UCLV
Universidad Central
"Marta Abreu" de Las Villas

FQF
Facultad de
Química y Farmacia

Departamento Lic. Química

TRABAJO DE DIPLOMA

Título: Análisis multivariado en la modelación antimalárica de sustancias orgánicas.

Autor: Osvaldo Delgado González

Tutores: Lic. Viviana Roche Llerena

Dr.C. Oscar Martínez Santiago

Consultante:

Dr.C. Juan Alberto Castillo Garit

Santa Clara, Junio 2019

UCLV
Universidad Central
"Marta Abreu" de Las Villas

FQF
Facultad de
Química y Farmacia

Departamento Lic. Química

DIPLOMA THESIS

Title: Multivariate analysis in the antimalarial modeling of organic substances.

Author: Osvaldo Delgado González

Thesis Director: Lic. Viviana Roche Llerena

Dr.C. Oscar Martínez Santiago

Consultant:

Dr.C. Juan Alberto Castillo Garit

Santa Clara, June 2019

Pensamiento:

Si Jehová no edificare la casa,

En vano trabajan los que la edifican:

Si Jehová no guardare la ciudad,

En vano vela la guardia.

Salmos 127:1

Dedicatoria:

A mi Dios de quién procede la sabiduría y la inteligencia.

A mi familia que me brindó su apoyo aún sin entenderme muchas veces; en especial a mi hermosa esposa y compañera fiel, a mi mamá y a mi papá quiénes me enseñaron a trabajar duro por mis sueños y a mi hermano por ser uno de los mejores regalos que Dios me ha dado.

Estoy agradecido:

A mi Dios y buen padre celestial y a Jesús su hijo en quien he depositado muchas veces mis cargas y ha sido mi socorro y salvación.

A mi esposa ***Beatriz Monteagudo Jiménez*** *por regalarme de su tiempo, paciencia, apoyo, ternura, nobleza, generosidad, compañía y amor sincero desde el primer año de la carrera.*

A mis padres ***Osvaldo Delgado Leal*** *y* ***Rosa María González Pérez*** *porque me han apoyado desmedidamente y se han preocupado por mi preparación, a ellos debo quien soy.*

A mi hermano ***Alberto Delgado González*** *por comprenderme, apoyarme y preocuparse por mí.*

A mis suegros ***Manuel Monteagudo Ulacia*** *y* ***Ana Ibis Jiménez Montes*** *por su cariño, generosidad, apoyo y preocupación.*

A mi tutora ***Viviana Roche Llerena*** *que siendo joven me ha guiado y ha acompañado en la realización de este trabajo.*

A ***Juan Alberto Castillo Garit*** *y* ***Naiví Flores Balmaseda*** *por su cariño, preocupación, desinterés, por impulsarme y ayudarme a hacer el estudio de clasificación.*

*A **Evys Ancede Gallardo** por introducirme en el lenguaje de programación Python, por ser un buen ejemplo de altruismo.*

*A **Aliuska Morales Helguera** por su comprensión, apoyo y aliento.*

*A los **amigos**, a los **compañeros de** aula de **Lic. Química** y de **Farmacia** y al **Grupo Cristiano Universitario** por acompañarme en la travesía de estos 5 años, porque junto a ustedes viví tiempos agradables; son 5 años que recordaré con ternura.*

*Al claustro de **Lic. Química** por sus enseñanzas, preparación, exigencias y por hacer germinar en mí el amor por la química.*

A todas las personas que han aportado su granito en mi formación y en la realización de este trabajo.

Resumen

La malaria es una enfermedad causante de más de medio millón de muertes anuales; los fármacos utilizados convencionalmente presentan fenómenos de resistencia. Surge así, la necesidad de desarrollar nuevos antimaláricos. Los estudios *in silico* han demostrado ser capaces de disminuir los altos costos de los procesos de síntesis y bioensayos, así como el tiempo empleado en la identificación de compuestos efectivos contra diversas dianas terapéuticas. En la presente investigación se realizan estudios de regresión y clasificación sobre un conjunto químico (Malaria Box); donde se emplean técnicas de Inteligencia Artificial y programas hechos en casa. Se modeló la actividad antimalárica (expresada como IC_{50}) de 317 entidades químicas, frente al parásito *Plasmodium falciparum*. Las estructuras químicas fueron codificadas mediante los descriptores moleculares 0D-2D implementados en el software Dragon7. Se empleó los programas hechos en casa con la metodología implementada en el software Weka para realizar la selección de atributos. Se obtuvieron 6 modelos de clasificación y 1 de regresión empleando Redes Neuronales Artificiales y Máquinas de Vectores de Soporte respectivamente. Los modelos de clasificación alcanzaron porcentajes predictivos entre 66 % - 81 %. El modelo de regresión es capaz de predecir con un coeficiente de determinación de 0.55, más del 50 % de la serie de prueba, después de extraídos compuestos *outliers*. Los modelos obtenidos permiten la

realización de posteriores estudios de cribado virtual para la identificación y propuesta de potenciales antimaláricos.

Abstract

Malaria is a disease that causes more than half a million deaths annually. Conventionally used drugs exhibit resistance phenomena. That suggests, the need to develop new antimalarials. *In silico* studies have shown to be able to reduce the high costs of the synthesis and bioassay processes, as well as the time spent in the identification of effective compounds against different therapeutic targets. In the present investigation, regression and classification studies are carried out on a chemical group (Malaria Box); where Artificial Intelligence techniques and homemade programs are used. The antimalarial activity (expressed as IC_{50}), of 317 chemical entities, was modeled against the *Plasmodium falciparum* parasite. The chemical structures were encoded using the 0D-2D molecular descriptors implemented in the Dragon7 software. Homemade programs were used with the methodology implemented in Weka software to perform the selection of attributes. Was obtained 6 classification models and 1 regression model using Artificial Neural Networks and Support Vectors Machines respectively. The classification models reached predictive percentages between 66 % - 81 %. The regression model is able to predict with a coefficient of determination of 0.55, more than 50 % of the test series, after extracted outliers. The models obtained allow the realization of subsequent virtual screening studies for the identification of new antimalarials.

Tabla de contenido

Glosario

CV	Crivado Virtual
QSAR	Quantitative Structure-Activity Relationship (Relaciones Cuantitativas de Estructura-Actividad)
AMs	antimaláricos
OMS	Organización Mundial de la salud
TCA	terapias combinadas con artemisina
DMs	descriptores moleculares
SVM	Support Vector Machine (Máquinas de Vectores de Soporte)
VC	Validación Cruzada
AC	Análisis de Conglomerados
MLP	MultilayerPerceptron

Introducción

Dentro de las enfermedades parasitarias, el paludismo o malaria es la más importante si se tienen en cuenta el número de individuos que enferman anualmente y su impacto socioeconómico. (Komba et al., 2009)

Los parásitos se transmiten generalmente por la picadura de la hembra del mosquito *Anopheles* y la forma más frecuente y grave de la enfermedad se debe a *Plasmodium falciparum*. El cuadro clínico clásico consiste en escalofrío, fiebre y sudoraciones repetidas. La infección por *Plasmodium falciparum* puede complicarse con un cuadro de malaria grave, caracterizada por acidosis metabólica, anemia severa, hipoglicemia, falla renal aguda, edema agudo del pulmón; en este estado, si se recibe tratamiento, la letalidad es de 15 – 20 % y si no se aplica es casi siempre fatal. (Colombiana, 2012-2013)

La diversidad genética le confiere a *Plasmodium* la capacidad para evadir la respuesta inmune del hospedador y producir variantes resistentes a medicamentos y vacunas, esto es en gran parte, responsable del éxito de la supervivencia de este parásito en la historia evolutiva, así como del fracaso de las medidas empleadas con el objetivo de erradicarlo. (Jimenez-Diaz et al., 2009)

El tratamiento ha sido posible durante muchos años gracias a la existencia de un número restringido de fármacos que presentan cada uno de ellos una

serie de limitaciones de tipo farmacológico, aunque el mayor problema es la aparición de resistencias. (Dockrell and Playfair, 1983)

Se ha promovido la implementación de combinaciones terapéuticas con derivados de artemisinina (la artemisina y sus derivados son una familia de fármacos que poseen la acción más rápida contra el paludismo provocado por *falciparum*), además de la asociación de sulfadoxina-pirimetamina (para tratar o prevenir la malaria), estas combinaciones presentan beneficios limitados y en determinados casos son de una eficacia cuestionable debido a los fenómenos de resistencia (Colombiana, 2012-2013) (Adhanom Ghebreyesus, 2018). Todo esto pone de manifiesto la imperiosa necesidad mundial de desarrollar nuevos antimaláricos (AMs). (Llerena, 2017)

En la actualidad se han realizado grandes esfuerzos para lograr más eficiencia en la selección de compuestos antimaláricos. Hay pocos ejemplos de la utilización previa de los métodos *in silico* con este propósito. Dentro de estas técnicas, el cribado virtual (CV) tiene la ventaja de ser más económica (ahorro en compra de reactivos y robotización), rápida, y permite tener en cuenta una cantidad de compuestos del orden de billones, cifra impensable experimentalmente. Por otra parte, los estudios de Relaciones Cuantitativas de Estructura-Actividad (por sus siglas en inglés, QSAR) se han utilizado ampliamente en la modelación de disímiles propiedades moleculares de naturaleza física, química y biológica, son actualmente, el enfoque más utilizado en el diseño de nuevos fármacos (Lazarou, 2009). Este tipo de

análisis es muy útil y generalmente se utiliza como principal herramienta en la selección de compuestos durante el protocolo de CV.

Se han desarrollado esfuerzos importantes para optimizar la búsqueda de nuevos antimaláricos mediante el uso de herramientas de quimioinformática; para describir mediante procedimientos matemáticos la relación que existe entre las estructuras y sus respectivas actividades y que las mismas sirvan para posteriores estudios en el desarrollo de nuevos antimaláricos. Este estudio se basa, fundamentalmente, en construir modelos de regresión y clasificación que permitan realizar, sobre los mismos, cribados virtuales y continuar con el estudio de nuevas entidades químicas que presenten mayor actividad antimalárica que los limitados fármacos que actualmente se utilizan en el tratamiento de esta enfermedad.

Problema científico:

¿Cómo determinar modelos de clasificación y de regresión que sean capaces de predecir satisfactoriamente la actividad antimalárica de sustancias orgánicas?

Hipótesis:

Mediante el uso de herramientas de inteligencia artificial es posible determinar modelos de clasificación y de regresión que predigan satisfactoriamente la actividad antimalárica de sustancias orgánicas.

Objetivo general:

Determinar modelos de clasificación y de regresión que predigan satisfactoriamente la actividad antimalárica de sustancias orgánicas contra el parásito *Plasmodium falciparum* y que puedan ser usados en posteriores estudios de cribado virtual y desarrollo de nuevas entidades químicas como candidatos prometedores para combatir la malaria.

Objetivos específicos:

- Identificar los descriptores moleculares del software DRAGON que describan adecuadamente la estructura de 317 entidades químicas mediante el uso de metodologías de selección de atributos implementadas en el software Weka.
- Determinar modelos de clasificación basados en técnicas de IBk, árboles y redes neuronales artificiales que relacionen las estructuras químicas con sus respectivas actividades biológicas.
- Determinar modelos de regresión no lineal mediante la aplicación de técnicas de SVM que predigan satisfactoriamente la actividad antimalárica de sustancias orgánicas contra el parásito *Plasmodium falciparum*.

CAPÍTULO I: REVISIÓN BIBLIOGRÁFICA

Capítulo I: Revisión Bibliográfica

1.1- Malaria

1.1.1- Historia de la enfermedad

Las fiebres palúdicas fueron descritas por Hipócrates 400 años antes de J.C. No solamente se diagnosticaba la enfermedad, sino que se realizaban pronósticos acerca de su evolución a pesar del desconocimiento de su etiología. Los términos empleados más comúnmente, malaria y paludismo, también conservaban algunas imprecisiones. La palabra malaria (mal aire) resultaba incorrecta al señalar como origen de la enfermedad, la transmisión por el aire; el término paludismo (de "*palus*", terreno pantanoso) podía hacer pensar que sólo se producía en aguas estancadas (Andriantsoanirina, 2009). La teoría bacteriana del paludismo apareció a raíz de las investigaciones de Klebs y Tommasi-Crudeli que describieron el "*Bacillus malariae*" en 1879 (Bell, 2009). El protozoo fue descrito por primera vez en la sangre de un paciente por Charles Laveran en 1880, que observó al parásito en un frotis sin teñir de sangre fresca (Machado-Tugores, 2012).

1.1.2- Definición

Malaria es una enfermedad causada por protozoarios del género *Plasmodium* (Figura 1):

- *Plasmodium falciparum*
- *Plasmodium vivax*

- *Plasmodium malariae*
- *Plasmodium ovale*
- *Plasmodium knowlesi* (en los últimos años en países del Asia)

Figura 1. Parásito *Plasmodium*.

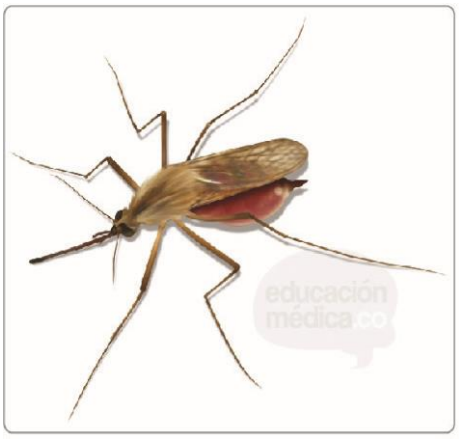


Figura 2. Mosquito hembra del género *Anopheles.*

Los parásitos del género *Plasmodium* son transmitidos al hombre generalmente por la picadura de la hembra del mosquito del género *Anopheles* (Figura 2).

Figura 3. Diferencias entre distintos tipos de *Plasmodium*.

1.1.3- Epidemiología

En el 2017, se estimaron 219 millones de casos de paludismo en todo el mundo.

La Región de África representa alrededor del 92 % (200 millones) de los casos de paludismo y muertes en todo el mundo. África subsahariana y la India soportaron casi el 80 % de la carga mundial de malaria.

El número estimado de muertes por malaria en 2017 fue prácticamente igual al del año anterior, 435 000.

Plasmodium falciparum es el parásito de la malaria más prevalente en la Región de África de la OMS, representa el 99,7 % de los casos estimados de malaria en 2017, así como en las Regiones de la OMS del Sudeste Asiático (62.8 %), Mediterráneo Oriental (69 %) y Pacífico Occidental (71,9 %). *Plasmodium vivax* es el parásito predominante en la Región de las Américas de la OMS; representa el 74,1 % de los casos de malaria (OMS, 2018).

