.395	0.242	
IBK_GS_7	0.291	0.256	
IBK_GEN_35	0.431	0.228	
LR_BF_68	0.944	0.079	
LR_GS_71	0.944	0.079	
LR_GEN_44	0.798	0.149	

SMOR_BF_22	0.671	0.176	
SMOR_GS_20	0.654	0.177	
SMOR_GEN_49	0.744	0.166	
RF_BF_15	0.337	0.259	
RF_GS_6	0.332	0.264	
RF_GEN_28	0.268	0.267	
QSARINS_20	0.724	0.170	
GP_BF_112	0.930	0.046	Activa (B)
GP_GS_117	0.922	0.048	
IBK_BF_17	0.554	0.083	
IBK_GS_6	0.356	0.102	
LR_BF_44	0.840	0.054	
LR_GS_44	0.840	0.054	
SMOR_BF_30	0.663	0.072	
SMOR_GS_28	0.650	0.075	
RF_BF_7	0.260	0.103	
RF_GS_4	0.259	0.107	
GP_GEN_507	0.206	0.116	
IBK_GEN_412	0.158	0.108	
LR_GEN_419	0.318	0.111	
SMOR_GEN_546	0.223	0.119	
RF_GEN_327	0.020	0.118	
Segunda búsqueda GP_BF_112			
GP_BF_92	0.925	0.047	
GP_GS_100	0.919	0.048	
GP_GEN_72	0.756	0.069	
IBK_BF_10	0.291	0.108	
IBK_GS_2	0.104	0.118	
IBK_GEN_54	0.302	0.103	
LR_BF_28	0.735	0.070	
LR_GS_28	0.735	0.070	
LR_GEN_58	0.620	0.086	
SMOR_BF_23	0.535	0.086	
SMOR_GS_18	0.504	0.088	
SMOR_GEN_72	0.803	0.060	
RF_BF_11	0.335	0.102	
RF_GS_4	0.279	0.107	
RF_GEN_44	0.190	0.109	
QSARINS_20	0.674	0.079	
Segunda búsqueda GP_GS_117			
GP_BF_112	0.927	0.047	

GP_GS_117	0.922	0.048	
GP_GEN_77	0.775	0.066	
IBK_BF_8	0.295	0.104	
IBK_GEN_61	0.319	0.102	
LR_BF_33	0.749	0.069	
LR_GS_34	0.749	0.069	
LR_GEN_65	0.706	0.073	
SMOR_BF_21	0.561	0.086	
SMOR_GS_21	0.561	0.086	
SMOR_GEN_73	0.757	0.067	
RF_BF_16	0.374	0.101	
RF_GS_15	0.354	0.102	
RF_GEN_60	0.180	0.110	
QSARINS_20	0.674	0.079	
Segunda Búsqueda LR_BF_44			
GP_BF_44	0.805	0.065	
GP_GS_42	0.797	0.066	
GP_GEN_27	0.693	0.077	
IBK_BF_8	0.370	0.101	
IBK_GS_9	0.334	0.100	
IBK_GEN_23	0.399	0.099	
LR_BF_44	0.840	0.054	
LR_GS_44	0.840	0.054	
LR_GEN_29	0.707	0.074	
SMOR_BF_27	0.720	0.068	
SMOR_GS_27	0.695	0.070	
SMOR_GEN_32	0.708	0.072	
RF_BF_6	0.303	0.109	
RF_GS_6	0.303	0.109	
RF_GEN_24	0.232	0.109	
QSARINS_20	0.687	0.078	
Segunda búsqueda LR_GS_44			
GP_BF_44	0.805	0.065	
GP_GS_42	0.797	0.066	
GP_GEN_27	0.693	0.077	
IBK_BF_8	0.37	0.101	
IBK_GS_9	0.334	0.1	
IBK_GEN_23	0.399	0.099	
LR_BF_44	0.84	0.054	
LR_GS_44	0.84	0.054	
LR_GEN_29	0.707	0.074	
SMOR_BF_27	0.72	0.068	

SMOR_GS_27	0.695	0.07	
SMOR_GEN_32	0.708	0.072	
RF_BF_6	0.303	0.109	
RF_GS_6	0.303	0.109	
RF_GEN_24	0.232	0.109	
QSARINS_20	0.687	0.078	

