UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS

FACULTAD MATEMÁTICA, FÍSICA Y COMPUTACIÓN

DEPARTAMENTO CIENCIAS DE LA COMPUTACIÓN

Evaluación del perfil neurotóxico de líquidos iónicos basado en técnicas de aprendizaje automático supervisado.

TESIS EN OPCIÓN AL TÍTULO DE MÁSTER EN BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

Rotceh Dominguez López

Santa Clara, 2015

UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS

FACULTAD MATEMÁTICA, FÍSICA Y COMPUTACIÓN

DEPARTAMENTO CIENCIAS DE LA COMPUTACIÓN

TESIS EN OPCIÓN AL TÍTULO DE MÁSTER EN BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

Título*:* Evaluación del perfil neurotóxico de líquidos iónicos basado en técnicas de aprendizaje automático supervisado**.**

Autor: Ing**.** Rotceh Dominguez López

Tutor: Dr.C Maykel Cruz Monteagudo**.**

Santa Clara, 2015

“Todo lo puedo en Cristo que me fortalece”

Filipenses 4:13

Dedicatoria:

A Dios por ser fuente de fuerzas para alcanzar metas en la vida, porque nunca falla.

A mi Rocío porque sus besitos cada mañana son el impulso de mi día a día.

A mi esposa porque su apoyo y amor fueron piedra angular en el desarrollo de este trabajo.

A mis padres y hermanos por ser imprescindibles en mi vida.

Agradecimientos

A Dios porque cuando vacilé, me dio fuerzas para seguir adelante, no me permitió fallar.

A Maykel Cruz por su certera tutoría. Espero poder retribuir tu ayuda algún día.

A Gladys Casas y María del Carmen Chávez porque su apoyo fue fundamental en este resultado.

A los amigos del CBQ, Evys, Yunierkis y Aliuska, gracias por su aporte en mi formación.

A los profesores de la maestría por su excelente labor.

A mis compañeros de la maestría Jose, Dailé, Yadira, Juan Carlos.

A la tropa de tantos viajes, Yamisleydi y Jose.

A todos los que de una forma u otra influyeron en la obtención de este resultado.

RESUMEN

La enzima Acetilcolinesterasa (AChE) juega un papel imprescindible en la hidrólisis del neurotransmisor Acetilcolina, el cual es el responsable de la transmisión de los impulsos nerviosos. Desde la década de los 30, especialistas en las ciencias químicas han producido compuestos que son capaces de inhibir esta enzima y por tanto afectar el proceso de transmisión de los impulsos nerviosos, lo cual provoca consecuencias graves para el organismo afectado. En la actualidad se ha comenzado la producción de compuestos llamados “líquidos iónicos”; que poseen características físico-químicas atractivas para la producción de solventes que son utilizados en la sustitución de solventes moleculares tóxicos para el medio ambiente. Estudios han demostrado que algunos líquidos iónicos pueden inhibir el funcionamiento de la enzima AChE y provocar daños al sistema nervioso central, surgiendo la necesidad de evaluar el perfil neurotóxico de los líquidos iónicos utilizando la enzima AChE como indicador de neurotoxicidad. Los estudios sobre la cuantificación de la relación entre la estructura y la actividad (QSAR) surgen para realizar análisis basados en la cuantificación de dicha relación presente en compuestos químicos. En el presente trabajo fue usada la habilidad de los líquidos iónicos para inhibir la AChE y de esta manera evaluar el perfil neurotóxico de los líquidos iónicos y consecuentemente su perfil no neurotóxico. En este sentido estudios QSAR pueden proveer información útil sobre el perfil químico de los líquidos iónicos. En el desarrollo del trabajo se aplicaron multiclasificadores, como técnicas de aprendizaje automático supervisado, y como resultado se obtuvieron modelos capaces de predecir si un nuevo líquido iónico es capaz o no de inhibir la AChE. Los multiclasificadores Bagging, Boosting, Stacking y Vote fueron utilizados en la experimentación con el fin de identificar modelos QSAR predictivos. Fueron calculadas cinco medidas de diversidad para los clasificadores bases utilizados en los multiclasificadores Stacking y Vote. Finalmente se obtuvieron dos modelos que superaron el desempeño de los clasificadores individuales utilizados, razón por la que fueron seleccionados para solucionar el problema. El multiclasificador AdaBoostM1, que utiliza una red neuronal MultilayerPerceptron como clasificador base y el multiclasificador Stacking, que utiliza la combinación de clasificadores FDLA, Jrip, Kstar, NaiveBayes y SMO como clasificadores bases, fueron los multiclasificadores seleccionados.

Palabras claves: Acetilcolinesterasa, líquidos iónicos, QSAR, multiclasificadores, medidas de diversidad

ABSTRACT

Acetylcholinesterase (AChE) enzyme plays an important role in the hydrolysis of the neurotransmitter known as Acetylcholine, an important enzyme in transmission of nerve impulses. Since the early 30s, specialists in chemical sciences have produced compounds able to inhibit this enzyme and thus affect the process of transmission of nerve impulses, with serious consequences over the affected organism. Ionic liquids (IL) represent a chemical family with a high number of industrial applications due to their attractive physical and chemical properties, mainly for the production of safer solvents that can replace standard solvents usually toxic for humans and the environment. However, previous studies have shown that ILs can inhibit AChE, with the potential neurotoxic effects associated. In this work, we used the ability of IL to inhibit AChE to evaluate the neurotoxic profile of IL and consequently their safety profile. In this sense, quantitative structure-activity/property relationships (QSAR/QSPR) can provide useful information on the property/activity profile of chemicals based on the quantification of the relationship between chemical structure and biological activity. Specifically, ensembles classifiers were applied as part of the QSAR studies, obtaining models able to predict whether a new IL inhibit or not AChE and so, providing an in silico profiling of their neurotoxic potential. Bagging, Boosting, Stacking and Vote ensembles approaches were used to find predictive QSAR models. Five measures of diversity for base classifiers used in Stacking and Vote ensembles were calculated. Finally, two models outperformed the individual classifiers and therefore were suitable to solve the problem. The following two models were obtained: an ensemble AdaBoostM1 using a MultilayerPerceptron network as base classifier and an ensemble Stacking using the combination of the FDLA, Jrip, Kstar, SMO and NaiveBayes as base classifiers.

Keywords: Acetylcholinesterase, ionic liquids, QSAR, ensembles, diversity measures.

ÍNDICE

INTRODUCCIÓN 1
CAPÍTULO 1: REVISIÓN BIBLIOGRÁFICA. 8
1.1. Estudios relación cuantitativa de estructura- actividad (QSAR) 8
1.2. Descriptores moleculares. 8
1.2.1. Descriptores basados en fragmentos 9
1.3. Extracción de conocimiento en base de datos. 10
1.3.1. Preprocesamiento. 10
1.3.2. Selección de atributos. 11
1.3.2.1. Selección de atributos utilizando el criterio de información mutua. 12
1.3.3. Balance de clases. 13
1.3.3.1. Balanceo de clases utilizando el algoritmo *SMOTE* como método Oversampling. 14
1.4. Minería de datos 15
1.4.1. Algoritmos de clasificación. 15
1.4.1.1. Aprendizaje basado en casos (perezosos). 15
1.4.1.2. Árboles de decisión. 16
1.4.1.3. Redes bayesianas. 16
1.4.1.4. Redes neuronales artificiales. 17
1.4.1.5. Máquinas de soporte vectorial. 18
1.4.1.7. Regresión logística. 18
1.4.1.8. Reglas de asociación. 19
1.5. Multiclasificadores. 19
1.5.1. Bagging. 20
1.5.2. Boosting. 21
1.5.3. Stacking. 22
1.5.4. Vote. 22
1.5.4.1. Combinación de salidas. 22
1.5.5. Medidas de diversidad para la determinación de clasificadores diversos. 24
1.5.6. Análisis de las medidas de diversidad 26
1.6. Medidas para evaluar el desempeño de los clasificadores. 27
CAPÍTULO 2: EXPERIMENTACIÓN 31
2.1 Descripción de los datos utilizados. 31
2.2 Preprocesamiento. 31
2.2.1. MRMR como algoritmo de reducción de dimensionalidad. 33
2.2.2. SMOTE como algoritmo de balanceo de datos. 33
2.3 Identificación de clasificadores base con desempeño aceptable. 34

2.4 Utilización de los algoritmos multiclasificadores. 35
2.4.1. Bagging 35
2.4.2. Boosting 35
2.4.3. Determinación de clasificadores diversos para los multiclasificadores. 36
2.4.4. Stacking 36
2.4.5. Vote 37
2.5 Medidas de evaluación de desempeño de clasificadores. 37
2.6 Consideraciones finales. 38
CAPÍTULO 3: ANÁLISIS DE RESULTADOS. 39
3.1 Preprocesamiento. 39
3.1.1 mRMR como algoritmo de reducción de dimensionalidad 41
3.1.2 SMOTE como algoritmo de balanceo de datos. 41
3.2 Identificación de clasificadores base con desempeño aceptable. 42
3.3 Resultados de los algoritmos multiclasificadores utilizados. 43
3.3.1 Resultados de la experimentación con el multiclasificador Bagging 43
3.3.2 Resultados de la experimentación con el multiclasificador AdaBoost.M1 44
3.3.3 Determinación de los clasificadores diversos. 44
3.3.4 Resultados de la experimentación con el multiclasificador Stacking 46
3.3.5 Resultados de la experimentación con el multiclasificador Vote. 48
3.4 Selección del multiclasificador para la solución del problema. 50
3.5 Conclusiones parciales. 51
CONCLUSIONES 52
RECOMENDACIONES 53
BIBLIOGRAFÍA 54
ANEXOS 59

INTRODUCCIÓN

Las neuronas son las células más caracterizadas y estudiadas del sistema nervioso. Se componen de tres partes: las dendritas, situadas en torno al citoplasma; el cuerpo celular o soma, y el axón. El axón tiene una doble misión: por una parte, une a las neuronas entre sí en el proceso denominado sinapsis y, por otra, al reunirse con cientos o miles de otros axones, da origen a los nervios que conectan al sistema nervioso con el resto del cuerpo.

Las neuronas son un tipo de célula altamente especializada cuya principal característica es su incapacidad para reproducirse. Esto significa que el ser humano nace con una cantidad determinada de neuronas, las que, si bien no pueden duplicarse, han demostrado ser unidades muy plásticas y capaces de generar reacciones en situaciones bastante desfavorables.

El sistema nervioso posee además un tipo de células nerviosas de apoyo, llamadas células gliales. Estas células desempeñan funciones como el mantenimiento del ambiente neuronal, eliminando el exceso de neurotransmisores; la destrucción de microorganismos; el aislamiento de los axones neuronales y la circulación del líquido cefalorraquídeo que recubre los principales órganos de este sistema.(Tejedor 2005)

Las neuronas son las encargadas de transmitir los impulsos nerviosos en todo el organismo**.** Para que el impulso eléctrico se transmita, los iones positivos de sodio, que en estado de descanso están presentes fuera de la neurona, deben traspasar la membrana celular. En estado de reposo el interior de la neurona tiene carga eléctrica negativa (membrana repolarizada). Cuando los iones positivos de sodio ingresan a la neurona, cambian la carga interna de negativa a positiva (membrana despolarizada) lo que provoca que los receptores se abran al paso de los iones y de esta forma se transmitan las cargas eléctricas y con esto los impulsos nerviosos. En la medida que el impulso avanza por la membrana, su interior recobra la carga negativa. De esta forma, el impulso va pasando a través de los axones de las neuronas y mediante la acción de los neurotransmisores desde una neurona a otra.(Lehninger 2005)

La acetilcolina (ACh) es un neurotransmisor en muchas sinapsis colinérgicas del sistema nervioso autónomo y en las conexiones neuromusculares. Los receptores de ACh (AChR) en las membranas postsinápticas pueden ser de dos tipos: receptores muscarínicos y receptores nicotínicos. Concretamente estos actúan en algunas sinapsis del sistema nervioso central y en las inervaciones del músculo esquelético; igualmente en las del músculo liso, corazón y glándulas endocrinas del sistema autonómico. Es una molécula con estructura química de éster y con cierta polaridad, pues tiene un nitrógeno cuaternario y por tanto con carga positiva, lo que le facilitará su interacción con proteínas, como su receptor (en la membrana postsináptica) o la enzima que la degrada, la acetilcolinesterasa.(Koolman et al. 2005)

La acetilcolinesterasa (AChE) es una esterasa que hidroliza a la acetilcolina. Es una enzima situada en las hendiduras sinápticas donde hidroliza la ACh, lo que conlleva a la repolarización de la membrana postsináptica y la preparación para la llegada de un nuevo impulso nervioso. La función normal de la ACh depende de su rápida hidrólisis por la AChE, que permite la brevedad y unidad de los impulsos nerviosos propagados sincrónicamente. (Mathews. et al. 2005)

La inhibición de la enzima AChE provoca que la ACh unida a sus receptores presentes en la membrana postsináptica no sea liberada rápidamente mientras que la ACh sigue siendo producida como parte de la transmisión de nuevos impulsos nerviosos. Esta situación provoca que los receptores no sean capaces de recibir los nuevos impulsos eléctricos emitidos por las neuronas vecinas, debido a que la AChE no hidroliza los neurotransmisores ACh que están presentes en las membranas postsinápticas, provocando serios efectos en el sistema nervioso central.(Metzler 2003)

La AChE ha sido el blanco principal para el diseño de insecticidas, (organofosforados y carbamatos) y en la producción de gases paralizantes utilizados en la guerra, debido a que la AChE está presente en todos los organismos que posee un sistema nervioso. La extrema toxicidad de los compuestos organofosforados fue identificada en Alemania e Inglaterra, en la década de los años 30, y utilizada con los fines anteriormente mencionados en la guerra. Entre los compuestos organofosforados más notorios está el *diisopropylphosphonofluoridate* el cual, al ser aplicado en dosis altas en animales, causa una muerte rápida; también existen otros identificados que son utilizados específicamente para eliminar insectos.(Metzler 2003) De igual forma se determinó que la inhibición de la AChE a nivel cerebral, hasta un 70% en el pez estuario, está bien correlacionada con la muerte.(Stock et al. 2004)

El centro activo de la enzima es bien conocido, por lo tanto, permite tener un conocimiento detallado de la relación que existe entre su estructura y su actividad biológica.(Stock et al. 2004) El centro activo se caracteriza por ser muy conservador entre los organismos complejos, situación que permite hacer una extrapolación de la relación estructura actividad (SAR) entre los organismos.(Stock et al. 2004, Arning et al. 2008).

La inhibición de la enzima resulta en una variedad de efectos tales como: enfermedades del corazón, miastenia gravis, debilidad muscular y en ocasiones la muerte. (Metzler 2003, Stock et al. 2004).

Por otra parte, los líquidos iónicos (LI) constituyen un área en las ciencias químicas que ha estado atrayendo la atención entre los investigadores. Los estudios en torno a este campo están justificados por las características físico-químicas únicas de estos compuestos, que proveen un amplio rango de aplicaciones (Störmann et al. 2011). El número de posibles combinaciones de aniones y cationes que pudieran componer un LI está por el orden de los 10^{18}(Arning et al. 2008). Sus amplias posibilidades estructurales, en contraposición a las variaciones estructurales limitadas de solventes moleculares, permite definir a los LI como "diseñadores de solventes"(Couling et al. 2006). También han sido calificados como "solventes verdes"(Lacrama et al. 2007, Luis et al. 2007, Putz et al. 2007).

La razón para llamar a los LI como "verdes" consiste, generalmente, en tres argumentos: i) poseen una presión de vapor muy baja que proporciona una reducción de la exposición por inhalación en el personal que trabaja con ellos, en comparación con disolventes moleculares convencionales; ii) poca inflamabilidad, que reduce fuertemente el riesgo de oxidaciones exotérmicas rápidas y iii) que se dice que son relativamente no tóxicos.

Existen investigadores que están de acuerdo con las primeras características, en tanto otros expresan su desacuerdo con la última característica y la justificación de su discrepancia se encuentra plasmada en varios artículos.(D. Zhao 2007, García et al. 2008, R. F. Frade 2010)

Tal preocupación sobre el perfil neurotóxico de los LI ha devenido de un esfuerzo de las industrias químicas, gobiernos, instituciones y academias para direccionar los principios de la "química verde"(Sierra et al. 2006); fundamentalmente hacia aquellos que disminuyen el peligro potencial de sustancias y productos químicos. Respetando siempre los siguientes principios: (Anastas et al. 1998)

Principio No 3 de la Química Verde: Siempre que sea posible, los métodos de síntesis deberán ser diseñados para ser utilizado y generar sustancias que tengan poca o ninguna toxicidad para la salud humana y el medio ambiente.

Principio No 4 de la Química Verde: Los productos químicos sintetizados deben ser diseñados para efectuar las funciones deseadas mientras sea minimizada su toxicidad.

En este sentido las investigaciones realizadas por la industria química, específicamente relacionadas con la producción y uso de los LI, se centran en la determinación de los perfiles físicos y químicos de los LI y sus posibles usos. Los principales usos y aplicaciones de los LI en la actualidad pueden ser consultados en el Anexo 1.

La aplicación de los LI es vital en la industrial actual. Su utilización está sustentada en la sustitución de productos químicos que tienen un efecto muy tóxico sobre el medio ambiente, comparado con los niveles de toxicidad detectados en los LI hasta el día de hoy. De ahí la importancia de determinar los LI que propician un diseño óptimo para un objetivo específico y que a la vez sean inocuos para el medio ambiente.

Por tal razón diferentes estudios han sido realizados teniendo en cuenta las características únicas que poseen estos compuestos químicos. La inmensa cantidad de combinaciones que pueden ser realizadas, con la utilización de los componentes que integran un compuesto de este tipo, permiten virtualmente diseñar y producir LI con un propósito específico. Siendo este uno de los objetivos fundamentales de las investigaciones en esta área.

