Original article[1]

Análisis de clustering subspace de pesticidas químicos

Heriberto Castañeta Maroni[1], Alex Quispe[1], Ebbe Yapu Tapia[1,*], Pablo Duchowicz[2] and Sergio Peignier[3]

[1]Instituto de Investigaciones Químicas IIQ, Universidad Mayor de San Andrés UMSA, Av. Villazón N° 1995, La Paz, Bolivia, 0201-0220, iiq@umsa.bo; [2]Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, 1900 La Plata, Argentina; [3]INSA Lyon, INRAE, BF2I, UMR0203, F-69621, Villeurbanne, France,

Keys: *Subspace clustering, Molecular Descriptors, Vapor pressure, Pesticides, PPDB database, Quantitative Structure-Property Relationships, QSPR;* **Claves:** *Subspace clustering, Descriptores moleculares, presión de vapor, Pesticidas, Base de datos PPDB, Relaciones cuantitativas estructura-propiedad, QSPR.*

ABSTRACT

Subspace clustering analysis of chemical pesticides. The "subspace clustering" algorithm was applied to analyse molecules with pesticidal properties. 1509 molecules from the PPDB (Pesticides Properties DataBase - AERU Hertfordshire University) database were analysed; 1005 molecules presented experimental vapour pressure data. Descriptors were calculated with the PaDEL-Descriptor program (v. 2.20) with a total of 14464 0D-2D molecular descriptors and fingerprint types. Subspace clustering allowed us to group molecules into clusters and simultaneously detect the descriptor subspaces that characterize each cluster. This technique allowed us to analyse the structure of a dataset by examining the similarity between groups of objects described in different subspaces, demonstrating its ability to study high-dimensional data..

RESUMEN

Se aplicó el algoritmo «clustering subespacial» para analizar moléculas con propiedades pesticidas. Se analizaron 1509 moléculas de la base de datos PPDB (Pesticides Properties DataBase - AERU Hertfordshire University); 1005 moléculas presentaron datos experimentales de presión de vapor. Los descriptores se calcularon con el programa PaDEL-Descriptor (v. 2.20) con un total de 14464 descriptores moleculares 0D-2D y tipos de huellas. El clustering de subespacios nos permitió agrupar moléculas en clusters y, simultáneamente, detectar los subespacios de descriptores que caracterizan a cada cluster. Esta técnica nos permitió analizar la estructura de un conjunto de datos examinando la similitud entre grupos de objetos descritos en diferentes subespacios, demostrando su capacidad para estudiar datos de alta dimensión

Revista Boliviana de Química, 2024, 41, 155-172
ISSN 0250-5460, Rev. Bol. Quim. *Paper edition*
ISSN 2078-3949, Rev. boliv. quim. *e-edition*, *Sep-Dec*
30 diciembre 2024, https://doi.org/10.34098/2078-3949.41.3.4

https://bolivianchemistryjournaliiq.umsa.bo; https://bolivianchemistryjournal.org

[1]Received December 2, 2024, accepted December 13, 2024, published December 30, 2024. *Mail to: eyapu@fcpn.edu.bo

INTRODUCCIÓN

Los pesticidas son muy utilizados para el control de plagas y enfermedades en la agricultura. Sin embargo, el uso indiscriminado y la manipulación inadecuada pueden desencadenar problemas en la salud humana y el ambiente[1]. La toxicidad de los pesticidas constituye un peligro para las personas que se exponen a dichas sustancias debido a su volatividad, unos más que otros[2 3], y afecta directamente a la salud de las personas principalmente por inhalación. Es importante conocer la presión de vapor de los pesticidas, puesto que una sustancia con alta presión de vapor resulta altamente toxica[4 5 6 7].

Existen pesticidas cuyos valores de presión de vapor no han sido determinadas experimentalmente por varias razones, entre ellas, su alta toxicidad, el alto costo económico y las condiciones de operabilidad[8]. Sin embargo, ésta puede estimarse mediante la obtención de modelos matemáticos predictivos QSPR (relaciones cuantitativas estructura propiedad)[9]. Es necesario subclasificar los pesticidas por similitud estructural para obtener modelos más precisos, trabajar con grupos de pesticidas reduce la complejidad y la dimensionalidad del problema y contribuye a una comprensión más profunda de los factores que influyen en la presión de vapor y mejora la interpretación de los modelos[10]. Comprender las similitudes estructurales permite el diseño racional de nuevos pesticidas. La subclasificación ayuda a identificar patrones estructurales que están asociados con propiedades específicas e induce al desarrollo de compuestos más seguros y efectivos; además, mejora la comprensión de los riesgos asociados con la exposición a estos compuestos químicos.

En el presente trabajo se emplea un conjunto de datos de la presión de vapor de 1005 pesticidas estructuralmente diversos. Para la subclasificación en función de la similitud estructural y presión de vapor, se utilizan los datos de presión de vapor experimentales conocidos a 20°C de una base de datos de propiedades de pesticidas (PPDB) disponible en la literatura.

Una etapa crucial en el análisis de moléculas químicas, mediante técnicas de quemo-informática es el cálculo de descriptores moleculares. Estas metodologías permiten representar cada estructura molecular como un vector numérico en un espacio de descriptores. Existen distintos programas que permiten calcular descriptores moleculares.

Entre los más conocidos se tiene al *PaDEL-Descriptor* (v. 2.20) que es una herramienta de software de código abierto que proporciona una amplia gama de descriptores moleculares, incluyendo descriptores 1D, 2D y 3D; se utiliza ampliamente en el campo de la quemo-informática y es fácil de usar. *ISIDA/Fragmentor* es un descriptor que cuenta tipos de átomos y fragmentos de subestructuras lineales que van desde 1 a 5 átomos de longitud. *RDKit* es una colección de herramientas de código abierto para la quemo-informática y la minería de datos químicos**.** *Dragon* **es** un software comercial que ofrece una amplia gama de descriptores moleculares, incluyendo descriptores 1D, 2D y 3D, es conocido por su amplia variedad de descriptores y su capacidad para manejar grandes conjuntos de datos**.**

En el presente trabajo se utilizó el programa *PaDEL*-Descriptor por ser una herramienta versátil y accesible que permite a los investigadores calcular una amplia gama de descriptores moleculares para análisis químico y modelado predictivo[11].

Las estructuras moleculares son elementos complejos, y los diferentes programas ya mencionados hacen uso de una gran cantidad de descriptores para poder caracterizar cada molécula de una manera más completa. Sin embargo, el análisis de datos descritos en espacios de múltiples dimensiones tiende a engendrar diferentes problemas que llevan el nombre de “maldición de la dimensión”[12 13]. En este contexto, uno de los problemas más importantes es el de la perdida de utilidad de diferentes medidas de distancia. En efecto, si uno desea caracterizar la estructura de un conjunto de datos mediante el cálculo de las distancias entre distintos pares de objetos, la tendencia general es que la diferencia entre dichas distancias tiende a disminuir cuando aumenta la dimensionalidad del espacio de descriptores. En este caso, la calidad de técnicas clásicas de análisis de datos, como el agrupamiento de objetos o clustering, tienden a disminuir, y diferentes métodos han sido propuestos para poder contrarrestar los efectos de la maldición de la dimensión.

El subspace clustering es una técnica de data-mining que tiene por finalidad agrupar objetos descritos numéricamente en un espacio de descriptores para formar grupos llamados clústeres de objetos similares, y para detectar simultáneamente los subespacios de descriptores que caracterizan cada clúster. Esta técnica permite analizar la estructura de un conjunto de datos examinando la similitud entre grupos de objetos descritos en diferentes subespacios[14]. Diferentes estudios han mostrado que el subspace clustering está particularmente adaptado al estudio de datos en altas dimensiones[15], y por ende es un excelente candidato para analizar descriptores moleculares. En

efecto, estas técnicas han sido empleadas con éxito para el análisis de moléculas químicas según propiedades fisicoquímicas como la adsorción[16].

En este artículo aplicamos con éxito SubCMedians[17], un algoritmo de subspace clustering que ha mostrado buenos resultados con respecto a otros algoritmos para agrupar descriptores de diferentes estructuras moleculares pesticidas y analizar la volatilidad de las moléculas contenidas en cada clúster. El resto del artículo se organiza de la siguiente manera: En la sección siguiente se describe el algoritmo SubCMedians así como el conjunto de datos analizados, su preprocesamiento y el protocolo experimental que hemos empleado. En la subsiguiente sección, se describen los resultados obtenidos. Finalmente, se concluye el trabajo resumiendo los puntos centrales del mismo.

EXPERIMENTAL

Materiales y Métodos

El trabajo se inició con 1509 moléculas de pesticidas en formato MDL mol (V2000), mismas que fueron cambiadas de formato de archivo con el programa Open Babel para Windows[18]. Del total de moléculas, 1005 pesticidas cuentan con valores de presión de vapor y son estructuralmente, diversos[19]. Los datos de presión de vapor se han obtenido de la base de datos PPDB, que muestra identidad química, fisicoquímica, de salud humana, ecotoxicológicos de plaguicidas y otros. Ha sido desarrollado por la Unidad de Investigación de Agricultura y Medio Ambiente (AERU) de la Universidad de Hertfordshire para una variedad de usuarios finales.

Se puede apreciar que las características del conjunto de datos involucran estructuras muy heterogéneas con composición química variada incluyendo una amplia variedad de elementos químicos en moléculas orgánicas y sales de diferente tamaño. Los compuestos no considerados en el presente análisis son: compuestos poliméricos, como, mancozeb, maneb, metiram. Para el caso de mezclas de isómeros que conducen a una misma estructura topológica, se considera solo uno de ellos, es decir, diclobutrazol.

Las estructuras de los pesticidas que no contaban con su canonical smile se dibujaron con el programa ACDLabs/ChemSketch[20] en formato MDL mol (V2000). Los descriptores moleculares se calcularon utilizando PaDEL, un software de código abierto y disponible gratuitamente. El Laboratorio de Exploración de Datos Farmacéuticos (PaDEL-Descriptor (v. 2.20)[21] calcula 14464 descriptores moleculares 0D-2D y tipos de huellas dactilares. Se analizaron los 14464 descriptores moleculares para eliminar descriptores colineales con información redundante.