ANEXO F. PARÁMETROS ESTADÍSTICOS PARA LA VALIDACIÓN DE LOS MEJORES MODELOS INDIVIDUALES CON LA PARTICIÓN ENTRENAMIENTO/PRUEBA PARA LA CLASE MUY ACTIVA (A)

Parámetro	LR_BF_70_SMOR_BF_22	GP_GS_131_QSARINS_20
R^2	0.780	0.808
MAE	0.154	0.144
RMSE	0.199	0.186
$Q^2_{10\text{-fold}}$	0.648	0.693
$MAE_{10\text{-fold}}$	0.197	0.184
$RMSE_{10\text{-fold}}$	0.254	0.237
Q^2_{LOO}	0.639	0.721
MAE_{LOO}	0.203	0.173
$RMSE_{LOO}$	0.258	0.224
Q^2_{Ext}	0.545	0.493
MAE_{Ext}	0.204	0.199
$RMSE_{Ext}$	0.267	0.262
$Q^2_{yScrambling}$	0.0251	0.0259

Parámetro	LR_GS_71_QSARINS_20	GP_GS_131_SMOR_BF_29
R^2	0.827	0.818
MAE	0.137	0.153
RMSE	0.176	0.199
$Q^2_{10\text{-fold}}$	0.719	0.666
$MAE_{10\text{-fold}}$	0.171	0.193
$RMSE_{10\text{-fold}}$	0.225	0.254
Q^2_{LOO}	0.724	0.699
MAE_{LOO}	0.172	0.186
$RMSE_{LOO}$	0.224	0.244
Q^2_{Ext}	0.638	0.547
MAE_{Ext}	0.186	0.187
$RMSE_{Ext}$	0.235	0.249
$Q^2_{yScrambling}$	0.0249	0.0237

Parámetro	GP_BF_101_QSARINS_20
R^2	0.815
MAE	0.127
RMSE	0.184
$Q^2_{10\text{-fold}}$	0.729
$MAE_{10\text{-fold}}$	0.174
$RMSE_{10\text{-fold}}$	0.221
Q^2_{LOO}	0.775
MAE_{LOO}	0.151
$RMSE_{LOO}$	0.202
Q^2_{Ext}	0.621
MAE_{Ext}	0.175
$RMSE_{Ext}$	0.230
$Q^2_{yScrambling}$	0.0254

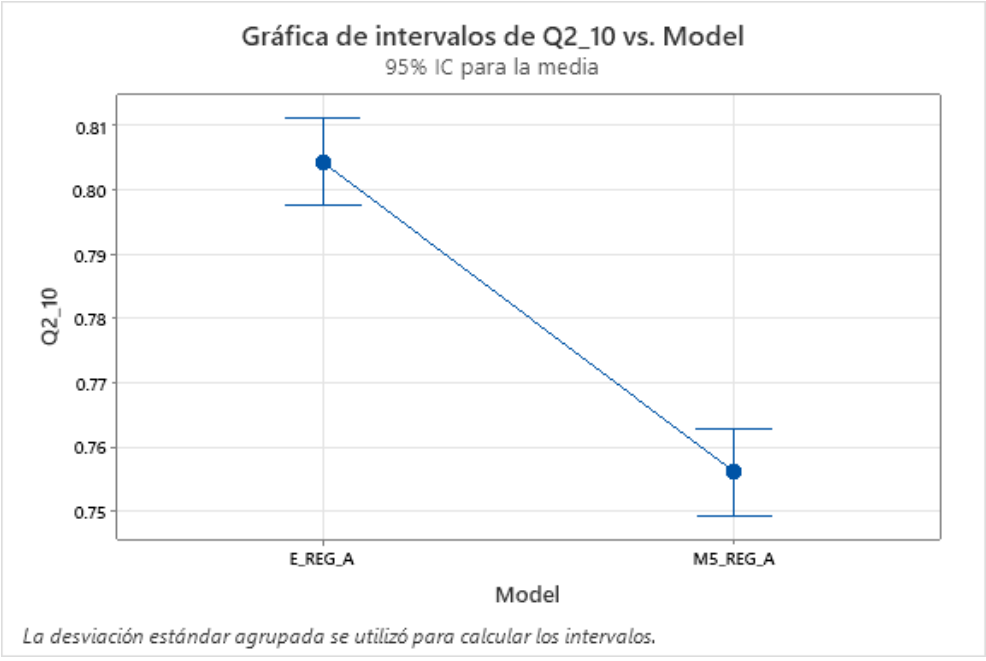
ANEXO G. PARÁMETROS ESTADÍSTICOS PARA LA VALIDACIÓN DE LOS MEJORES MODELOS INDIVIDUALES CON LA PARTICIÓN ENTRENAMIENTO/PRUEBA PARA LA CLASE ACTIVA (B)