Como parte de la estimación del perfil neurotóxico de los LI un creciente número de estudios, que analizan el peligro de diferentes LI, han sido realizados y determinados en diferentes sistemas de pruebas biológicas (Sierra et al. 2006, Torrecilla et al. 2010, Torrecilla et al. 2010).

Empresas como Merck KGaA, en unión con el "Grupo de investigación de los Líquidos Iónicos" del Centro de investigaciones y tecnologías medioambientales (UFT), han realizado investigaciones para determinar el perfil neurotóxico de los LI. El peligro potencial ecotoxicológico de las estructuras químicas de los LI así como los LI con una aplicabilidad técnica óptima y con un reducido peligro potencial para el hombre y el medioambiente han sido objetivos en sus investigaciones.

En el marco de esta cooperación la UFT ha investigado a cientos de LI, donde múltiples estructuras han sido caracterizadas ecotoxicológicamente en detalles, utilizando sistemas de pruebas ecotoxicológicas que comprenden diferentes niveles de organización y complejidad biológica. Estas pruebas son realizadas con la utilización de enzimas, líneas de células animales, microorganismos y organismos tanto marinos como terrestres.

A partir de los resultados de estas investigaciones se determinó que existen LI que poseen un peligro potencial para el hombre y para el medioambiente. Fueron identificados LI con niveles de bajo peligro potencial hasta aquellos que son potencialmente tóxicos.

La UFT realiza un conjunto de pruebas que comprenden ensayos tanto *in vitro* como *in vivo* que permiten identificar el perfil químico de estos productos y los efectos biológicos que estos producen. Dentro del conjunto de sistemas de pruebas efectuadas, por este centro, se realizó el estudio de neurotoxicidad ya que es una propiedad importante a tener en cuenta al evaluar la seguridad de cualquier compuesto químico. Con este objetivo se consumó un estudio para evaluar el indicador de neurotoxicidad de los LI denominado "Ensayos sobre inhibición de la enzima Acetilcolinesterasa", realizado a partir de la enzima AChE, extraída de la manta eléctrica *Electrophorus electricus* (Störmann et al. 2011). Esta prueba se realizó teniendo en cuenta que esta enzima es parte esencial del sistema nervioso central de los organismos complejos(Stock et al. 2004). Los resultados de este estudio permitieron, a la UFT, determinar que los LI pueden inhibir la enzima y por consecuencia aparecer los efectos no deseados antes mencionados.

Evaluar la importancia de encontrar una extrapolación desde los resultados obtenidos de los ensayos *in vitro* hacia los ensayos *in vivo* es necesario. Los resultados *in vitro* solo dan indicaciones acerca del peligro potencial de los compuestos químicos en la inhibición de la enzima, pero no necesariamente indican que esos resultado se generalicen al resto de los organismos. Los resultados de los ensayos *in vivo* (*Electrophorus electricus*) son un indicador de que la AChE de otros organismos también pueden ser inhibidas. (Stock et al. 2004)

A partir de que la gran mayoría de los LI aún no han sido sintetizados, es necesario desarrollar métodos para predecir el peligro potencial de LI desconocidos. Estos métodos deben ser basados en el indicador de neurotoxicidad antes mencionado, elaborados con el objetivo de facilitar el diseño de nuevos materiales seguros reduciendo la necesidad de realizar las síntesis basadas en el método de prueba - error para disminuir los costos y el tiempo de su producción.

El uso de la computación ha mostrado ser un factor fundamental en investigaciones realizadas en múltiples especialidades. Ha sido utilizada como herramienta para realizar simulaciones matemáticas y de otras índoles, con el objetivo de realizar experimentos más certeros y cercanos a la realidad. Posibilitándose la reducción de la cantidad de experimentos a realizar en una investigación, experimentos que por su naturaleza resultan muy costosos.

Como resultado de las investigaciones realizadas sobre el efecto inhibitorio de los LI sobre la AChE ha sido obtenido un gran volumen de registros que describen las características estructurales de los LI ya sintetizados y la actividad biológica que tienen sobre la AChE.

La computación, como herramienta, complementa la ejecución de estudios que son realizados con la utilización de simulaciones o procesamiento de grandes volúmenes de datos que de forma manual sería inviable realizar.

El aprendizaje automático es un área de la inteligencia artificial en la que el hombre ha obtenido buenos resultados en el procesamiento de grandes volúmenes de información. En la actualidad está siendo

utilizada con éxito en la solución a problemas similares en la rama de la química. En este sentido pocos artículos, relacionados con herramientas computacionales y su utilización para la predicción de la toxicidad de los LI, han sido reportados en la literatura (Couling et al. 2006, Lacrama et al. 2007, Luis et al. 2007, Putz et al. 2007, Arning et al. 2008, García et al. 2008, Luis et al. 2010, Torrecilla et al. 2010, Hossain et al. 2011, Mohammad et al. 2011, Mohammad et al. 2011) y específicamente la predicción del potencial inhibitorio de los LI sobre la enzima AChE, como un problema de clasificación, no es la excepción.

Las tareas de clasificación constituyen una rama del aprendizaje automático, denominado "aprendizaje automático supervisado". Estas tareas pueden ser realizadas debido a la existencia de disímiles algoritmos, ya desarrollados con estos fines, cuya clasificación es realizada a partir de ejemplos previamente clasificados por especialistas. Estos algoritmos deben ser capaces de asociar los patrones de entradas recibidos, con las clasificaciones dadas por los especialistas mediante una función que describa esta relación.

Los multiclasificadores son variantes de clasificadores que utilizan su propia heurística para determinar la relación existente entre los patrones de entradas y las clasificaciones ya establecidas. En este caso se apoyan en las clasificaciones emitidas por varios clasificadores individuales, de forma que se utilizan el conocimiento inducido por estos para generar su propia clasificación. Los multiclasificadores utilizan diversas maneras para inducir el conocimiento a partir de los datos. Algunos multiclasificadores se entrenan con pequeñas partes del conjunto original de entrenamiento, hasta cubrir todo el conjunto de entrenamiento, posibilitando al algoritmo enfocarse en las instancias difíciles de clasificar. Otros aprenden de los criterios emitidos por varios clasificadores bases y otros, utilizan mecanismos para combinar las salidas emitidas por los clasificadores individuales utilizados en el aprendizaje.

El uso de los multiclasificadores se fundamenta en las ventajas que estos pudieran ofrecer. Al respecto en (Kuncheva 2004) se sugieren razones estadísticas, computacionales y de representación que justifican el por qué su uso pudiera ser mejor que el uso de un clasificador individual.

Como resultado de la revisión realizada, se determina que los resultados registrados en la literatura, respecto a la evaluación del perfil neurotóxico de los LI sobre la AChE, solo se limitan a la utilización de técnicas de clasificación, basadas en la utilización de un solo algoritmo, o técnicas de regresión. No siendo posible encontrar artículos donde fueran utilizados multiclasificadores con el mismo objetivo.

Teniendo en cuenta lo anteriormente planteado se determina como **objetivo de la investigación**:

Identificar un modelo, basado en un multiclasificador, que permita predecir si un nuevo LI será capaz de inhibir o no la AChE utilizando como base la información estructural de las moléculas.

Para lograr el objetivo de la investigación se proponen los siguientes **objetivos específicos**:

- Cuantificar la estructura molecular de los líquidos iónicos.
- Identificar los clasificadores individuales posibles a utilizar en la confección del multiclasificador.
- Identificar los multiclasificadores a utilizar en la experimentación.

- Diseñar y ejecutar experimentos que permitan evaluar el desempeño de los multiclasificadores.
- Elegir el modelo multiclasificador que mejor desempeño tenga en la predicción de la neurotoxicidad de los LI.

A partir de los objetivos específicos establecidos podemos plantear las siguientes **preguntas de la investigación**:

- ¿Cómo cuantificar la estructura molecular de los líquidos iónicos?
- ¿Cuáles clasificadores son utilizados en la solución de problemas similares?
- ¿Cuáles multiclasificadores son utilizados en la solución de problemas similares?
- ¿Será posible obtener un multiclasificador que alcance mejores resultados que el clasificador individual que mejor desempeño demuestre en la predicción de la neurotoxicidad de los LI?

Las **tareas de la investigación** que permiten alcanzar los objetivos propuestos se plantean a continuación:

- Caracterización de los estudios sobre cuantificación de la relación estructura-actividad (QSAR).
- Caracterización del proceso de extracción del conocimiento en base de datos.
- Caracterización de la etapa de preprocesamiento aplicada a estudios quimioinformáticos.
- Caracterización de algoritmos de clasificación a utilizar en el desarrollo del trabajo.
- Caracterización de multiclasificadores a utilizar en el desarrollo del trabajo.
- Ejecución de experimentos que permitan evaluar los algoritmos de clasificación a utilizar en el desarrollo del trabajo.
- Ejecución de experimentos que permitan evaluar los multiclasificadores a utilizar en el desarrollo del trabajo.

A partir de la situación anteriormente planteada se formuló la siguiente **hipótesis de investigación**:

La utilización de multiclasificadores garantiza la obtención de modelos confiables, que poseen desempeños superiores a los alcanzados por clasificadores individuales, en la evaluación del perfil neurotóxico de los líquidos iónicos.

Estructuración del trabajo:

El trabajo se encuentra estructurado en los siguientes capítulos:

Capítulo 1: Revisión bibliográfica

En este capítulo se recoge información sobre la importancia y el rol desempeñado por la enzima Acetilcolinesterasa en el sistema nervioso central. Se especifica cómo la enzima es el centro de atención en la producción de insecticidas y armas químicas debido que juega un papel imprescindible en la transmisión de los impulsos nerviosos. Se recoge información sobre los compuestos químicos llamados “Líquidos iónicos” que, en la actualidad, están siendo muy utilizados en la industria química, debido a ciertas características que lo hace potencialmente útil en la producción de solventes y otros productos. Se documenta sobre estudios de la influencia negativa que tienen algunos líquidos iónicos sobre la AChE

y de los esfuerzos realizados por identificar mecanismos que permitan orientar la industria hacia la producción de líquidos iónicos seguros para el medio ambiente.

Se analiza la utilización de técnicas computacionales como método de apoyo a investigaciones, con el objetivo de reducir el costoso proceso de prueba y error aplicado hasta el día de hoy en muchas ramas de la investigación.

Se documenta sobre técnicas de minería de datos útiles para el procesamiento de grandes volúmenes de datos, específicamente sobre clasificadores y multiclasificadores. Se investiga sobre las medidas utilizadas para realizar la evaluación de los resultados obtenidos, a partir del uso de estas técnicas computaciones, en la experimentación documentada en el capítulo 2.

Capítulo 2: Experimentación

En este capítulo se especifica el procedimiento seguido en la experimentación, con el objetivo de validar la hipótesis propuesta. Se inicia con la etapa de preprocesamiento realizada sobre los datos que incluye una limpieza manual, siguiendo los principios propuestos por Tropsha y su equipo, se continúa con la reducción de la dimensionalidad del conjunto de datos para evitar problemas de sobreajuste en el aprendizaje de los clasificadores y multiclasificadores y por último se efectúa un balance en la distribución de las clases presentes en el problema. A continuación se documenta la configuración establecida a los multiclasificadores Bagging y Boosting, se procede con la documentación del cálculo de las medidas de diversidad, para la determinación de las combinaciones de clasificadores diversos utilizadas en la configuración de los multiclasificadores Stacking y Vote.

Los clasificadores y multiclasificadores utilizados se encuentran incluidos en el software de análisis de datos utilizado. Finalmente se evalúa el desempeño de los modelos multiclasificadores obtenidos para determinar cuáles pueden ser utilizados en la solución del problema.

Capítulo 3: Análisis de los resultados.

En este capítulo se recogen todos los resultados obtenidos de la experimentación diseñada en el capítulo anterior. Se realiza el análisis de la información obtenida, producto de la experimentación ejecutada, y se determinan los modelos multiclasificadores que poseen un desempeño superior al mejor desempeño alcanzado por los clasificadores individuales utilizados y que por lo tanto justifica su uso en la solución del problema. Lográndose de esta manera darle cumplimiento al objetivo propuesto en la investigación y permitiendo comprobar la veracidad de la hipótesis planteada.

CAPÍTULO 1: REVISIÓN BIBLIOGRÁFICA.

Introducción.

En el capítulo se mencionan los inicios de los estudios sobre la relación cuantitativa de estructura-actividad (*QSAR*) y cómo la computación se ha convertido en una herramienta útil en la modelación de estos problemas en las ciencias químicas. Se documenta cómo la extracción del conocimiento en base de datos (*KDD*) es una rama de la computación que es utilizada para predecir el comportamiento de compuestos químicos a partir del conocimiento extraído de grandes bases de datos donde se registran las características de las moléculas, denominadas "Descriptores moleculares", y su actividad biológica. Se hace un estudio de etapas del *KDD* y dentro de ellas, las actividades a realizar para poder determinar un modelo capaz de predecir la clasificación de un compuesto químico nuevo del cual se desconoce su actividad biológica.

1.1. Estudios relación cuantitativa de estructura- actividad (QSAR)

Durante años los especialistas en las ciencias químicas han buscado la manera de relacionar la estructura molecular presente en las moléculas con respecto a su actividad biológica. Desde los inicios de estos estudios, a finales del siglo XIX, Richet logró establecer una fórmula mediante la cual se dieron los primeros pasos en el establecimiento de la relación antes mencionada. Seguido se registraron nuevas propuestas en este mismo ámbito como la expresada por Louis Hammet en 1933(Molina 2010), Hans y Fujita (Hansh 1964) y Free Wilson (Bort 2001)

Pero no es hasta la década de los años 80 que se comienza a incorporar las herramientas computacionales a los procesos de investigación asociados a las ciencias químicas. La necesidad de representar las moléculas de manera que fuera posible determinar, matemáticamente, la relación existente entre su estructura y su actividad biológica permitió el surgimiento de los denominados "descriptores moleculares", siendo estos un número calculado que permite representar a la molécula de interés a partir de sus fragmentos. Desde el momento en que la estructura de una molécula puede ser representada mediante números, los estudios QSAR se reducen a la correlación entre conjuntos de números que tienen un sentido químico. Estas relaciones permiten cuantificar cómo los cambios en la estructura de una molécula inducen cambios en la actividad biológica de la misma.(Molina 2010)

1.2. Descriptores moleculares.

Los descriptores moleculares (DMs) son una serie de números que caracterizan una molécula y contienen la información derivada de la estructura atómica covalente (Randi 1998). La información de la actividad no está explícitamente incluida en la estructura molecular (Randi 1991) y por tanto los descriptores moleculares son basados en diferentes propiedades moleculares que van desde características físico-químicas, química-cuánticas, geométricas (Randi 1995) hasta las topológicas (Ivanciuc 2001).

Los DMs se clasifican en dos grupos (2D y 3D). Los DMs pertenecientes al grupo 2D son los siguientes:

- Descriptores constitucionales: Son aquellos que permite describir el peso molecular, el número de átomos en la molécula, el número de enlaces aromáticos simples, dobles y triples, así como el número de anillos aromáticos.
- Descriptores químicos-cuánticos y electrostáticos: Capturan la información de la naturaleza eléctrica de la molécula, polarizabilidad molecular y cargas parciales.
- Descriptores topológicos: Permiten tratar la estructura de la molécula como un grafo donde los nodos son átomos y las aristas son los enlaces químicos. Tales como los índices Radic, Weiner, Balabán, Kier and Hall, Gálvez entre otros (Aguiar et al. 2013)
- Descriptores geométricos: Captura la información de la superficie de la molécula tal como área atómica de *van der Waals* y su superposición, volumen molecular, principales momentos de inercia, índices gravitacionales, áreas de sombra y su área total de la superficie solvente-accesible.
- Descriptores basados en fragmentos y huellas moleculares: Se basan en la identificación de subestructuras o fragmentos dentro de una molécula.

Información sobre los DMs pertenecientes al grupo 3D pueden ser consultados en (Aguiar et al. 2013).

1.2.1.Descriptores basados en fragmentos

Los DMs basados en fragmentos representan subestructuras seleccionadas de un grafo molecular en 2D y su ocurrencia en la molécula. Ellos constituyen uno de los más importantes tipos de DMs (Todeschini 2000.). Su principal ventaja está relacionada con la simplicidad de su cálculo, almacenamiento e interpretación. Los fragmentos subestructurales constituyen descriptores basados en información los cuales codifican la información almacenada en las estructuras moleculares. Cada molécula es representada por un vector de ocurrencia de cada fragmento, que permite separar a los compuestos activos de los inactivos.

Estos descriptores, como regla, se representan como cadenas que pueden ser almacenadas y procesadas por computadoras. Además su rol primario es proveer capacidad de búsqueda en las subestructuras de forma eficiente en largos conjuntos de moléculas que constituyen una base de datos. Constituyendo esto una ventaja bien aprovechada en la realización de búsquedas de similaridad, identificación de conglomerados entre otras actividades en la realización de estudios basados en el enfoque QSAR (Vaskin et al. 2008).

Existen diversas aplicaciones que pueden generar cientos o miles de descriptores diferentes, tales como Dragon (Dragon 2006) o ISIDA (ISIDA 2013), de ahí la posibilidad de aplicar técnicas de aprendizaje automático para determinar modelos que correlacionen la estructura de la molécula con la actividad biológica que estas poseen. Las técnicas de aprendizaje automático son un área de la inteligencia artificial que permiten extraer conocimientos a partir de grandes volúmenes de datos, siendo conocido

este proceso como extracción de conocimiento en base de datos o como *Knowledge Discover in Database* (KDD).

1.3. Extracción de conocimiento en base de datos.

Procesar grandes volúmenes de datos constituye una operación inviable para la capacidad humana, por tal razón han sido desarrolladas disímiles técnicas computacionales, para automatizar el proceso de identificar patrones o extraer conocimientos útiles a partir de grandes volúmenes de datos.