Pre-procesamiento de los datos

Antes de aplicar la técnica de subspace clustering al conjunto de datos presentado anteriormente, se procedió a aplicar diferentes técnicas de pre-procesamiento: En primer lugar, se procedió a filtrar los descriptores para los cuales la mayoría de las estructuras moleculares no tenían valores. El programa de descriptores moleculares empleado organiza en orden creciente, los descriptores en función del número de valores faltantes que poseen, como se puede evidenciar en la figura 1. En este trabajo se eliminaron todos aquellos descriptores para los cuales menos del 20% de moléculas tenían un valor. Después de esta etapa, solo se conservaron los primeros 126 descriptores.

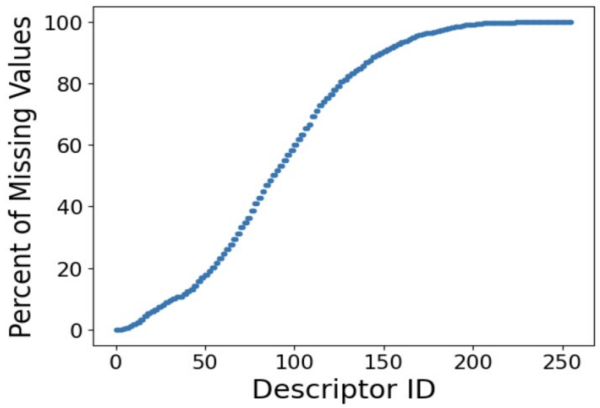


Figura 1: Porcentaje de valores faltantes en cada descriptor. Se filtraron todos los descriptores con más de 80% de valores faltantes.

Posteriormente se procedió a rellenar los valores faltantes con ceros, en este punto, es importante resaltar que las coordenadas de las diferentes estructuras moleculares en el espacio de descriptores corresponden únicamente a valores reales superiores o iguales a cero. Posteriormente se aplicó una transformación logarítmica a los datos, empleando la función $logScale: \Re^+ \rightarrow \Re^+$ tal que $\forall x \in \Re^+$, $logScale(x) = log_{10}(x+1)$. Finalmente se aplicó una normalizacion $Zscore$ a estos datos, con el fin de que en cada dimensión, las coordenadas de los datos tengan una media igual a cero y una dispersión estándar igual a uno: $Zscore: \Re \rightarrow \Re^+$ tal que $\forall x \in \Re, Zscore(x) = (x - \mu)/\sigma$. Donde μ y σ representan respectivamente la media y la desviación estándar en la dimensión considerada.

Subspace clustering

SubCMedians

SubCMedians[17] tiene por objetivo elaborar un modelo de subspace clustering basado en medianas descritas en subespacios específicos. El uso de medianas provee a este método propiedades interesantes, como ser una intrínseca robustez al ruido y a valores anómalos. Por otra parte, la utilización de subespacios le permite hacer frente a los efectos de la maldición de la dimensión.

El algoritmo recibe como entrada un conjunto de objetos $S = \{s \in \Re^m\}$ descritos en un espacio de m dimensiones, y sea $D = \{d_1, \ldots, d_m\}$ el conjunto de dimensiones, y tiene por objetivo elaborar un modelo de subspace clustering, definido como un conjunto de medianas candidatas M, en el que cada mediana $m_i \in M$ está asociada a un subespacio $D_i \subseteq D$. Cada objeto $s \in S$ es asociado al clúster C_i que posee la mediana candidata m_i más cercana, es decir a la que minimiza la distancia $dist(s, m_i) = \sum_{d \in D_i} |s_d - m_{i,d}| + \sum_{d \in D \setminus D_i} |s_d - \mu_d|$, donde $m_{i,d}$ representa la coordenada de la mediana m_i en la dimensión d, y μ_d es la media de las coordenadas de todos los objetos en S en la dimension d (en el caso de un conjunto de datos normalizado por la tecnica de z-score $\mu_d = 0, \forall\, d \in D$). El Error Absoluto de un modelo M con respecto a un objeto s se define como $AE(s, M) = Min_{m \in M}\, dist(s, m)$ la distancia entre el objeto y la mediana más cercana, y por consiguiente la Suma de Errores Absolutos de un conjunto de datos S con respecto a un modelo M se define como $SAE(S, M) = \sum_{s \in S} AE(s, M)$. Por otra parte, el tamaño de un modelo M, denominado $Size(M) = \sum_i |D_i|$, es la suma de las dimensionalidades de todos los subespacios asociados a las medianas en M, y corresponde al nivel de detalle capturado por el modelo. SubCMedians tiene por objetivo construir el conjunto de medianas M que minimice la SAE y tal $Size(M) \leq SD_{max}$, donde SD_{max} es un parámetro que caracteriza el maximo nivel de detalles que debe poseer el modelo M.

Con el fin de reducir los tiempos de cálculo SubCMedians emplea una muestra dinámica del conjunto de datos $\hat{S}$, de tamaño N, para estimar la SAE del modelo M. El algoritmo inicializa $\hat{S}$ tomando N objetos aleatorios en S, y a cada iteracion, un objeto elegido al azar en $\hat{S}$ es substituido por otro elemento extraido uniformemente en $S \setminus \hat{S}$, con el fin de modificar la muestra. La SAE del modelo se actualiza de manera incremental substrayendo la AE del objeto eliminado y adicionando la AE del nuevo objeto. Si el error actualizado es superior al anterior, se procede a realizar un intento de optimización del modelo actual, en caso contrario, el algoritmo no realiza ninguna acción.

Con el fin de minimizar la SAE manteniendo la restricción $Size(M) \leq SD_{max}$, SubCMedians actualiza iterativamente un modelo M, combinando coordenadas de los objetos del conjunto de datos, mediante una tecnica de optimización estocástica de hill-climbing durante Nb_{iter} iteraciones. SubCMedians comienza con un modelo inicial, el modelo vacío (sin medianas) $M_\emptyset$ cuyo error absoluto a un objeto s es $AE(s, M_\emptyset) = \sum_{d \in D_i} |s_d - \mu_d|$. A cada iteración, a partir de un modelo M SubCMedians se produce una variante aleatoria M', la cual reemplaza al modelo actual M en el caso en el que $SAE(S, M') \leq SAE(S, M)$. Con el fin de generar una variante M', SubCMedians emplea una estrategia basada en ponderaciones para guiar su búsqueda local hacia subespacios prometedores. Sea $W_{i,d}$ el peso asociado a la dimensión d para la mediana m_i, y sea $w = \sum_{i,d} W_{i,d}$ el peso total del modelo actual. Si $w \geq SD_{max}$ el algoritmo procede a elegir al azar un par de indices (i, d) con una probabilidad proporcional a $W_{i,d}$, se reduce el peso correspondiente $W_{i,d} \leftarrow W_{i,d} - 1$ y si $W_{i,d} = 0$ entonces se elimina la coordenada correspondiente en la mediana seleccionada $m_{i,d} \leftarrow 0$. A continuación, SubCMedians selecciona aleatoriamente un objeto $s \in \hat{S}$ y una dimensión $d' \in D$, y selecciona una mediana c con una probabilidad proporcional a su suma de pesos correspondientes $\sum_d W_{c,d}$, o genera una nueva mediana con una probabilidad igual a $1/w$, finalmente se modifica la coordinada de la mediana en cuestión $m_{c,d'} \leftarrow s_{d'}$ y se incrementa el peso asociado $W_{c,d'} \leftarrow W_{c,d'} + 1$

Una vez que el algoritmo ha completado Nb_{iter} iteraciones, se genera un modelo de subspace clustering asociando cada objeto $s \in S$ al clúster C_i que posee la mediana candidata m_i más cercana (la que minimiza la AE). Si varias

medianas minimizan la AE, entonces se elige una de forma no determinista. Si una mediana no posee ningún objeto asociado, entonces no da lugar a un clúster.

Modelo de fusión

Con el fin de tener un modelo más robusto y de mejor calidad, se procedió a utilizar SubCMedians para producir un conjunto de 100 modelos de subspace clustering independientes $\{M_1, M_2, \ldots, M_{100}\}$. En cada ejecución se utilizaron los mismos parámetros $SD_{max} = 300$, $N = |S|$, y $Nb_{iter} = 10{,}000$. Posteriormente se fusionaron los diferentes modelos en uno solo $M = \cup_{i=1}^{100} M_i$, se procedió a formar clústeres agrupando los objetos en el conjunto de datos entorno a las medianas de M y se contabilizó el número de objetos $|C_i|$ en cada clúster C_i asociado a cada mediana $m_i \in M$. Finalmente se conservaron las k medianas con mayor número de objetos asociados para formar un nuevo modelo M^* que se optimizo durante 10,000 nuevas iteraciones. Este procedimiento permite una mejor exploración del espacio de modelos posibles, seleccionando las k medianas que mejor describen los datos, esto es entre un conjunto de medianas generadas por 100 modelos independientes. Con el fin de determinar el número k de medianas que debían conservarse, representamos el número de objetos contenidos en cada clúster en orden decreciente, y utilizamos el método de codo para determinar el número de medianas a partir del cual la cantidad de objetos en los clústeres se estanca con valores más bajos. En la figura 1 podemos apreciar que a partir de 13 medianas los clústeres tienden a poseer menos objetos y a ser por ende menos representativos. Por consiguiente, decidimos conservar las $k = 13$ medianas con más objetos asociados para crear el modelo de fusión.

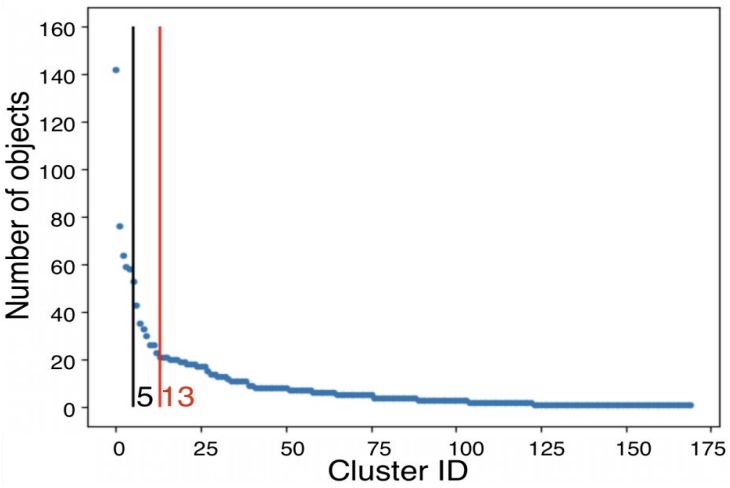


Figura 2: Numero de objetos asociados a cada clúster, en el modelo que fusióna 100 modelos independientes de SubCMedians. El método del codo sugiere conservar los 13 clústeres más grandes con el fin de formar el nuevo modelo de subspace clustering.