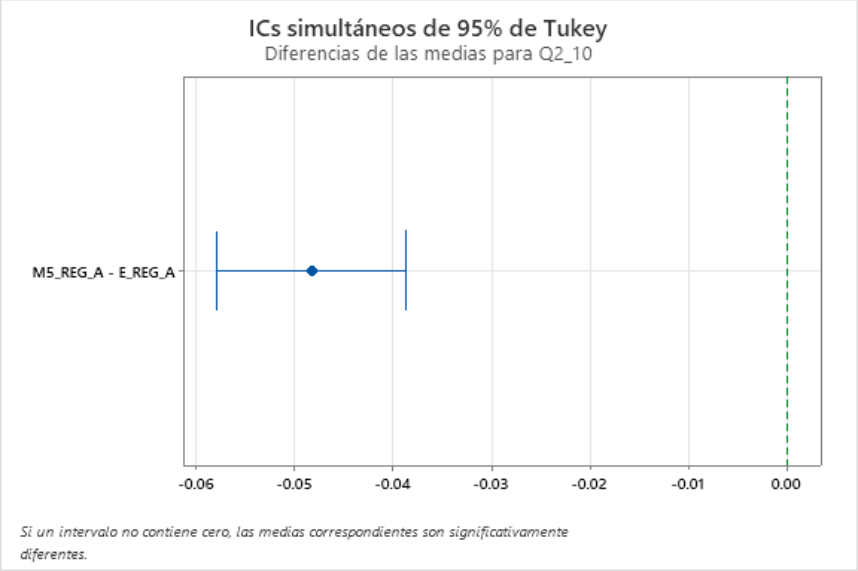
Parámetro	GP_BF_112_LR_BF_28	LR_BF_44_QSARINS_20
R^2	0.843	0.769
MAE	0.056	0.067
RMSE	0.069	0.084
$Q^2_{10\text{-fold}}$	0.708	0.655
$MAE_{10\text{-fold}}$	0.075	0.083
$RMSE_{10\text{-fold}}$	0.096	0.104
Q^2_{LOO}	0.725	0.658
MAE_{LOO}	0.075	0.082
$RMSE_{LOO}$	0.093	0.103
Q^2_{Ext}	0.681	0.726
MAE_{Ext}	0.075	0.066
$RMSE_{Ext}$	0.093	0.081
$Q^2_{yScrambling}$	0.026	0.026

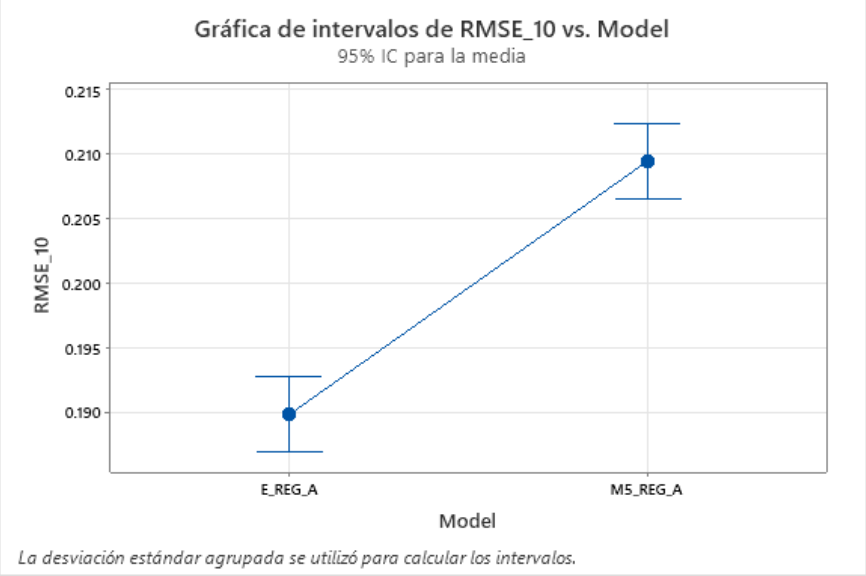
Parámetro	GP_BF_112_QSARINS_20	LR_BF_44_SMOR_BF_27
R^2	0.782	0.777
MAE	0.066	0.065
RMSE	0.082	0.083
$Q^2_{10\text{-fold}}$	0.691	0.606
$MAE_{10\text{-fold}}$	0.079	0.089
$RMSE_{10\text{-fold}}$	0.097	0.113
Q^2_{LOO}	0.678	0.608
MAE_{LOO}	0.081	0.086
$RMSE_{LOO}$	0.100	0.112
Q^2_{Ext}	0.665	0.827
MAE_{Ext}	0.077	0.050
$RMSE_{Ext}$	0.092	0.068
$Q^2_{yScrambling}$	0.027	0.027