1.1.4- Impacto económico de la Malaria

El paludismo produce pérdidas económicas importantes, que, a largo plazo, han llevado a diferencias considerables entre los valores del Producto Interno Bruto (PIB) de los países con y sin paludismo (sobre todo en África). Los costos sanitarios de esta enfermedad incluyen gastos tanto personales como públicos en prevención y tratamiento. En algunos países con gran carga de paludismo, la enfermedad es responsable de: (Machado-Tugores, 2012)

- Hasta un 40 % del gasto sanitario público.
- Un 30 % a 50 % de los ingresos en hospitales.
- Hasta un 60 % de las consultas ambulatorias.

En la actualidad el gasto en el tratamiento de la enfermedad es muy superior al de las pruebas de diagnóstico rápido (PDR), pero se espera que disminuya por la estrategia de ampliar la prueba parasitológica a todos los casos sospechosos de malaria antes de usar el tratamiento. Con los precios actuales de las PDR y las terapias combinadas con artemisina (TCA) y el estricto cumplimiento del protocolo de tratamiento, el ahorro en materias primas podría llegar a 68 millones de dólares en el sector público de la región africana (M., 2012). De igual manera el acceso universal a las redes mosquiteras tratadas con insecticida en África en 2015, podría reducir entre 31 y 48 millones el número de casos de paludismo que acuden a centros de salud pública. (M., 2012)

1.2.- Descriptores o Índices Moleculares

Los descriptores moleculares (DMs) juegan un papel fundamental en el desarrollo de la Química, las Ciencias Farmacéuticas, las investigaciones de nuevos materiales, *etc*. (Todeschini and Consonni, 2009)

Los Descriptores o Índices Moleculares (DMs o IMs) definidos como el resultado final de un procedimiento lógico-matemático el cual transforma la información química, codificada dentro de una representación simbólica de

una molécula, en un número (o conjunto de estos). Juegan un papel fundamental en los estudios *in silico* pues pueden brindar una visión más amplia en la interpretación de las propiedades moleculares y/o son capaces de tomar parte en un modelo para la predicción de notables propiedades moleculares. (Flores Balmaseda, 2015)

Los descriptores moleculares pueden ser agrupados en dos clases generales:

1. Los derivados de medidas experimentales como: logP, refractividad molar, momento dipolo, polarizabilidad y otras propiedades químico-físicas en general.
2. Descriptores Moleculares Teóricos, los cuales son derivados de representaciones simbólicas de las moléculas y estos a su vez pueden ser clasificados acorde con diferentes formas de representación molecular.

Los DMs pueden ser clasificados de acuerdo a la naturaleza en su definición y la complejidad de los rasgos moleculares estructurales que codifican, de forma general según las dimensiones que abarcan en:

- DMs-0D (Descriptores Constitucionales) se obtienen directamente de la fórmula molecular y son independientes de cualquier conocimiento sobre la estructura molecular, ej. número de átomos (A), el peso molecular (MW), conteo de átomos-tipo (Nx) o cualquier función de las propiedades atómicas.

- DMs-1D (Descriptores Unidimensionales), que están basados en la representación unidimensional de la molécula (representación que consiste en una lista de fragmentos estructurales de la molécula), aunque no requieren del conocimiento completo de la estructura molecular, por ejemplo, descriptores de búsqueda y análisis subestructural, como los Descriptores de Conteo de Fragmentos.
- DMs-2D (Descriptores Bidimensionales, Invariantes de Grafos), que son índices basados en la representación bidimensional o topológica de la molécula, o sea, que consideran la conectividad de los átomos (vértices) en la molécula (pseudografo) en términos de la presencia y naturaleza de los enlaces químicos (aristas), por ejemplo, Perfiles Moleculares, Índices de Información Topológica.

  Un grafo no es más que un conjunto de vértices interconectados por aristas en el cual cada vértice representa un objeto y la arista que conecta dos vértices representa la relación entre estos dos objetos (Gorbátov, 1988). En la química grafo-teórica los objetos del grafo pueden representar orbitales, átomos (o sus núcleos), enlaces, grupos de átomos, moléculas o colecciones de moléculas (Marrero-Ponce et al., 2012). De esta forma, los vértices del grafo podrían representar los átomos y las aristas las interacciones entre objetos químicos (ej. átomos), por lo cual estas últimas se usan para definir enlaces químicos, reacciones, mecanismos de reacciones, modelos cinéticos,

u otra relación o transformación de los objetos químicos. (Martínez Santiago et al.)

- DMs-3D (Descriptores Tridimensionales), que constituyen índices basados no solo en la naturaleza y conectividad de los átomos, sino también en la configuración espacial de la molécula, por ejemplo, Descriptores WHIM, MoRSE-3D, Índices de Similaridad, Descriptores de Superficie/Volumen, Químico-Quánticos.
- DMs-4D (Descriptores Tetradimensionales), que son descriptores basados no solo en la configuración espacial de la molécula, sino también en los campos escalares de interacción que se originan como consecuencia de la distribución electrónica en dicha entidad química, por ejemplo, Valores de la Energía de Interacción. (Flores Balmaseda, 2015)

1.3- Elementos de quimiometría

La quimiometría se define como la disciplina química que combina herramientas matemáticas y estadísticas con procedimientos para el análisis e interpretación de los datos químicos. (Brereton, 1990, Van de Waterbeemd, 1995)

1.3.1- Estudio QSAR

Desarrollar relaciones cuantitativas estructura-actividad es el paso final de un complejo proceso que comienza con una determinada descripción de la

estructura molecular y finaliza con algunas inferencias, hipótesis y predicciones del comportamiento (biológico, químico-físico, medioambiental, *etc*.) de las moléculas en un sistema analizado. Un estudio QSAR se basa en el supuesto de que en la estructura molecular (su conectividad, sus características geométricas, estéricas y sus propiedades electrónicas) están contenidas las características responsables de las propiedades físicas, químicas y biológicas que muestran las sustancias y que esta información puede ser capturada en uno o más DMs. (Todeschini and Consonni, 2009)

$$Actividad = f(propiedades\ fisicoqu\acute{i}micas\ y/o\ propiedades\ estructurales)$$

Figura 4. La forma matemática más general de QSAR.

El método QSAR asigna parámetros a cada grupo químico, de forma que, al modificar la estructura química puede valorarse la contribución de cada grupo funcional a la actividad del fármaco o del tóxico en cuestión y a partir de ahí, cómo variará la actividad de esa sustancia.(JT and k, 2006)

La mayoría de las estrategias QSAR son enfocadas hacia la construcción de modelos basados fundamentalmente en métodos de clasificación o regresión, aunque de manera general muchos métodos quimiométricos son usados, en dependencia del problema bajo estudio.

1.3.2 Clasificación

Estas técnicas se utilizan para la asignación de objetos a una de varias clases basado en una regla de clasificación. Son métodos de aprendizaje

supervisado, pues se “aprende” a partir de una serie de casos con variables predictivas y función “objetivo” o variable dependiente (esta serie de entrenamiento es el “maestro o supervisor”).

El objetivo de tales técnicas es calcular una regla de clasificación (y, posiblemente, límites de clases, o probabilidades de pertenencia a una clase), basados en los objetos de la serie de entrenamiento y aplicar esta regla para asignar una de estas clases, a objetos de clases previamente desconocidas.

Los Métodos de Clasificación son apropiados para modelar varias respuestas QSAR, como por ejemplo: compuestos activos/no-activos, compuestos de toxicidad baja/mediana/alta, compuestos mutagénicos/no-mutagénicos, etc. (Marrero-Ponce, 2011).

Los Métodos de Clasificación de origen estadístico más populares son: Análisis Discriminante Lineal o LDA, de sus siglas en inglés, Linear Discriminant Análisis; K-ésimos Vecinos más Cercanos o KNN, de sus siglas en inglés, Kth Nearest Neighbours; Métodos de Árboles de Clasificación o CTM, de sus siglas en inglés, Classification Tree Methods (también conocidos en la literatura como DT, acrónimo de Decision Trees); además de que se pueden encontrar métodos de clasificación o de análisis supervisado dentro de la Inteligencia Artificial, en particular las redes neuronales, con muchas posibilidades de aplicación en nuestro campo por su capacidad de tratar problemas con niveles mucho más libres de las variables predictivas y

cuya función objetivo representa combinaciones esencialmente no lineales de ellas difíciles de representar por una ecuación de regresión, por complicado que se plantee el modelo no lineal de la misma.

La calidad de los modelos de clasificación se evalúa por los parámetros de clasificación, para propósitos de ajuste y predicción. (González, 2013)

Redes neuronales

Estructura inspirada en un modelo simplificado de las neuronas biológicas. Se forma de un conjunto de elementos sencillos (neuronas) que tiene varias entradas y una salida. Estos elementos se interconectan entre sí para formar redes (red neuronal artificial (RNA)). Las RNA se entrenan para aprender relaciones de entrada-salida mediante la presentación de ejemplos, modificando los pesos.

Una vez entrenada, la red neuronal se puede utilizar para diversas tareas:

- ✓ Clasificación
- ✓ Clasificación no-supervisada
- ✓ Asociación
- ✓ Complementar patrones (26)

"Una red neuronal artificial es un conjunto de algoritmos matemáticos que procesan información y encuentran relaciones no lineales entre el conjunto de datos, y cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona" (González Salcedo et al., 2012). Las RNA tratan de emular ciertas características propias de los

cerebros humanos, como la capacidad de memorizar y de asociar hechos por tanto son modelos artificiales y simplificados del cerebro humano, un sistema que es capaz de adquirir conocimiento a través de la experiencia.

Perceptrón multicapa: posibilita la solución de problemas complejos. A diferencia de las anteriores arquitecturas de RNA, la estructura de red multicapa (*MultiLayerPerceptrón (*MLP)), posee al menos tres niveles de neuronas: una capa de entrada, una capa oculta y una capa de salida. (González Salcedo et al., 2012)

KNN

Significa algoritmo del vecino más cercano (*K-Nearest-Neighbor* por sus siglas en inglés). Es proporcionado por Weka como "IBK" (Delgado Castillo et al., 2016). Este método permite clasificar información y obtener predicciones. Básicamente se basa en tomar la información que deseamos clasificar, calcular las distancias de todos los miembros de la base de datos (comúnmente un conjunto de entrenamiento) y se toma un número particular de vecinos (k) que son los que tuvieron distancias menores, luego se visualiza la clase de los mismos, de allí se determina que tipo es el dato que estamos buscando.

Con este tipo de algoritmos es necesario conocer bien la información y como parte medular, decidir el tipo de cálculo para las distancias y el número de vecinos a considerar. Otro punto importante es el factor de búsqueda, también es común establecer un rango de búsqueda.

A continuación, se comentan sus características principales:

 No requiere la construcción de un modelo.

 Trabaja con los datos originales (no requiere representaciones especiales de los mismos).

 Realiza su predicción en base a información local (esto cuando se aplican heurísticas de búsqueda).

 Puede producir resultados equivocados si el cálculo de distancia no es el adecuado.

 En general, mientras K sea mayor (sin ser esto una regla), es decir, mientras más vecinos sean tomados en cuenta, la probabilidad de una predicción correcta aumenta. (Pitol, 2014)

Árboles de clasificación

Los métodos basados en árboles (o árboles de decisión) son bastante populares en minería de datos, pudiéndose usar para clasificación y regresión. Estos métodos se derivan de una metodología previa denominada detección de interacción automática (*automatic interaction detection*). Son útiles para la exploración inicial de datos y apropiados cuando hay un número elevado de datos, y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo. Sin embargo, no constituyen una herramienta demasiado precisa de análisis. En conjuntos pequeños de datos es poco probable que revelen la estructura de ellos, de modo que su mejor aplicación se encuentra en grandes masas de datos donde pueden

revelar formas complejas en la estructura que no se pueden detectar con los métodos convencionales de regresión.

Problemas donde los árboles de clasificación se pueden usar:

1. Regresión con una variable dependiente continua.

2. Regresión binaria.

3. Problemas de clasificación con categorías múltiples ordinales.

4. Problemas de clasificación con categorías múltiples nominales.

Ventajas de los árboles de clasificación

1. Los resultados son invariantes por una transformación monótona de las variables explicativas.

2. La metodología se adapta fácilmente en situaciones donde aparecen datos perdidos (*missing*), sin necesidad de eliminar la observación completa.

3. Están adaptados para recoger el comportamiento no aditivo, de manera que las interacciones se incluyen de manera automática.

4. Incluye modelos de regresión, así como modelos de clasificación generales que se pueden aplicar de manera inmediata para diagnosis. (Garcia and Lozano, 2007)

Dentro del aprendizaje automático los árboles de decisión han tenido gran difusión, en esto han influido varios factores: accesibilidad a diferentes implementaciones, la explicación que aporta a la clasificación, la posibilidad de ser representados gráficamente, y la rapidez de clasificar nuevos patrones. Están dentro de los métodos de clasificación supervisada, teniendo

una variable dependiente o clase. Su construcción se realiza mediante un proceso de inducción, se basan en una estructura en forma de árbol, donde las ramas representan conjuntos de decisiones, las cuales generan reglas para la clasificación de un conjunto de datos en subgrupos de datos. Las ramificaciones se generan de forma recursiva hasta que se cumplan ciertos criterios de parada. (Ramírez and Solano, 2009)

1.3.3 Regresión

Este término representa un conjunto de métodos que sirven para modelar y predecir la relación entre la variable respuesta de tipo continuo y una o más variables predictoras o dependientes. Además de la bien conocida regresión lineal de Mínimos Cuadrados Ordinarios, también son Importantes los modelos de regresión sesgada, regresión no-lineal y de regresión robusta.

Dentro de los métodos no lineales se encuentran algunas técnicas de la Inteligencia Artificial como las Redes Neuronales, con el método de propagación hacia atrás y las Máquinas con Soporte Vectorial.

Máquinas de Vectores Soporte

Las Máquinas de Vectores de Soporte (SVM) son una moderna y efectiva técnica de inteligencia artificial, que ha tenido un formidable desarrollo en los últimos años. Estas son sistemas de aprendizaje que usan un espacio de hipótesis de funciones lineales en un espacio de rasgos de mayor dimensión (Betancourt, 2005), entrenadas por un algoritmo proveniente de la teoría de

optimización (Abril, 2003). De forma general el algoritmo se enfoca en el problema de aprender a discriminar entre miembros positivos y negativos de vectores n-dimensionales. Mediante una función matemática denominada *kernel*, los datos originales se redimensionan para buscar una separabilidad lineal de los mismos. De manera general, las SVM permiten encontrar un hiperplano óptimo que separe las clases (Betancourt, 2005).