El aprendizaje automático forma parte de estas técnicas computacionales. Permite crear mecanismos para extraer el conocimiento subyacente en la información. El aprendizaje automático puede clasificarse como descriptivo o predictivo. El primero se encarga de describir las características de los datos de forma concisa y resumida. El segundo constituye uno o varios modelos que realizan inferencia[1] sobre un conjunto de datos que se denomina "datos de entrenamiento". Luego, a partir de estos modelos, se intenta predecir el comportamiento de nuevos datos recopilados en el problema. En la bibliografía se utiliza el término "aprendizaje automático supervisado" para las técnicas predictivas y "aprendizaje automático no supervisado" para las técnicas descriptivas.(Ruiz 2005)

El proceso de extracción de conocimiento en base de datos es un proceso que consta de varias etapas. Las etapas que componen la ejecución del KDD son i) determinación del objetivo, ii) preprocesamiento, iii) minería de datos, iiii) análisis de los resultados y iiiii) aplicación de los resultados. Ver Anexo 2.

1.3.1. Preprocesamiento.

Según fue mencionado con anterioridad existen disímiles softwares para el cálculo de descriptores moleculares. Cada uno de ellos es capaz de generar entre cientos y miles de ellos. Solo una parte de los DMs identificados están relacionados con la actividad biológica del compuesto, mientras que el resto pueden estar intercorrelacionados constituyendo información no relevante a utilizar en el KDD. Debido a que las técnicas de aprendizaje supervisado requiere grandes volúmenes de información para obtener un modelo apropiado en el problema específico que se requiera, la incorporación de información no relevante, redundante o atributos altamente correlacionados en el modelo, guían hacia un aprendizaje con sobreajuste[2], situación que atenta contra la interpretabilidad del modelo QSAR final (Vaskin et al. 2008).

[1] Evaluación realizada a partir de varias proposiciones.

[2] Término utilizado para expresar que un algoritmo de aprendizaje automático supervisado fue entrenado con datos únicos o "extraños" pudiendo quedar ajustado a características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo. En presencia de sobreajuste, el éxito al responder ante las muestras de entrenamiento es elevado mientras que su actuación ante muestras de prueba está deteriorado.

Para evitar este problema se debe prestar atención al conjunto de datos a tratar, garantizando que los datos a utilizar estén normalizados y libres de información perdida así como de instancias que constituyan *outliers*[3] (Larose 2005).

Para el desarrollo de esta etapa Alexander Tropsha y su equipo (Tropsha 2010, Tropsha et al. 2010) proponen una metodología para preprocesar conjuntos de datos con información sobre compuestos químicos. En esta metodología se propone aplicar los principios generales del KDD enfocado a las especificidades de esta área la cual propone los siguientes pasos:

- Remover todos los compuestos metálicos y organometálicos pues no son manejados apropiadamente por las herramientas que calculan los descriptores moleculares.
- Realizar una conversión estructural utilizando una herramienta que permita la conversión de la representación de la molécula en un grafo molecular. Con el objetivo de normalizar toda la información.
- Realizar la estandarización de las moléculas, utilizando una herramienta que permita transformar las moléculas contenidas en una base datos. Debido a que una misma molécula puede estar representada por distintos patrones estructurales pudiendo parecer dos moléculas distintas, por lo tanto, el resultado de esta operación permitirá estar listos para identificar si existen duplicados de una molécula.
- Eliminar las moléculas duplicadas.
- Inspeccionar manualmente los datos en busca de otros problemas estructurales.

La limpieza en los datos es un paso importante en el KDD debido a que la obtención de un buen modelo depende directamente de la calidad de los datos.

Las actividades antes mencionadas solo garantizan que toda la información se encuentre en un mismo formato y se eliminen las instancias que pudieran brindar información errónea durante la elaboración del modelo.

1.3.2. Selección de atributos.

Parte del preprocesamiento es la reducción de la dimensionalidad, mediante la selección de los atributos que aportan información relevante sobre la instancia y que garantizan una correcta distinción entre nuevos ejemplos identificados.

La selección de atributos consiste en la determinación de un subconjunto de atributos que aportan mayor información sobre las instancias de un problema y que permiten discriminar la instancia que describen, según sus clases.

[3] Se define para datos que constituyan anomalías o casos raros.

Existen tres enfoques al enfrentar este problema(Davis et al. 2006):

Los algoritmos clasificados como “filtros” se basan en una medida específica para determinar la relevancia de un atributo dentro de un conjunto grande de atributos con respecto al atributo dependiente. La medida utilizada para determinar la relevancia de un atributo es independiente del algoritmo de clasificación a utilizar para construir el modelo deseado. Los algoritmos de filtro se basan en la entropía, información mutua entre otras relaciones existentes entre los datos. (Kohavi et al. 1997, Aguiar et al. 2013)

Los algoritmos clasificados como “envolventes” utilizan como medida, para la selección de atributos, un algoritmo de clasificación. Para la selección de los atributos relevantes se apoyan en la información obtenida tanto de las variables independientes como en la dependiente. En la literatura se sugiere que, para mejores resultados, se utilice el mismo algoritmo utilizado en la selección de atributos, en la tarea de clasificación a realizar debido a que este seleccionará los atributos que le son relevantes según su heurística (Peng et al. 2005).

Los algoritmos híbridos son aquellos que toman ventajas de los enfoques anteriores explotando los diferentes criterios de evaluación en las diferentes etapas de búsqueda. Estos algoritmos pueden ser clasificados como envolventes pero, la característica que los diferencia es que, antes de aplicar el método envolvente aplican un método filtro para así reducir el número de variables a utilizar en la aplicación del envolvente.

1.3.2.1. Selección de atributos utilizando el criterio de información mutua.

En (Peng et al. 2005) se establece que utilizando el criterio de información mutua se puede determinar la relevancia entre conjuntos de atributos y la clase. El objetivo, en todo algoritmo de selección de atributos, es encontrar un subconjunto de atributos que mejor caracterice a la variable dependiente.

Unos de los enfoques más utilizados es la máxima relevancia. Este enfoque determina los atributos con la máxima relevancia con respecto a la variable dependiente. La relevancia se caracteriza en términos de correlación o información mutua. La información mutua entre dos variables aleatorias x y y está definida en términos de las funciones de densidad probabilística $p(x)$; $p(y)$; $p(x,y)$:

$$I(x;y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$

Este autor propone una heurística denominada mínima-redundancia-máxima-relevancia (mRMR). Esta heurística utiliza medidas de relevancia y redundancia para seleccionar los atributos prometedores, tanto para atributos continuos como discretos.

En términos de información mutua la máxima relevancia D se determina a continuación:

$$\max D(S,c), \quad D = \frac{1}{|S|}\sum_{xi\,\epsilon\,S} I(x_i,c)$$

Donde S es el conjunto de m atributos de xi que tiene la mayor dependencia con la clase objetivo c. Se enfatiza que en ocasiones los atributos con mayor relevancia respecto a la clase son redundantes entre ellos, lo que provoca que la combinación de atributos más relevantes no sea la mejor combinación de atributos respecto a la clase objetivo. A partir de este supuesto se determina la mínima redundancia R entre ellos. La mínima redundancia se determina de la siguiente forma:

$$\min R(S),\ \ R = \frac{1}{|S|^2} \sum_{xi,xj \in S} I(xi; xj)$$

A continuación se combina D y R de la forma:

$$\max\ \theta(D, R)\,, \qquad \theta = D - R$$

El objetivo la de búsqueda incremental aplicada por el algoritmo es seleccionar los atributos que maximicen θ.

1.3.3. Balance de clases.

Extraer conocimientos utilizando el aprendizaje automático a partir de datos no balanceados es en la actualidad un desafío, debido al mal funcionamiento de los algoritmos frente a este tipo de situación. El problema de desbalance ocurre cuando el número de instancias que representa una clase es mayor que el número de instancias que representa la otra clase. Esta característica puede ser representada por la razón entre ambas cantidades, según la siguiente fórmula.

$$IR = \frac{A}{B}$$

Donde A representa la cantidad de instancias de la clase mayoritaria y B la cantidad de instancias de la clase minoritaria.

En numerosas situaciones aparece desbalance en la información, debido a la ocurrencia de sucesos raros o debido al alto costo que implica la obtención de la información. Los clasificadores, ante esta situación, obtienen buenos resultados con la clase más representada mientras que respecto a la otra ocurre todo lo contrario. Un punto importante a destacar es que, en conjuntos de datos desbalanceados, la información importante se encuentra en los datos menos representados, sin embargo los clasificadores pueden considerarlos como “ruido” pues no consideran la distribución de las clases. (Chavéz 2008)

Producto de las investigaciones se han creado dos enfoques para lidiar con esta situación. Estos se basan en algoritmos que se centran en el proceso de aprendizaje y en aquellos que se basan en modificar la distribución de los datos.(García et al. 2008)

Debido a que los algoritmos que modifican la distribución de las clases son independientes del algoritmo de clasificación seleccionado, trabajar con estos algoritmos parece ser una decisión acertada debido a que, al igual que los especialistas, manipulan las instancias que pueden ser consideradas ruidosas o con poca ocurrencia entre las instancias de su tipo y de esta manera contribuir a la modificación la distribución de las clases.

Los algoritmos que modifican la distribución de los datos se clasifican en tres categorías: (Richert et al. 2013)

- *Undersampling*: Son aquellos que obtienen subconjuntos de instancias eliminando algunos ejemplos de la clase mayoritaria.
- *Oversampling*: Son aquellos que aumentan el tamaño del conjunto menos representado creando nuevas instancias sintéticas a partir de la información de las instancias pertenecientes a la clase de interés.
- Métodos híbridos: Son aquellos que combinan los enfoques anteriores, eliminando (*undersampling*) algunos de los ejemplos expandidos de la clase minoritaria, como resultados de haber sido aplicado algún método para crear nuevos ejemplos (*oversampling*), buscando eliminar el posible sobreajuste en la clasificación a realizar.

1.3.3.1. Balanceo de clases utilizando el algoritmo *SMOTE* como método Oversampling.

Como fue establecido con anterioridad, los métodos *oversampling* son aquellos que introducen instancias sintéticas, producidas a partir de la información existente, en la clase que está menos representada. La creación de instancias similares, aunque ficticias, ayudan en el entrenamiento de los clasificadores. En (Chawla et al. 2002) se fundamenta la creación de un algoritmo llamado *Synthetic Minority Oversampling Technique* (SMOTE). Este algoritmo utiliza cada instancia de la clase minoritaria para producir e introducir instancias sintéticas en el segmento que une a una instancia con las k instancias más cercanas pertenecientes a la clase minoritaria.

Dependiendo de la cantidad de nuevas instancias requeridas, los vecinos utilizados que pertenecen a los k vecinos más cercanos son seleccionados aleatoriamente.

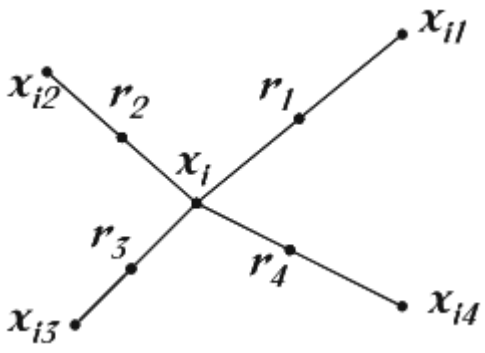


FIGURA 1: PROCESO DE CREACIÓN DE LAS INSTANCIAS SINTÉTICAS

En la Figura 1 xi representa la instancia seleccionada, $xi1, \ldots . xi4$ son los vecinos más cercanos seleccionados y $r1, \ldots, r4$ las instancias sintéticas creadas por una interpolación aleatoria.

Las instancias sintéticas son generadas tomando la diferencia entre la instancia xi y los vecinos más cercanos. A continuación se multiplica esa diferencia por un número aleatorio entre 0 y 1 y se agrega este valor al vector xi. Esta operación provoca la creación de una nueva instancia colocada en un punto aleatorio sobre el segmento que une dos instancias específicas. Este enfoque refuerza la región de

decisión de la clase minoritaria, convirtiéndola en un espacio más general. Lográndose así igualar la cantidad de instancias entre las clases de un problema.

1.4. Minería de datos

La tercera etapa del KDD implica la aplicación de técnicas de aprendizaje automatizado para la obtención de modelos que permitan identificar patrones dentro de la información procesada. Estas técnicas utilizan como entrada vectores que contienen, dentro de sus componentes, variables que describen los sucesos o ejemplos que se desean analizar. Estos datos pueden o no tener asociado una variable que se denomina variable dependiente o predictora. En caso de que los datos no posean esa variable el problema se cataloga como “Aprendizaje Automático no Supervisado” y en caso contrario entonces se cataloga como “Aprendizaje Automático Supervisado”. Esta variable determina el enfoque del análisis a aplicar, en caso de que sea continua entonces se aplica un enfoque de regresión y en caso de que sea discreta se aplica un enfoque de clasificación.

1.4.1. Algoritmos de clasificación.

Este trabajo se centra el caso del Aprendizaje automático supervisado, donde los algoritmos utilizados son denominados “algoritmos de clasificación” o “clasificadores”.

Estos algoritmos se centran en determinar funciones que permiten discriminar los objetos que son utilizados para determinar estas funciones. Denominados “objetos de entrenamiento” o “datos de entrenamiento”.

Los algoritmos de clasificación son mecanismos de aprendizaje, donde el proceso se desarrolla tomando cada instancia u objeto y asignándola a una clase o categoría en particular.

Entre los métodos de clasificación más utilizados están los que realizan un aprendizaje basados en casos, también denominados “perezosos”, los árboles de decisión, las redes bayesianas, redes neuronales artificiales, análisis discriminante, regresión logística, las máquinas de soporte vectorial, las reglas de asociación y otros que constituyen, en su mayoría, variantes de los antes mencionados, para los cuales se puede encontrar bibliografía en múltiples revistas(Cabrera et al. 2013).

1.4.1.1. Aprendizaje basado en casos (perezosos).

Estos algoritmos se basan en la premisa de usar experiencias para resolver un problema. Estos algoritmos necesitan de una serie de ejemplos que serán utilizados como entrenamiento y al momento de enfrentarse a una nueva situación lo hacen buscando semejanzas entre la nueva situación y los ejemplos utilizados en el entrenamiento. En este caso, cada nueva instancia se compara con el resto de la base de casos a través de alguna medida de similaridad, ya sea la distancia euclideana u otras. La clase con la que será catalogada la nueva instancia dependerá de la clase de la instancia que más similar sea a la nueva instancia, esta similitud se infiere por la cercanía entre ambas instancias, debido a que se asume que estos objetos pueden ser representados como puntos en un espacio y por tanto objetos similares estarán más cercanos en dicho espacio. A este proceso se le denomina método del “vecino

más cercano". Si en lugar de usar el caso más cercano se utilizan los K casos más cercanos, entonces se denomina los k-vecinos más cercanos (KNN). (Mitchel 2007, Cabrera et al. 2013)

1.4.1.2. Árboles de decisión.

Los árboles de decisión son uno de los métodos más usados y prácticos para la inferencia inductiva. Implementan un método de aproximación de funciones que es robusto a los ruidos. Las funciones aprendidas por el árbol pueden ser representadas como un conjunto de reglas que mejoran la interpretabilidad por parte de los humanos.

La clasificación de las instancias se realiza a partir de la estructuración de un árbol obtenido a partir de los datos de entrenamiento. Cada nodo del árbol provee una prueba a algún atributo de la instancia, donde cada rama generada a partir de esa prueba, corresponde a cada uno de los posibles valores del atributo probado. La clasificación corresponde a la hoja ubicada en la rama tomada a partir de la última prueba realizada a la instancia.

En (Quinlan 1986) se establece, con el ID3, el paradigma de los métodos para identificar reglas a partir de los árboles de decisión, pero tiene limitaciones. La primera limitación es cuando el conjunto de entrenamiento posee ruidos y/o cuando el tamaño del conjunto es pequeño; siendo estas razones que pueden llevar a la creación de modelos no generalizadores debido a la aparición del sobreajuste. Esta situación es resuelta mediante la detención del crecimiento del árbol antes de alcanzar la clasificación perfecta o realizar podas al árbol sobreajustado. Otra de las limitantes es que ID3 trabaja con atributos pertenecientes al dominio discreto, por lo tanto, no puede ser aplicado a problemas con atributos pertenecientes al dominio continuo. Esta situación puede ser resuelta si se discretizan los atributos continuos del problema.

El algoritmo C4.5 es una variante del ID3 que puede ser aplicado a problemas con datos pertenecientes al dominio continuo, usa puntos de corte en las pruebas hechas a los atributos e introduce medidas para evitar el sobreajuste, en particular los criterios de parada en la división y poda del árbol. Utiliza el criterio razón de ganancia para así evitar que las variables con mayor cantidad de valores sean beneficiadas en la selección. C4.5 realiza una poda al árbol inducido basado en pruebas de hipótesis para determinar si es recomendable extender o no una rama del árbol.(Quinlan 1993, Mitchel 2007, Cabrera et al. 2013)

1.4.1.3. Redes bayesianas.

Son redes probabilistas que permiten representar gráficamente las variables y las relaciones entre las variables de un problema. Proveen información sobre las relaciones de dependencia e independencia condicional existentes entre las variables. La inclusión de las relaciones de independencia en la propia estructura de la red permite representar conocimiento de forma compacta pues reduce el número de parámetros necesarios. (Chavéz 2008)

Las redes bayesianas se definen como un grafo que cumple con las siguientes características:

- Los nodos de la red son variables aleatorias que pueden representar los atributos en un problema de clasificación.

- Cada par de nodos se conectan entre sí mediante arcos dirigidos. El significado de que un enlace va de un nodo a otro es que el primero tiene una influencia directa sobre el segundo.