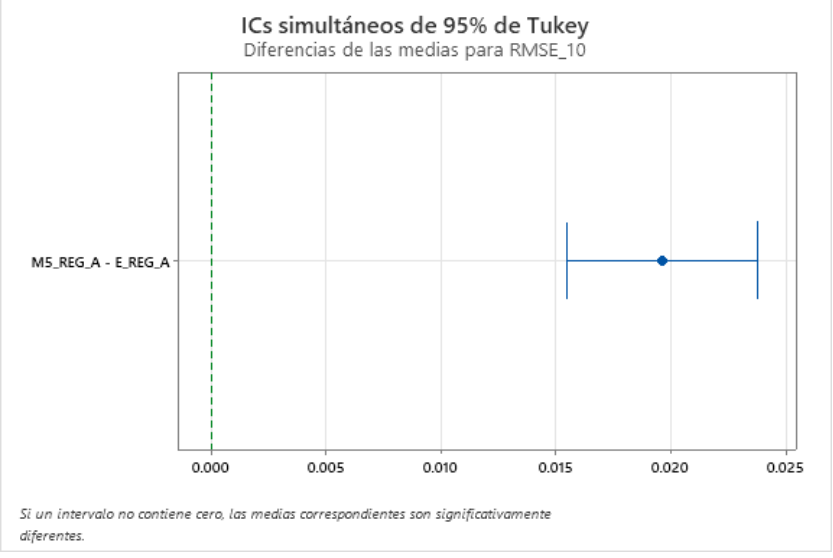
ANEXO H. ANÁLISIS DE ANOVA ENTRE EL MEJOR MODELO INDIVIDUAL DE REGRESIÓN (M5_REG_A) Y EL MODELO TIPO ENSAMBLE (E_REG_A) PARA LA CLASE MUY ACTIVA (A)

Nombre del modelo	$Q^2_{10\text{-fold}}$	$RMSE_{10\text{-fold}}$
E_REG_A	0.810	0.188
E_REG_A	0.801	0.192
E_REG_A	0.807	0.188
E_REG_A	0.803	0.190
E_REG_A	0.806	0.189
E_REG_A	0.800	0.191
E_REG_A	0.803	0.190
E_REG_A	0.800	0.191
E_REG_A	0.799	0.193
E_REG_A	0.814	0.186
M5_REG_A	0.729	0.221
M5_REG_A	0.764	0.206
M5_REG_A	0.770	0.203
M5_REG_A	0.769	0.204
M5_REG_A	0.744	0.215
M5_REG_A	0.748	0.213
M5_REG_A	0.771	0.203
M5_REG_A	0.753	0.211
M5_REG_A	0.761	0.207
M5_REG_A	0.752	0.211