Funciones Kernel

Las funciones *kernel* son funciones matemáticas que se emplean en las SVM (Cortes et al., 1995). Estas funciones son las que le permiten convertir lo que sería un problema de regresión no lineal en el espacio dimensional original, a un problema más sencillo de regresión lineal en un espacio dimensional mayor. El tipo de *kernel* determina la transformación o mapeo que se le realizará a los datos. Entre los *kernels* más empleados por su implementación en diversos programas de modelación como Weka (García Morate, 2008), se encuentran:

El *kernel* Polinómico:

$$K(x \cdot z) = (x \cdot z + 1)^p \quad (1.1)$$

El *kernel* Gaussiano:

$$K(x \cdot z) = e^{\left(\frac{-\|x-z\|}{2\sigma^2}\right)} \quad (1.2)$$

Y el *kernel* Universal de Pearson:

$$K(x_i, x_j) = \frac{1}{\left[1 + \left(\frac{2\sqrt{2^{\frac{1}{\omega}} - 1}\sqrt{\|x_i - x_j\|^2}}{\sigma^2}\right)\right]^{\omega}} \quad (1.3)$$

Donde ω y σ controlan la altura y amplitud del pico de la función.

1.3.4 Validación interna y externa de modelos

Una condición necesaria para que sea válido un modelo de clasificación o regresión es que el coeficiente de determinación (R^2) esté cercano, tanto como sea posible, a uno y que el error estándar estimado (s) sea pequeño (capacidad de ajuste a los datos); sin embargo la consideración de estos parámetros estadísticos no es suficiente, pues los valores de los mismos no necesariamente están relacionados con la capacidad del modelo de realizar buenas predicciones de una data futura (Todeschini, 2009). Las técnicas de validación constituyen herramientas fundamentales a la hora de evaluar la capacidad predictiva de los modelos obtenidos por métodos multivariados de regresión y clasificación (Diaconis and Efron, 1983) (Cramer et al., 1988) (Golbraikh and Tropsha, 2002). Si se utiliza como método de evaluación de la selección supervisada de Weka el **Use training set,** se entrenará el modelo con todos los datos disponibles y luego lo aplicará otra vez sobre los mismos. Debido a esto, para comprobar la capacidad predictiva del mismo es necesario hacer validaciones tanto internas como externas.

Validación interna

La validación cruzada (VC) opera mediante la realización de un número de reducidas modificaciones al conjunto de compuestos del conjunto de datos

original y calcula la precisión de las predicciones de cada uno de los resultados de los modelos. El método de evaluación de la selección supervisada **Cross-validation** (validación cruzada) es la opción más elaborada y costosa. Se realizan tantas evaluaciones como se indica en el parámetro *Folds* (carpetas). Se dividen las instancias en tantas carpetas como indica este parámetro y en cada evaluación se toman las instancias de cada carpeta como datos de prueba, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados son el promedio de todas las ejecuciones.

Validación externa

La validación externa permite evaluar si los modelos obtenidos son generalizables a nuevos compuestos químicos y el "verdadero" poder predictivo de los mismos (Tropsha et al., 2003). Para esto se divide el conjunto de datos en 2 subconjuntos: la serie de entrenamiento (sirve para construir el modelo) y la serie de predicción (no utilizada en la selección de variables ni en el desarrollo del modelo, pero usada exclusivamente para evaluar el modelo tras su formación). El método de evaluación de la selección supervisada **Supplied test set** del Weka**,** le indica la ruta del fichero arff de prueba sobre el cuál va a predecir según los datos aprendidos en el entrenamiento.

1.3.5 Compuestos outliers y técnicas de selección

Los *outliers* son compuestos que se desvían significativamente del comportamiento típico del modelo desarrollado (no se ajustan al modelo) o son pobremente predichos por estos, que afectan los parámetros estadísticos del mismo (Gonzalez Diaz, 2002). Generalmente, la identificación de *outliers* busca un mejoramiento cualitativo del modelo. Existen varias técnicas para detectar la presencia de *outliers*, tales como: los análisis de los residuales estandarizados, los residuales studentizados, el método de Leverage, la estadística DFITS, la distancia de Cook y el método de dejar “varios” fuera. (Pyka and Planar., 1993) (Fernández-Oliva et al., 2019)

CAPÍTULO II: MATERIALES Y MÉTODOS

Capítulo II: Materiales Y Métodos

2.1- Conjunto de datos

La base de datos que se utiliza en este trabajo ha sido nombrada *Malaria Box*, (Spangenberg et al., 2013), ésta fue creada a partir de una recopilación de la información contenida en tres bases de datos diferentes (*St Jude´s, Novartis y GSK*), las cuales fueron sometidas a varios análisis. Los compuestos en cada base de datos fueron procesados para:

1. eliminar sales.
2. eliminar pequeños fragmentos.
3. desprotonar bases/ protonar ácidos.
4. generar tautómeros canónicos. (generar isómeros que difieren en la posición del grupo funcional, ocurre la migración de un grupo o átomo)
5. eliminar duplicados.

Dadas las limitaciones en el número de compuestos que pueden ser puestos a prueba en detalle, para maximizar el potencial impacto de *Malaria Box*, fue importante para los creadores del conjunto químico, maximizar la diversidad estructural en los compuestos seleccionados, *Malaria Box* cuenta con 400 compuestos, que cubren la diversidad química del conjunto comercial y de estos, solo 317 compuestos tienen referida la actividad biológica. La actividad antimalárica fue determinada sobre la cepa 3D7 de *Plasmodium falciparum*, el estudio fue llevado a cabo por el mismo experimentador, utilizando el

mismo protocolo para las 400 moléculas que incluye la base de datos, la misma es reportada en IC_{50} (la concentración que aniquila el 50 % de la población del parásito) expresada en µM (10^{-6} mol/L), en un rango entre 0 - 4 µM, en el desarrollo de este trabajo son evaluados los 317 compuestos que tienen referida la actividad. (Spangenberg et al., 2013)

2.2- Trabajo con Malaria Box

2.2.1 Preparación de la base de datos

Para el diseño de las series de entrenamiento, predicción o externa, tanto para el estudio de clasificación como para el de regresión se utiliza el software STATISTICA, el mismo, es un paquete estadístico usado en investigación, minería de datos y en el ámbito empresarial. Dentro sus opciones cuentan el análisis de conglomerados, el cual se utilizará para el desarrollo de este las series anteriormente mencionadas.

Análisis de conglomerados (AC): es un grupo de métodos para el reconocimiento de similitudes entre los casos (objetos) o entre las variables y destaca algunas categorías como un conjunto de casos similares (o variables). Estos métodos representan un caso especial del análisis exploratorio de datos, comprendidos de una serie de diferentes "algoritmos de clasificación" y permiten organizar los datos en subsistemas (conglomerados). Los métodos más populares son los Métodos Aglomerantes Jerárquicos (de Unión Promedio, de Unión Completa, de Unión

Simple, de Unión Promedio Ponderado, etc.), y otros métodos populares son los Métodos no-Jerárquicos, tales como el Método de las K-Medias y el Método de Jarvis-Patrick. En los primeros, cada objeto comienza dentro de su propio conglomerado y en los pasos posteriores, se unen a los pares más cercanos hasta que se obtiene uno solo que agrupe todas las observaciones. A diferencia de estos, los Métodos no-Jerárquicos carecen de la construcción de árboles. En su lugar, asignan los objetos a conglomerados una vez que el número de estos este especificado. (Tugores, 2014)

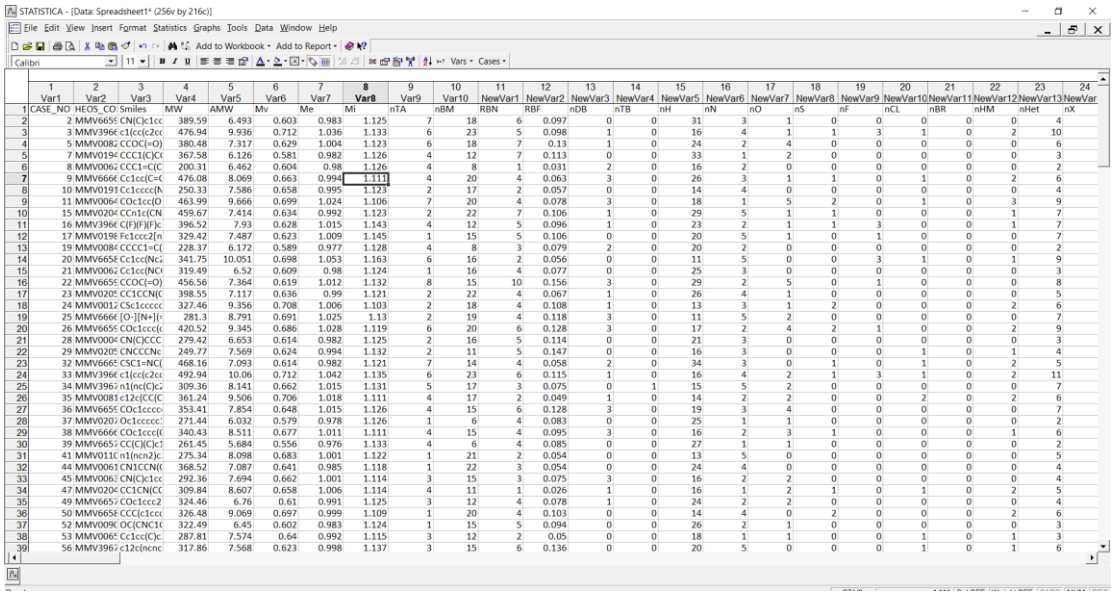


Figura 5. Ventana del software STATISTICA 8.0.

2.2.2- Cálculo de los descriptores moleculares

Para calcular los descriptores moleculares fue utilizado el software profesional DRAGON 7. En él se pueden encontrar más de 5000 descriptores moleculares en familias 0D, 1D, 2D, así como 3D. La naturaleza y magnitud

de los descriptores incluidos en este software computacional lo convierten en una poderosa herramienta de la química computacional para propiciar la generación de sistemas con una amplia aplicación en la investigación experimental, tanto para la interpretación de los resultados obtenidos y la planificación de futuros, así como para deducir información no asequible experimentalmente. Es ampliamente utilizado a nivel mundial como parte de numerosos estudios de modelación QSAR y ha demostrado resultados satisfactorios (Flores Balmaseda, 2015). Para la realización de este estudio se calcularon los descriptores desde 0D hasta 2D.

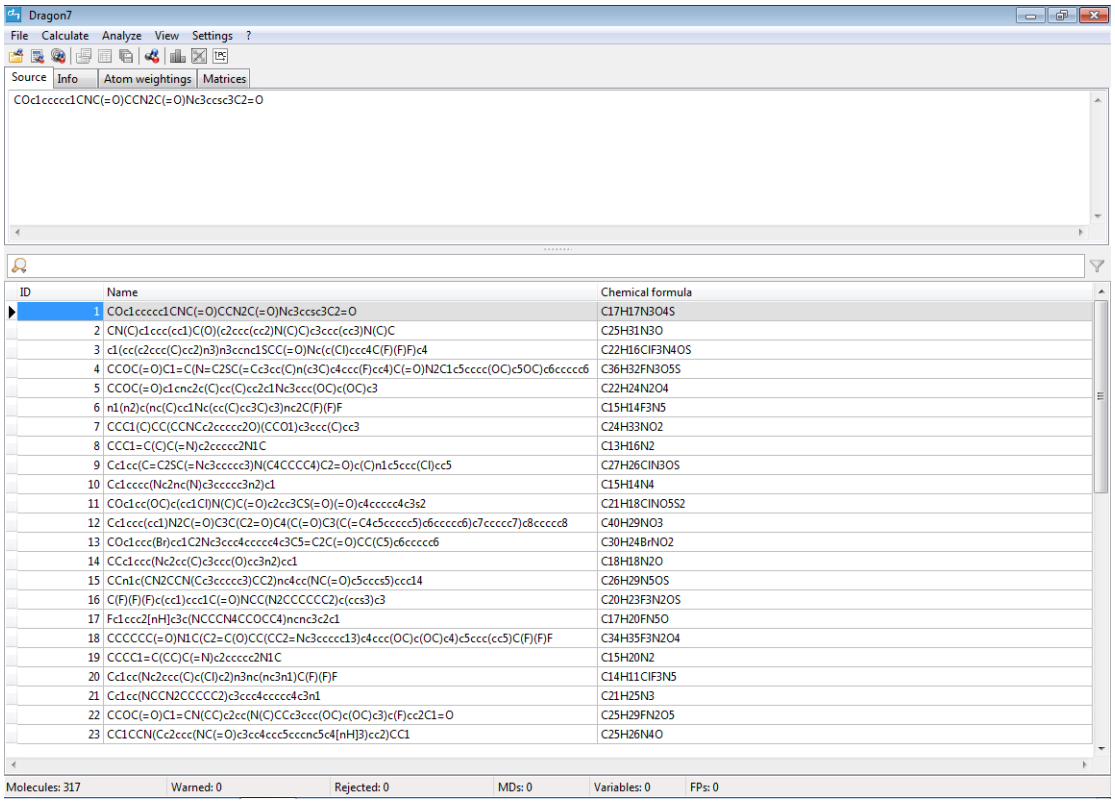


Figura 6. Interfaz gráfica del programa DRAGON 7.0.

2.2.3- Selección de atributos, estudios de regresión y clasificación

El software llamado Weka es un conocido programa para aprendizaje automatizado (Morate, 2008) y minería de datos escrito en Java (Munteanu, 2013) y desarrollado en la Universidad de Waikato.

Weka es una colección de algoritmos de aprendizaje automático para la selección de datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamar desde su propio código Java.

Selección del número óptimo de predictores. Principio de la Parsimonia

La exactitud de un modelo aumenta en la medida en que se añaden variables a la ecuación; pero a partir de cierto punto, el incremento de esta para cada nueva variable que se añade, es insignificante. Un buen modelo no debe presentar ni demasiadas variables, ni debe olvidar las que sean verdaderamente relevantes. Es decir, debe cumplir el principio de la *parsimonia*, según el cual un fenómeno debe ser descrito con el número mínimo de elementos posibles.