- Por cada nodo hay una tabla de probabilidad condicional que sirve para cuantificar los efectos de los padres sobre el nodo.

- El grafo no tiene ciclos dirigidos por lo que se denomina Grafos dirigidos acíclicos.

Una red bayesiana puede calcular la distribución de probabilidad para cualquier subconjunto de variables de la red, dado los valores o distribuciones de las variables restantes (Mitchell 1997).

Este tipo de clasificador no es muy sensible a los cambios de sus parámetros, ya que se basa en información de toda la base, lo cual hace que pequeños cambios en la base no sean necesariamente significativos. (Cabrera et al. 2013)

1.4.1.4. Redes neuronales artificiales.

Las redes neuronales artificiales (RNA) son algoritmos que tratan de simular el cerebro biológico. No obstante las RNA son muy simples con respecto a un cerebro biológico. Una RNA es un sistema computacional paralelo que consiste en muchos elementos de procesamiento conectados entre ellos en un sentido específico con el objetivo de resolver una tarea. Las redes neuronales artificiales han ganado popularidad debido a que son adaptativos, robustos y tolerantes a fallos y/o ruido.(Yao et al. 2005)

Las redes neuronales artificiales son construidas bajos diferentes estructuras, según su topología y tipo de aprendizaje, Según (Aguiar et al. 2013) las redes más frecuentes son las redes encaminadas hacia adelante (*feed-forward networks*) o también llamadas redes unidireccionales tales como las basadas en el *perceptron*[4] y las de función de base radial. En estas redes la información fluye en una sola dirección: desde las neuronas de entrada hasta las neuronas de salida pasando a través de un conjunto de capas ocultas.

Redes perceptron multicapas: Compuestas por perceptrones interconectados distribuidos en capas. Cada *perceptron* es capaz de combinar los valores de entrada, devolviendo como salidas valores tanto continuos como binarios. Es importante destacar que las entradas de los perceptrones tienen asociados pesos que representan la relevancia de la entrada.(Yao et al. 2005)

Redes de base radial: Compuestas por tres capas (una capa de entrada, una capa oculta y una capa de salida). Operan en base a la distancia que separa el vector de entradas respecto al vector sináptico que cada neurona almacena, cantidad a la que es aplicada una función radial con forma gaussiana. En esta arquitectura, a diferencia de la anterior, las neuronas son de respuesta localizada ya que solo responden con una intensidad apreciable cuando el vector de entrada presentado a cada neurona y el centroide almacenado por esta pertenece a una zona próxima en el espacio de las entradas (Martin et al. 2001).

[4] Modelo neuronal unidireccional introducido por Rosenblatt a finales de los años cincuenta.

1.4.1.5. Máquinas de soporte vectorial.

Se basan en el principio de minimización del riesgo estructural a partir de la teoría de aprendizaje estadístico. Es un clasificador que es poco afectado por los datos duplicados y tiene poco riesgo de sufrir sobreajuste. (Dehmer et al. 2012)

En un problema de dos clases, el objetivo del clasificador es encontrar la mejor función de clasificación para distinguir miembros entre clases en un conjunto de datos de entrenamiento. Para una función de clasificación lineal, la función corresponde a un hiperplano que pasa por el medio de las dos clases separándolas. Una vez que esta función es determinada, una nueva instancia puede ser clasificada simplemente probando el signo de la función. La nueva instancia pertenecerá a la clase positiva si el signo de la función es positivo o pertenecerá a la clase negativa si el signo de la función resulta negativo.

Existen muchos hiperplanos, por lo tanto, las máquinas de soporte vectorial garantizan adicionalmente que la mejor función es encontrada maximizando el margen entre las dos clases. Intuitivamente el margen es definido como la cantidad de espacio o la separación entre las dos clases. (Wu et al. 2008)

1.4.1.6. Análisis discriminante.

Es una técnica multivariada que soluciona problemas de clasificación. Funciona con datos que ya están clasificados en grupos para de esta forma derivar reglas y así clasificar nuevos individuos. La sugerencia de Fisher es buscar una transformación lineal a las variables predictoras de forma que la separación entre los grupos sea maximizada.(Casas et al. 2013)

El objetivo del análisis es encontrar la combinación lineal de las variables independientes que mejor permita diferenciar los grupos. Una vez encontrada esta función podrá ser utilizada para clasificar nuevos casos.

El análisis discriminante aprovecha la información contenida en las variables independientes para crear una función lineal de la forma:

$$D = b_1 * X_1 + b_2 * X_2$$

Donde $b1$ y $b2$ son ponderaciones de las variables que consiguen hacer que las instancias de uno de los grupos obtengan puntuaciones máximas o mínimas para los sujetos del otro grupo.

La función *D* debe ser tal que la distancia entre los centroides de cada grupo sea la máxima posible, consiguiendo de esta forma que los grupos se encuentren lo más distantes posibles.(Inc 2006)

1.4.1.7. Regresión logística.

La regresión logística es una técnica estadística multivariada, útil para situaciones en las cuales se desea predecir la presencia o ausencia de características o salidas basadas en los valores de las variables

independientes. Los coeficientes pueden ser utilizados para estimar la razón de ventaja para cada variable independiente en el modelo. En este tipo de problemas la variable dependiente tiene que ser dicotómica (cuya ausencia o presencia se ha establecido con los valores cero y uno respectivamente).(Casas et al. 2013).

Es usado para modelar la probabilidad de la ocurrencia de algún evento como una función lineal de un conjunto de rasgos o variables independientes. De esta forma se calcula la probabilidad de que una instancia desconocida pertenezca a un grupo. Pudiendo ser descrita la influencia de una variable independiente (según el valor del coeficiente determinado) sobre la variable dependiente. Cuando el valor del coeficiente es alto implica que esta variable independiente afecta fuertemente a la variable dependiente.(Dehmer et al. 2012). Esta técnica es ampliamente utilizada para realizar predicciones utilizando una base de entrenamiento categorizada en dos grupos.(Ding et al. 2005)

1.4.1.8. Reglas de asociación.

Son algoritmos que extraen reglas directamente del conjunto de entrenamiento suministrado, siendo estas reglas más expresivas que las reglas propuestas por otros algoritmos como los árboles de decisión, los algoritmos genéticos u otros que puedan ser utilizados para proponer reglas de clasificación.(Mitchel 2007).

Estos algoritmos se agrupan por el tipo de enfoque que utilizan:

Cubrimiento secuencial: Son la familia de algoritmos que utilizan como estrategia la inducción de una regla. A continuación eliminan las instancias que son cubiertas por esa regla y nuevamente inducen otra regla a partir de las instancias que no fueron cubiertas por la regla anterior y así sucesivamente hasta que se logre dar un cubrimiento total a las instancias de entrenamiento. La inducción de la regla tiene gran efectividad, pero no necesariamente un alto cubrimiento, es decir, son reglas que identifican muy bien a las instancias pero no necesariamente a todas las instancias. El conjunto final de reglas obtenidas es ordenado de modo que las reglas más efectivas serán consideradas primero cuando una nueva instancia necesite ser clasificada.

Cubrimiento simultáneo: Son la familia de algoritmos que tienen un funcionamiento parecido al ID3, funcionan similar a la estrategia anterior pero organizan el espacio de búsqueda similar al ID3. La búsqueda inicia considerando la regla más general, entonces agrega el atributo de prueba que mejore el desempeño de la regla medido sobre los ejemplos de entrenamiento. A continuación el proceso se repite agregando nuevos atributos de prueba hasta que se alcance un nivel aceptable de desempeño.

1.5. Multiclasificadores.

Cuando como resultado de la aplicación de diferentes clasificadores no se llega a un resultado definitivo producto que, mediante la utilización de las diferentes métricas para evaluar el desempeño, no se encuentra algún clasificador que satisfaga las necesidades del investigador, la combinación de clasificadores puede ser un área interesante a explorar. Resultados sobre como la combinación de

clasificadores demuestran ventajas sobre los clasificadores individuales han sido registrados en la literatura. (Kuncheva et al. 2003, Kuncheva 2007, Bonet et al. 2012)

Razones estadísticas, computacionales y de representación son expuestas como razones que pudieran justificar que utilizar un multiclasificador puede ser mejor que utilizar un clasificador simple. La razón estadística se basa en que dado un conjunto de datos y un conjunto de clasificadores, con generalizaciones diferentes sobre el conjunto de datos, generan incertidumbre sobre la elección de clasificador adecuado para su utilización. En cambio, en vez de seleccionar un clasificador de forma segura, se pueden utilizar todos los clasificadores y promediar sus salidas. El modelo obtenido puede no sea mejor que el mejor clasificador individual utilizado, pero su uso disminuirá o eliminará el riesgo de haber seleccionado un clasificador inadecuado para el problema. La razón computacional se expone debido que algunos algoritmos pueden caer en óptimos locales a consecuencia de su heurística. Se asume que el entrenamiento de cada clasificador comienza en algún lugar del espacio de soluciones y que terminan cerca del clasificador óptimo. Alguna forma de agregación guiará al multiclasificador que es la mejor aproximación al clasificador óptimo antes que seleccionar un clasificador individual que más se aproxime al clasificador óptimo. La razón referente a la representación se basa en que es posible que el espacio considerado no contenga el óptimo clasificador, debido a que el espacio fue restringido a clasificadores lineales y el óptimo buscado no pertenece a él. Un multiclasificador de clasificadores lineales puede aproximar cualquier frontera de decisión. En este caso puede ser justificado el uso de un multiclasificador que directamente entrenar un clasificador complejo.

En (Segrera et al. 2006) se especifica que la combinación de clasificadores muestra mayor precisión que cualquiera de los clasificadores individuales. Una condición necesaria y suficiente para que esto se cumpla es que los clasificadores base sean precisos y diversos. Los clasificadores se consideran precisos si su error es inferior a 0.5 y son considerados diversos aquellos cuyos errores de salida no son correlacionados.

Los algoritmos multiclasificadores a analizar en este trabajo son: Bagging(Breiman 1996), Boosting(Schapire 1990), Stacking(Wolper 1992) y Vote, todos implementados en la plataforma Weka(Hall et al. 2009), aunque existen muchos otros.

Los algoritmos Bagging y Boosting garantizan la diversidad pues realizan el entrenamiento logrando generar distintas muestras a partir de una selección de instancias con reemplazo de manera que propician la obtención de muestras *bootstrap* donde pueden haber ejemplos repetidos. Ambos combinan las salidas utilizando la técnica del voto mayoritario y el voto mayoritario pesado respectivamente, que será explicado más adelante. Los algoritmos Stacking y Vote esperan la diversidad a partir de los clasificadores base utilizados, por lo tanto un buen resultado depende de cuan diversos sean las salidas de los clasificadores base utilizados.(Segrera et al. 2006, Bonet et al. 2012)

1.5.1. Bagging.

Bagging es un algoritmo propuesto por Breiman, se basa en la generación de conjuntos de entrenamiento aleatorios con reemplazo. En este caso una muestra se crea al realizar un muestreo uniforme de forma

aleatoria sobre el conjunto de entrenamiento original de manera que es posible que se tengan instancias repetidas en la creación de estas muestras. Por tal motivo se crean distintas muestras, de ahí que el clasificador se entrena con las diferentes muestras y por lo tanto se garantiza la diversidad propuesta anteriormente. El funcionamiento de este algoritmo se muestra en la Figura 2. Para clasificar una instancia desconocida se predice la clase por cada clasificador obtenido a partir de las distintas muestras obtenidas y para emitir la decisión final se utiliza el método de voto mayoritario.

Este algoritmo también puede ser utilizado para realizar predicciones numéricas, es decir, que la variable dependiente posee un valor continuo. En este caso para realizar la predicción de este valor en una nueva instancia se promedian los valores obtenidos en cada clasificador obtenido. (Cabrera et al. 2013)

1.5.2. Boosting.

Propuesto por Schapire, se basa en la asignación de un peso a cada ejemplo del conjunto de entrenamiento. En cada iteración este algoritmo aprende un modelo que minimiza la suma de los pesos de los ejemplos mal clasificados. Los errores de cada iteración se utilizan para actualizar los pesos de los ejemplos errados y los ejemplos bien clasificados, de forma que una instancia mal clasificada tendrá mayor preferencia para ser utilizada al momento de confeccionar el conjunto de entrenamiento. Situación que permite al clasificador centrarse en las instancias difíciles de clasificar.

AdaBoost.M1 es una variante bien conocida y se encuentra implementada en la plataforma Weka para problemas de clasificación. Para este algoritmo, en cada ciclo, se selecciona una muestra y se aprende un modelo a través de la evidencia ponderada, se estima el error del modelo y dependiendo del valor del error se detiene el algoritmo o se continúa el proceso de aprendizaje. A diferencia del Bagging este método toma como criterio de parada si el error >=0.5 o no poder seguir disminuyendo el error. (Segrera et al. 2006, Cabrera et al. 2013)

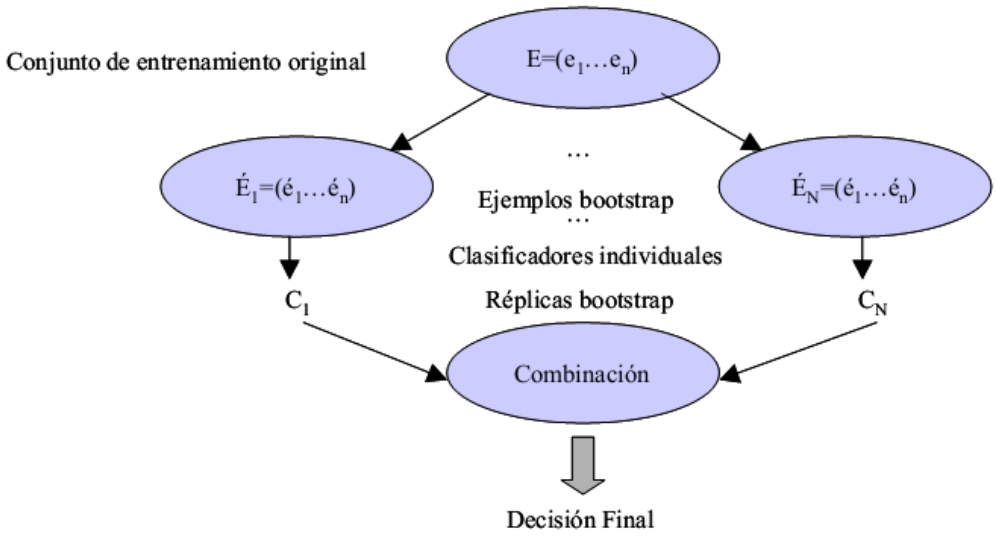


FIGURA 2: ESQUEMA DEL ALGORITMO BAGGING

1.5.3. Stacking.

Propuesto por Wolper (Wolper 1992), a diferencia de los algoritmos explicados anteriormente, la diversidad debe ser garantizada por los clasificadores que utiliza como base. Combina múltiples modelos obtenidos a partir de diferentes clasificadores para un mismo conjunto de datos. Para la combinación de las salidas no utilizan el método del voto mayoritario sino que introduce un metaclasificador que aprende de la relación entre las salidas de los clasificadores de base y la clase original.

Este metaclasificador utiliza como base de entrenamiento las predicciones realizadas por los clasificadores bases utilizados, donde por cada instancia se tiene un vector de atributos compuestos por la clase de salida de cada clasificador base y como atributo dependiente la clase original de la instancia (Segrera et al. 2006, Bonet et al. 2012, Cabrera et al. 2013). El funcionamiento de este algoritmo se presenta en la Figura 3.

1.5.4.Vote.

Al igual que el Stacking, este algoritmo espera la diversidad en la selección de los algoritmos de clasificación utilizados como base. Las salidas de estos clasificadores están dadas por vectores con una distribución de probabilidad para cada clase. Vote combina estas probabilidades para emitir un criterio final.

Posee la capacidad de combinar las salidas utilizando los métodos propuestos en (Kuncheva 2004) tales como: Voto mayoritario, promedio de las probabilidades, entre otros.

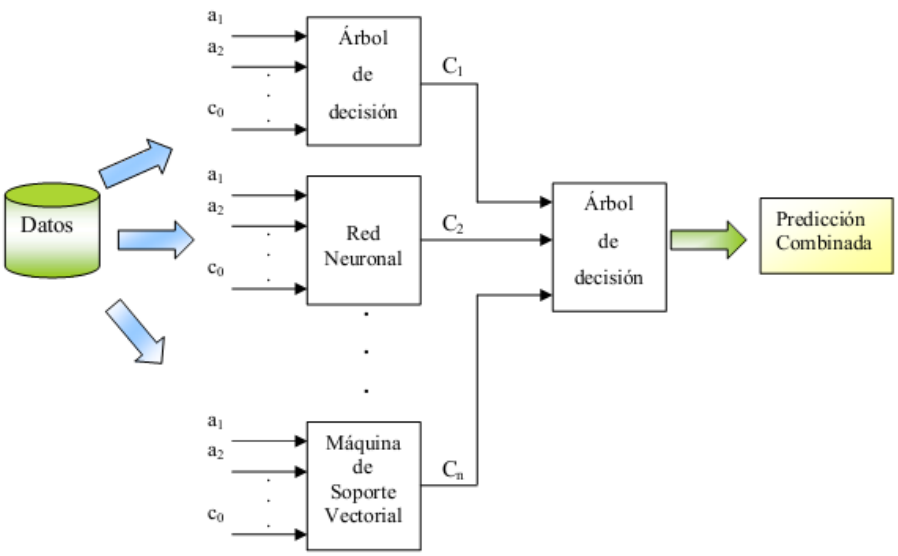


FIGURA 3: ESQUEMA DEL ALGORITMO STACKING

1.5.4.1. Combinación de salidas.

En ocasiones es oportuno combinar las salidas de clasificadores individuales y de diferentes grados de complejidad entrenados sobre un único conjunto de entrenamiento, aprovechando que esta estrategia puede disminuir el tiempo de entrenamiento y aumentar la robustez y el desempeño de la clasificación

(Tax et al. 2003). Otra de las ventajas de la combinación de clasificadores es que de esta manera se aprovecha el conocimiento obtenido a partir de la búsqueda realizada por cada clasificador sobre diferentes áreas del espacio de soluciones, lo que garantiza un cubrimiento de espacio más amplio que el realizado de forma individual por cada clasificador.