Visualización

Para visualizar la estructura del conjunto de datos analizados, empleamos el algoritmo T-distributed stochastic neighbor embedding (t-SNE)[22]. Este algoritmo de reducción de dimensionalidad permite proyectar de manera no lineal, datos de alta dimensión a un plano bidimensional, de tal manera que objetos similares en el espacio de alta dimensionalidad tienden a ser representados mediante puntos cercanos en el plano bidimensional, mientras que objetos disímiles tienden a representarse distantes entre sí en el plano. Antes de aplicar este método, como lo sugieren sus autores, empleamos el método de reducción de dimensionalidad lineal llamado Análisis en Componentes Principales o ACP[23] para proyectar primero el espacio de descriptores en un espacio con 50 dimensiones, y así facilitar el trabajo del algoritmo t-SNE. En la práctica utilizamos la implementación de dicho algoritmo presente en la librería de Python scikit-learn[24], estableciendo una perplejidad igual a 70 y fijando los demás parámetros a sus valores por defecto[2]

Clasificación y SHAP-values

[2] Los demás parámetros de t-SNE no tienen un impacto importante en el resultado, y se evaluaron diferentes valores de perplejidad para optimizar la representación final.

Para poder analizar las dimensiones más importantes que caracterizan cada clúster, se procedió a entrenar un modelo de Random Forest[25], que a partir de los objetos descritos en el espacio de descriptores químicos predice el clúster al cual el objeto pertenece. En la práctica se utilizó la implementación del algoritmo de Random Forest disponible en la librería scikit-learn de Python[24], y se fijó a 1,000 el número de árboles de clasificación, con el fin de tener una mejor predicción y contrarrestar posibles problemas de sobre entrenamiento, mientras que los demás parámetros fueron establecidos a sus valores estándar. Se evaluó la calidad de este modelo por medio de una estrategia de validación cruzada estratificada y repetida, el conjunto de datos fue dividido en 10 subconjuntos de mismo tamaño y con la misma proporción de objetos de cada clúster, cada uno de ellos se empleó a su vez como conjunto de validación, mientras que los restantes subconjuntos se emplearon como conjunto de entrenamiento; este procedimiento se repitió un total de 5 veces, lo cual generó un total de 50 estimaciones de la calidad en diferentes subconjuntos de datos. En este trabajo utilizamos la medida $Accuracy$, que contabiliza la proporción de objetos correctamente, para evaluar la calidad del modelo.

Finalmente, para poder analizar la importancia de cada descripción en la clasificación de cada clúster, procedimos a calcular las llamadas SHAP-values[26], que permiten cuantificar el grado de importancia de cada descriptor en la tarea predictiva de un modelo dado. En la práctica, dado que empleamos un modelo de Random Forest, basado en arboles de clasificación, empleamos el algoritmo TreeSHAP con el fin de estimar dichos valores, de manera más rápida.

RESULTADOS Y DISCUSIÓN

Como se puede observar en la figura 2, el número de iteraciones $Nb_{iter} = 10{,}000$ fue suficiente para que los 100 modelos individuales llegaran a converger hacia niveles similares en términos de SAE, entorno a valores cercanos a $90{,}000$. Por otra parte, en la misma figura podemos apreciar que la fusión de modelos produjo un modelo con una SAE muy inferior, por debajo de $80{,}000$. Esto sugiere que el modelo fusión describe de manera más precisa los datos. Por otra parte, el modelo de fusión formado por las 13 medianas más importantes de los 100 modelos independientes, comenzó directamente con un valor muy bajo de SAE (dado que fue inicializado con las mejores medianas de los modelos independientes finales), y solo optimizo ligeramente su SAE. Después de $10{,}000$ iteraciones, el modelo fusión conservó sus 13 medianas, y los clústeres asociados lograron capturar un número bastante variable de objetos que van de 22 objetos para el clúster más pequeño (clúster 12), a 318 objetos para el clúster más grande (clúster 3).

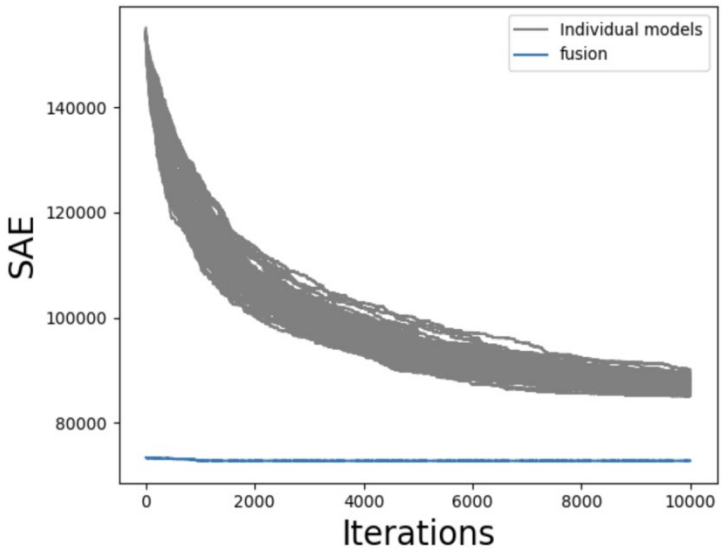


Figura 3: Evolución de la suma de errores absolutos (SAE) de 100 modelos independientes generados por SubCMedians (gris) y evolución del modelo de subspace clustering de fusión (azul).

Revista Boliviana de Química, 2024, Vol. 41, N°3, 155-172
https://doi.org/10.34098/2078-3949.41.3.4

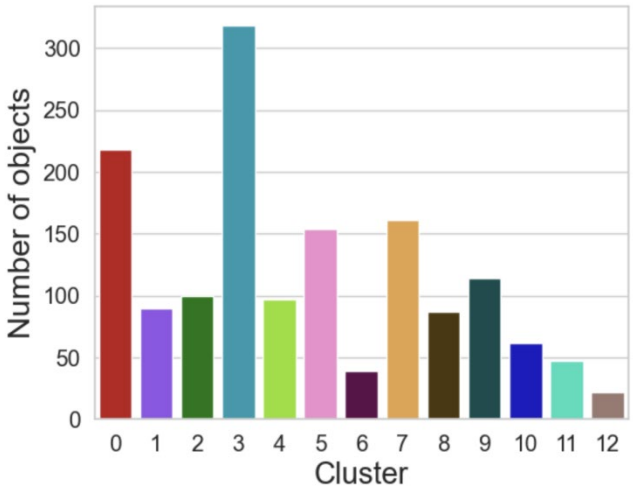


Figura 4: Cantidad de objetos pertenecientes a cada clúster en el modelo de subspace clustering de fusión. El número de objetos por clúster varía de manera importante (de cerca de 20 objetos para el clúster a más de 300 para el clúster más grande).

La representación en 2D generada por el algoritmo t-SNE muestra una clara cohesión de los clústeres formados por SubCMedians, como puede apreciarse en la figura 4. Por otra parte, en esta figura se evidencia que algunos clústeres (por ejemplo el clúster 0) poseen un mayor número de moléculas con volatilidad alta ($log_{10}(P_v + 1) > 5$), las cuales se ven representadas con círculos de mayor diámetro en la imagen 4. Esta observación se ve corroborada por el boxenplot representado en la figura 5 y la tabla ilustrada en la figura 6, que representa la distribución de la volatilidad en escala logarítmica ($log_{10}(P_v + 1)$) de las moleculas contenidas en cada clúster. Los clústeres 0, 7, 5 y 1 poseen claramente más estructuras moleculares con volatilidad elevada, mientras que los clústeres 3, 8, 6 y 10 están claramente poblados por moléculas con volatilidad baja.

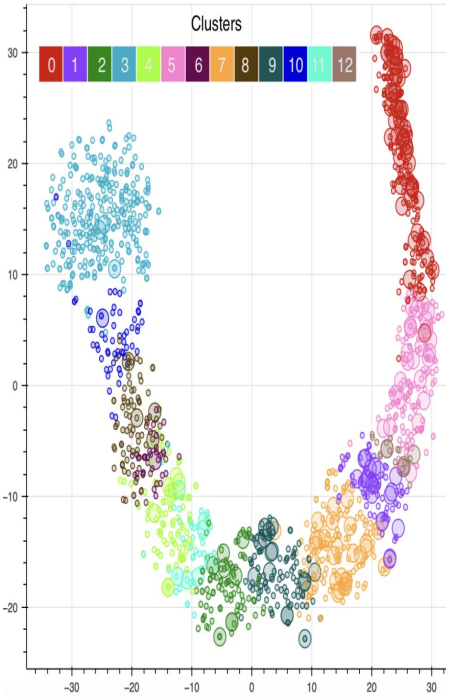


Figura 5: Proyección 2D generada por el método t-SNE de cada objeto. El color de cada punto evidencia el clúster de pertenencia del objeto representado. Los puntos grandes representan moléculas volátiles.

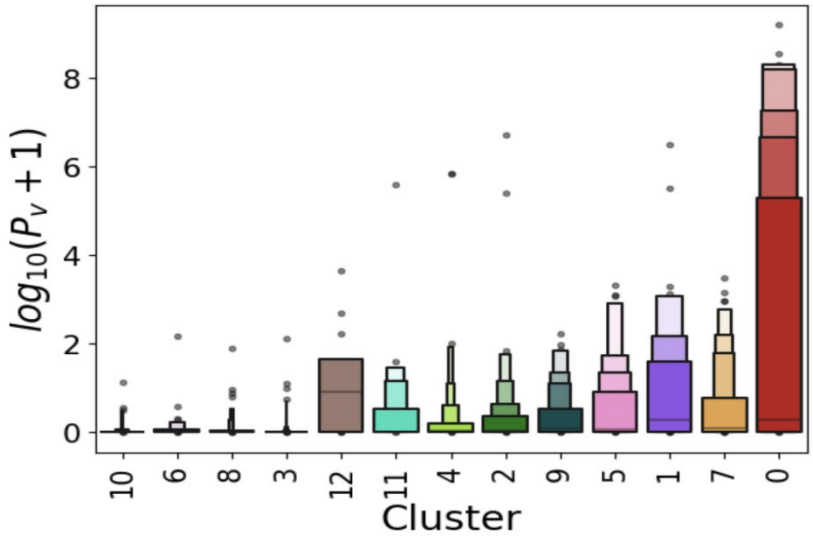


Figura 6: Distribución de la volatilidad de las moléculas contenidas en cada cluster.