Diversos procedimientos se han propuesto para seleccionar el número óptimo de variables a incluir en la ecuación, para este estudio, dentro de los métodos de selección de atributos incorporado al software Weka 3.6.2 se utilizaron como evaluadores de atributos en este trabajo:

- WrapperSubsetEval (Esta técnica evalúa conjuntos de atributos mediante el uso de una serie de entrenamiento. La validación cruzada

se usa para estimar la exactitud de la serie de entrenamiento para un conjunto de atributos determinado. Es el clasificador más utilizado para estimar la precisión de subconjuntos). (Kohavi Ron, 1997)

- ClassifierSubsetEval (Evalúa los subconjuntos de atributos sobre los datos de entrenamiento o un conjunto de pruebas por separado)

Y como métodos de búsqueda:

- BestFirst, busca en el espacio de los subconjuntos de atributos mediante una escalada codiciosa aumentada con una función de retroceso.
- GeneticSearch, son algoritmos de búsqueda basados en los mecanismos de selección natural y genética natural. Combinan la supervivencia de los más compatibles entre las estructuras de cadenas, con la información ya aleatorizada e intercambiada para construir un algoritmo de búsqueda con algunas de las capacidades de innovación de la búsqueda humana.
- LinearForwardSelection, es una extensión del BestFirst.
- RankSearch, utiliza un evaluador atributo / subconjunto para clasificar todos los atributos. Si se especifica un subconjunto evaluador, a continuación, se utiliza una búsqueda de la selección hacia adelante para generar una lista clasificada. Desde la lista clasificada de atributos, se evalúan los subconjuntos de tamaño creciente. Se informa el mejor conjunto de atributos. RankSearch es lineal en el

número de atributos si se utiliza un sencillo evaluador atributo como GainRatioAttributeEval.

- SubsetSizeForwardSelection, es una extensión de LinearForwardSelection; la búsqueda realiza una validación cruzada interior (semilla y número de pliegues se pueden especificar) y se realizan algoritmos LinearForwardSelection sobre cada pliegue hasta encontrar el tamaño óptimo de subconjunto de todos los datos.
- GreedyStepwise realiza una búsqueda codiciosa hacia adelante o hacia atrás a través del espacio de los subconjuntos de atributos. Puede comenzar sin atributos / con todos los atributos o desde un punto arbitrario en el espacio. Se detiene cuando la adición / eliminación de cualquier atributo restante produce una disminución en la evaluación. También puede producir una lista clasificada de atributos si se recorre el espacio de un lado a otro y se registra el orden en que se seleccionan los atributos.

Modelos de clasificación y regresión

Weka contiene herramientas para los datos de pre-procesamiento, *clustering* (Morate, 2008), reglas de asociación, y la visualización (Munteanu, 2013), así como clasificación y regresión con el uso de las Máquinas de Soporte Vectorial, redes neuronales, entre otros.

El paquete Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz

gráfica de usuario para acceder fácilmente a sus funcionalidades. Para determinar los modelos de clasificación se utilizan en este trabajo, los clasificadores J48, IBk y Multilayer Perceptron, mientras que para construir los modelos de regresión se utilizan las Máquinas de Vectores de Soporte.

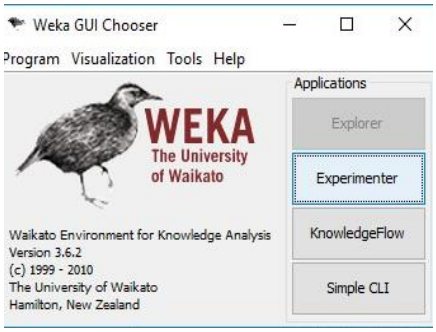


Figura 7. Ventana de entrada a la interfaz gráfica (Weka GUI Chooser).

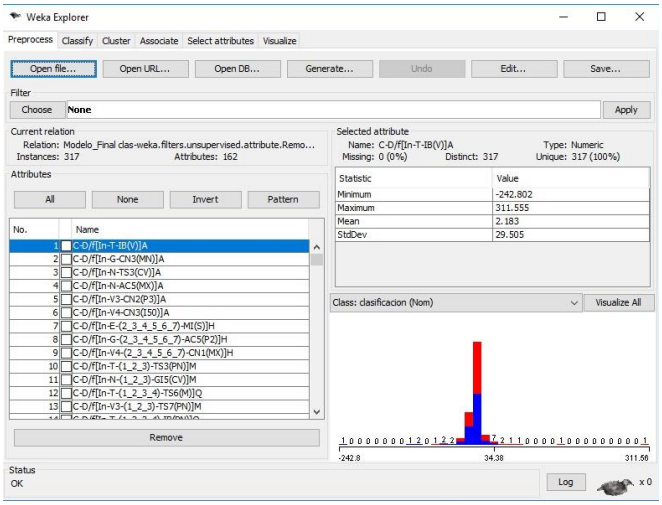


Figura 8. Ventana de la opción de trabajo Explorer.

2.2.4 Automatización del proceso de obtención de modelos

La automatización del proceso de obtención de modelos se hizo a través de softwares hechos en casa, los cuales están implementados en el lenguaje de programación Python versión 3.6.8.

Python es un lenguaje de programación creado en 1991 por Guido Van Rossum; es interpretado, multiparadigma, usa tipado dinámico, es multiplataforma y posee licencia Python Software Foundation License (licencia de software libre permisiva, Open Source, copycenter).(Downey et al., 2016)

After_Select_Attributes

Este programa es un binario (ejecutable (.exe)) creado con la librería de Python llamada pyinstaller. Toma como entrada la carpeta contenedora de los ficheros de salida buffers de la selección de atributos y devuelve un fichero con extensión .csv, que muestra:

1. La ruta de ejecución del programa.
2. La cantidad de veces que se repite el atributo.
3. La cantidad de atributos que se repiten.
4. Los atributos numerados.
5. Los atributos nombrados.
6. La cantidad de ficheros que se tuvieron en cuenta para la construcción del fichero con extensión .csv.
7. Los ficheros que se tuvieron en cuenta.

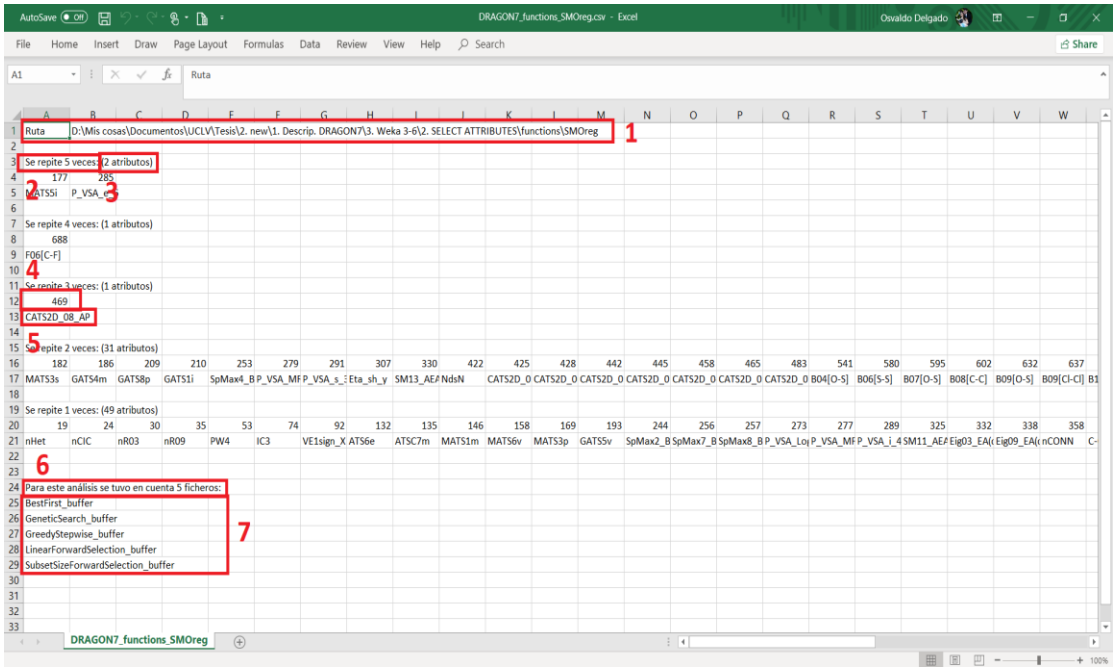


Figura 9. Fichero con extensión .csv construido por After_Select_Atributes.exe.

Figura 10. Icono de After_Select_Atributes.

Pasos del 1 al 4.

Paso 1:

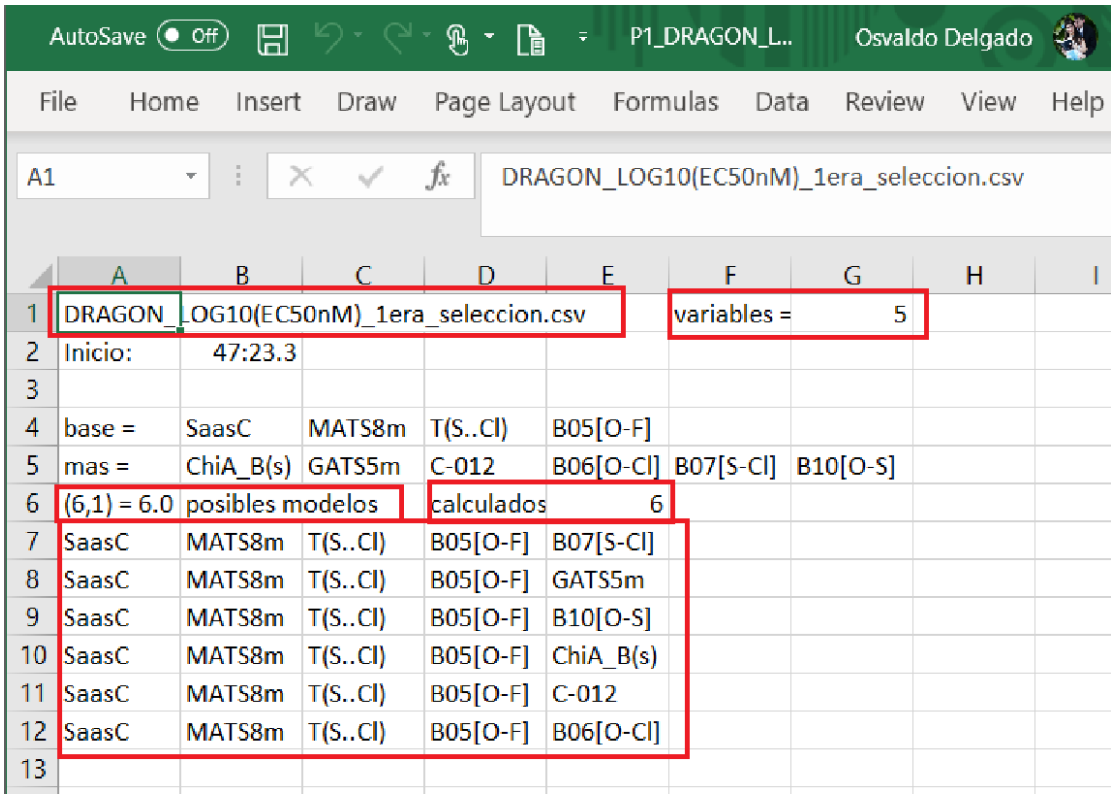
Toma como entrada ficheros con extensión .csv generados por After_Select_Atributes. Para luego construir modelos según la entrada del usuario.

Entrada:

1. Ruta.
2. La cantidad de atributos que desea que sus modelos tengan (se puede insertar rangos).
3. La cantidad de modelos que desea obtener (si le inserta el 0 le devuelve todas las posibles combinaciones).

Salida:

Crea una carpeta llamada “Modelos” que contiene ficheros con extensión .csv uno por cada cantidad de atributos.

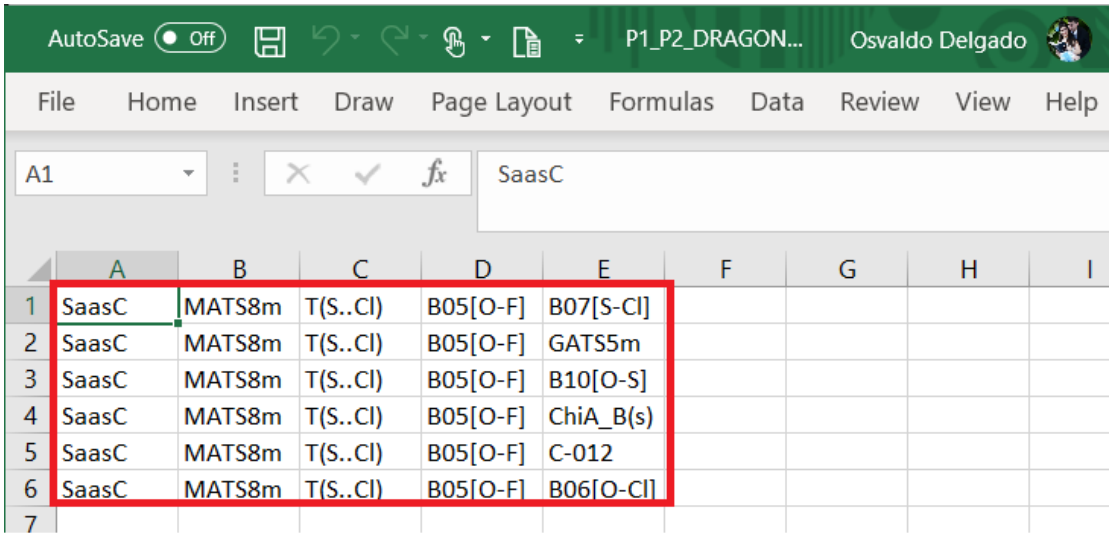


	A	B	C	D	E	F	G	H	I
1	DRAGON_LOG10(EC50nM)_1era_seleccion.csv					variables =	5		
2	Inicio:	47:23.3							
3									
4	base =	SaasC	MATS8m	T(S..Cl)	B05[O-F]				
5	mas =	ChiA_B(s)	GATS5m	C-012	B06[O-Cl]	B07[S-Cl]	B10[O-S]		
6	(6,1) = 6.0	posibles modelos		calculados	6				
7	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B07[S-Cl]				
8	SaasC	MATS8m	T(S..Cl)	B05[O-F]	GATS5m				
9	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B10[O-S]				
10	SaasC	MATS8m	T(S..Cl)	B05[O-F]	ChiA_B(s)				
11	SaasC	MATS8m	T(S..Cl)	B05[O-F]	C-012				
12	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B06[O-Cl]				
13									

Figura 11. Visualización de la salida para **Paso 1**.