La combinación de las salidas de los clasificadores bases se realiza a partir de la utilización de los vectores de distribución de probabilidades emitidos por cada clasificador base. Para comprender algunas de las funciones de combinación de las salidas considerar *N* clasificadores para un problema de clasificación de *M* clases. Cada clasificador emite un vector con una distribución de probabilidad para cada clase.

La probabilidad de un conjunto de clasificadores para cada una de las clases de un caso X se puede representar por la siguiente matriz.

$$MP = \begin{bmatrix} P(N_1, M_1) & \dots & P(N_1, M_j) & \dots & P(N_1, M_n) \\ \dots & \dots & \dots & \dots & \dots \\ P(N_i, M_1) & \dots & P(N_i, M_j) & \dots & P(N_i, M_n) \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

Donde las filas representan las salidas de clasificador Ni para todas las clases Mj. Donde el clasificador Ni tiene para la clase Mj un estimado de probabilidad $P(Ni, Mj),\ j = 1,2,3, \dots, n$.

Por lo tanto: $\sum_{j=1}^{n} P(Ni, Mj) = 1$

Las columnas representan el estimado de los clasificadores $N_1, N_2, \dots, N_i$ para las clases $M_1, M_2, \dots, Mj$.

A partir de la información de la matriz *MP* se definen las funciones de combinación de interés para este trabajo.

En (Kittler et al. 1998) se establece que existen varios esquemas para combinar las salidas emitidas por varios clasificadores por separado. Estas combinaciones se realizan utilizando la probabilidad posterior emitida en los vectores de distribución de probabilidad obtenida de cada clasificador base.

El algoritmo Vote propone que la combinación de las salidas de sus clasificadores base puede hacerse según el voto mayoritario, la suma, el producto, el máximo, el mínimo, el promedio y la mediana de las probabilidades de cada clase para cada una de las salidas de los clasificadores bases.

En este trabajo se utilizan dos de esos esquemas. El voto mayoritario y el promedio de las probabilidades.

1.5.4.1.1. Voto mayoritario.

Esta estrategia de combinación establece que una nueva instancia será clasificada según el criterio mayoritario entre todos los clasificadores.

En (Kittler et al. 1998) se establece este esquema de clasificación se obtiene de la siguiente manera:

Las probabilidades posteriores $P(Ni, Mj)$ para producir salidas binarias se determina como:

$$MPij = \begin{cases} 1, & si\ P(Ni, Mj) = max_{j=1}^{m} P(Ni, Mj) \\ 0, & de\ otra\ manera \end{cases}$$

Y partir de esta información el voto mayoritario (VM) se obtiene como:

$$VM = max_{j=1}^{m} \sum_{i=1}^{n} MPij$$

Donde *m* es el total de clases y *n* es el total de clasificadores, por lo tanto no es más que la selección de la máxima votación entre las votaciones hechas por todos los clasificadores sobre cada clase.

1.5.4.1.2. Promedio de las probabilidades

El promedio de las probabilidades (*PP*) se basa en que la decisión final se toma como la media aritmética entre las probabilidades posteriores para una clase.

En (Kittler et al. 1998) establece que, basándose en la matriz *MP*, este esquema se obtiene de la siguiente manera:

$$PP = max_{j=1}^{m} \frac{1}{N} \sum_{i=1}^{N} MPij$$

El promedio de las probabilidades se establece tomando el máximo valor de entre todos los promedios de las probabilidades por clases. En caso de que una de las salidas fuera un *outlier* entonces esta medida de tendencia central se ve afectada por sus extremos. En este caso (Kittler et al. 1998) propone sustituir el promedio por la mediana como medida de tendencia central.

1.5.5. Medidas de diversidad para la determinación de clasificadores diversos.

En (Kuncheva 2004, Bonet et al. 2012) se establece que los mejores resultados en la clasificación se obtienen con la combinación de clasificadores más diversos. La diversidad se establece como criterios que permiten evaluar si las salidas de los clasificadores muestran los resultados de una exploración diferente sobre el espacio de soluciones.

En (Kuncheva 2004) se recopilan varias medidas para calcular la diversidad. Estas medidas se pueden clasificar en medidas por pares de clasificadores y medidas respecto a todos los clasificadores (ver Tabla 2). En (Brown et al. 2005) se refiere a las medidas por pares de clasificadores como el promedio de una medida de distancia en particular entre todos los posibles pares de clasificadores. De igual manera se refiere a las medidas respecto a todos los clasificadores como medidas que usan la entropía o la correlación de cada miembro de los clasificadores base respecto al resto. En (Brown et al. 2005) enfatiza que no se puede establecer que existe un vínculo prominente entre la diversidad de los clasificadores base de un multiclasificador y su desempeño.

Las medidas en forma de pares de clasificadores se calculan utilizando las salidas de dos clasificadores bases (Ci,Cj). Donde las salidas son datos binarios (1,0).

	Cj correcto(1)	Cj incorrecto(0)
Ci correcto(1)	A	B
Ci incorrecto(0)	C	D
Total de casos= N = A+B+C+D		

TABLA 1: TABLA DE CLASIFICACIÓN DE RESULTADOS ENTRE CLASIFICADORES

Donde A es la cantidad de instancias que ambos clasificadores clasificaron correctamente, B es la cantidad de instancias que Ci clasificó correctamente y Cj clasificó incorrectamente, C es la cantidad de instancias que Ci clasificó incorrectamente y Cj clasificó de forma correcta y D es el total de instancias que ambos clasificadores clasificaron incorrectamente.

Las medidas de diversidad por pares de clasificadores se calculan por cada par de clasificadores de un total de L clasificadores. Los L clasificadores producen un total de T pares de valores según la siguiente fórmula:

$$T = \frac{L(L-1)}{2}$$

Para obtener un único valor de diversidad entre todos los T pares de valores se hace necesario promediar estos T pares de valores (Kuncheva 2004).

En (Kuncheva 2004, Cabrera et al. 2013) se recogen varias formas de determinar la diversidad. Tal como se muestra en la tabla 2:

Medidas de diversidad	
Por pares(*pairwise*)	**Grupales(*non-pairwise*)**
Coeficiente de correlación	Entropía
Estadístico Q	Varianza de Kohavi-Wolpert
Medidas de diferencia	Desacuerdo de expertos
Medidas de doble fallo	Medida de dificultad
Combinación de la medida de diferencia y de doble fallo.	Diversidad generalizada
	Diversidad de coincidencia de fallos
	Diversidad de distintos fallos
	Medida de variabilidad

TABLA 2: RELACIÓN DE MEDIDAS DE DIVERSIDAD REPORTADAS.

La selección de las medidas de diversidad, utilizadas en este trabajo, está basado en el teorema "*No free lunch*" debido a que no existe alguna medida que sea la mejor opción en la solución de algún tipo de problema, por tal razón su elección realizada se fundamenta en la selección en aquellas medidas que han sido utilizadas en la solución de problemas similares. En (Bonet et al. 2012, Bonet et al. 2013), son utilizadas las medidas pareadas las cuales se describen a continuación:

- Coeficiente de correlación: El coeficiente de correlación puede ser calculado para pares de clasificadores. Existen *C* coeficientes por cada par de clasificador y la media aritmética de

estos es una manera simple de calcular esta medida. Por el contrario al ser calculada a partir de la Tabla 1, dado que las entradas de esta tabla son las probabilidades para los pares de salidas, mientras menor sea el valor del coeficiente, mayor será la diversidad (Kuncheva 2007, Cabrera et al. 2013). Su valor oscila entre -1 y 1. Se calcula según la siguiente ecuación:

$$\rho i,j = \frac{(A*D)-(B*C)}{\sqrt{(A+B)*(C+D)*(A+C)*(B+D)}} \quad -1 \leq \rho \leq 1$$

- Estadístico Q: Para un par de clasificadores estadísticamente independientes, su valor de $Q_{i,j}$ va a ser 0. En general, el valor de Q va a oscilar entre −1 y 1. Aquellos clasificadores que tienden a reconocer los mismos objetos correctamente tendrán un valor positivo de Q, y aquellos que comentan errores en diferentes objetos poseerán un valor negativo de Q (Kuncheva 2007, Cabrera et al. 2013). Mientras menor sea su valor mayor diversidad expresa. El estadístico Q calcula según la siguiente ecuación:

$$Qi,j = \frac{(A*D)-(B*C)}{(A*D)+(B*C)} \quad -1 \leq Q \leq 1$$

- Medidas de diferencia: Esta medida es igual a la probabilidad de que dos clasificadores discrepen en sus clasificaciones (Kuncheva 2007, Cabrera et al. 2013). Su valor oscila entre 0 y 1. Mientras mayor sea su valor mayor diversidad expresa. Se calcula según la siguiente ecuación:

$$Di,j = \frac{B+C}{N} \quad 0 \leq D \leq 1$$

- Medidas de doble fallo: Esta medida nos permite conocer la probabilidad de que dos clasificadores estén equivocados en sus clasificaciones, ya que es más importante conocer cuando ambos clasificadores cometen errores simultáneos que cuando ambos clasificadores están correctos (Kuncheva 2007, Cabrera et al. 2013). Su valor oscila entre 0 y 1. Mientras menor sea el valor mayor diversidad expresa. Se calcula según la siguiente ecuación:

$$DFi,j = \frac{D}{N} \quad 0 \leq DF \leq 1$$

- Combinación de Doble Fallo y Medida de diferencia: Es una propuesta donde se combinan la medida de diferencia y la medida de doble fallo. Mientras más alto sea el valor de esta medida mayor será la diversidad (Bonet et al. 2013). Se calcula según la siguiente ecuación:

$$R = \frac{Di,j}{DFi,j}$$

1.5.6. Análisis de las medidas de diversidad.

El análisis a realizar sobre el resultado del cálculo de las medidas de diversidad es complejo debido a la gran cantidad de valores a considerar. En el análisis a realizar es necesario tener en cuenta la orientación de las medidas de diversidad pues el valor alcanzado por estas permite identificar si una combinación es diversa o no. Las medidas de diversidad pueden ser agrupadas según la orientación que posean sus

valores: i) medidas de orientación ascendente y ii) medidas de orientación descendentes. Las medidas con orientación ascendentes son aquellas que expresan mayor diversidad cuanto mayor sea el valor alcanzado por ellas. Por el contrario aquellas que poseen una orientación descendente son aquellas que expresan mayor diversidad cuanto menor sea el valor alcanzado por ellas. Seleccionar las combinaciones de clasificadores más diversos implica que estas combinaciones cumplan con el criterio de mayor diversidad, tanto en las medidas con orientación ascendente como en las medidas con orientación descendente. Las combinaciones más diversas serán aquellas que alcanzan la mayor diversidad en la mayor cantidad de medidas de diversidad calculadas. Ver orientación de las medidas en la Tabla 3.

Medida de diversidad	Nomenclatura	Orientación
Coeficiente de correlación	ρ	Mientras menor sea el valor alcanzado más diversa es la combinación de clasificadores.
Estadístico Q	Q	
Medidas de doble fallo	DF	
Medidas de diferencia	D	Mientras mayor sea el valor alcanzado más diversa es la combinación de clasificadores.
Combinación de "Medida de diferencia" y "Doble Fallo"	R	

TABLA 3: ORIENTACIÓN DE LAS MEDIDAS DE DIVERSIDAD

1.6. Medidas para evaluar el desempeño de los clasificadores.

Las medidas de evaluación del desempeño de algoritmos de clasificación permiten determinar, en un ambiente de competencia, cuál de los clasificadores se comporta mejor frente al problema que se está investigando. Estas medidas se obtienen a partir de la matriz de confusión. En un problema de dos clases con clases "positiva" y "negativa" la matriz de confusión se representa como en la Tabla 4.

Matriz de confusión		**Clases predichas**	
		positiva	**negativa**
Clases reales	**positiva**	TP	FN
	negativa	FP	TN

TABLA 4: MATRIZ DE CONFUSIÓN

TP: Representa la cantidad de casos positivos identificados correctamente.

FP: Representa la cantidad de casos negativos identificados erróneamente como positivos.

FN: Representa la cantidad de casos positivos identificados erróneamente como negativos.

TN: Representa la cantidad de casos negativos identificados correctamente.

A partir de esta información se pueden calcular las medidas utilizadas para la evaluación de clasificadores (Chavéz 2008), que son mencionados a continuación.

Error: Es el porciento que representan los errores cometidos en la clasificación respecto al total de casos analizados.

$$Error = (FP + FN)/(TP + FP + TN + FN)$$

Efectividad: Es el porciento que representan los aciertos realizados en la clasificación respecto al total de casos analizados.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} = 1 - Error$$

Kappa: Utilizado para medir el acuerdo entre las clases predichas y las clases reales, es análogo al Coeficiente de correlación de Matthews. El valor que arroja esta medida varía en el intervalo de [-1,1], donde un valor cercano a -1 indica que existe un desacuerdo total entre las clases predichas y las clases observadas. Cuando el valor obtenido es cercano a 1 indica que existe un acuerdo perfecto entre las clases predichas y las clases observadas por el clasificador. Esta medida cuando es utilizada en un problema de dos clases no es más descriptiva que la curva ROC(García et al. 2008).

$$kappa = \frac{n\sum_{I=1}^{c} x_{ii} - \sum_{i=1}^{c} x_{i.}x_{.i}}{n^2 - \sum_{i=1}^{c} x_{i.}x_{.i}}$$

Existen otras cuatro medidas que directamente evalúan la clasificación de las clases de forma independiente.

TP_{rate}: Permite calcular el porciento de instancias positivas clasificadas correctamente como perteneciente a la clase positiva. También conocida como Recall y Sensibilidad.

$$TPrate = TP\ /(TP + FN)$$

TN_{rate}: Permite calcular el porciento de instancias negativas clasificadas correctamente como perteneciente a la clase negativa. También conocida como Especificidad.

$$TNrate = TN\ /(TN + FP)$$

FP_{rate}: Permite calcular el porciento de instancias negativas clasificadas incorrectamente como perteneciente a la clase positiva.

$$FPrate = FP\ /(TN + FP)$$

FN_{rate}: Permite calcular el porciento de instancias positivas clasificadas incorrectamente como pertenecientes a la clase negativa.

$$FNrate = FN\ /(TP + FN)$$

Precisión: Permite calcular el porciento que representan las instancias clasificadas como positivas que verdaderamente son positivas(Davis et al. 2006).

$$Precision = TP/(TP + FP)$$

MCC: El coeficiente de correlación de Matthews permite medir la calidad de los clasificadores (Witten et al. 2005). Devuelve un valor entre [-1,1] donde un valor cercano a -1 indica que la predicción no tiene relación entre lo observado y lo predicho por el clasificador. En cambio un valor cercano a 1 indica que si existe dicha relación. Un predicción perfecta permitirá obtener un MCC=1.

$$MCC = (TP * TN) - (FP * FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Medida F: Es la medida de precisión de un clasificador. Se calcula combinando la Precisión y el Recall (Richert et al. 2013). Donde el valor de β está relacionado con la importancia que se le asigna a la Precisión o al Recall. En el caso de aprendizaje automático supervisado nos interesa que ambos parámetros tengan la misma importancia por lo tanto $\beta = 1$. Esta medida nos permite evaluar el desempeño de un clasificador para la clase de interés por lo tanto este deberá ser calculado para ambas clases con el objetivo de saber si el clasificador es bueno en cada clase. Esta función devuelve un valor que varía entre [0,1]; mientras más cercano a uno sea este valor mejor es el clasificador (Guo et al. , Billsus et al. 1998).

$$F = \frac{(1 + \beta^2) * Precision * Recall}{(\beta^2 * Precision) + Recall}$$

La utilización de las medidas TP_{rate}, FP_{rate}, TN_{rate}, Acc y el Error, frente a problemas en dominios imbalanceados, no es recomendable pues son medidas sensibles a la distribución de los datos (Richert et al. 2013) y por lo tanto no son útiles en ese caso. Para ello se recomienda utilizar como medida para evaluar el desempeño de un clasificador el área debajo de la curva del gráfico *Receiver Operating Characteristic* (ROC), debido a que esta medida no es sensible a la distribución de los datos. Este gráfico visualiza la relación entre los beneficios (TP_{rate}) y el costo (FP_{rate}) en la clasificación.

Espacio ROC: La forma de graficar un espacio ROC es mediante un gráfica de dos dimensiones donde en el eje de las Y estará el porciento de instancias positivas correctamente clasificadas (TP_{rate}) y en el eje de las X estará el porciento de instancias negativas clasificadas como positivas (FP_{rate}).

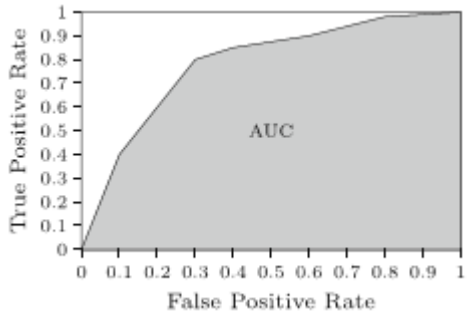


FIGURA 4: REPRESENTACIÓN DEL AUC

En este gráfico los puntos (0,0) y (1,1) son clasificadores triviales en los cuales el 100% de una de las clases es etiquetada correctamente y la otra no. El punto de especial atención es el (0,1) pues significa que todos los ejemplos positivos fueron clasificados correctamente y ningún ejemplo negativo fue etiquetado como positivo.