Figura 7: Cantidad de moléculas de cada clase de volatilidad, i.e., alta (high), baja (low) o sin datos (NaN), para cada uno de los cluster.

El modelo de Random Forest que fue entrenado para poder clasificar las moléculas en función de sus clústeres a partir de descriptores químicos, ha demostrado buenos valores de *Accuracy* (cuya media sobrepasa 0.93), como puede evidenciarse en el boxplot representado en la figura 8.

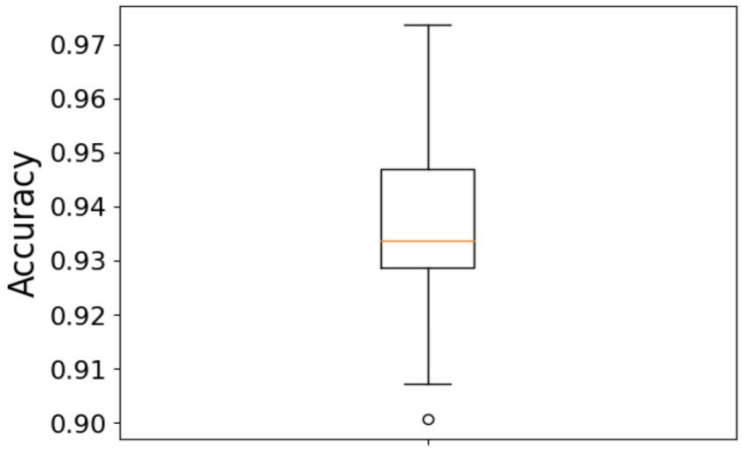


Figura 8: Distribución de la medida de calidad (Accuracy) del modelo Random Forest que tiene por objetivo clasificar las moléculas en función de su clúster de pertenencia

Esto demuestra una vez más la cohesión de los clústeres generados pro SubCMedians, lo cual hace que la tarea de clasificación asociada sea relativamente sencilla. Este modelo no se apoya solo en unos cuantos descriptores y tiene

más bien tendencia a emplear los diferentes descriptores con una importancia similar como puede evidenciarse en el gráfico de barras representado en la figura 9, que ilustra la medida de importancia (Feature Importance) asociada a los 20 descriptores más importantes para el modelo de Random Forest en la tarea de clasificación. La suma de las medidas de Feature Importance atribuidas a cada descriptor es igual a uno, y por ende, de ser igualmente importantes, los 126 descriptores del conjunto de datos deberían exhibir un Feature Importance igual a $1/126 \approx 0.0079$. Sin embargo, es importante notar que la importancia de cada descriptor para clasificar las moléculas depende de cada clúster.

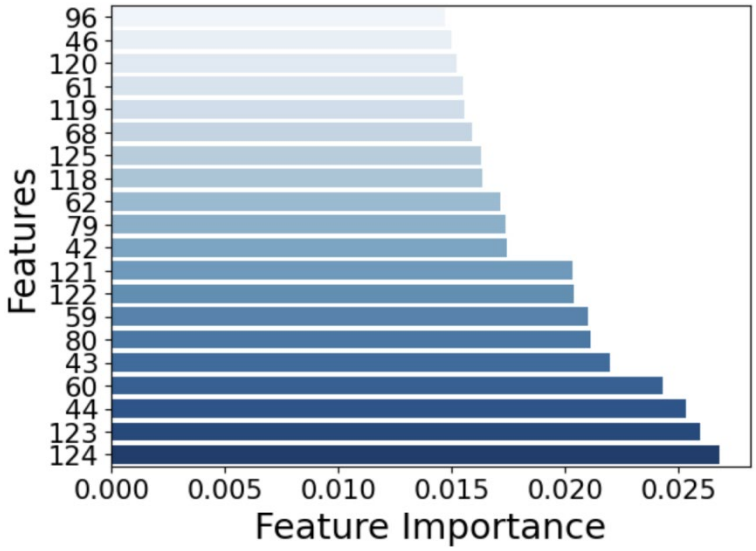
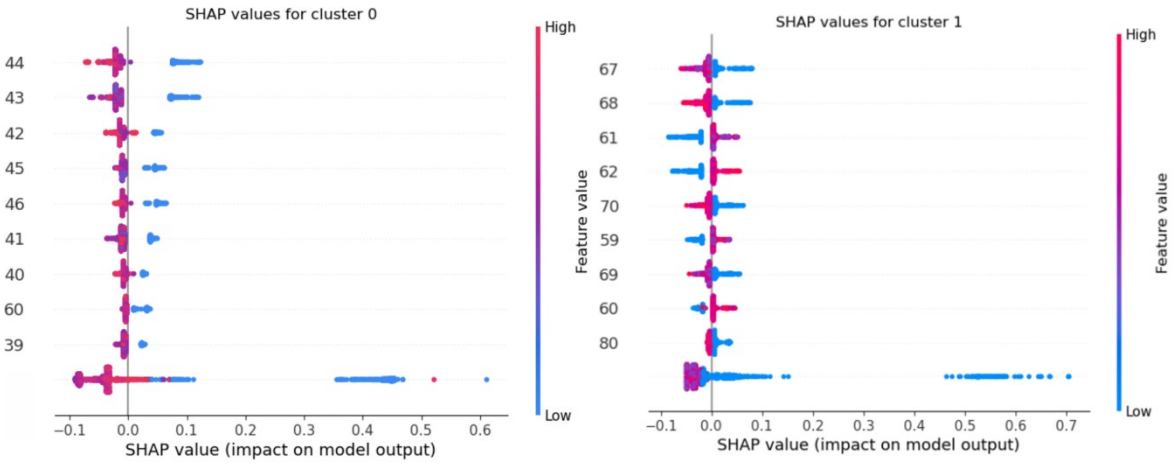
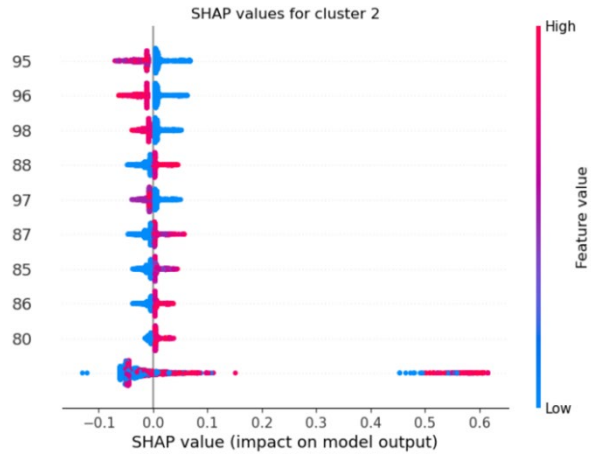


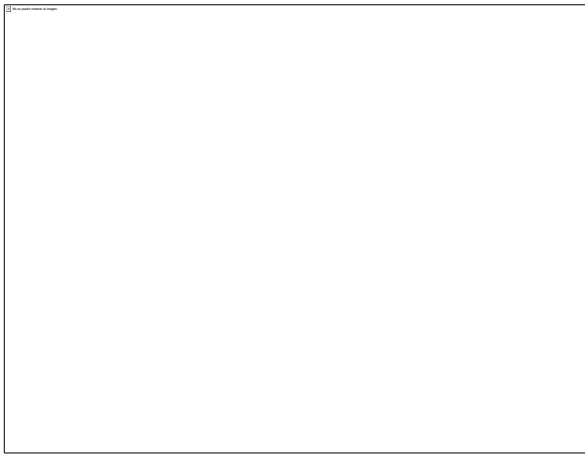
Figura 9: Veinte descriptores más importantes para el modelo de Random Forest y sus valores de Feature Importance correspondientes.

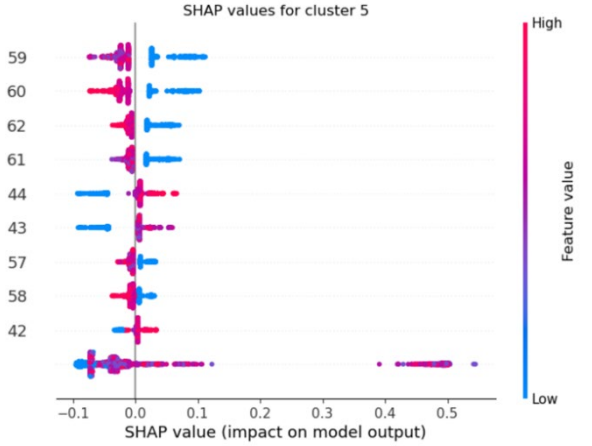
En efecto, por ejemplo, el clúster 0, enriquecido en moléculas con alta volatilidad (Tabla 1), están caracterizado por los descriptores 39, 40, 41, 42, 43, 44, 45, 46 y 60 (los valores bajos en estos descriptores tienen a caracterizar los objetos que pertenecen en este clúster). Mientras que el clúster 3, enriquecido en moléculas con baja o nula volatilidad, esta caracterizado por los descriptores 117, 118, 119, 120, 121, 122, 124 y 125 (los valores altos en estos descriptores tienden a caracterizar los objetos que pertenecen en este clúster).