Paso 2:

El Paso 1 devuelve una entrada legible por las personas, pero no útil para la máquina, por lo que el Paso 2 abre cada uno de los ficheros y elimina todas las líneas que no son la combinación de descriptores, los datos son guardados en una carpeta con nombre “Modelos_solo”.



	A	B	C	D	E	F	G	H	I
1	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B07[S-Cl]				
2	SaasC	MATS8m	T(S..Cl)	B05[O-F]	GATS5m				
3	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B10[O-S]				
4	SaasC	MATS8m	T(S..Cl)	B05[O-F]	ChiA_B(s)				
5	SaasC	MATS8m	T(S..Cl)	B05[O-F]	C-012				
6	SaasC	MATS8m	T(S..Cl)	B05[O-F]	B06[O-Cl]				
7									

Figura 12. Visualización de la salida para **Paso 2**.

Paso 3:

Este es el encargado de construir para cada combinación dentro del .csv un fichero con extensión .arff de entrenamiento y de prueba; para hacerlo se basa en dos ficheros con extensión .arff madres (entrenamiento y prueba) que hay que copiarlos dentro de una carpeta llamada arff.

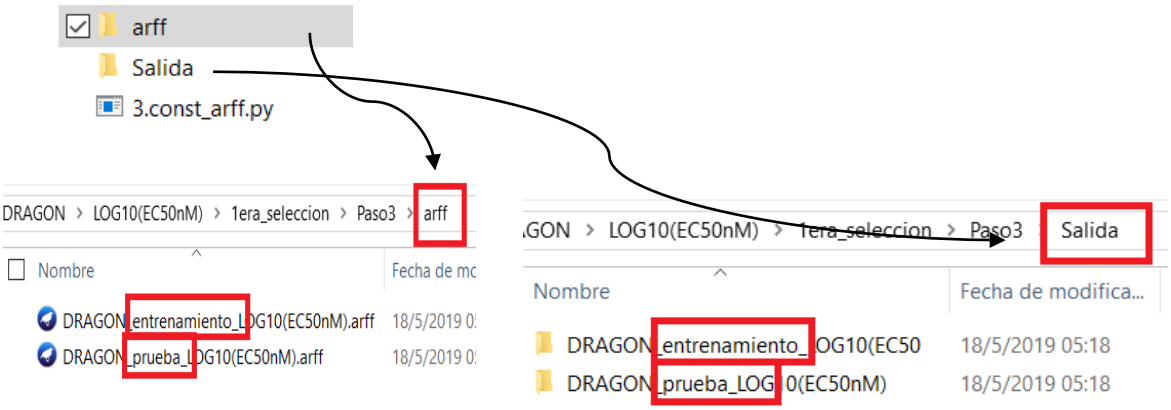


Figura 13. Esquema para **Paso 3**.

Para la regresión hay 3 scripts (en este caso es un programa del que hace uso el intérprete de Python para ejecutar sus órdenes) a ejecutar.

Para la clasificación son 4.

De ellos los 2 primeros son iguales y el 3ero es muy similar, en el caso de la clasificación se utiliza otro más llamado 4- tabla.py.

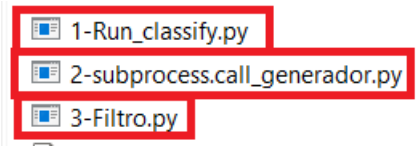


Figura 14. Los 3 scripts que se utilizan en **Paso 4** para regresión.

1- *Run_classify.py:*

Entrada:

- Ruta de la carpeta de entrenamiento generada en el paso 3 que está dentro de la carpeta “Salida”
- Ruta de la carpeta prueba generada en el paso 3 que está dentro de la carpeta “Salida”
- La opción deseada (1- IBk, 2- J48, 3- MultilayerPerceptron, 4- SMO, 5- SMOreg)
- Ruta de la carpeta donde el usuario desee guardar los ficheros con extensión .bat que serán generados.

Salida:

Por cada fichero de entrenamiento y prueba genera un fichero con extensión .bat que contiene la información para llamar la máquina virtual de java y ejecutar la orden para la opción elegida; esto es el equivalente a utilizar el Simple CLI del Weka.

2- *subprocess.call_generador.py*

Entrada:

- Ruta de entrada de la carpeta contenedora de los ficheros con extensión .bat.
- La cantidad de ejecuciones independientes que desea hacer.

Recomendación: no debe utilizarse una mayor cantidad de ejecuciones independientes que cantidad de núcleos posea la computadora donde lo ejecutará.

Salida:

- Crea una carpeta llamada Subprocess con la cantidad de ficheros con extensión .py igual al valor ingresado en la opción de cantidad de ejecuciones independientes.

Luego de ejecutar estos ficheros con extensión .py, en la misma ruta de los .bat se generan ficheros de texto plano sin extensión con el mismo nombre de los .bat que contiene la información de interés real.

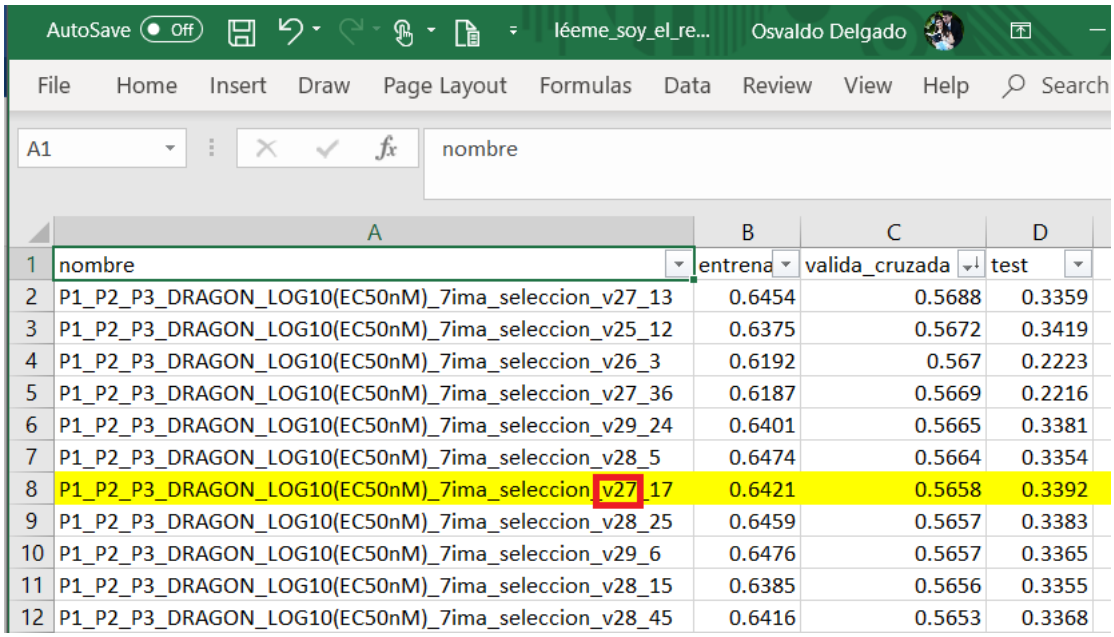
3- Filtro

Entrada:

- La ruta donde se encuentran los ficheros sin extensión.

Salida:

- Un fichero llamado léeme_soy_el_resumen.csv que contiene el nombre del archivo que lo generó, el valor de correlación de Pearson para la serie de entrenamiento y test; además de la validación cruzada.

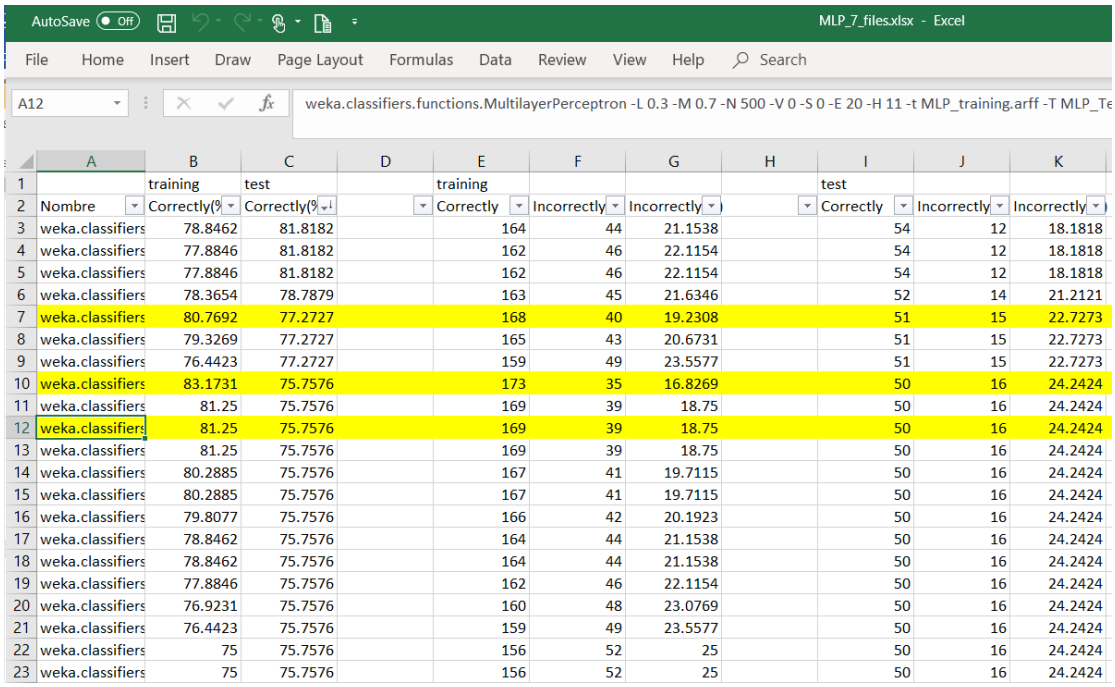


	nombre	entrena	valida_cruzada	test
2	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v27_13	0.6454	0.5688	0.3359
3	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v25_12	0.6375	0.5672	0.3419
4	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v26_3	0.6192	0.567	0.2223
5	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v27_36	0.6187	0.5669	0.2216
6	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v29_24	0.6401	0.5665	0.3381
7	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v28_5	0.6474	0.5664	0.3354
8	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v27_17	0.6421	0.5658	0.3392
9	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v28_25	0.6459	0.5657	0.3383
10	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v29_6	0.6476	0.5657	0.3365
11	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v28_15	0.6385	0.5656	0.3355
12	P1_P2_P3_DRAGON_LOG10(EC50nM)_7ima_seleccion_v28_45	0.6416	0.5653	0.3368

Figura 15. Visualización de la salida de Filtro en **Paso 4** de regresión.

El número después de la “v” indica la cantidad de variables que tiene el modelo y el número después del guion bajo que aparece a continuación es un número genérico, que permite concretamente ubicar cuales fueron los descriptores que dieron origen a esos resultados.

Para el caso de la clasificación en el Paso 4, el script 3- Filtro.py devuelve un fichero con extensión .txt donde contiene detalles como las instancias correctamente clasificadas, las instancias correctamente clasificadas en por ciento, las instancias incorrectamente clasificadas, las instancias incorrectamente clasificadas en por ciento y la matriz de confusión para el *training* y para el *test* de cada uno de los modelos. Y se utiliza además otro script llamado 4- tabla.py que construye un fichero con extensión .csv y una salida cómoda al analista.



	A	B	C	D	E	F	G	H	I	J	K
1		training	test		training				test		
2	Nombre	Correctly(%	Correctly(%		Correctly	Incorrectly	Incorrectly		Correctly	Incorrectly	Incorrectly
3	weka.classifiers	78.8462	81.8182		164	44	21.1538		54	12	18.1818
4	weka.classifiers	77.8846	81.8182		162	46	22.1154		54	12	18.1818
5	weka.classifiers	77.8846	81.8182		162	46	22.1154		54	12	18.1818
6	weka.classifiers	78.3654	78.7879		163	45	21.6346		52	14	21.2121
7	weka.classifiers	80.7692	77.2727		168	40	19.2308		51	15	22.7273
8	weka.classifiers	79.3269	77.2727		165	43	20.6731		51	15	22.7273
9	weka.classifiers	76.4423	77.2727		159	49	23.5577		51	15	22.7273
10	weka.classifiers	83.1731	75.7576		173	35	16.8269		50	16	24.2424
11	weka.classifiers	81.25	75.7576		169	39	18.75		50	16	24.2424
12	weka.classifiers	81.25	75.7576		169	39	18.75		50	16	24.2424
13	weka.classifiers	81.25	75.7576		169	39	18.75		50	16	24.2424
14	weka.classifiers	80.2885	75.7576		167	41	19.7115		50	16	24.2424
15	weka.classifiers	80.2885	75.7576		167	41	19.7115		50	16	24.2424
16	weka.classifiers	79.8077	75.7576		166	42	20.1923		50	16	24.2424
17	weka.classifiers	78.8462	75.7576		164	44	21.1538		50	16	24.2424
18	weka.classifiers	78.8462	75.7576		164	44	21.1538		50	16	24.2424
19	weka.classifiers	77.8846	75.7576		162	46	22.1154		50	16	24.2424
20	weka.classifiers	76.9231	75.7576		160	48	23.0769		50	16	24.2424
21	weka.classifiers	76.4423	75.7576		159	49	23.5577		50	16	24.2424
22	weka.classifiers	75	75.7576		156	52	25		50	16	24.2424
23	weka.classifiers	75	75.7576		156	52	25		50	16	24.2424

Figura 16. Visualización de la salida para tabla en **Paso 4**.

CAPÍTULO III: RESULTADOS Y DISCUSIÓN

Capítulo III: Resultados y discusión

3.1 Cálculo de los descriptores moleculares

Con la ayuda del software DRAGON 7 se realizó el cálculo de los descriptores moleculares 0D, 1D y 2D implementados en el mismo (3839 DMs / 5270 DMs) a las 317 estructuras químicas reportadas en *Malaria Box*. Fueron excluidos aquellos descriptores con un valor de desviación estándar menor que 0.0001 y que tuvieran una correlación superior al 0.9500; lo que resultó en una salida de 812 descriptores moleculares para cada una de las 317 moléculas, los cuales codifican información estructural no ortogonal de las mismas, lo que asegura una correcta descripción.