Los puntos ROC son fracciones del beneficio y el costo total. Permiten comparar el desempeño de los clasificadores a través del rango completo de la distribución de las clases y costo del error. En ocasiones no se puede establecer una relación dominante entre las curvas ROC de dos clasificadores en su rango completo. En esta situación, para realizar la comparación entre clasificadores, se convierte el desempeño basado en la curva ROC en un escalar simple que representa el desempeño mostrado en la curva ROC. El método para lograrlo es calculando el área bajo la curva (AUC). El AUC es un valor que siempre estará entre 0 y 1. El AUC tiene un valor estadístico importante, pues representa la probabilidad de que el clasificador favorecerá mucho más a una instancia positiva seleccionada al azar que a una instancia negativa seleccionada al azar (ver Figura 4). El clasificador perfecto tendrá un AUC=1 mientras uno que no discrimine bien entre las clases de un problema tendrá un AUC=0.5.

El clasificador con mayor AUC es aquel que tiene el punto más alto en el eje donde se reflejan el beneficio (TP_{rate}) y un valor menor en el eje donde se refleja el costo (FP_{rate}) (Billsus et al. 1998). Para calcular el AUC se utiliza la siguiente fórmula:

$$AUC = \frac{1 + TPrate - FPrate}{2}$$

Para la selección de un clasificador, teniendo en cuenta estas medidas, se debe tener presente que el clasificador seleccionado es aquel que posea mejores resultados en las medidas de evaluación seleccionadas para realizar el análisis. En caso de que no encuentre alguno que supere al resto de los clasificadores en las medidas seleccionadas para su análisis, entonces se debe seleccionar aquel clasificador cuyos resultados satisfagan las necesidades del especialista que realiza el estudio.

Consideraciones finales.

En el capítulo se hizo referencia a cada uno de los pasos a realizar en las distintas etapas del KDD en la solución de problemas de clasificación. Se especificó las actividades a realizar cuando se está en la etapa de preprocesamiento de los datos, iniciando con la ejecución de un análisis manual de los datos, continuando con la selección de atributos y el balance de los datos. Seguidamente se realizó la caracterización de ocho categorías de clasificadores: perezosos, árboles de decisión, redes bayesianas, redes neuronales, máquinas de soporte vectoriales, análisis discriminante, regresión logística y reglas de asociación. También se especificó que al no alcanzar resultados satisfactorios con clasificadores individuales, la combinación de estos pueden ser una solución interesante siempre que los clasificadores bases utilizados sean precisos y diversos. Las razones que justifican el uso de multiclasificadores están recogidas en la literatura. Seguidamente se analizaron cuatro multiclasificadores, Bagging, Boosting, Stacking y Vote implementados en la herramienta Weka. Los dos primeros multiclasificadores garantizan la diversidad, generando muestras *boostrap* a partir del conjunto de datos original. Los restantes multiclasificadores no garantizan la diversidad de sus clasificadores bases, por lo tanto se especifican medidas para calcular la diversidad de los clasificadores bases a utilizar en la solución del problema. Finalmente se establecieron las medidas para evaluar el desempeño de los multiclasificadores. A partir de todo lo descrito en el capítulo se cuenta con la teoría necesaria para enfrentar un problema de clasificación con la aplicación de multiclasificadores para su solución.

CAPÍTULO 2: EXPERIMENTACIÓN

Introducción.

En el presente capítulo se describe la experimentación a realizar, con el objetivo de determinar el modelo multiclasificador que mejor desempeño posea ante la identificación de los líquidos iónicos que sean "tóxicos" o "seguros" sobre la enzima Acetilcolinesterasa. Se describe el conjunto de datos utilizados. Se establece la metodología a seguir en la limpieza manual de los datos, el método de selección de atributos relevantes para reducir la dimensionalidad del problema y el proceso de balanceo de la distribución de las clases presentes en el conjunto de datos a utilizar. Se analizan individualmente los clasificadores a utilizar y se conforman los multiclasificadores descritos en el epígrafe 1.5. Finalmente se determinan los modelos multiclasificador, con mejor desempeño, ante el problema investigado.

2.1 Descripción de los datos utilizados.

Producto a las discrepancias respecto a la toxicidad de los LI algunas empresas y centros de investigación desarrollaron una batería de pruebas, en organismos unicelulares, plantas e incluso células de organismos complejos, con el objetivo de determinar cuál es el resultado ante una exposición a distintas concentraciones de LI.

Los resultados obtenidos fueron acompañados de una caracterización molecular de cada uno de los LI y su actividad biológica. Información publicada por (Stock et al. 2004, Arning et al. 2008) y recopilada en una base de datos pública de amplio acceso disponible en (Störmann et al. 2011).

La base de datos de interés para el trabajo es aquella donde se recoge la información antes mencionada respecto a la enzima AChE. Esta base de datos recoge información de 288 LI, con sus respectivas actividades biológicas. Según se describe en la Tabla 5.

Base de datos	Cantidad de clases	Distribución por clases	Total de casos
AChE_inhibitor	2	236-52	288

TABLA 5: CARACTERÍSTICAS DE LA BASE DE DATOS UTILIZADA EN EL EXPERIMENTO.

Alexander Tropsha y su equipo (Tropsha 2010, Tropsha et al. 2010) proponen una metodología para la limpieza manual de los datos, siguiendo los principios generales del KDD enfocado a las especificidades de esta área.

2.2 Preprocesamiento.

La metodología propuesta por Tropsha propone eliminar aquellas moléculas que poseen elementos metálicos, organometálicos y poco representados. Dejando fuera del análisis las moléculas que contienen elementos que no son manejados por las herramientas que calculan descriptores moleculares,

y aquellas que poseen elementos poco representados que pudieran constituir ruidos en la fase de aprendizaje. Los elementos poco representados son determinados con la utilización del software JChem(Jchem 2013) para Excel, Este software posee funciones para determinar los elementos que están contenidos en una molécula y con esta información determinar el total de elementos que se encuentran contenidos en la base completa.

La estructura química de cada molécula, que se encuentra codificada con SMILES[5], debe ser estandarizada debido a que en ocasiones un mismo grupo funcional puede ser representado por diferentes patrones estructurales en un conjunto de datos dado. Para la quimioinformática, esta situación conduce hacia serios problemas, debido a que los descriptores moleculares calculados a partir de las representaciones diferentes, pueden ser significativamente distintos (Tropsha et al. 2010), de forma que estos se calculan como si fueran moléculas diferentes. La estandarización estructural de las moléculas se puede realizar utilizando la herramienta *ChemAxon´s Standarizer* (ChemAxon 2012), pues permite realizar normalizaciones de forma rápida y eficiente (Tropsha et al. 2010).

El resultado de esta operación permite identificar si existen duplicados de una molécula y por lo tanto realizar la eliminación de las moléculas duplicadas. Los duplicados pueden afectar la frecuencia observada de una molécula y la distribución de los compuestos acorde a su similaridad estructural. Como consecuencia los duplicados deben ser eliminados antes de realizar cualquier estudio de modelación. Para la eliminación de duplicados se utiliza el software ISIDA's EdiSDF(ISIDA 2013), debido a que es una herramienta libre para fines académicos (Tropsha et al. 2010).

Finalmente un chequeo manual debe ser realizado por un especialista en busca de otros problemas estructurales que pudieran haber quedado luego de la aplicación de esta metodología.

Luego de eliminar las moléculas identificadas en los pasos anteriores se procede a calcular los descriptores moleculares utilizando la herramienta ISIDA. El conjunto de datos resultante recoge la información de 261 moléculas, clasificadas en dos grupos: 226 moléculas tóxicas y 35 no tóxicas o seguras. Cada molécula está representada por 693 descriptores moleculares.

Según se establece en (Tropsha et al. 2010) para la utilización de los datos, en problemas de clasificación, se debe particionar el conjunto y seleccionar solo el 75% de las instancias de cada clase para ser utilizada como conjunto de entrenamiento y dejar el 25% de cada clase para validar el aprendizaje (ver Anexo 3) en lo que se denomina conjunto de validación externa. Este conjunto de validación externa es utilizado para evaluar la calidad del aprendizaje inducido a partir del conjunto de entrenamiento. Se utilizó el SPSS (SPSS 2010) para realizar un muestreo aleatorio y de esta forma seleccionar el 75% de las moléculas de cada clase para el entrenamiento y dejar el resto para probar el aprendizaje.

El conjunto de datos a utilizar en el entrenamiento posee dos características que obligan a continuar realizando acciones como parte de la etapa de preprocesamiento. Presenta una alta dimensionalidad

[5] Notación lineal usada para representar la estructura molecular computacionalmente procesable.

con respecto al total de instancias que posee, pudiendo propiciar problemas de sobreajuste en la clasificación. Presenta además un desbalance entre la cantidad de instancias que conforman los dos grupos en los que se clasifican estas moléculas, teniendo un $IR = 6.45$ que indica que el criterio de los clasificadores sea parcializado a favor de las moléculas más representadas. En lo adelante se continúa en el preprocesamiento del conjunto de entrenamiento y se mantiene el conjunto de validación externa con las características originales, de modo que el modelo obtenido pueda ser evaluado en un conjunto de datos que posee características similares al problema real.

2.2.1. MRMR como algoritmo de reducción de dimensionalidad.

En (Peng et al. 2005) se pone a disposición de los interesados un software que permite determinar los atributos más relevantes respecto al atributo dependiente y menos redundantes entre ellos, con el objetivo de reducir la dimensionalidad en conjuntos de datos. Además se demuestra la efectividad de dicha propuesta, mostrando que el desempeño de un grupo de clasificadores mejora cuando se realiza el aprendizaje utilizando el subconjunto de atributos seleccionados luego de aplicada su propuesta.

En (Cruz-Monteagudo et al. 2013) se utiliza este algoritmo para realizar una reducción de la dimensionalidad de un conjunto de datos donde se recoge los resultados del efecto de aplicar de distintas concentraciones de LI sobre células de ratas (ICP-81). El reducto obtenido es utilizado en la aplicación del enfoque de aprendizaje automatizado supervisado para la determinación de un modelo que predijo la toxicidad de los LI respecto a la ICP-81. Permitiendo determinar las características específicas que determinan si una molécula es tóxicas frente al ICP-81, y de esta forma poder trabajar en la producción de nuevos LI garantizando que estos sean no tóxicos.

En (Peng et al. 2005) se enfatiza que en la utilización del software hay que establecer un conjunto de parámetros tales como: el conjunto de datos a reducir, el número de atributos a seleccionar, entre otros.

Para el establecimiento de la cantidad de atributos a seleccionar, el autor realizó múltiples pruebas donde realizó varias clasificaciones variando la cantidad de atributos a seleccionar. Determinó que los clasificadores alcanzaban su mejor desempeño con la selección de 50 atributos en adelante. Por tal razón se hace la selección de 50 atributos en este trabajo, apoyado además, de que en (Cruz-Monteagudo et al. 2013) también se utilizó esa misma cantidad de atributos a seleccionar.

2.2.2. SMOTE como algoritmo de balanceo de datos.

Producto a que el conjunto de datos de entrenamiento posee un desbalance con un $IR = 6.45$ se hace necesario la aplicación de una técnica que permita modificar la distribución de las clases para lograr un IR igual o cercano a 1. La herramienta Weka posee el algoritmo llamado SMOTE que se clasifica dentro de las técnicas *oversampling.*

La distribución inicial de los datos, en el conjunto de entrenamiento, es de 169 moléculas tóxicas y 26 moléculas no tóxicas, por tal razón hay que configurar el SMOTE de manera que la clase menos representada alcance la misma cantidad que la clase más representada.

Entre los parámetros de configuración del SMOTE, el parámetro a cambiar es el porcentaje de instancias a crear. Al establecer este valor al 550% permitió lograr un IR=1, significando que se alcanzó igual cantidad de instancias en ambas clases.

2.3 Identificación de clasificadores base con desempeño aceptable.

El procedimiento seguido en esta actividad fue propuesto en (Tropsha 2010). En este trabajo se propone la evaluación de clasificadores y de ellos seleccionar aquellos clasificadores que posean un Acc>0.6, TPrate>0.6, TNrate>0.6, Kappa>0.6 en la evaluación del aprendizaje tanto sobre la matriz de entrenamiento, como sobre el conjunto de validación externa, utilizando una validación cruzada con 5 grupos y dejando una molécula fuera, que es la misma estrategia que la validación cruzada pero en este caso estableciendo la misma cantidad de grupos que de moléculas. De esta forma es posible determinar los clasificadores que tuvieron un desempeño confiable y que cuyo aprendizaje no estuvo influenciado por la forma en la que se particionaron los datos.

Los clasificadores individuales evaluados son aquellos que se encuentran en los grupos de clasificadores que posee la herramienta Weka. En el Anexo 6 se recoge los resultados del desempeño de los clasificadores que se relacionan en la Tabla 6 y que cumplen con las condiciones expresadas anteriormente.

Grupo	**Clasificadores**
Trees	ExtraTree
	HoeffdingTree
	RandomTree
	LMT
Bayes	BayesianLogisticRegression
	BayesNet
	ComplementeNaiveBayes
	NaiveBayes
	NaiveBayesMultinominal
Lazy	IBK estableciendo k=5
	Kstar
Rules	Jrip
	Nnge
Functions	FDLA
	KernelLogisticRegression
	Logistic
	MultilayerPerceptron,
	MultilayerPerceptronCS
	SimpleLogistic,
	SMO

TABLA 6: RELACIÓN DE CLASIFICADORES SELECCIONADOS

A continuación se analiza que diseñar un multiclasificador utilizando 20 clasificadores como base, sería un proceso muy exhaustivo de realizar. Se valora que estos algoritmos, en su mayoría, son variantes de ocho tipos de clasificadores, por lo que los resultados a obtener entre los clasificadores de un mismo tipo no serán muy diferentes. En (Kuncheva 2004) se establece que una forma de garantizar diversidad es basándose en la naturaleza propia del clasificador. Por tal razón se seleccionan los clasificadores que mejor Acc tuvieron en su respectiva categoría de clasificadores. Reduciendo a ocho los clasificadores a utilizar.

2.4 Utilización de los algoritmos multiclasificadores.

Weka posee un conjunto de multiclasificadores como parte de sus clasificadores. Algunos de los enfoques principales utilizados en el desarrollo de estos multiclasificadores fueron mencionados en el epígrafe 1.5. Para la realización de este trabajo se decidió utilizar los multiclasificadores descritos en los epígrafes del 1.5.1 al 1.5.4. Para la configuración de cada multiclasificador se utilizaron los parámetros por defecto que propone Weka, con algunas excepciones que serán especificadas en cada caso.

2.4.1. Bagging

Bagging es un algoritmo que garantiza la diversidad de sus clasificadores. Utiliza para entrenar distintos conjuntos de entrenamiento obtenidos a partir del conjunto original mediante la realización de una selección aleatoria con reemplazo. En este caso se realiza la clasificación utilizando como algoritmo base cada uno de los clasificadores individuales seleccionados, con el objetivo de evaluar el desempeño alcanzado y compararlo con el resto de los multiclasificadores, y así determinar el modelo que mejor desempeño muestre. En la experimentación se utiliza la configuración por defecto de los parámetros del multiclasificador.

2.4.2. Boosting

Boosting es un algoritmo que al igual que Bagging garantiza la diversidad utilizando distintos conjuntos de entrenamientos generados con una selección aleatoria con reemplazo a partir de conjunto de datos original. A diferencia del Bagging, el entrenamiento se hace de forma secuencial aplicando una estrategia que garantiza que las instancias mal clasificadas tengan más probabilidades de pertenecer al próximo conjunto de entrenamiento a utilizar. En este caso se utiliza el algoritmo AdaBoostM1 implementado para problemas de clasificación. Se cambia el parámetro "*useResampling"* en la configuración del algoritmo. Se establece el valor del parámetro en "*True*" para cambiar la estrategia mediante la cual garantiza que las instancias mal clasificadas pasen a formar parte del próximo conjunto de entrenamiento. Nuevamente se realiza la clasificación utilizando todos los clasificadores individuales seleccionados con el objetivo de evaluar el desempeño de estos y compararlos con el resto de los multiclasificadores para determinar el multiclasificador que mejor desempeño muestre.

2.4.3. Determinación de clasificadores diversos para los multiclasificadores.

Los algoritmos Stacking y Vote no garantizan la diversidad en sus clasificadores por lo tanto es necesario garantizar que las entradas recibidas sean las salidas más diversas emitidas por los clasificadores bases utilizados. Garantizando a los multiclasificadores las salidas de los clasificadores que, entre todos, garanticen un mayor cubrimiento del espacio químico.

En la determinación de los clasificadores individuales diversos a emplear, se utilizarán las medidas de diversidad explicadas en el epígrafe 1.5.5 y que fueron utilizadas en (Bonet et al. 2012, Bonet et al. 2013, Cabrera et al. 2013) para la determinación de clasificadores diversos.

Al contar con ocho clasificadores la diversidad deberá ser calculada para un total de 248 combinaciones determinado por la combinatoria mostrada a continuación:

$$combinaciones = \binom{8}{2} + \binom{8}{3} + \binom{8}{4} + \cdots + \binom{8}{8} = 248$$

Las medidas de diversidad a utilizar en este problema consideran un par de clasificadores a la vez. Un multiclasificador de *L* clasificadores produce $N = \frac{L(L-1)}{2}$ pares de medidas, las cuales son promediadas para obtener un valor global para el multiclasificador.

El cálculo de las medidas de diversidad será realizado con la utilización de un software llamado "Diversidad", implementado sobre plataforma Java y desarrollado en (Cabrera et al. 2013).