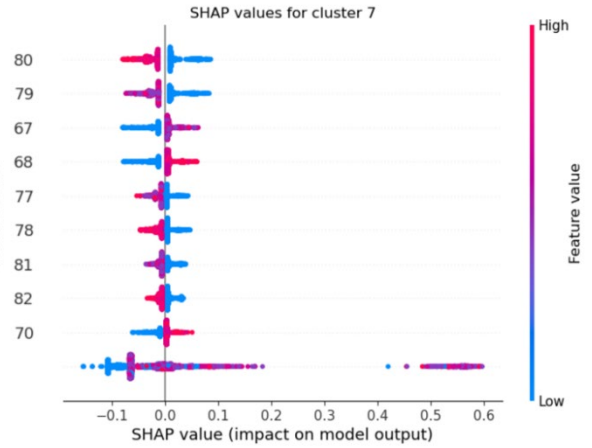

SHAP values for cluster 2
95
96
98
88
97
87
85
86
80
−0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6
SHAP value (impact on model output)
Feature value
High
Low


SHAP values for cluster 3
124
123
121
122
125
119
120
118
117
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
Feature value
High
Low




SHAP values for cluster 5
59
60
62
61
44
43
57
58
42
−0.1 0.0 0.1 0.2 0.3 0.4 0.5
SHAP value (impact on model output)
Feature value
High
Low


SHAP values for cluster 6
106
105
107
108
109
110
116
115
113
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
SHAP value (impact on model output)
Feature value
High
Low


SHAP values for cluster 7
80
79
67
68
77
78
81
82
70
−0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6
SHAP value (impact on model output)
Feature value
High
Low

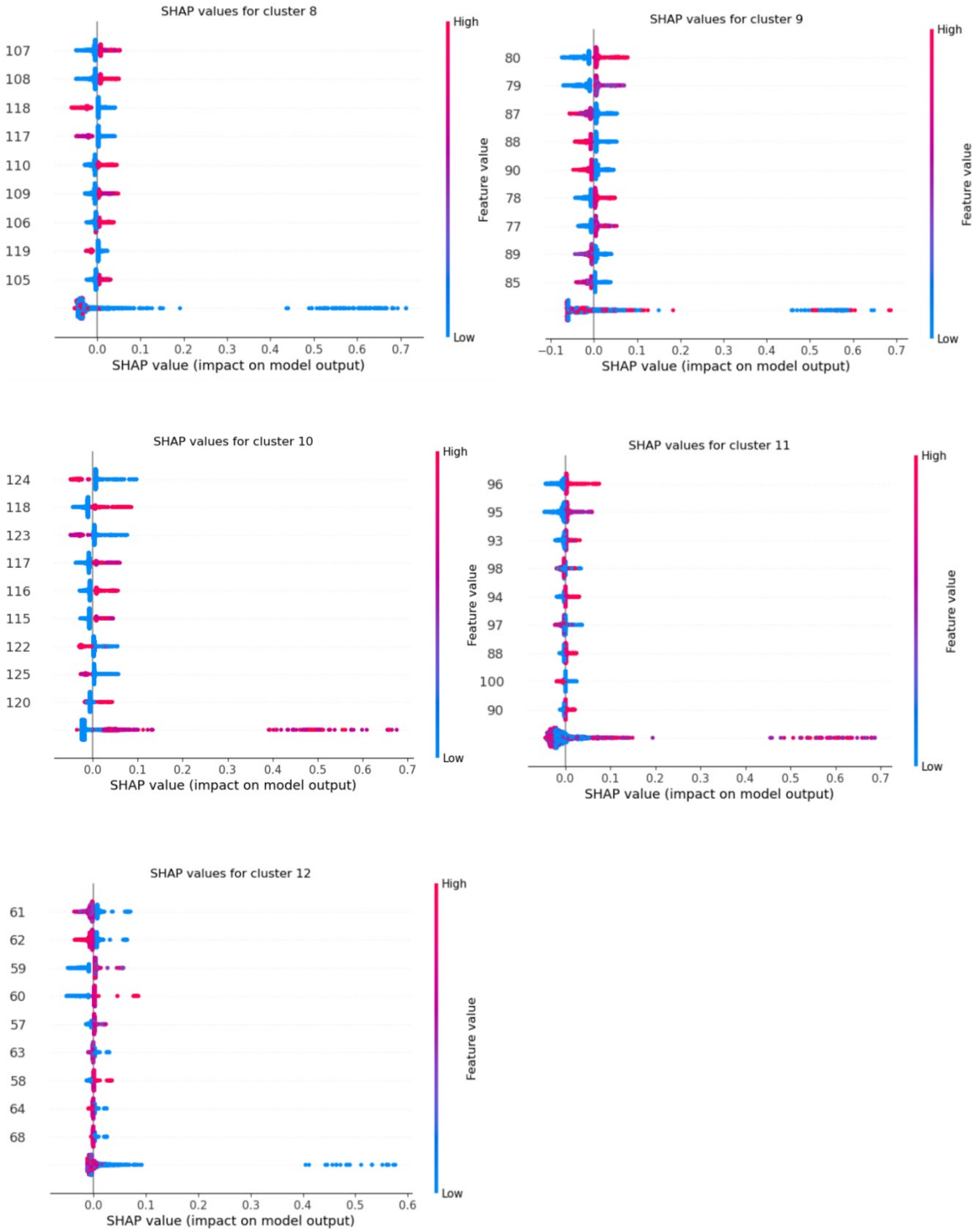


Figura 10: SHAP-values (importancia) de los nueve descriptores más importantes la predicción de cada uno de los clústeres por parte del modelo de Random Forest.

Dada la gran heterogeneidad de las estructuras moleculares (Tabla 1), los grupos clasificados tienen ciertas tendencias con características genéricas, por ejemplo, en el grupo 0, se encuentran compuestos de hidrocarburos halogenados, fenoles, ácidos y derivados carboxílicos, algunas sales orgánicas, alcoholes y algunos nitrogenados con

pesos moleculares entre 70 y 400 g/mol. En el grupo 1, se encuentran compuestos orgánicos clorados, nitrogenados, herbicidas, ácidos, ésteres y otros compuestos, sus pesos moleculares varían entre 30 y 500 g/mol. En el grupo 2, se encuentran carbamatos, organofosforados, piretroides y otros, con pesos moleculares entre 170 y 500 g/mol. En el grupo tres, Piretroides, herbicidas, fungicidas con pesos moleculares entre 250 y 550 g/mol. Los otros grupos también son clasificados casi del mismo modo.

CONCLUSIONES

En este estudio se aplicó SubCMedians, un algoritmo de subspace clustering para analizar un conjunto de datos de estructuras moleculares aplicadas en sustancias utilizadas en la agricultura como pesticidas. El modelo de clustering obtenido consta de 13 clústeres (Tabla 1), cada uno de ellos posee diferentes tipos de moléculas con distintas propiedades. Por ejemplo, se ha podido evidenciar que algunos de estos clústeres contenían moléculas con volatilidad más elevada, mientras que otros solo contenían moléculas con baja volatilidad. Por otra parte, cada clúster está caracterizado por un subconjunto de descriptores químicos característicos. Se ha intentado dar una explicación de los clústers desde un punto de vista de la estructura molecular y su composición, sin embargo, dada la gran heterogeneidad de las mismas, los grupos clasificados muestran una cierta tendencia de carácter genérico a agruparse en tipos de compuestos y estructuras con características similares. Finalmente, los resultados del presente trabajo, abren la posibilidad de establecer modelos cuantitativos de relaciones estructura-propiedad a objeto de predecir el valor de la presión de vapor de los compuestos en cada grupo subclasificado mediante modelos matemático-estadísticos o de redes neuronales para todos aquellos compuestos que no cuentan con datos experimentales.

Tabla 1

C_i	Moléculas
0	1,2-dibromoethane, 1,2-dichloropropane, cis 1,3-dichloropropene, trans 1,3-dichloropropene, 1,4-dimethylnaphthalene, 1-decanol, 1-methylcyclopropene, 2-(octylthio)ethanol, 2,4,5-trichlorophenol, 2,4-dimethylphenol, 24-epibrassinolide, 2-aminobutane, 2-phenylphenol, 4-aminopyridine, 8-hydroxyquinoline, acetic acid, acrolein, acrylonitrile, allyl alcohol, alpha-chlorohydrin, aluminium ammonium sulphate, aluminium phosphide, aluminium sulphate, amitrole, ammonium acetate, ammonium carbonate, ammonium hydroxide, ammonium sulphamate, ammonium thiocyanate, nthracene oil, anthraquinone, benzoic acid, biphenyl, bromomethane, butanethiol, calcium phosphide, camphechlor, carbon dioxide, carbon disulphide, carbon tetrachloride, chloral hydrate, chlordecone, chlormequat chloride, chloroneb, chloropicrin, choline chloride, copper (1) oxide, copper II acetate, copper II carbonate, copper II chloride, copper II hydroxide, copper oxychloride, copper sulphate, cyanamide, cyhexatin, cyromazine, dalapon, dalapon-sodium, dazomet, DDD, DDT, defenuron, dibromochloropropane, dichlobenil, dicyclopentadiene, didecyldimethylammonium chloride, dimefox, dimethyl disulfide, dimexano, diphenylamine, disodium octaborate tetrahydrate, disodium phosphonate, endothal, ethanethiol, ethephon, ethylene dichloride, fentin chloride, fentin hydroxide, fenuron, ferbam, fluoroacetamide, formaldehyde, furfural, glutaraldehyde, hexachlorobenzene, hydrogen peroxide, hymexazol, iodomethane, iron sulphate, isobutyric acid, lindane, magnesium phosphide, maleic hydrazide, mepiquat chloride, merphos, metaldehyde, metam-sodium, methyl isothiocyanate, methylarsonic acid, nabam, naphthalene, N-methylneodecanamide, N-nitrosodimethylamine, paraquat dichloride, paraquat, pentachlorophenol, peroxyacetic acid, phenylmercury chloride, phosphine, piproctanyl bromide, prohexadione, quartz sand, silica, sodium carbonate, sodium chlorate, sodium chloride, sodium hypochlorite, sulfuryl fluoride, sulphur, sulphuric acid, TCA-sodium, thiocyclam, thiourea, thiram, zinc phosphide, zineb, 2-naphthyloxyacetic acid, alpha-hexachlorocyclohexane, silver thiosulphate, 2-hydrazinoethanol, 2-imidazolidone, 2-methoxyethylmercury chloride, aluminium silicate, ANTU, arsenic acid, arsenous oxide, azobenzene, barium carbonate, barium polysulphide, bis(2-chloroethyl)ether, bis(methylmercury) sulphate, bisthiosemi, bis-trichloromethyl sulfone, boric acid, cacodylic acid, calciferol, calcium acid methanearsonate, calcium arsenate, calcium carbide, calcium carbonate, calcium chloride, calcium cyanamide, calcium hydroxide, calcium phosphate, chloropon, chloroxylenol, cholecalciferol, copper acetoarsenite, copper naphthenate, cycloprate, cyperquat, DCIP, diammonium phosphate, dioctyldiethylenetriamine, disodium methanearsonate, ethanedial, ethylicin, ethylmercury bromide, ferric phosphate, ferric pyrophosphate, fosetyl, gliftor, hexachloroacetone, hexadecanoic acid,