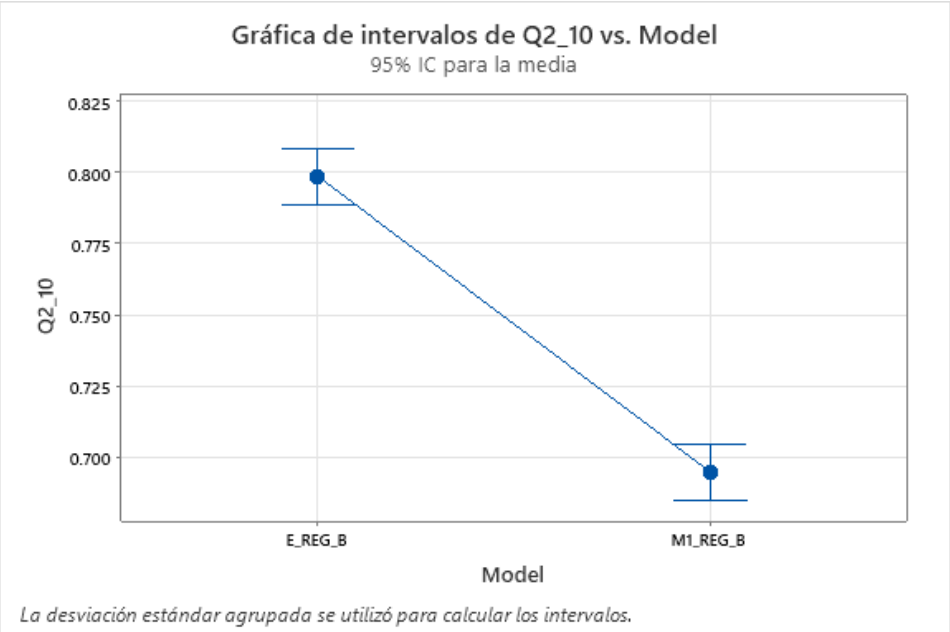

ICs simultáneos de 95% de Tukey
Diferencias de las medias para Q2_10
M5_REG_A - E_REG_A
-0.06
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.

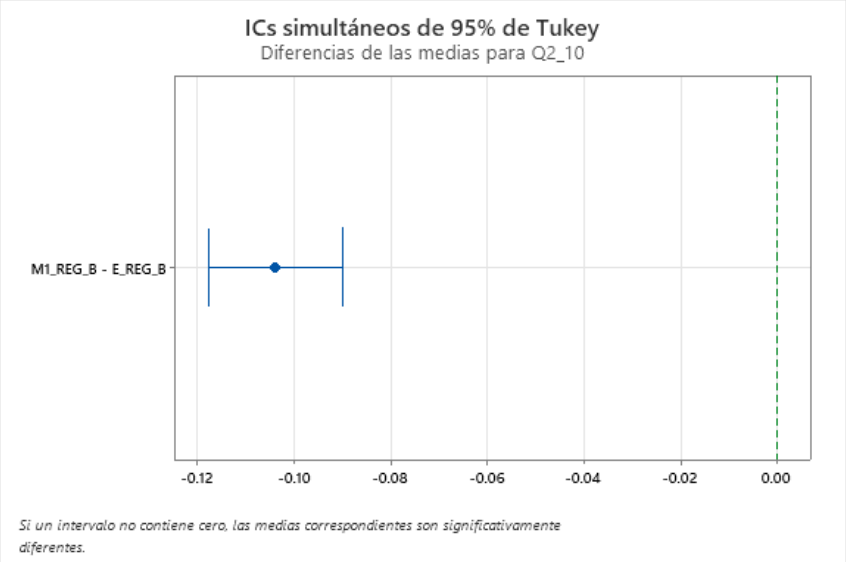

Gráfica de intervalos de RMSE_10 vs. Model
95% IC para la media
RMSE_10
0.215
0.210
0.205
0.200
0.195
0.190
E_REG_A
M5_REG_A
Model
La desviación estándar agrupada se utilizó para calcular los intervalos.