El procedimiento seguido en la determinación de las combinaciones de clasificadores más diversas se relaciona a continuación:

Se estableció seleccionar las combinaciones de clasificadores que presentan mayor diversidad en las cinco medidas de diversidad planteadas en el epígrafe 1.5.5. Se establece también seleccionar aquellas combinaciones que presentan mayor diversidad en cuatro de las cinco medidas de diversidad calculadas. De esta forma es posible poder identificar las combinaciones que mejores condiciones presentan para alcanzar un desempeño superior al desempeño alcanzado por los clasificadores individuales utilizados, siguiendo el criterio emitido por Kuncheva en sus artículos. Permitiendo acotar el espacio de búsqueda a solo 31 combinaciones entre las 248 posibles a explorar.

También se establece, que la selección de los clasificadores diversos se debe realizar ordenando los valores de las medidas de diversidad en forma descendente o ascendente, según la orientación de las medidas y seleccionar las combinaciones cuyos valores de diversidad se encuentren por encima o por debajo de la media aritmética de la medida analizada respecto a todos los valores alcanzados por dicha medida en el conjunto. Garantizando así la identificación de 31 combinaciones de clasificadores con la mayor diversidad posible entre las 248 posibles combinaciones de clasificadores a explorar.

2.4.4. Stacking

Este algoritmo utiliza las salidas obtenidas por diferentes clasificadores sobre un único conjunto de datos. Como método de combinación utiliza un algoritmo de clasificación denominado metaclasificador. Este

metaclasificador utiliza en su aprendizaje una base de casos conformada por instancias conformadas por un vector de rasgos compuesto por las clases de salida emitidas por cada clasificador base y como atributo dependiente, la clase original de la instancia.

La configuración de los parámetros de este algoritmo se mantuvo con sus valores por defecto, excepto la definición del metaclasificador que en este caso se estableció la utilización del algoritmo *Extratree* debido a que fue el clasificador individual que mejor desempeño tuvo entre los clasificadores individuales seleccionados.

La experimentación se realizó utilizando las 31 combinaciones más diversas determinadas utilizando las cinco medidas de diversidad especificadas en el epígrafe 1.5.5.

2.4.5.Vote

Al igual que el algoritmo Stacking, Vote no garantiza la diversidad entre sus clasificadores de base. Vote utiliza el criterio emitido por varios clasificadores para emitir un criterio único por cada instancia y para ello se apoya en el vector de probabilidades que cada clasificador emite sobre cada instancia. Vote propone varias formas para combinar las salidas de los clasificadores bases las cuales están documentadas en el epígrafe 1.5.4.1.

La experimentación, con este multiclasificador, se realizó utilizando las mismas combinaciones propuestas para el multiclasificador Stacking, expuesto en el epígrafe anterior.

2.5 Medidas de evaluación de desempeño de clasificadores.

Luego de realizar el proceso de clasificación se pasa a evaluar el desempeño de los clasificadores. En la literatura se encuentran registradas varias métricas utilizadas en esta tarea, las cuales están enfocadas principalmente a problemas de dos clases. Información sobre las métricas implementadas en Weka y que fueron documentadas para este trabajo se encuentran en el epígrafe 1.6.

La métricas utilizas para este trabajo son la efectividad (Acc), el estadístico kappa (Kappa) y el área bajo la curva ROC (AUC).

La selección de los clasificadores individuales fue basada en aquellos con valores, en estas métricas, superiores a los 0.6, debido a que clasificadores con valores por debajo de este valor son considerados “no confiables” por parte de los especialistas.

Luego de recopilados los resultados de los clasificadores individuales y de los multiclasificadores, se pasa a contrastar si existe algún multiclasificador que posea un desempeño superior al mejor desempeño alcanzado por alguno de los clasificadores individuales.

Para realizar esta actividad se determina el clasificador individual con el mejor desempeño y luego se identifican los multiclasificadores que superan este desempeño. En caso de que existan varios multiclasificadores que superen este desempeño se analiza si existen diferencias significativas entre ellos. En caso positivo se determina el multiclasificador que mejor desempeño posee en la solución del

problema que se pretende solucionar. En caso contrario se seleccionan todos como modelos multiclasificadores que resuelven el problema.

2.6 Consideraciones finales.

En el capítulo se especificó el procedimiento a seguir para determinar si existe un multiclasificador con las condiciones para ser utilizado en la clasificación de los LI que pueden ser “tóxicos” o “seguros” ante la enzima AChE. Se establece que el término “tóxicos” es referente a las moléculas que inhiben el funcionamiento de la enzima y “seguros” se establece para aquellos compuestos que no inhiben el funcionamiento de la enzima. En el cuerpo del capítulo se pasa de la etapa de preprocesamiento, donde se propone como primera actividad la eliminación manual de algunos compuestos, siguiendo la metodología propuesta en el epígrafe 2.2. Se continuó con la selección de un subconjunto de atributos que poseen una máxima relevancia con respecto al atributo dependiente y una mínima redundancia entre ellos, como parte de la selección de atributos. Se continuó con la actividad de balancear la distribución de las clases utilizando un algoritmo conocido en la literatura como SMOTE. Luego fueron evaluados los clasificadores individuales y seleccionados aquellos que mejor desempeño tuvieron siguiendo las métricas propuestas y el criterio de los especialistas. Seguidamente se conformaron los multiclasificadores analizados en el epígrafe 1.5 y fue evaluado su desempeño. Finalmente se determinaron los multiclasificadores que mejor desempeño tuvieron para el problema investigado.

CAPÍTULO 3: ANÁLISIS DE RESULTADOS.

Introducción.

En el presente capítulo se hace un análisis de los resultados obtenidos en la experimentación propuesta en el capítulo 2. Finalmente se emite un criterio sobre la elección de los multiclasificadores que mejores resultados proporcionaron en la clasificación de los LI en el problema investigado.

3.1 Preprocesamiento.

Luego de aplicar la metodología establecida por Tropsha en (Tropsha 2010, Tropsha et al. 2010) se identificaron, de forma manual, moléculas que poseían elementos metálicos, organometálicos y elementos poco representados. Moléculas que debían ser eliminadas ya que no son manejadas por las herramientas que calculan los descriptores moleculares.

Los elementos metálicos, organometálicos y poco representados identificados fueron: Sodio (Na), Litio (Li), Potasio (K), Antimonio (Sb) y el Hierro (Fe) tal como se representa en la Tabla 7:

Nombre de la molécula	Código SMILES
Potassium 1-(N,N-dihexyl-1-hexanamine)-2,3,4,5,6,7,8,9,10,11,12-undecahydrododecaborate(1-)	[K+].CCCCCC[N](CCCCCC)(CCCCCC)[B-]1234[BH]567[BH]89%10[BH]%11%12%13[BH]585[BH]%118%11[BH]%12%12%14[BH]9%139[BH]16%10[BH]2%129[BH]38%14[BH]475%11
Lithium 1-(N,N-dihexyl-1-hexanamine)-2,3,4,5,6,7,8,9,10,11,12-undecahydrododecaborate(1-)	[Li+].CCCCCC[N](CCCCCC)(CCCCCC)[B-]1234[BH]567[BH]89%10[BH]%11%12%13[BH]585[BH]%118%11[BH]%12%12%14[BH]9%139[BH]16%10[BH]2%129[BH]38%14[BH]475%11
3,7-Dihydro-1,3,7,9-tetramethyl-2,6-dioxo-1H-purinium trifluoromethanesulfonate	[O-]S(=O)(=O)C(F)(F)F.CN1C=[N+](C)C2=C1C(=O)N(C)C(=O)N2C
1-Butylpyridinium μ-chlorohexachlorodialuminate	Cl[Al](Cl)(Cl)[Cl-][Al](Cl)(Cl)Cl.CCCC[n+]1ccccc1
1-Butyl-3-methyl-1H-imidazolium (OC-6-11)-hexafluoroantimonate(1-)	[N+]1(C)=CN(C=C1)CCCC.[Sb-@OH1](F)(F)(F)(F)(F)F
1-Hexyl-3-methyl-1H-imidazolium, salt with 1,2-benzisothiazol-3(2H)-one 1,1-dioxide (1:1)	CCCCCCN1C=C[N+](C)=C1.O=C1[N-]S(=O)(=O)c2ccccc12
1-Butyl-3-methyl-1H-imidazolium tetrachloroferrate(1-)	Cl[Fe-](Cl)(Cl)Cl.CCCCN1C=C[N+](C)=C1
Sodium hexafluorophosphate(1-)	*[Na+].F[P-](F)(F)(F)(F)F*
Sodium (OC-6-11)-hexafluoroantimonate(1-)	*[Na+].[Sb-@OH1](F)(F)(F)(F)(F)F*
Sodium dodecyl sulfate	[Na+].CCCCCCCCCCCCOS([O-])(=O)=O
2-Hydroxy-N,N,N-trimethylethanaminium hydroxide	[OH-].C[N+](C)(C)CCO
Lithium (T-4)-bis[2-hydroxybenzoato(2-)-O1,O2]borate(1-)	[Li+].O=C1O[B-]2(OC(=O)c3ccccc3O2)Oc2ccccc12

Lithium bis[1,2-benzenediolato(2-)-O1,O2]borate(1-)	[Li+].O1c2ccccc2O[B-]11Oc2ccccc2O1
Lithium bis[ethanedioato(2-)-O1,O2]borate(1-)	[Li+].O=C1O[B-]2(OC1=O)OC(=O)C(=O)O2
Lithium bis[propanedioato(2-)-O1,O3]borate(1-)	[Li+].O=C1CC(=O)O[B-]2(O1)OC(=O)CC(=O)O2
Lithium 1,1,1-trifluoro-N-[(trifluoromethyl)sulfonyl]methanesulfonamide	[Li+].FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F
Sodium tetrafluoroborate(1-)	[Na+].F[B-](F)(F)F
Sodium methanesulfonate	[Na+].CS([O-])(=O)=O
Sodium 4-methylbenzenesulfonate	[Na+].Cc1ccc(cc1)S([O-])(=O)=O
Sodium 1-butanesulfonate	[Na+].CCCCS([O-])(=O)=O
Sodium 1-hexanesulfonate	[Na+].CCCCCCS([O-])(=O)=O
Sodium 1-octanesulfonate	[Na+].CCCCCCCCS([O-])(=O)=O
Sodium trifluoromethanesulfonate	[Na+].[O-]S(=O)(=O)C(F)(F)F
Sodium trifluoroacetate	[Na+].[O-]C(=O)C(F)(F)F
Sodium methyl sulfate	[Na+].COS([O-])(=O)=O
Sodium thiocyanate	[Na+].[S-]C#N
Sodium octylsulfate	[Na+].CCCCCCCCOS([O-])(=O)=O

TABLA 7: TABLA DE MOLÉCULAS ELIMINADAS EN EL PREPROCESAMIENTO.

Para la identificación de los elementos poco representados se utilizó el software JCHEM para Excel, el cual posee las funciones *JCComposition*, *JCMolFormula* y *JCAtomCount* mediante las cuales se pudo determinar el total de elementos que estaban contenidos en las moléculas. A partir de esta información se pudo determinar que el elemento poco representado era el metaloide Antimonio (Sb). Tal como se muestra en la Figura 5.

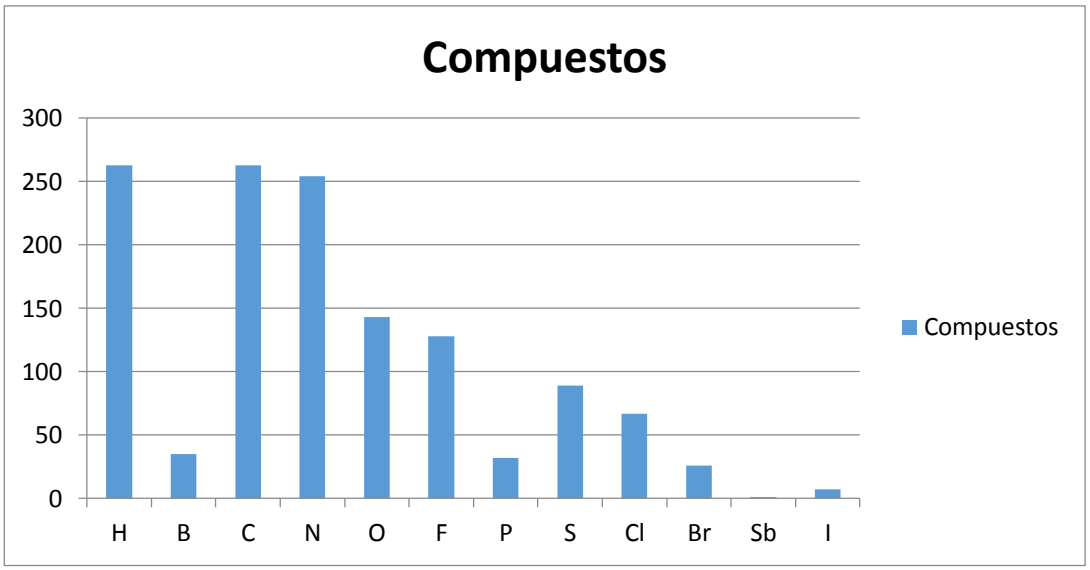


FIGURA 5: REPRESENTATIVIDAD DE LOS ELEMENTOS.

La estandarización de los compuestos fue realizada con la utilización del software ChemAxon´s Standarizer. La aplicación de este procedimiento garantiza que los grupos funcionales presentes en la

estructura de todas las moléculas queden con la misma representación, debido a la aplicación de múltiples filtros que incorpora la herramienta y que fueron aplicados en (Cruz-Monteagudo et al. 2013). El próximo paso en la limpieza manual es la identificación de estructuras duplicadas, la cual se realiza con el software EdiSDF y de esta forma eliminar los compuestos duplicados. Luego de la aplicación de este paso se determinó que en el conjunto de datos no se contaba con estructuras duplicadas.

A partir de este momento se realizó al cálculo de los descriptores moleculares con la utilización del ISIDA. El resultado de esta operación fue la obtención de 693 atributos para la base de casos. Para consultar los descriptores moleculares obtenidos ver el Anexo 4.

La metodología propuesta en el epígrafe 2.2 establece que el conjunto de datos debe ser particionado de forma que el 75% de las instancias, que pertenecen a ambas clases, sean utilizadas para realizar el entrenamiento de los clasificadores y el restante 25% de ellas sean utilizadas como método de validación del aprendizaje inducido por los clasificadores. El particionamiento fue realizado con la utilización del software SPSS. Las características de los conjuntos de datos generados a partir del particionamiento propuesto pueden ser consultadas en el Anexo 8 y Anexo 9.

Según se especificó en el epígrafe 2.2, las herramientas que realizan el cálculo de los descriptores moleculares generan cientos y en ocasiones miles de ellos para un compuesto. En ocasiones muchos de los atributos obtenidos no aportan información relevante sobre la instancia. Por tal razón se debe pasar a realizar una selección de los descriptores moleculares más relevantes para el estudio a realizar.

3.1.1 mRMR como algoritmo de reducción de dimensionalidad

En (Peng et al. 2005) se propone un algoritmo para la selección de atributos relevantes basándose en la relevancia de cada atributo con su atributo clase, luego se evalúa la redundancia de cada atributo respecto al resto de los atributos buscando disminuir la redundancia entre los atributos que mayor relevancia tienen respecto a la clase.

Luego de aplicado el algoritmo, sobre el conjunto de entrenamiento, se obtuvo el subconjunto de los 50 descriptores más relevantes y menos redundantes del problema. Los descriptores relevantes pueden ser consultados en el Anexo 5.

3.1.2 SMOTE como algoritmo de balanceo de datos.

Producto que el conjunto de entrenamiento presenta además un problema de desbalance en la distribución de las clases se procede a aplicar una técnica para balancear la distribución de estas. El algoritmo SMOTE fue propuesto por (Chawla et al. 2002). Debido a la aceptación del algoritmo en la comunidad científica se han realizado múltiples propuestas con mejoras al mismo, tal como las presentadas en (Ramentol et al. 2009, Ramentol et al. 2011, Ramentol et al. 2012), aunque es importante especificar que no son las únicas.

El algoritmo SMOTE propone la creación de instancias artificiales que no difieren mucho de las instancias reales y que ayudan a balancear la distribución de los datos respecto a la clase menos representada.

El conjunto de datos entrenamiento original presentaba la distribución que se representa en la Figura 6.

Luego de aplicado el algoritmo SMOTE se alcanza el balance deseado entre las clases para el problema a resolver. Tal como se muestra en la Figura 7.