	hexaflurate, lead arsenate, lime sulphur, limestone, mancopper, mepiquat, mercuric oxide, mercurous chloride, metam-potassium, metepa, methiotepa, methyl nonyl ketone, methylarsenic sulphide, methylene bisthiocyanate, mirex, monosodium methylarsonate, nickel bis(dimethyldithiocarbamate), p,p'-DDT, perthane, piproctanyl, polybutene, potassium bicarbonate, potassium iodide, potassium phosphonates, potassium thiocyanate, prohydrojasmon, propionic acid, sodium aluminium silicate, sodium arsenite, sodium hydrogen carbonate, sodium monochloroacetate, sodium monofluoroacetate, sodium pentaborate, sodium tetraborate pentahydrate, strobane, tiaojiean, triacontanol, trioxymethylene, urea sulphate, urea, zinc borate, zinc oxide, hexachlorocyclohexane, alpha-naphthylthiourea
1	2,4,5-trichlorophenoxyacetic acid, 2,4-D, 2,5-dichlorobenzoic acid methyl ester, allethrin, azocyclotin, bioallethrin, chlordimeform, chlorfenac, chlorotoluron, chlorthal-dimethyl, cyanophos, cycloate, demeton-S-methyl, dichlorprop, dichlorprop-P, dimetachlone, dipropetryn, DNOC, ethion, etofenprox, fenchlorphos, fenpiclonil, fentin acetate, fluometuron, fluorbenside, fosamine, fuberidazole, glyphosate, glyphosine, hexachlorophene, indanofan, kelevan, kinoprene, MCPA, mebenil, methoprene, metolcarb, nitrapyrin, phthalide, propachlor, propamocarb hydrochloride, propamocarb, propanil, propham, propoxur, secbumeton, siduron, S-methoprene, temephos, terallethrin, thiometon, tridiphane, trimethacarb, vernolate, XMC, xylylcarb, lavandulyl senecioate, 10,10'-oxybisphenoxarsine, 2,4-D-dimethylammonium, alorac, chlorfluazole, chlorfluren methyl, chlorflurenol, ciobutide, clofibric acid, clofop, cloprop, credazine, cyanatryn, cyclethrin, cypromid, demephion-S, demeton-O-methyl sulfone, diamidaphos, dibutyl phthalate, dimethametryn, dipymetitrone, ethanamine, fenaminosulf, glyphosate, isopropylamine salt, glyphosate, potassium salt, isobenzan, isopamphos, malonoben, parinol, propamidine, saisentong, thiosultap, thiosultap-disodium, thiosultap-monosodium
2	acephate, acetamiprid, acethion, acibenzolar-S-methyl, alachlor, aramite, asulam sodium, asulam, barban, benfluralin, bispyribac-sodium, bromoxynil butanoate, bromoxynil octanoate, butoxycarboxim, carbetamide, carpropamid, chlorthion, chlorthiophos, clomazone, clopyralid, coumatetralyl, cyhalofop, dehydroacetic acid, diafenthiuron, diclocymet, dimefluthrin, dimethachlor, dimethirimol, dinitramine, dinotefuran, ethofumesate, ethoxyquin, fenazaquin, fenbuconazole, fenhexamid, fenitrothion, fenpropathrin, flufenprox, fluothiuron, flurprimidol, indaziflam, ioxynil octanoate, ipconazole, mepronil, metalaxyl, metalaxyl-M, metconazole, methazole, metofluthrin, myclobutanil, orbencarb, permethrin, phenisopham, piperalin, prodiamine, profluralin, profluthrin, propargite, pyrifenox, quinalphos, quinoclamine, silafluofen, sulcotrione, sulfluramid, tebuconazole, tetramethrin, thiazafluron, thiodicarb, trifluralin, zarilamid, zoxamide, epsilon-metofluthrin, 1-(4-chloro-1,3-dihydro-1,3-dioxo-2H-isoindole-2-yl)-cyclohexanecarboxamide , 1-(4-chlorophenyl)-3-(2,6-dichlorobenzoyl)urea , allyxycarb, ametridione, amidochlor, biopermethrin, bromophos-ethyl, cispermethrin, cyprofuram, denatonium benzoate, difenopenten ethyl, dinex-diclexine, dioxathion, formparanate, heterophos, iodobonil, menadione, menazon, methasulfocarb, morfamquat dichloride, morfamquat, morphothion, phosnichlor, quinclorac-dimethylammonium, quinothion, thiocarboxime, thiophanate, transpermethrin
3	acrinathrin, afidopyropen, alanycarb, amidosulfuron, amisulbrom, azafenidin, azamethiphos, azimsulfuron, azinphos-ethyl, azinphos-methyl, azoxystrobin, benazolin ethyl, benfuracarb, benomyl, bensulfuron-methyl, bensulide, benthiavalicarb isopropyl, benzobicyclon, benzofenap, benzovindiflupyr, benzoximate, beta-cyfluthrin, bixafen, bromofenoxim, bupirimate, butafenacil, cafenstrole, carbosulfan, carfentrazone-ethyl, chlorantraniliprole, chlorfenapyr, chlorfluazuron, chlorimuron-ethyl, chlorsulfuron, cinidon-ethyl, cinosulfuron, clodinafop-propargyl, clofencet, cloransulam-methyl, cyantraniliprole, cyazofamid, cyclaniliprole, cyclosulfamuron, cyflufenamid, cyfluthrin, cyhalothrin, dialifos, diclosulam, difenoconazole, diflufenican, diflufenzopyr, diflumetorim, dimefuron, dimoxystrobin, diniconazole, ethaboxam, ethametsulfuron-methyl, ethiozin, ethoxysulfuron, fenamidone, fenoxaprop-ethyl, fenoxaprop-P-ethyl, fenpicoxamid, fenpyrazamine, fenpyroximate, fentrazamide, fipronil, flamprop-M-isopropyl, flazasulfuron, flocoumafen, florasulam,

	florpyrauxifen-benzyl, fluacrypyrim, fluazinam, fluazolate, flubendiamide, flucarbazone-sodium, flucetosulfuron, flucycloxuron, flufenacet, flufenoxuron, flufenpyr-ethyl, flumetsulam, flumiclorac-pentyl, flumioxazin, flumipropyn, fluopyram, fluoroglycofen, fluoxastrobin, flupoxam, flupyrsulfuron-methyl-sodium, fluquinconazole, fluroxypyr, fluroxypyr-meptyl, flusulfamide, fluthiacet methyl, flutianil, fluvalinate, fluxapyroxad, fomesafen, foramsulfuron, fosmethilan, fosthiazate, furametpyr, furconazole-cis, gamma-cyhalothrin, griseofulvin, halauxifen-methyl, halosulfuron-methyl, haloxyfop-etotyl, haloxyfop-P-methyl, hexaflumuron, imazamox, imazapic, imazapyr, imazaquin, imazethapyr, imazosulfuron, imibenconazole, imicyafos, indoxacarb, iodosulfuron-methyl-sodium, isofenphos, isofetamid, isoflucypram, isomethiozin, isopyrazam, isotianil, isoxaben, isoxaflutole, lactofen, lambda-cyhalothrin, lufenuron, mefentrifluconazole, mesosulfuron-methyl, metaflumizone, metamifop, methidathion, metosulam, metsulfuron, metsulfuron-methyl, nicosulfuron, norflurazon, novaluron, noviflumuron, orthosulfamuron, oxadiargyl, oxadiazon, oxasulfuron, oxathiapiprolin, oxaziclomefone, oxpoconazole fumarate, oxycarboxin, penflufen, penoxsulam, penthiopyrad, pentoxazone, phosalone, picolinafen, picoxystrobin, pinoxaden, pirimiphos-ethyl, primisulfuron methyl, primisulfuron, profoxydim, propaquizafop, propoxycarbazone-sodium, proquinazid, prosulfuron, pydiflumetofen, pyraclofos, pyraclostrobin, pyraflufen, pyraflufen-ethyl, pyrazolynate, pyrazophos, pyrazosulfuron-ethyl, pyrazoxyfen, pyribencarb, pyribenzoxim, pyridaben, pyridafenthion, pyridalyl, pyridate, pyriftalid, pyriminobac-methyl, pyrithiobac-sodium, pyroxasulfone, pyroxsulam, quinonamid, rimsulfuron, saflufenacil, sedaxane, sintofen, S-metolachlor, spirotetramat, sulfentrazone, sulfometuron-methyl, sulfosulfuron, tau-fluvalinate, tembotrione, thiamethoxam, thiapronil, thiazopyr, thidiazimin, thiencarbazone-methyl, thifensulfuron-methyl, thifluzamide, tolfenpyrad, topramezone, triadimefon, triafamone, triasulfuron, triazamate, tribenuron-methyl, trifloxysulfuron-sodium, triflumizole, triflusulfuron-methyl, tritosulfuron, uniconazole, acetoprole, athidathion, bencarbazone, bensulfuron, benthiavalicarb, benzfendizone, bethoxazin, bicyclopyrone, broflanilide, brucine, butamifos, carfentrazone, chloretazate, chlorimuron, chlorprazophos, coumethoxystrobin, coumithoate, coumoxystrobin, cyclopyrimorate, cyenopyrafen, cyhalodiamide, cypendazole, cyprosulfamide, dichlobentiazox, dicloromezotiaz, dimethyl(4-piperidinocarbonyloxy-2,5-xylyl)sulfonium toluene-4-sulfonate, diniconazole-M, dufulin, endothion, enoxastrobin, ethiprole, fenaminstrobin, fenasulam, fenoxasulfone, fenpirithrin, fenquinotrione, fenridazon, fenthiaprop ethyl, florpyrauxifen, fluazaindolizine, flufenerim, flufenoxystrobin, flufiprole, flupropacil, flupyrsulfuron, fluxametamide, fuphenthiourea, furconazole, gliotoxin, halauxifen, halosafen, iodosulfuron, iofensulfuron sodium, iofensulfuron, ipfencarbazone, ipfentrifluconazole, isamidofos, isoxachlortole, isoxapyrifop, lythidathion, mecarbinzid, mesosulfuron, metazosulfuron, methiozolin, methyl (((1-(5-(2-chloro-4-(trifluoromethyl)phenoxy)-2-nitrophenyl)-2-methoxyethylidene)amino)oxy)acetate, monosulfuron, monosulfuron-methyl, norbormide, oxapyrazon, oxapyrazon-dimolamine, oxapyrazon-sodium, paichongding, picarbutrazox, profluazol, propoxycarbazone, propyrisulfuron, prothidathion, pyflubumide, pyraclonil, pyrafluprole, pyrametostrobin, pyraoxystrobin, pyraziflumid, pyribambenz-isopropyl, pyribambenz-propyl, pyrifluquinazon, pyrimisulfan, tefuryltrione, thifensulfuron, tiafenacil, tolpyralate, tribenuron, triclopyricarb, trifloxysulfuron, triflumezopyrim, triflusulfuron, guadipyr, SYP-1924
4	acequinocyl, acetochlor, alloxydim, ametoctradin, amidithion, aminopyralid, benalaxyl, benalaxyl-M, bentazone, buminafos, carbophenothion, chlorpyrifos-methyl, coumachlor, cycloprothrin, cycloxydim, cyhalofop-butyl, cyphenothrin, cyproconazole, diazinon, dichlofluanid, diclobutrazol, diethatyl ethyl, difenacoum, diflubenzuron, dimethenamid, dimethenamid-P, dinoseb acetate, ditalimfos, drazoxolon, esfenvalerate, ethirimol, fenfluthrin, fenoxanil, fenvalerate, fluchloralin, flutolanil, furalaxyl-M, imazalil, iminoctadine tris(albesilate), imiprothrin, iprovalicarb, isazofos, isoxathion, leptophos, malathion, mandestrobin, mefluidide, meperfluthrin, metolachlor, metsulfovax, ofurace, oryzalin, oxyfluorfen, pethoxamid, phosmet, pretilachlor, propisochlor, pyracarbolid, renofluthrin, sethoxydim, sulprofos, tepraloxydim, terbacil, tetrachlorvinphos, tralkoxydim, transfluthrin, triclopyr, vamidothion, vinclozolin, anisuron, benzamacril isobutyl, benzoylprop, bromfenvinfos, bromobonil, brompyrazon, carbasulam, chlorazifop, chlorprocarb, cyanthoate, delachlor, dimidazon, dinocton, dinofenate, fluazifop-P, fluoridamid, fospirate, isocarbophos, isocil, methocrotophos, momfluorothrin, myclozolin,