3.2 Separación por clústeres y diseño de las series

Tanto para la regresión como para la clasificación se tuvo en cuenta las 317 estructuras químicas que tenían reportado un valor de IC 50. En el caso de la clasificación se tomó como valor de corte 1 µM, se clasificaron los que tenían un valor de IC 50 por debajo de 1 µM como muy_activos (211 compuestos) y por encima de 1 µM como activos (106 compuestos); además de que se hizo el estudio de AC a las moléculas clasificadas como activo y muy_activo de forma independiente; o sea, se realizaron 3 estudios de AC (1 para regresión y 2 para clasificación).

Se evaluó la distribución y diversidad estructural dentro de los grupos de observaciones del total de la base de datos, mediante un AC jerárquicos

según el procedimiento k-NNCA (siglas en inglés de Algoritmo de clúster de k-Vecinos más Próximos) implementado en el paquete de procesamiento STATISTICA 8.0. En el caso de la clasificación; previo al estudio de clúster fueron estandarizadas las matrices de DMs (no se estandarizó en la regresión) anteriormente calculados. A cada uno de los dos grupos formados se les realizó un segundo AC del tipo k-MCA (k-means clúster analysis).

Como resultado del estudio de AC se sacaron las moléculas que se unían al final de los clústeres por poseer la menor similitud estructural del resto de las móleculas (outliers).

Para la regresión en el AC K-Means se dividió la base de datos de 317 moléculas en 8 clústers y se dectectaron 30 compuestos *outliers*.

Para la clasificación del AC K-Means se obtuvo 5 clústers para las moléculas clasificadas como activo y se detectaron 6 moléculas outliers; mientras que para las moléculas clasificadas como muy_activo se dividió en 5 clústeres y se detectó 8 moléculas *outliers*.

Todos los compuestos *outliers* fueron removidos.

3.2.1 Diseño de las series de entrenamiento, predicción y external para clasificación

Para los compuestos clasificados como activo y muy_activo se implementó la misma metodología según la clasificación:

Realizado el AC se procedió a separar cada uno de los clústeres en 3 subconjuntos (entrenamiento, prueba y *external*). Para la serie *external* se tomó el 10 % del total de compuestos; fue tomado el 90 % de los compuestos para las series de entrenamiento y prueba.

Para la serie de prueba se tomó el 25 % de los compuestos restantes; de esta forma, los restantes compuestos formaron parte de la serie de entrenamiento.

Resultado del AC se ordenó cada uno de los clústeres según el valor de la distancia entre el compuesto y la media del clúster y según el juicio del analista se separó en entrenamiento, prueba y *external* de tal manera que, por cada dos compuestos con una distancia similar, con uno se entrenara y el otro se tomara para la serie de predicción o *external*. Esto permite asegurar la representatividad de elementos del mismo dominio en cada subconjunto obtenido. Las series obtenidas, así como el por ciento que representan de la data son representadas en la Tabla # 1.

Tabla # 1. Series de entrenamiento, predicción y *external* para la clasificación.

Serie / Clasificación	activo	Por ciento	muy_activo	Por ciento
Entrenamiento	71 moléculas	71	137 moléculas	67,49
Prueba	20 moléculas	20	46 moléculas	22,66
Externa	9 moléculas	9	20 moléculas	9,85
TOTAL	100 moléculas	100	203 moléculas	100

3.2.2 Diseño de las series de entrenamiento y predicción para regresión

Realizado el AC se procedió a separar cada uno de los clústeres en 2 subconjuntos (entrenamiento, prueba).

Para la serie de prueba se tomó el 25 % del total de compuestos y el 75 % de los compuestos para la serie de entrenamiento.

Resultado del AC se ordenó cada uno de los clústeres según el valor de la distancia entre el compuesto y la media del clúster y según el juicio del analista se separó en entrenamiento y prueba de tal manera que, por cada dos compuestos con una distancia similar, con uno se entrenara y el otro se tomara para la serie de predicción o *external*. Esto permite asegurar la representatividad de elementos del mismo dominio en cada subconjunto obtenido. Las series obtenidas, así como el por ciento que representan del conjunto de datos son representadas en la Tabla # 2.

Tabla # 2. Series de entrenamiento y predicción para la regresión.

Serie / Clasificación	No. de compuestos	Por ciento
Entrenamiento	215 compuestos	74,91
Prueba	72 compuestos	25,08
Total	287 compuestos	100

3.3 Estudio de clasificación

3.3.1 Selección de atributos para la clasificación

Se implementó para cada atributo evaluador (ClassifierSubsetEval y WrapperSubsetEval) cinco métodos de búsqueda (BestFirst, GeneticSearch,

LinearForwardSelection, RankSearch y SubsetSizeForwardSelection) sin variar sus parámetros por defecto:

Se implementó el ejecutable After_Select_Attributes.exe para:

- Los 5 buffers correspondientes a WrapperSubsetEval.
- Los 5 buffers correspondientes a ClassifierSubsetEval.
- Los 10 buffers correspondientes a la combinación de WrapperSubsetEval y ClassifierSubsetEval.

Primero se determinó cuáles eran las variables comunes que luego serían utilizadas para construir todos los modelos posibles resultantes de la combinación no repetida de los DMs para 6, 7 y 8 variables.

Se encontró que las variables que más se repetían y por ende con las que se construirían los modelos fueron:

Para IBk:

Se analizaron 129 modelos.

Coeficiente binomial (9,6) = 9*8*7*6*5*4 / 6! = 84

Coeficiente binomial (9,7) = 9*8*7*6*5*4*3 / 7! = 36

Coeficiente binomial (9,8) = 9*8*7*6*5*4*3*2 / 8! = 9

Total = 84 + 36 + 9 =129

Tabla # 3. DMs comunes para los modelos de IBk.

Número del descriptor	Nombre del descriptor
665	B08[O-Cl]
673	B09[N-N]

360	nCb-
374	nN=C-N<
447	H-049
515	CATS2D_09_AA
517	CATS2D_03_AP
700	B10[O-S]
294	P_VSA_i_2

Para J48:

Se analizaron 129 modelos.

Coeficiente binomial (9,6) = 9*8*7*6*5*4 / 6! = 84

Coeficiente binomial (9,7) = 9*8*7*6*5*4*3 / 7! = 36

Coeficiente binomial (9,8) = 9*8*7*6*5*4*3*2 / 8! = 9

Total = 84 + 36 + 9 =129

Tabla # 4. DMs comunes para los modelos de J48.

Número del descriptor	Nombre del descriptor
158	MATS6m
447	H-049
2	AMW
6	nTA
214	GATS1i
665	B08[O-Cl]
712	F03[C-S]
757	F08[C-F]

223	GATS2s

Para MLP se utilizaron 7 métodos de búsqueda repartido en 2 atributos seleccionadores los cuales se muestran en la Tabla # 5:

Tabla # 5. Métodos de búsqueda utilizados para los atributos seleccionadores WrapperSubsetEval y ClassifierSubsetEval.

Atributo seleccionador / método de búsqueda	BestFirst	GeneticSearch	LinearForwardSelection	RankSearch	SubsetSizeForwardSelection
WrapperSubsetEval	X		X		X
ClassifierSubsetEval	X	X	X		X

Se analizaron 32 318 modelos.

Coeficiente binomial (16,6) = 16*15*14*13*12*11 / 6! = 8 008

Coeficiente binomial (16,7) = 16*15*14*13*12*11*10 / 7! = 11 440

Coeficiente binomial (16,8) = 16*15*14*13*12*11*10*9 / 8! = 12 870

Total = 8 008 + 11 440 + 12 870 = 32 318

Tabla # 6. DMs comunes para los modelos de MLP.

Número del descriptor	Nombre del descriptor
158	MATS6m
665	B08[O-Cl]
214	GATS1i

453	N-072
750	F07[N-N]
428	C-027
795	DLS_07
54	S3K
105	WiA_D/Dt
178	MATS1i
224	GATS3s
323	SpMax_AEA(bo)
515	CATS2D_09_AA
711	F03[C-N]
320	SpMax_EA(ri)
425	C-020

A continuación, para cada uno de estos 3 clasificadores (IBk, J48, MLP) se corrieron los 7 scripts según vienen descritos en 4 pasos y se realizó una selección de los mejores modelos a los que posteriormente se les varió los parámetros del clasificador.

3.3.2 Modelos de clasificación

3.3.2.1- Clasificador IBk

Tabla # 7. Por ciento de clasificación de entrenamiento y prueba para los mejores modelos de IBk.

Modelo	Entrenamiento Clasificados	Prueba Clasificados

	correctamente (%)	correctamente (%)
V6_45	100	65.15
V6_62	100	65.15
v6_29	75.97	72.72

Donde V#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

Luego de realizar la variación de los parámetros del clasificador no se obtuvo ningún modelo con un valor de entrenamiento y prueba por encima del 75 % y que la diferencia entre ambos por cientos fuera menor que el 10 %.

3.3.2.2- Clasificador J48

Tabla # 8. Por ciento de clasificación de entrenamiento y prueba para los mejores modelos de J48.

Modelo	Entrenamiento Clasificados correctamente (%)	Prueba Clasificados correctamente (%)
V6_2	74.52	68.18
V6_7	74.04	69.69
v6_28	74.04	69.69
V6_51	81.73	56.06

Donde V#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

Luego de realizar la variación de los parámetros del clasificador no se obtuvo ningún modelo con un valor de entrenamiento y prueba por encima del 75 % y que la diferencia entre ambos por cientos fuera menor que el 10 %.

3.3.2.3- Clasificador MLP

Tabla # 9. Por ciento de clasificación de entrenamiento y prueba para los mejores modelos de MLP.

Modelo	Entrenamiento Clasificados correctamente (%)	Prueba Clasificados correctamente (%)
V6_5887	77.88	81.81
V7_2133	87.01	71.21
V7_7545	77.88	68.18
V8_4801	84.61	77.27
V8_9042	90.38	66.66
V8_10424	78.36	77.27

Donde V#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

Luego de realizar la variación de los parámetros del clasificador se obtuvo:

Tabla #10. Parámetros variados, por ciento de clasificación para las series de entrenamiento, prueba y *external* de los mejores modelos de MLP.

Modelo	Parámetros	Entrenamiento Clasificados	Prueba Clasificados	*External* Clasificados

		correctamente	correctamente	correctamente
v6_5887	-L 0.3 -M 0.6 -N 500 -V 0 -S 0 -E 20 -H 8 -t	80.77 %	77.27 %	68.97 %
v6_5887	-L 0.3 -M 0.3 -N 500 -V 0-S 0-E 20 -H 11 -t	83.17 %	75.76 %	72.41 %
v6_5887	-L 0.3 -M 0.7 -N 500 -V 0 -S 0-E 20-H 11 -t	81.25 %	75.76 %	79.31 %
_v7_2133	-L 1.0 -M 0.7 -N 500 -V 0-S 0 -E 20-H 10 -t	86.06 %	77.27 %	58.62 %
v7_7545	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 4 -t	81.25 %	77.27 %	68.97 %
v7_7545	-L 0.3 -M 0.2 -N 500 -V 0 -S 0-E 20 -H 10-t	80.76 %	77.27 %	65.52 %
v7_7545	-L 1.0 -M 0.5 -N 500 -V 0 -S 0 -E 20 -H 9 -t	81.25 %	75.76 %	65.52 %
v8_4801	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 5 -t	84.61 %	77.27 %	58.62 %
v8_4801	-L 0.9 -M 0.8 -N 500 -V 0 -S 0 -E 20 -H 7 -t	85.09 %	75.76 %	68.97 %
v8_9042	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 6 -t	86.05 %	78.79 %	65.52 %
v8_9042	-L 0.3 -M 0.9 -N 500 -V 0 -S 0 -E 20 -H 5 -t	82.69 %	78.79 %	65.52 %
v8_9042	-L 0.3 -M 0.3 -N 500 -	86.53 %	77.27 %	65.52 %

	V 0 -S 0 -E 20 -H 6 -t			

Donde V#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

3.3.3 Discusión de los resultados del estudio de clasificación

Cuando fueron empleados clasificadores como IBk y J48 incluso luego de realizar la variación de los parámetros del clasificador no se encontraron modelos con un valor de por ciento de correctamente clasificados superior del 75 % para los correctamente clasificados en por ciento para la serie de entrenamiento y que la diferencia de los correctamente clasificados en por ciento de la serie de predicción y entrenamiento no supere el 10 %. Con el uso de estos dos clasificadores, para esta base de datos química, no se obtienen modelos capaces de clasificar correctamente la actividad antimalárica en un posterior estudio de CV.

Para los modelos de MultilayerPerceptron se seleccionaron los 6 mejores modelos que oscilan aproximadamente entre 78 % - 90 % para las instancias correctamente clasificadas en por ciento para la serie de entrenamiento y en un rango de 66 % - 81 % para las moléculas correctamente clasificadas en por ciento para la serie de predicción. Luego de realizar la variación de los parámetros del clasificador se seleccionaron 12 modelos de MLP a los que luego de realizar la predicción sobre la serie *external* arrojó los siguientes resultados:

El modelo v6_5887 fue el que mejor métrica tiene al implementarlo con los parámetros de MultilayerPerceptron -L 0.3 -M 0.7 -N 500 -V 0 -S 0 -E 20 -H 11 con valores de instancias correctamente clasificadas en por ciento por encima de 75 % en la serie de entrenamiento (81.25 %), prueba (75.76 %) y *external* (79.31 %). Por lo que puede ser utilizado para posteriores estudios de cribado virtual, ya que predice más del 79 % de los casos de una serie externa, se asegura así, una correcta clasificación de la actividad antimalárica.

Los descriptores moleculares que intervienen fundamentalmente en el estudio de clasificación para la modelación de la actividad antimalárica son los basados en heteroátomos y enlaces de los mismos con carbono o entre ellos, además influyen algunas propiedades químico-físicas como potencial de ionización, entre otras, lo que significa que estas partes o propiedades de las moléculas son las más influyentes en su actividad frente al *Plasmodium falciparum*.

3.4- Estudio de regresión

Durante el desarrollo de este trabajo se utilizó la regresión no lineal, específicamente SVM aplicadas a la regresión; además, se aplicó una transformación logarítmica de base 10 sobre la variable dependiente (el valor de IC_{50} con una concentración en μM) con el objetivo de que los datos experimentales no estuvieran numéricamente tan distantes entre sí.

3.4.1 Selección de atributos para la regresión

Se utilizó la versión 3.6.2 del software Weka para realizar la selección de atributos para el estudio de regresión.

El objetivo de los métodos de selección de atributos es identificar, mediante un conjunto de datos que poseen unos ciertos atributos, aquellos atributos que tienen más peso a la hora de determinar si los datos aportan información a la construcción del modelo; o sea, eliminación de atributos redundantes e irrelevantes. Si hay un número excesivo de atributos, esto puede hacer que el modelo sea demasiado complejo y se produzca sobreajuste (*overfitting*).