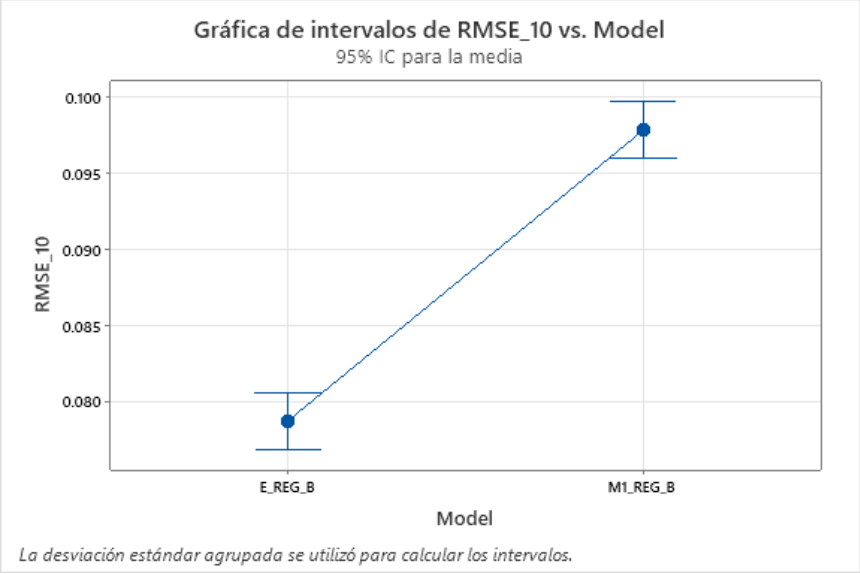

ICs simultáneos de 95% de Tukey
Diferencias de las medias para RMSE_10
M5_REG_A - E_REG_A
0.000
0.005
0.010
0.015
0.020
0.025
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.

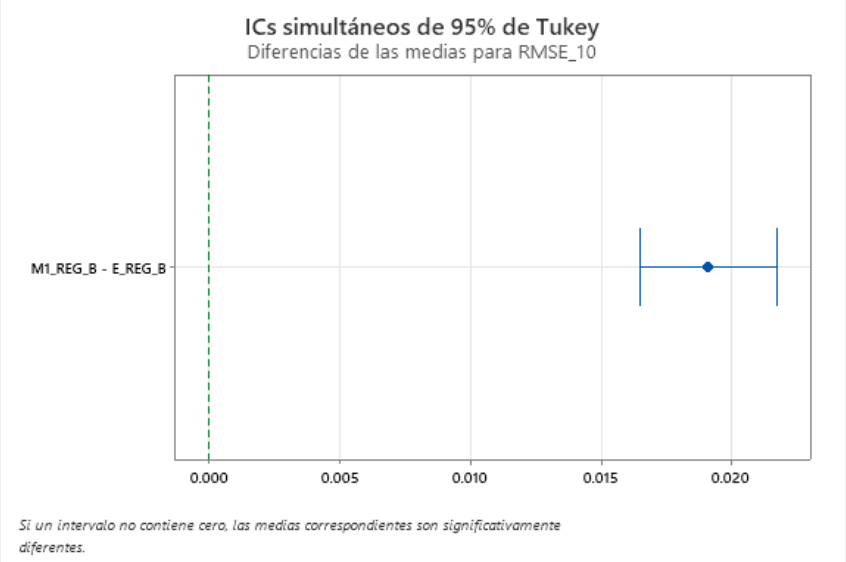
ANEXO I. ANÁLISIS DE ANOVA ENTRE EL MEJOR MODELO INDIVIDUAL DE REGRESIÓN (M1_REG_B) Y EL MODELO TIPO ENSAMBLE (E_REG_B) PARA LA CLASE ACTIVA (B)

Nombre del modelo	$Q^2_{10\text{-fold}}$	$RMSE_{10\text{-fold}}$
E_REG_A	0.810	0.188
E_REG_A	0.801	0.192
E_REG_A	0.807	0.188
E_REG_A	0.803	0.190
E_REG_A	0.806	0.189
E_REG_A	0.800	0.191
E_REG_A	0.803	0.190
E_REG_A	0.800	0.191
E_REG_A	0.799	0.193
E_REG_A	0.814	0.186
M5_REG_A	0.729	0.221
M5_REG_A	0.764	0.206
M5_REG_A	0.770	0.203
M5_REG_A	0.769	0.204
M5_REG_A	0.744	0.215
M5_REG_A	0.748	0.213
M5_REG_A	0.771	0.203
M5_REG_A	0.753	0.211
M5_REG_A	0.761	0.207
M5_REG_A	0.752	0.211




ICs simultáneos de 95% de Tukey
Diferencias de las medias para Q2_10
M1_REG_B - E_REG_B
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.


Gráfica de intervalos de RMSE_10 vs. Model
95% IC para la media
RMSE_10
0.100
0.095
0.090
0.085
0.080
E_REG_B
M1_REG_B
Model
La desviación estándar agrupada se utilizó para calcular los intervalos.


ICs simultáneos de 95% de Tukey
Diferencias de las medias para RMSE_10
M1_REG_B - E_REG_B
0.000
0.005
0.010
0.015
0.020
Si un intervalo no contiene cero, las medias correspondientes son significativamente diferentes.