	phenkapton, pyrisoxazole, tolprocarb, trifopsime, epsilon-momfluorothrin, huanjunzuo
5	(4-chlorophenoxy)acetic acid, 1,2-benzisothiazolin-3-one, 1-naphthylacetamide, 1-naphthylacetic acid, 2,3,6-TBA, aldimorph, aldrin, allidochlor, ametryn, amitraz, ampropylfos, aspon, atraton, atrazine, azoxybenzene, benzyl benzoate, bromoxynil, bromuron, bronopol, carbaryl, cartap hydrochloride, cartap, chloralose, chloranil, chlorbenside, chlordane, chlorophacinone, chlorothalonil, chlorthiamid, cinmethylin, clofentezine, cybutryne, cycluron, daminozide, demeton, desmetryn, dichlorophen, dichlorvos, dicloran, dicofol, dieldrin, dienochlor, difenoxuron, dimethipin, dimethyl phthalate, diofenolan, diphacinone, diphenamid, diuron, dodemorph acetate, dodemorph, dodine, endrin, fenbutatin oxide, fenpropidin, flupropanate-sodium, flurenol, fosetyl-aluminium, guazatine, heptachlor, hydroprene, iminoctadine triacetate, ioxynil, isodrin, isoproturon, isoval, medetomidine, methamidophos, methoxychlor, monuron, prohexadione-calcium, prometon, prometryn, propazine, propineb, simazine, simetryn, sodium o-nitrophenolate, sodium p-nitrophenolate, spiroxamine, sulfotep, tebutam, terbumeton, terbuthylazine, terbutryn, tetraethyl pyrophosphate, thiabendazole, thiocyclam oxalate, tribufos, tributyltin oxide, tridemorph, trietazine, trifenmorph, triforine, ziram, 2,2-dibromo-3-nitrilopropionamide, 2,3,5-tri-iodobenzoic acid, 2-allyphenol, 2-methoxyethylmercury acetate, asomate, azithiram, benclothiaz, benzipram, bithionol, bromocyclen, carbamorph, chlorazine, chlorbicyclen, chlorethoxyfos, chloroxynil, chlorquinox, cisanilide, copper abietate, copper bis(3-phenylsalicylate), cycloheximide, cyperquat chloride, cyprazine, demephion-O, demeton-O-methyl, diammonium ethylenebis(dithiocarbamate), diethyltoluamide, dikegulac, dikegulac-sodium, dixanthogen, epocholeone, epofenonane, etem, ethylene bisisothiocyanate sulphide, furalane, heptopargil, iminoctadine, ioxynil lithium, ioxynil sodium, isocarbamide, isonoruron, kaolin, mazidox, methiuron, methoxyphenone, mipafox, naphthalic anhydride, nornicotine, noruron, o,o'-DDT, o,p'-DDT, oxine-copper, phenyl mercuric acetate, phenylmercury nitrate, pindone, schradan, sebuthylazine, semiamitraz hydrochloride, tetrasul, 2-diethylaminoethyl hexanoate
6	amicarbazone, aminocyclopyrachlor, azaconazole, bromacil, buthidazole, difunon, furilazole, halacrinate, hexaconazole, hexythiazox, inabenfide, metamitron, metazachlor, metominostrobin, metoxadiazone, nitenpyram, oxadixyl, oxolinic acid, phosphamidon, picloram, prothiofos, simeconazole, thiacloprid, thicyofen, tolylfluanid, triazoxide, triticonazole, (R)-hexaconazole , alloxydim sodium, amidothioate, butonate, clomeprop, etaphos, imazamethabenz, isolan, picloram-dimethylammonium, quinofumelin, trifenofos, (S)-hexaconazole
7	2,4-DB, aclonifen, aldicarb, aminocarb, ancymidol, anilazine, aziprotryne, bendiocarb, benodanil, bromophos, bromopropylate, bufencarb, butocarboxim, buturon, cadusafos, carbendazim, chinomethionat, chloramben, chloraniformethan, chlorfenethol, chlorfenprop-methyl, chlorfenson, chlormephos, chlornitrofen, chlorobenzilate, chloropropylate, chloroxuron, chlorpropham, crimidine, crotamiton, cumyluron, cyprodinil, dichlofenthion, dichlone, diclofop, diclofop-methyl, diflovidazin, dimethrin, dinoseb, dinoterb, dioxabenzophos, dioxacarb, diquat dibromide, diquat, disulfoton, dithianon, edifenphos, empenthrin, ethoprophos, etridiazole, fenarimol, fenobucarb, fenoprop, fenoxycarb, fensulfothion, fenthion, fluridone, flusilazole, folpet, forchlorfenuron, formetanate, fosthietan, glyphosate trimesium, hexazinone, icaridin, iodofenphos, isoprocarb, isoprothiolane, isothioate, linuron, lithium perfluorooctane sulfonate, MCPB, mecoprop, mecoprop-P, methacrifos, methomyl, methoprotryne, metobromuron, metoxuron, mevinphos, monalide, monolinuron, naled, napropamide, napropamide-M, neburon, nitrofen, octhilinone, oxydemeton-methyl, parathion-methyl, pebulate, phorate, phosdiphen, phosfolan, prallethrin, procymidone, promecarb, propaphos, propyzamide, prosulfocarb, proximpham, pyrimethanil, pyriproxyfen, pyroquilone, quinoxyfen, tebufenozide, terbufos, thidiazuron, thiophanate-methyl, tolclofos-methyl, trichlorfon, tricyclazole, trinexapac-ethyl, bromoxynil heptanoate, 2,4-DEP, azothoate, benquinox, benzadox, chloranocryl, chlorfensulphide, clofop-isobutyl, cloxyfonac, daimuron, demeton-S-methyl sulfone, dipropalin, dipyrithione, disul, fenapanil, fenson, flucofuron, fluenetil, fluoromidine, fluotrimazole, flupropadine hydrochloride, flupropadine, flurenol-butyl, fosamine ammonium, glufosinate, glufosinate-P, glyodin, haloxydine, hexylthiofos, isopyrimol, m-cumenyl methylcarbamate, methyldymron, mexacarbate,