Los valores que toma la variable dependiente (actividad antimalárica) son numéricos, como consecuencia lo que realizará el algoritmo seleccionado como clasificador será una regresión.

Se utilizó el clasificador SMOreg (Máquinas de vectores de soporte asociada a la regresión).

Como atributo seleccionador se utilizó WrapperSubsetEval y 5 métodos de búsqueda. Los métodos de búsqueda empleados fueron BestFirst, GeneticSearch, LinearForwardSelection, SubsetSizeForwardSelection GreedyStepwise.

Luego de realizada la selección, se implementó el ejecutable After_Select_Attributes.exe para detectar cuáles eran los descriptores que sólo aparecían por un solo método de búsqueda; estos fueron excluidos. Se

siguió la metodología descrita anteriormente en 6 ocasiones más hasta que se mantuvo constante la salida de atributos seleccionados del software Weka.

Sobre la selección final de los atributos se implementó After_Select_Attributes.exe para detectar cuales eran los mejores descriptores para construir los modelos y luego se le aplicó los pasos del Paso 1 donde se definió que devolviera para un rango de 1 a 50 variables los primeros 20 modelos; se ejecutaron los Pasos 2, 3, 4 y se detectó 2 modelos de 27 variables que tenían un valor de coeficiente de Pearson similar con otros modelos que tenían muchas más variables. Por lo que se amplió el estudio de 20 modelos para 27 variables, a todos los posibles modelos de 27 variables (70 modelos).

3.4.2- Modelos de Regresión

3.4.2.1- Modelo de SMOreg

A continuación, se muestran los parámetros estadísticos (Tabla # 11) y el gráfico de correlación entre la actividad experimental y la predicha por el modelo para: el entrenamiento (Figura 17), la validación cruzada (Figura 18) y la predicción (Figura 19) del mejor modelo resultado luego de haber realizado 7 selecciones de variables y explorado todos los modelos de 27 variables, se encontró un modelo no lineal para la actividad antimalárica con el método de

las Máquinas de Vectores de Soporte para la regresión (SMOreg), se empleó una complejidad unitaria (C = 1) y la función PoliKernel con exponente 1:

Tabla # 11. Parámetros estadísticos para el mejor modelo SMOreg de 27 variables (v27_17).

Serie	N	R^2	Mean absolute error
Entrenamiento	215	0.4137	0.2228
Validación cruzada	215	0.3095	0.2545
Prueba	72	0.1206	0.3458

Donde N es el número de compuestos y R^2 es el coeficiente de determinación. v#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

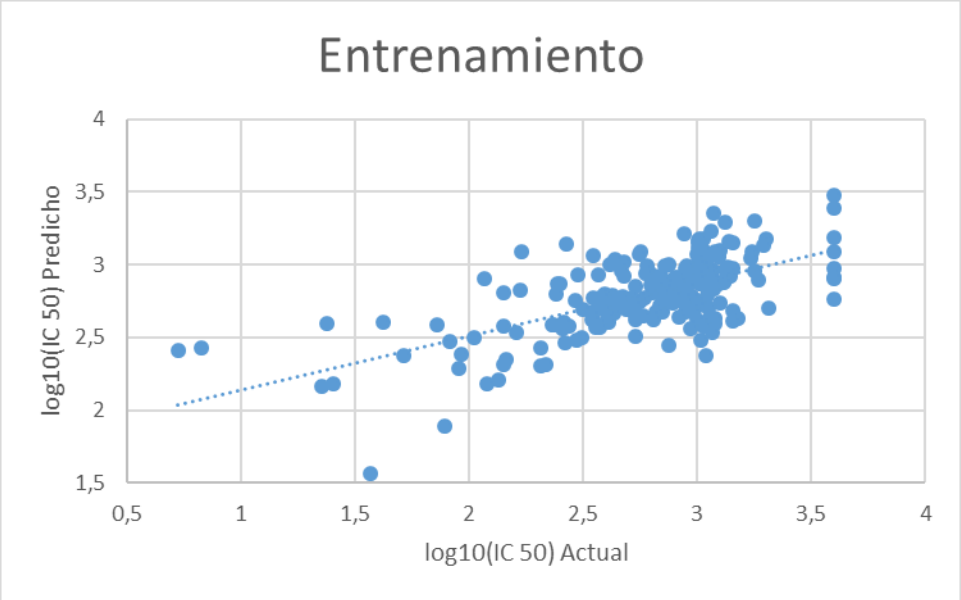


Figura #17. Gráfico de correlación para el entrenamiento del mejor modelo de 27 variables.

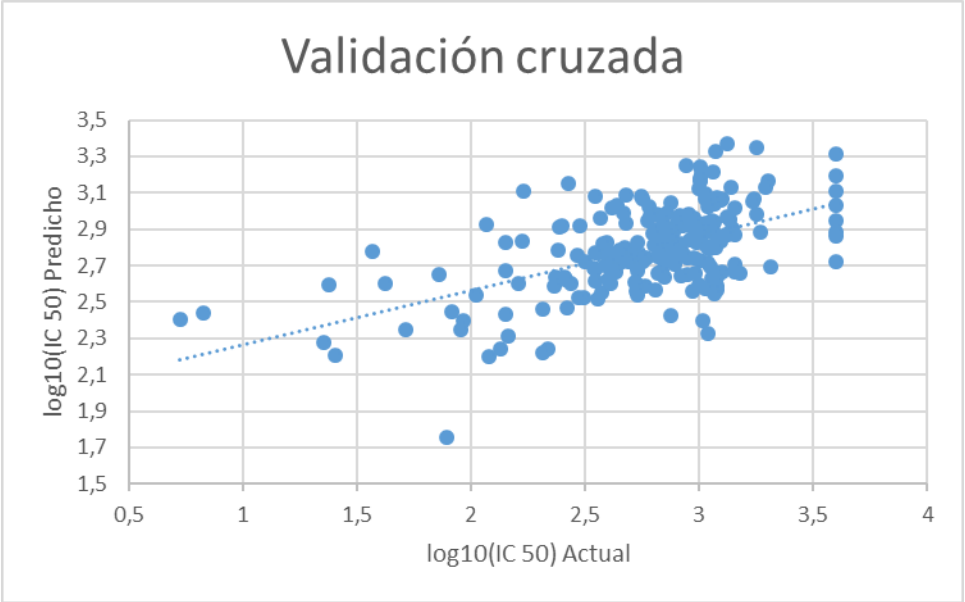


Figura #18. Gráfico de correlación para la validación interna del mejor modelo de 27 variables.

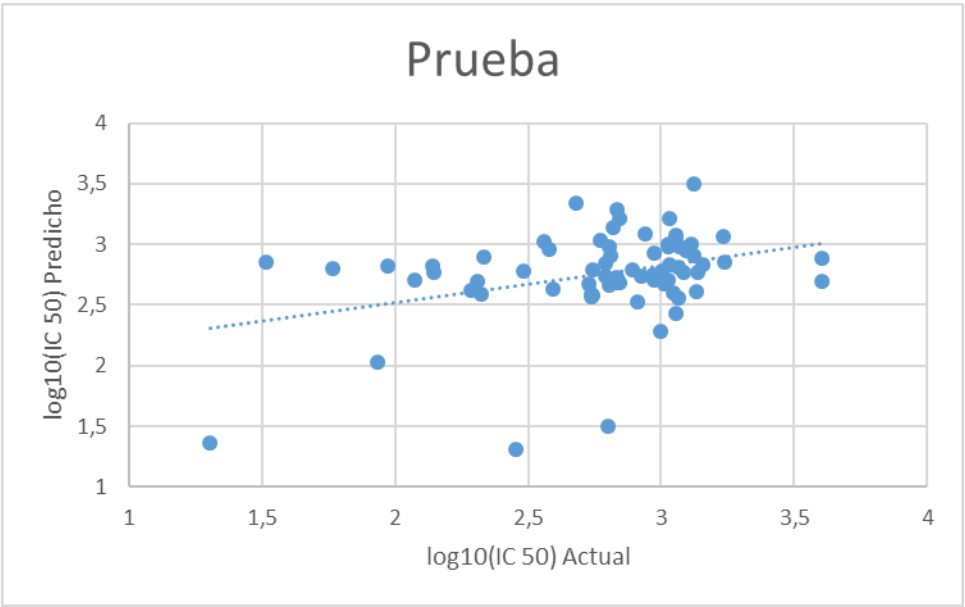


Figura #19. Gráfico de correlación para la predicción (validación externa) del mejor modelo de 27 variables.

Una de las principales fortalezas de la base de datos es la diversidad estructural, o sea, el conjunto de aplicación, sin embargo, esto a su vez constituye un obstáculo para realizar el estudio de regresión. Además, se

modela una actividad biológica, esta es una práctica muy compleja ya que depende, además de la estructura, de otros factores no controlables por el experimentador como, por ejemplo, el efecto de la matriz biológica sobre la capacidad individual de acción de cada una de las moléculas. Sin embargo, se asume que la estructura química posee el protagonismo en cuanto a influenciar el valor de la propiedad.

Los modelos desarrollados luego de ser entrenados, no son capaces de predecir sobre sí mismos ni siquiera el 50 % de los casos (moléculas) y sólo son capaces de predecir correctamente poco más de un 10 % de los casos de la serie externa.

Todo modelo es aplicable dentro de su dominio, por lo que se decidió hacer un estudio para aquellos valores que se desviaban del comportamiento estándar del modelo, reducir el conjunto de aplicación a costa de obtener un mejor modelo.

3.4.2.2- Modelo de SMOreg con la extracción de compuestos *outliers*

Se utilizó para la identificación de compuestos *outliers,* el método de los residuales estandarizados y como criterio se tomó ± 2 · DESV. S (desviación estándar) sucesivas veces.

Se extrajeron un total de 84 *outlier*s, 63 en la serie de entrenamiento y como consecuencia, 21 de la serie de prueba para mantener la proporción 75 % entrenamiento, 25 % prueba.

A continuación, se muestran los parámetros estadísticos (Tabla # 12) y los gráficos de correlación entre la actividad experimental y la predicha por el modelo para: el entrenamiento (Figura 20), la validación (Figura 21) y la predicción (Figura 22) del modelo no lineal para la actividad antimalárica con el método de las Máquinas de Vectores de Soporte para la regresión (SMOreg), se empleó una complejidad unitaria (C = 1) y la función PoliKernel con exponente 1, resultado luego de la extracción de *outliers*:

Tabla # 12 Parámetros estadísticos para modelo SMOreg de 27 variables (v27_17) sin *outliers*.

Serie	N	R^2	Mean absolute error
Entrenamiento	152	0.8241	0.0882
Validación cruzada	152	0.6487	0.1209
Prueba	51	0.5509	0.1951

Donde N es el número de compuestos y R^2 es el coeficiente de determinación. V#_* significa: modelo generado al que aleatoriamente fue asignado el número * de # variables.

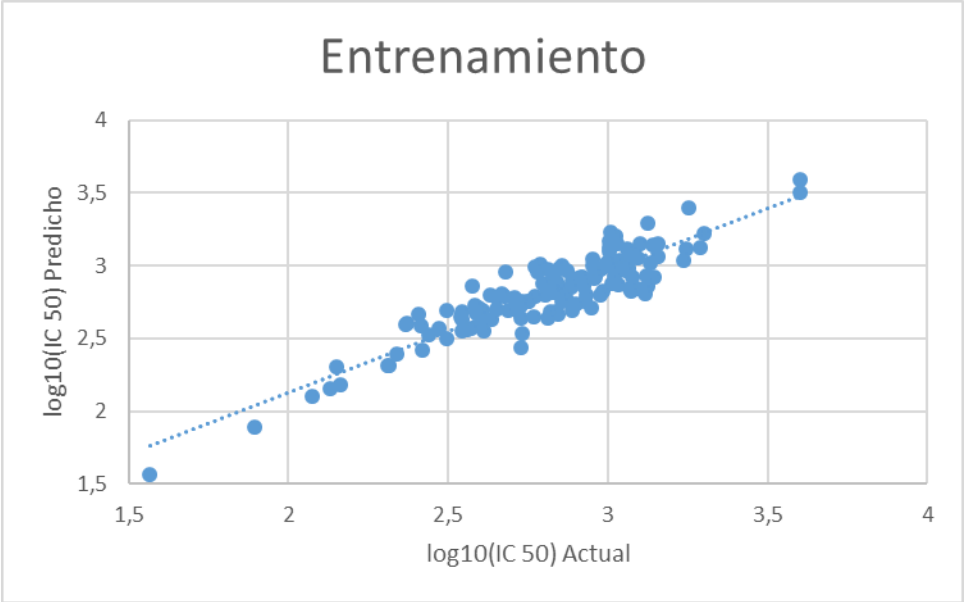


Figura # 20. Gráfico de correlación para el entrenamiento del mejor modelo de 27 variables sin *outliers.*

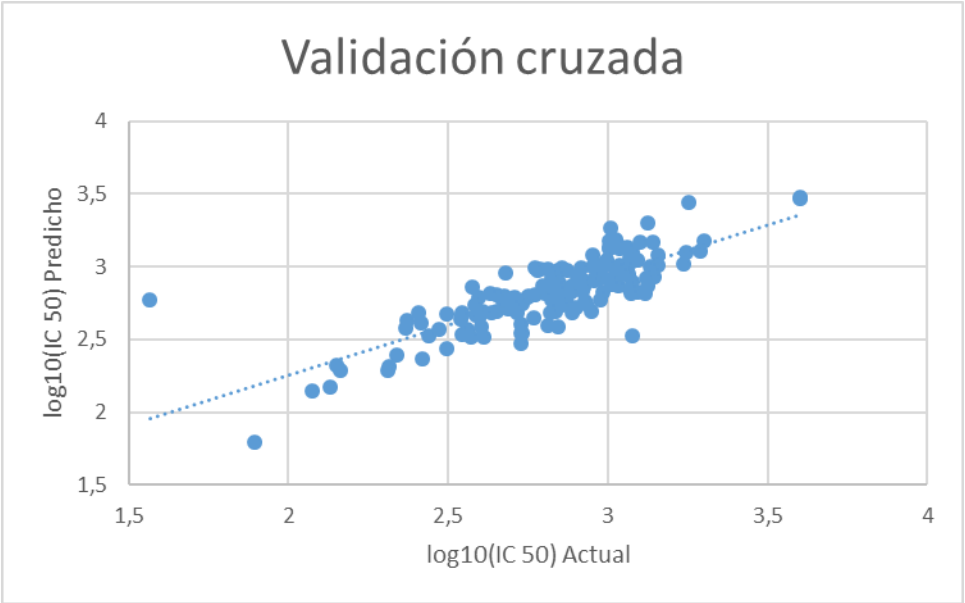


Figura # 21. Gráfico de correlación para la validación interna del mejor modelo de 27 variables sin *outliers.*