	mucochloric anhydride, nitrilacarb, piperonyl sulfoxide, plifenate, proglinazine, protrifenbute, prynachlor, pydanon, pyrazachlor, sulfallate, thiomersal, thionazin, triprene, verbutin, xylachlor
8	alpha-cypermethrin, beta-cypermethrin, bifenox, bifenthrin, boscalid, brodifacoum, bromadiolone, butachlor, butenachlor, butroxydim, carboxin, chlorpyrifos, chlozolinate, chromafenozide, clethodim, coumafuryl, cyflumetofen, cypermethrin, deltamethrin, difethialone, dimethomorph, dinobuton, EPN, epoxiconazole, etoxazole, etrimfos, fenamiphos, flamprop, fluazifop-butyl, fluazifop-P-butyl, flubenzimine, flucythrinate, flumetralin, fluopicolide, flupyradifurone, flurtamone, furathiocarb, hydramethylnon, imazamethabenz-methyl, kresoxim-methyl, lenacil, mesotrione, N-(3-chloro-2,6-dimethylphenyl)-2-methoxy-N-(tetrahydr-2-oxo-3-furanyl)acetamide, niclosamide, phenthoate, piperophos, pirimiphos-methyl, prochloraz, quizalofop-ethyl, quizalofop-P-ethyl, silthiofam, spiromesifen, tecloftalam, teflubenzuron, tefluthrin, tralomethrin, trifloxystrobin, triflumuron, zeta-cypermethrin, amibuzin, benzoylprop-ethyl, chlorphoxim, climbazole, clodinafop, cloproxydim, dinopenton, dinosulfon, dithiopyr, ethoxyfen ethyl, ethoxyfen, flumorph, fufenozide, hyquincarb, inezin, kappa-bifenthrin, kappa-tefluthrin, mecarphon, medinoterb acetate, pyramat, pyranocoumarin, pyrimorph, quintiofos, tebufloquin, terbuchlor, triamiphos, trichlamide, triclopyr triethylammonium
9	aldicarb sulfone, alpha-endosulfan, benfuresate, bensultap, benzthiazuron, bromobutide, butathiofos, butralin, captan, carbofuran, chlobenthiazone, chlomethoxyfen, chlorbromuron, chlorbufam, chlorflurenol methyl, cloethocarb, cyclanilide, cymiazol, cymoxanil, desmedipham, di-allate, dicamba, dichlorprop-P 2-ethylyhexyl ester, diclomezine, dicrotophos, diethofencarb, difenzoquat metilsulfate, dimepiperate, dimethoate, endosulfan, esprocarb, ethiofencarb, ethoate-methyl, fenfuram, fenothiocarb, fludioxonil, fluorodifen, fluoroimide, formetanate hydrochloride, formothion, glufosinate-ammonium, halfenprox, heptenophos, iprobenfos, iprodione, IPSP, isopropalin, karbutilate, MCPA-thioethyl, mepanipyrim, mephosfolan, methabenzthiazuron, methiocarb, metrafenone, monocrotophos, naproanilide, naptalam, nitrothal isopropyl, nuarimol, omethoate, oxabetrinil, oxydeprofos, parathion-ethyl, penconazole, pencycuron, pendimethalin, pentanochlor, phenmedipham, phenothrin, pyridafol, quinclorac, quinmerac, tebupirimfos, tebuthiuron, thiobencarb, thiofanox, tiocarbazil, tri-allate, 2-(thiocyanomethylthio)benzothiazole, 2,6-dichloro-N-((4-(trifluoromethyl)phenyl)methyl-benzamide, amiton, azoluron, benzadox ammonium, benzamacril, benzamorf, beta-endosulfan, butopyronoxyl, carbanolate, chloreturon, chloroprallethrin, DAEP, debacarb, demephion, difenopenten, dinex, dinosam, disul-sodium, eglinazine-ethyl, erbon, etinofen, fenoprop-butotyl, fenthion sulfoxide, halofenozide, isouron, medinoterb, methfuroxam, nithiazine, oxydisulfoton, proglinazine ethyl, pyridinitril, silatrane, tetcyclacis, tetramethylfluthrin, tioxazafen
10	acifluorfen, acifluorfen-sodium, amidoflumet, anilofos, beflubutamid, benazolin, binapacryl, bistrifluron, bitertanol, bromuconazole, clothianidin, dinocap, etaconazole, famoxadone, fenchlorazole-ethyl, fenoxaprop-P, furalaxyl, haloxyfop, haloxyfop-P, imidacloprid, mandipropamid, mecarbam, mefenacet, meptyldinocap, metribuzin, orysastrobin, pirimicarb, profenofos, propetamphos, propiconazole, prothioconazole, pyriofenone, quizalofop-P-tefuryl, spirodiclofen, sulfoxaflor, tebufenpyrad, tetraconazole, thenylchlor, tiadinil, triadimenol, valifenalate, cis-propiconazole, amiprofos-methyl, brofluthrinate, buthiuron, cambendichlor, chlorazifop propargyl, clacyfos, fenazaflor, fenthiaprop, flamprop-methyl, flometoquin, fluensulfone, furyloxyfen, imidaclothiz, isofenphos-methyl, kadethrin, metobenzuron, picloram-trolamine, pyriminostrobin, tralocythrin, triclopyr butotyl
11	bifenazate, bioresmethrin, bromethalin, buprofezin, buthiobate, captafol, chlorfenvinphos, chloridazon, dimethylvinphos, ethalfluralin, ethidimuron, flonicamid, flurochloridone, flutriafol, fluxofenim, fonofos, furmecyclox, methoxyfenozide, oxamyl, paclobutrazol, phosametine, phoxim, promacyl, prothoate, pymetrozine, resmethrin, spinetoram, triapenthenol, triazophos, trichloronate, warfarin, (R)-flutriafol, (S)-flutriafol , bentaluron, benzofluor, carboxazole, cliodinate, ethylchlozate, fenitropan, furethrin, heptafluthrin, metazoxolon, nitralin, perfluidone, phosacetim, probenazole, sulphaquinoxaline

12	butylate, chlorphonium chloride, cyanazine, EPTC, etacelasil, fenpropimorph, molinate, quintozene, sodium 5-nitroguaiacolate, tecnazene, tetradifon, triarathene, 1,1-bis(4-chlorophenyl)-2-ethoxyethanol, bismerthiazol, chlorfenazole, chlorfluren, chlorphonium, difenzoquat, isopolinate, prothiocarb, pyripropanol, thioquinox

REFERENCIAS

[1] Van Den Berg, F. et al., **1999**, Emisión de Pesticidas al Aire. En: Van Dijk, HFG, Van Pul, WAJ, De Voogt, P. (eds) Destino de los pesticidas en la atmósfera: implicaciones para la evaluación de riesgos ambientales. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1536-2_9

[2] Víctor H. Estellano, Karla Pozo, Tom Harner, Margot Franken y Mauricio Zaballa Ciencia y tecnología ambientales **2008** 42 (7), 2528-2534 DOI: 10.1021/es702754m

[3] Degrendele, C., Okonski, K., Melymuk, L., Landlová, L., Kukucka, P., Audy, O., Kohoutek, J., Cupr, P., & Klánová, J., **2016**, Pesticides in the atmosphere: a comparison of gas-particle partitioning and particle size distribution of legacy and current-use pesticides. *Atmos. Chem. Phys.*, *16*(3), 1531–1544. https://doi.org/10.5194/acp-16-1531-2016.

[4] Gila, Y., & Sinfort, C., **2005**, Emission of pesticides to the air during sprayer application: A bibliographic review. Atmospheric Environment, *39* (28), 5183-5193 .https://doi.org/10.1016/j.atmosenv.2005.05.019

[5] Nascimento, MM, da Rocha, GO & de Andrade, JB Pesticidas en partículas finas en el aire: de un método de análisis ecológico a la caracterización atmosférica y la evaluación de riesgos. Representante científico 7 , 2267 (**2017**). https://doi.org/10.1038/s41598-017-02518-1

[6] Bedos C., Cellier P., Calvet R. Barriuso E., **2002**, Occurrence of pesticides in the atmosphere in France. Agronomie, vol. 22, no 1, p. 35-49. DOI: 10.1051/agro: 2001004

[7] Boonupara, T., Udomkun, P., & Khan, E., **2023**, Airborne Pesticides from Agricultural Practices: A Critical Review of Pathways, Influencing Factors, and Human Health Implications. *Toxics, 11*(10), 858.(https://doi.org/10.3390/toxics11100858).

[8] Duchowicz PR., **2020**, QSPR studies on water solubility, octanol-water partition coefficient and vapour pressure of pesticides. SAR QSAR Environ Res. Feb;*31*(2):135-148. doi: 10.1080/1062936X.2019.1699602. Epub 2019 Dec 16. PMID: 31842624.

[9] Duchowicz PR, Fioressi SE, Bacelo DE, Quispe AQ, Yapu EL, Castañeta H. QSPR predicting the vapor pressure of pesticides into high/low volatility classes. Environ Sci Pollut Res Int. 2024 Jan;*31*(1):1395-1402. doi: 10.1007/s11356-023-31235-8. Epub 2023 Dec 1. PMID: 38038924.

[10] C. W. Yap, "PaDEL-descriptor: un software de código abierto para calcular descriptores moleculares y huellas dactilares". J Comput Chem, *32* (7),1466-1474. 2011.

[11] K. Beyer, J. Goldstein, R. Ramakrishnan, et U. Shaft, « When is “nearest neighbor” meaningful? », in *Database Theory—ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, Springer, 1999, p. 217-235.

[12] A. Zimek, E. Schubert, et H.-P. Kriegel, « A survey on unsupervised outlier detection in high-dimensional numerical data », *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 5, nº 5, p. 363-387, 2012.

[13] A. Patrikainen et M. Meila, « Comparing subspace clusterings », *IEEE Trans. Knowl. Data Eng.*, vol. 18, nº 7, p. 902-916, 2006.

[14] H.-P. Kriegel, P. Kröger, et A. Zimek, « Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering », *Acm Trans. Knowl. Discov. Data Tkdd*, vol. 3, nº 1, p. 1-58, 2009.

[15] S. Peignier et H. Castañeta, « Análisis de ‘subspace clustering’de moléculas utilizando Chameleoclust, un algoritmo evolutivo », *Rev. Boliv. Quím.*, vol. 32, nº 5, p. 110-120, 2015.

[16] S. Peignier, C. Rigotti, A. Rossi, et G. Beslon, « Weight-based search to find clusters around medians in subspaces », in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018, p. 471-480.

[17] Open Babel for Windows, http://openbabel.org/wiki/Category:Installation, last accessed March 2023

[18] PPDB: Pesticide Properties DataBase, https://sitem.herts.ac.uk/aeru/ppdb/en/atoz.htm, last accessed March 2023

[19] ACDLabs/ChemSketch, www.acdlabs.com, last accessed March 2023

[20] PaDEL 2.20 (Pharmaceutical Data Exploration Laboratory), http://www.yapcwsoft.com, last accessed March 2023

[21] Van der Maaten et G. Hinton, « Visualizing data using t-SNE. », *J. Mach. Learn. Res.*, vol. 9, nº 11, 2008.
[22] K. Peason, « On lines and planes of closest fit to systems of point in space », *Philos. Mag.*, vol. 2, nº 11, p. 559-572, 1901.
[23] O. Kramer et O. Kramer, « Scikit-learn », *Mach. Learn. Evol. Strateg.*, p. 45-53, 2016.
[24] L. Breiman, « Random forests », *Mach. Learn.*, vol. 45, p. 5-32, 2001.
[25] S. M. Lundberg et S.-I. Lee, « A unified approach to interpreting model predictions », *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[26] S. M. Lundberg *et al.*, « Explainable AI for trees: From local explanations to global understanding », *ArXiv Prepr. ArXiv190504610*, 2019.