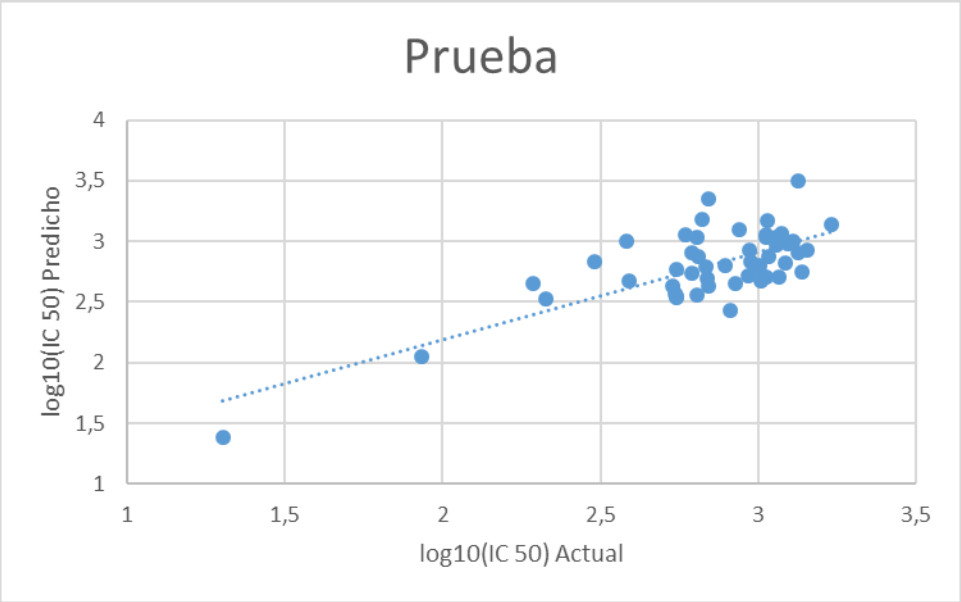


Figura # 22. Gráfico de correlación para la predicción (validación externa) del mejor modelo de 27 variables sin *outliers.*

Como puede observarse en los parámetros estadísticos y en los gráficos, el valor de coeficiente de determinación para la serie de entrenamiento indica que el modelo luego de entrenar es capaz de predecir sobre sí mismo más del 80 % de los casos, cuando se le introducen perturbaciones internas, predice prácticamente el 65 % de los casos (difieren en un 15 % aproximadamente los valores de R^2 del entrenamiento y la validación cruzada), lo que le confiere robustez al modelo. Cuando dicho modelo entrena con una serie externa es capaz de predecir el 55 % de los casos.

3.4.3 Discusión de los resultados del estudio de regresión

La aplicación de técnicas de regresión a *Malaria Box* es un proceso complejo, la diversidad estructural de la misma dificulta su modelación de la propiedad biológica. Sin embargo, al reducir el conjunto de aplicación a costa de

obtener un mejor modelo fue desarrollado uno robusto que predice la actividad antimalárica de una serie de prueba coherente con resultados obtenidos por otros autores (Gerardo et al., 2014), lo que le permite ser utilizado en posteriores estudios de CV.

Los descriptores moleculares que intervienen en el estudio de regresión para la modelación de la actividad antimalárica son los basados en heteroátomos y aromaticidad fundamentalmente; además de propiedades químico-físicas como electronegatividad y logP, lo que significa que estas partes o propiedades de las moléculas son las más influyentes en la actividad frente al parásito *Plasmodium falciparum*.

Conclusiones

- Fueron identificados, de los descriptores moleculares del software DRAGON, los que mejor describen la actividad antimalárica de 317 estructuras, para ello fueron utilizadas las metodologías de selección de atributos implementadas en el software Weka y automatizados los pasos para dichas selecciones con softwares hechos en casa, lo que permitió un análisis más exhaustivo de todas las posibles combinaciones de atributos.
- Se determinaron modelos de clasificación basados en técnicas de IBk, J48 y MLP, este último método arrojó los mejores modelos de clasificación para su posterior uso en estudios de CV.
- Se determinaron modelos de regresión mediante la aplicación de técnicas de SVM, extrayendo compuestos *outliers,* disminuyendo un tanto el dominio de aplicación del mismo; fue desarrollado un modelo que predice satisfactoriamente la actividad antimalárica de sustancias orgánicas frente al parásito *Plasmodium falciparum*, el cual puede ser utilizado en posteriores estudios de CV.

Recomendaciones

- Realizar un estudio de CV aplicando los modelos desarrollados.
- Extender la metodología utilizada con los softwares hechos en casa a estudios QSAR en general, mediante la creación de una interfaz gráfica de los mismos.
- Ampliar el análisis multivariado mediante el empleo de otras técnicas de la inteligencia artificial.

Referencias Bibliográficas

ABRIL, L. G. 2003. Modelos de clasificación basados en máquinas de vectores soporte. *Dialnet,(17)*.

ADHANOM GHEBREYESUS, D. T. 2018. *WORLD MALARIA REPORT 2017*.

ANDRIANTSOANIRINA, V., MENARD, D., TUSEO, L. & DURAND, R. 2009. History and current status of Plasmodium falciparum antimalarial drug resistance in Madagascar. *Scand J Infect Dis*.

BELL, C. E., SLUTSKER, L., BEACH, R. F., FOSTER, S. O., JIMENEZ, G. & SARMIENTO, M. E. 2009. Malaria control in the municipality of San Esteban, Honduras. *Rev Panam Salud Publica, 25, 213-7*.

BETANCOURT, G. A. 2005. Las máquinas de soporte vectorial (SVMs). *Scientia et Technica,* 1.

BRERETON, R. G. 1990. *Chemometrics*, Ellis Horwood, Chichester, UK,.

COLOMBIANA, F. M. 2012-2013. *MALARIA MEMORIAS*.

CORTES, C., VAPNIK, V. & AT&T BELL LABS. 1995. Support-Vector Networks. *Machine Learning,* 20**,** 273-297.

CRAMER, R. D. I., BUNCE, J. D., PATTERSON, D. E. & FRANK, I. E. 1988. Crossvalidation, bootstrapping and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.,* 7**,** 18–25.

DELGADO CASTILLO, D., MARTÍN PÉREZ, R., HERNÁNDEZ PÉREZ, L., OROZCO MORÁLEZ, R. & LORENZO GINORI, J. 2016. Algoritmos de aprendizaje automático para la clasificación de neuronas piramidales afectadas por el envejecimiento. *Revista Cubana de Informática Médica,* 8**,** 559-571.

DIACONIS, P. & EFRON, B. 1983. Computer intensive methods in statistics. *Sci. Am., 248, 96–108.*

DOCKRELL, H. M. & PLAYFAIR, J. H. 1983. Killing of blood-stage murine malaria parasites by hydrogen peroxide. *Infect Immun,* 39**,** 456-9.

DOWNEY, A., WENTWORTH, P., ELKNER, J. & MEYERS, C. 2016. How To Think Like A Computer Scientist: Learning with Python 3.

FERNÁNDEZ-OLIVA, A., ABREU-ORTEGA, M., FERNÁNDEZ-BAIZÁN, C. & MACIÁ PÉREZ, F. 2019. *Algoritmo para la detección de casos excepcionales basado en la Teoría de Conjuntos Aproximados*.

FLORES BALMASEDA, N. 2015. Identificación de nuevos compuestos con potencial actividad antileishmaniásica mediante estudios in silico.

GARCIA, E. & LOZANO, F. 2007. *Boosting Support Vector Machines*.

GARCÍA MORATE, D. 2008. Manual de Weka. morate@gmail. com.

GERARDO, M. C.-M., HUONG, L.-T.-T., YOVANI, M.-P., JUAN, A. C.-G., FRANCISCO, T., FACUNDO, P.-G. & CONCEPCION, A. 2014. Analysis of Proteasome Inhibition Prediction Using Atom-Based

Quadratic Indices Enhanced by Machine Learning Classification Techniques. *Letters in Drug Design & Discovery,* 11**,** 705-711.

GOLBRAIKH, A. & TROPSHA, A. 2002. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Div.,* 5**,** 231–243.

GONZÁLEZ, D. 2013. Nouveaux ligands de quadruplexes. Approches in silico. *et in vitro: Université de Bordeaux, France*.

GONZALEZ DIAZ, H., OLAZABAL, E., CASTANEDO, N., SANCHEZ, I. H., MORALES, A., SERRANO, H. S., GONZALEZ, J. & DE ARMAS, R. R. J. 2002. *Mol Model (Online). 8, 237*.

GONZÁLEZ SALCEDO, L., PATRICIA GUERRERO ZÚÑIGA, A., DELVASTO ARJONA, S. & LUIS ERNESTO WILL, A. 2012. *Exploración con redes neuronales artificiales para estimar la resistencia a la compresión, en concretos fibroreforzados con acero*.

GORBÁTOV, V. A. 1988. *Fundamentos de la Matemática Discreta,* Moscú, URSS: Mir.

JIMENEZ-DIAZ, M. B., MULET, T., VIERA, S., GOMEZ, V., GARUTI, H., IBANEZ, J., ALVAREZ-DOVAL, A., SHULTZ, L. D., MARTINEZ, A., GARGALLO-VIOLA, D. & ANGULO-BARTUREN, I. 2009. Improved murine model of malaria using Plasmodium falciparum competent strains and non-myelodepleted NOD-scid IL2Rgammanull mice engrafted with human erythrocytes. *Antimicrob Agents Chemother,* 53**,** 4533-6.

JT, L. & K, R. 2006. On selection of training and test sets for the development of predictive QSAR models. *QSAR & Combinatorial Science,* 25**,** 235-251.

KOHAVI RON, G. H. 1997. Wrappers for feature subset selection. Artificial Intelligence.**,** 273-324.

KOMBA, A. N., MAKANI, J., SADARANGANI, M., AJALA-AGBO, T., BERKLEY, J. A., NEWTON, C. R., MARSH, K. & WILLIAMS, T. N. 2009. Malaria as a cause of morbidity and mortality in children with homozygous sickle cell disease on the coast of Kenya. *Clin Infect Dis,* 49**,** 216-22.

LAZAROU, M., GUEVARA PATINO, J. A., JENNINGS, R. M., MCINTOSH, R. S., SHI, J., HOWELL, S., CULLEN, E., JONES, T., ADAME-GALLEGOS, J. R., CHAPPEL, J. A., MCBRIDE, J. S., BLACKMAN, M. J., HOLDER, A. A. & PLEASS, R. J. 2009. Inhibition of erythrocyte invasion and Plasmodium falciparum Merozoite Surface Protein 1 processing by human IgG1 and IgG3 antibodies. Infect Immun.

LLERENA, V. R. 2017. Modelación in silico de actividad antimalárica de sustancias orgánicas.

M., C. 2012. World Malaria Report:. W.G.M. Programme.

MACHADO-TUGORES, Y., MENESES-MARCEL, A., MARRERO-PONCE, Y., ARAN-REDO, V., ESCARIO, J. A., LE, T. T. H., GARCÍA-SANCHEZ, R. N. & GÓMEZ-BARRIO 2012. Descubrimiento de nuevos antimaláricos a partir de fármacos conocidos mediante cribado in silico e in vitro. *Anales de la Real Academia Nacional de Farmacia, 78, 401-434.*

MARRERO-PONCE, Y., MARTÍNEZ-SANTIAGO, O., LÓPEZ, Y. M. & S. J. BARIGYE, F. T. 2012. Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors I. Theory and QSPR application. *J Comput Aided Mol Des,* 26**,** 1907.

MARRERO-PONCE, Y. S.-M., D; GÁLVEZ-LLOMPART, M; RECIO MC, GINER RM, GARCÍA-DOMÈNECH, R 2011. Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and in vivo screening: The nitroindazolinone chemotype. *European journal of medicinal chemistry***,** 5736-5753.

MARTÍNEZ SANTIAGO, O., MARRERO PONCE, Y., MILLÁN CABRERA, R., BARIGYE, S. J., MARTÍNEZ LÓPEZ, Y., ARTILES MARTÍNEZ, L. M., GUERRA DE LEÓN, J. O. & PÉREZ GIMÉNEZ, F. Extending Graph Derivative Descriptors to N-Dimensional Atom-Relations. *MATCH (Commun. Math. Chem.), accepted for publications*.

MORATE, D. G. 2008. Manual DE Weka. *Disponível através do e-mail diego. garcia. morate@ mail. com*.

MUNTEANU, C. R. 2013. Técnicas de ingeniería informática e inteligencia artificial para clasificación: aplicaciones para el descubrimiento de fármacos y dianas moleculares.

OMS 2018. This year's World malaria report at a glance. *World Malaria Report***,** 210.

PITOL, F. 2014. K-NN (K Nearest Neighbor), el algoritmo del vecino mas cercano. *http://ferminpitol.blogspot.com/2014/03/k-nn-k-nearest-neighbor-el-algoritmo.html* [Online].

PYKA, A. & PLANAR., J. 1993. Chromatogr. Mod. TLC.

RAMÍREZ, V. & SOLANO, R. 2009. Inteligencia artificial avanzada árboles de clasificación.

SPANGENBERG, T., BURROWS, J. N., KOWALCZYK, P., MCDONALD, S., WELLS, T. N. C. W. & WILLIS, P. 2013. The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases. *PLoS One*.

TODESCHINI, R. & CONSONNI, V. 2009. *Molecular Descriptors for Chemoinformatics*, wiley-VCH.

TODESCHINI, R. C., V. 2009. Molecular Descriptors for Chemoinformatics. *wiley-VCH*.

TROPSHA, A., GRAMATICA, P. & GOMBAR, V. K. 2003. *QSAR Comb. Sci.,* 22**,** 69.

TUGORES, Y. M. 2014. Tamizaje farmacológico en la búsqueda de potenciales fármacos antimaláricos
integrando nuevos modelos in silico y corroboración experimental. Madrid: UNIVERSIDAD COMPLUTENSE DE MADRID.
VAN DE WATERBEEMD, H. 1995. Chemometric Methods in Molecular Design (Methods and Principles in Medicinal Chemistry). *John Wiley & Sons: New York*.