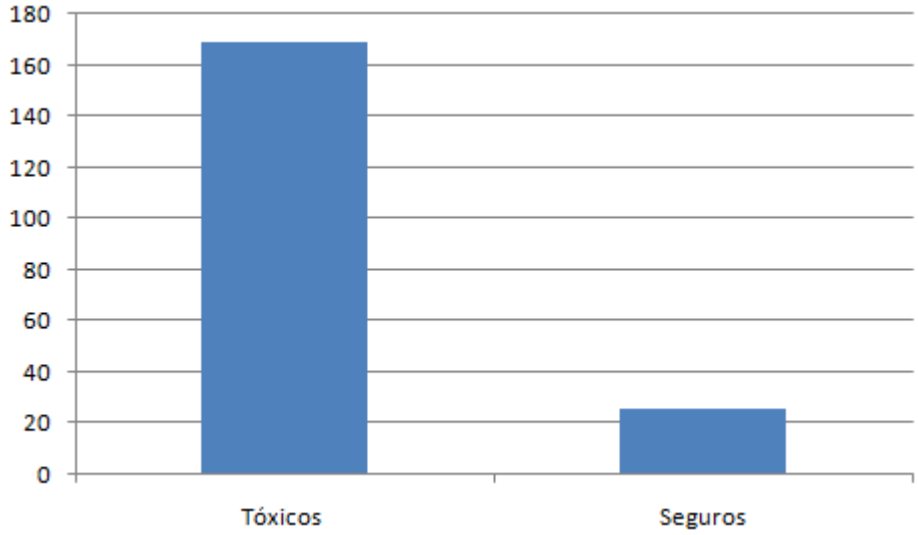


FIGURA 6: DISTRIBUCIÓN ORIGINAL DE LA MATRIZ DE ENTRENAMIENTO

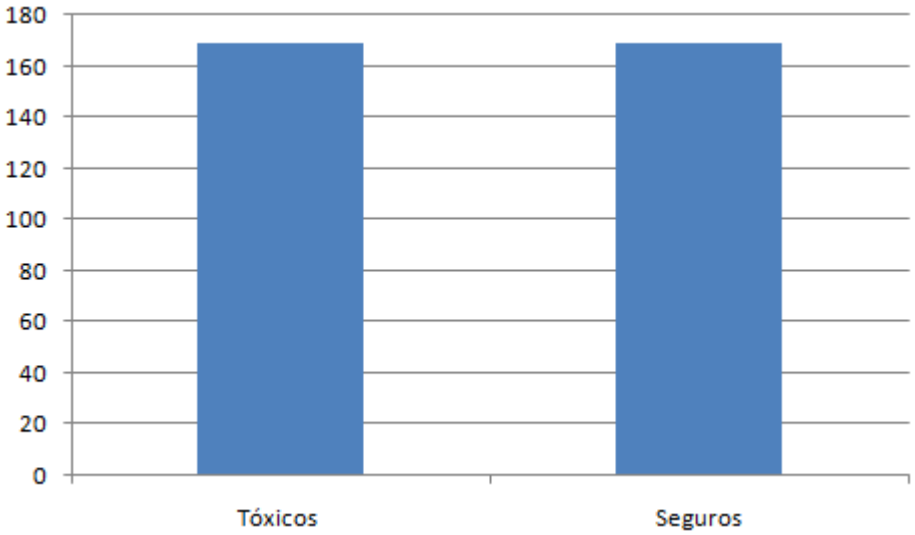


FIGURA 7: DISTRIBUCIÓN DE LAS CLASES LUEGO DE APLICADO EL ALGORITMO SMOTE

3.2 Identificación de clasificadores base con desempeño aceptable.

Para la selección de los clasificadores bases a utilizar, se evaluaron los clasificadores implementados en la herramienta Weka. Siguiendo las sugerencias de los especialistas, especificadas en el epígrafe 2.3, fueron seleccionados los clasificadores con un desempeño confiable. Información sobre los clasificadores seleccionados siguiendo la propuesta, puede ser consultada Anexo 6.

A continuación se seleccionó un clasificador por cada categoría de clasificadores, es decir, fueron seleccionados ocho clasificadores con buen desempeño, quedando reducido el problema a ocho clasificadores. La elección de los clasificadores se realizó seleccionando los clasificadores que mejor desempeño tuvieron por grupo y aquellos clasificadores que están agrupados dentro de la categoría

"functions", en la herramienta Weka, pero que utilizan heurísticas diferentes. Quedando reducido el problema a la utilización de los clasificadores registrados en la Tabla 8.

Grupo	Clasificadores
Árboles de decisión	ExtraTree
Red bayesiana	NaiveBayes
Basado en instancias(perezoso)	Kstar
Reglas	Jrip
Análisis discriminante	FDLA
Regresión logística	Logistic
Red neuronal	MultilayerPerceptron
Máquina de soporte vectorial	SMO

TABLA 8: ALGORITMOS SELECCIONADOS PARA EL MULTICLASIFICADOR

3.3 Resultados de los algoritmos multiclasificadores utilizados.

Para poder establecer un patrón de comparación y determinar si un multiclasificador es bueno para solucionar un problema, este debe tener un desempeño superior al mejor clasificador individual, de lo contrario no es posible justificar su uso en la solución del problema. Ver desempeño de los clasificadores individuales en la Tabla 9.

Clasificador	Acc	AUC	Kappa
ExtraTree	***0.95454545***	***0.928***	***0.7777***
NaiveBayes	0.92424242	0.922	0.7179
Kstar	0.95454545	0.916	0.7975
Jrip	0.93939394	0.827	0.7161
FLDA	0.92424242	0.934	0.7179
Logistic	0.90909091	0.92	0.6471
Multilayerperceptron	0.93939394	0.953	0.7647
SMO	0.92424242	0.909	0.7179

TABLA 9: DESEMPEÑO DE LOS CLASIFICADORES DE BASE.

Como se puede apreciar es el clasificador *ExtraTree* el que muestra mejor desempeño entre los clasificadores individuales. En este caso todos los clasificadores tienen un desempeño bueno pero en este caso se trata de maximizar la cantidad de instancias identificadas por el clasificador en cada clase. Por tal razón los parámetros del desempeño del *ExtraTree* serán el patrón de comparación a utilizar para elegir solo los multiclasificadores que superen este desempeño.

3.3.1 Resultados de la experimentación con el multiclasificador Bagging

Luego de aplicada la configuración planteada en el epígrafe 2.4.1, se procedió a realizar la experimentación. Los resultados obtenidos quedaron registrados en la Tabla 10.

Clasificador de base	ACC	AUC	KAPPA
ExtraTree	***0.95454545***	***0.916***	***0.7973***
NaiveBayes	0.92424242	0.96	0.7179
Kstar	***0.95454545***	***0.961***	***0.7975***
Jrip	0.92424242	0.972	0.6967

FLDA	0.92424242	0.945	0.7179
Logistic	0.89393939	0.936	0.5678
MultilayerPerceptron	0.93939394	0.982	0.7647
SMO	0.92424242	0.957	0.7179

TABLA 10: DESEMPEÑO DE LOS MULTICLASIFICADORES BAGGING

Como se aprecia en la tabla solo dos variantes de los multiclasificador Bagging igualan en efectividad al mejor clasificador individual. No encontrándose ningún modelo que supere el desempeño del mejor clasificador individual. Según (Bonet et al. 2013) solo se justifica el uso de un multiclasificador si el mismo supera en desempeño al mejor clasificador individual.

3.3.2 Resultados de la experimentación con el multiclasificador AdaBoost.M1

Luego de aplicada la configuración planteada en el 2.4.2, se procedió a realizar la experimentación a partir del cual se obtuvo los siguientes resultados (ver Tabla 11).

Clasificador de base	ACC	AUC	Kappa
ExtraTree	0.95454545	0.978	0.7975
NaiveBayes	0.95454545	0.978	0.7975
Kstar	0.95454545	0.976	0.7975
Jrip	0.95454545	0.988	0.7975
FLDA	0.90909091	0.955	0.6471
Logistic	0.90909091	0.951	0.6471
MultilayerPerceptron	***0.96969697***	***0.991***	***0.8713***
SMO	0.90909091	0.967	0.6471

TABLA 11: DESEMPEÑO DE LOS MULTICLASIFICADORES BOOSTING

Como se aprecia en la tabla, solo la variante del multiclasificador *AdaboostM1* con la red neuronal *MultilayerPerceptron* supera en efectividad al mejor clasificador individual, por lo que estamos en presencia de un candidato a utilizar como solución al problema. La razón que justifica su uso es debido a que es capaz de identificar una instancia más que el clasificador individual utilizado como patrón de comparación.

En (Freund et al. 1996, Freund et al. 1999) se especifica que el algoritmo Boosting se utiliza para mejorar la clasificación de un clasificador inestable[6]. En este caso la red neuronal se comportó como un clasificador inestable.

3.3.3 Determinación de los clasificadores diversos.

En el epígrafe 2.4.3 se especifica que para la conformación de los multiclasificadores Stacking y Vote se hace necesario determinar cuáles clasificadores bases son más diversos. Según (Kuncheva et al. 2003, Bonet et al. 2012, Bonet et al. 2013, Cabrera et al. 2013) la forma para determinar cuáles combinaciones de clasificadores son diversas es mediante la utilización de las medidas de diversidad. Pues propician la acotación del espacio, de posibles combinaciones de clasificadores, a explorar.

[6] Calificativo dado a clasificadores que con solo variar un poco sus parámetros emite resultados diferentes.

En el mismo epígrafe se especifican las medidas de diversidad que serán calculadas para determinar cuáles combinaciones serán seleccionadas como las más diversas. Para el cálculo de las medidas de diversidad se utilizó el software "Diversidad" propuesto en (Cabrera et al. 2013). En la Tabla 12, se muestra un extracto de los resultados obtenidos luego de calcular las medidas de diversidad. Para consultar esta información completa ver Anexo 7.

R	p	D	Q	DF	Subconjunto de Clasificadores
2.00	0.47	0.06	0.94	0.03	JRIP--MULTILAYERPERCEPTRON
2.00	0.49	0.06	0.95	0.03	EXTRATREE--SMO
2.00	0.49	0.06	0.95	0.03	FDLA--KSTAR
2.00	0.49	0.06	0.95	0.03	EXTRATREE--FDLA
2.00	0.49	0.06	0.95	0.03	EXTRATREE--NAIVEBAYES
2.50	0.41	0.07	0.91	0.03	JRIP--SMO
2.50	0.41	0.07	0.91	0.03	FDLA--JRIP
2.50	0.41	0.07	0.91	0.03	JRIP--NAIVEBAYES
2.00	0.49	0.06	0.95	0.03	KSTAR--NAIVEBAYES
2.00	0.49	0.06	0.95	0.03	KSTAR--SMO
1.50	0.56	0.04	0.97	0.03	EXTRATREE--MULTILAYERPERCEPTRON
1.50	0.56	0.04	0.97	0.03	KSTAR--MULTILAYERPERCEPTRON
1.33	0.58	0.06	0.97	0.04	JRIP--LOGISTIC

TABLA 12: RESULTADOS DE LAS MEDIDAS DE DIVERSIDAD

En la tabla se muestran las medidas de diversidad según la siguiente simbología: la columna ***R*** representa el valor de la medida de diversidad "razón entre medida de diferencia y medida de doble fallo", las combinaciones más diversas, según esta medida, son aquellas que tengan un valor alto. La columna ***p*** representa la medida "coeficiente de correlación", las combinaciones más diversas según esta medida son aquellas cuyo valor sea bajo. La columna ***Q*** representa la medida "estadístico Q", las combinaciones más diversas según esta medida, son aquellas cuyo valor sea bajo. Por último la columna ***DF*** representa la medida "doble fallo", las combinaciones más diversas según esta medida, son aquellas cuyo valor es bajo.

Teniendo en cuenta el criterio planteado en el epígrafe 2.4.3 fueron seleccionadas todas aquellas combinaciones cuyos valores de diversidad estaban por encima o por debajo de la media aritmética, de cada medida, según la orientación de la misma. De esta forma se seleccionaron y separaron en un grupo aquellas combinaciones que alcanzaron la mayor diversidad en las cinco medidas y en otro aquellas combinaciones que alcanzaron la mayor diversidad en cuatro de las cinco medidas. Tal y como se muestra en la Tabla 13 yTabla 14.

R	p	D	Q	DF	Subconjunto de Clasificadores
2.00	0.47	0.06	0.94	0.03	JRIP—MULTILAYERPERCEPTRON
2.00	0.49	0.06	0.95	0.03	EXTRATREE—SMO
2.00	0.49	0.06	0.95	0.03	FDLA—KSTAR
2.00	0.49	0.06	0.95	0.03	EXTRATREE—FDLA

2.00	0.49	0.06	0.95	0.03	EXTRATREE—NAIVEBAYES
2.50	0.41	0.07	0.91	0.03	JRIP—SMO
2.50	0.41	0.07	0.91	0.03	FDLA—JRIP
2.50	0.41	0.07	0.91	0.03	JRIP—NAIVEBAYES
2.00	0.49	0.06	0.95	0.03	KSTAR—NAIVEBAYES
2.00	0.49	0.06	0.95	0.03	KSTAR—SMO
1.61	0.59	0.05	0.95	0.03	EXTRATREE--JRIP—SMO
1.61	0.59	0.05	0.95	0.03	EXTRATREE--JRIP—NAIVEBAYES
1.61	0.59	0.05	0.95	0.03	EXTRATREE--FDLA—JRIP
1.61	0.59	0.05	0.95	0.03	FDLA--JRIP—KSTAR
1.61	0.59	0.05	0.95	0.03	JRIP--KSTAR—SMO
1.61	0.59	0.05	0.95	0.03	JRIP--KSTAR—NAIVEBAYES
1.64	0.57	0.05	0.95	0.04	EXTRATREE--FDLA--JRIP—NAIVEBAYES
1.64	0.57	0.05	0.95	0.04	FDLA--JRIP--KSTAR—NAIVEBAYES
1.64	0.57	0.05	0.95	0.04	EXTRATREE--JRIP--NAIVEBAYES—SMO
1.64	0.57	0.05	0.95	0.04	JRIP--KSTAR--NAIVEBAYES—SMO

TABLA 13: COMBINACIONES MÁS DIVERSAS EN LAS CINCO MEDIDAS DE DIVERSIDAD

R	p	D	Q	DF	Subconjunto de Clasificadores
1.83	0.53	0.06	0.94	0.04	JRIP--NAIVEBAYES—SMO
1.83	0.53	0.06	0.94	0.04	FDLA--JRIP—NAIVEBAYES
1.53	0.57	0.06	0.95	0.04	JRIP--LOGISTIC—NAIVEBAYES
1.58	0.59	0.05	0.95	0.04	FDLA--JRIP—MULTILAYERPERCEPTRON
1.67	0.61	0.05	0.94	0.04	FDLA--JRIP—SMO
1.58	0.59	0.05	0.95	0.04	JRIP--MULTILAYERPERCEPTRON—SMO
1.58	0.59	0.05	0.95	0.04	JRIP--MULTILAYERPERCEPTRON—NAIVEBAYES
1.56	0.61	0.05	0.95	0.04	EXTRATREE--FDLA--JRIP—SMO
1.56	0.61	0.05	0.95	0.04	FDLA--JRIP--KSTAR—SMO
1.48	0.61	0.05	0.96	0.04	FDLA--JRIP--KSTAR--NAIVEBAYES—SMO
1.48	0.61	0.05	0.96	0.04	EXTRATREE--FDLA--JRIP--NAIVEBAYES--SMO

TABLA 14: COMBINACIONES MÁS DIVERSAS EN CUATRO DE LAS CINCO MEDIDAS DE DIVERSIDAD

A partir de la información antes compilada se está en condiciones de evaluar el desempeño de los siguientes multiclasificadores utilizando las combinaciones diversas antes identificadas.

3.3.4 Resultados de la experimentación con el multiclasificador Stacking

En el epígrafe 2.4.4 se planteó que el algoritmo Stacking utiliza las salidas de los clasificadores utilizados para con ellas conformar un conjunto de entrenamiento. Formado por vectores que contienen, la clasificación emitida por cada clasificador base como variables independientes y como variable dependiente la clase original de cada instancia. Este conjunto de entrenamiento es utilizado por el

metaclasificador para realizar la clasificación. Este algoritmo, a diferencia de los algoritmos Bagging, Boosting y Vote, no utiliza ningún método de combinación descritos en el epígrafe 1.5.4.1.

Luego de aplicada la configuración planteada en el epígrafe 2.4.4, se procedió a realizar la experimentación con el Stacking. La experimentación se realizó para las combinaciones más diversas en las cinco medidas de diversidad y luego para las combinaciones diversas en cuatro de las cinco medidas de diversidad. Los resultados obtenidos fueron registrados en la Tabla 15 yTabla 16.

Subconjunto de Clasificadores	ACC	AUC	KAPPA
JRIP—MULTILAYERPERCEPTRON	0.92424242	0.909	0.7179
EXTRATREE—SMO	0.92424242	0.877	0.7179
FDLA—KSTAR	0.90909091	0.854	0.6471
EXTRATREE—FDLA	0.93939394	0.918	0.7647
EXTRATREE—NAIVEBAYES	0.95454545	0.88	0.7975
JRIP—SMO	0.92424242	0.805	0.6927
FDLA—JRIP	0.90909091	0.854	0.6471
JRIP—NAIVEBAYES	0.92424242	0.909	0.7179
KSTAR—NAIVEBAYES	0.90909091	0.854	0.6471
KSTAR—SMO	0.93939394	0.871	0.7417
EXTRATREE--JRIP—SMO	0.90909091	0.851	0.614
EXTRATREE--JRIP—NAIVEBAYES	0.89393939	0.845	0.6051
EXTRATREE--FDLA—JRIP	0.92424242	0.863	0.6927
FDLA--JRIP—KSTAR	0.95454545	0.88	0.7975
JRIP--KSTAR—SMO	0.92424242	0.909	0.7179
JRIP--KSTAR—NAIVEBAYES	0.92424242	0.863	0.6927
EXTRATREE--FDLA--JRIP—NAIVEBAYES	0.90909091	0.854	0.6471
FDLA--JRIP--KSTAR—NAIVEBAYES	0.93939394	0.871	0.7427
EXTRATREE--JRIP--NAIVEBAYES—SMO	0.87878788	0.883	0.5982
JRIP--KSTAR--NAIVEBAYES—SMO	0.95454545	0.927	0.8156

TABLA 15: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR STACKING

Subconjunto de Clasificadores	ACC	AUC	KAPPA
JRIP--NAIVEBAYES—SMO	0.86363636	0.921	0.563
FDLA--JRIP—NAIVEBAYES	0.90909091	0.901	0.674
JRIP--LOGISTIC—NAIVEBAYES	0.86363636	0.874	0.563
FDLA--JRIP—MULTILAYERPERCEPTRON	0.89393939	0.84	0.605
FDLA--JRIP—SMO	0.90909091	0.854	0.647
JRIP--MULTILAYERPERCEPTRON—SMO	0.90909091	0.854	0.647
JRIP--MULTILAYERPERCEPTRON—NAIVEBAYES	0.90909091	0.854	0.647
EXTRATREE--FDLA--JRIP—SMO	0.89393939	0.845	0.605
FDLA--JRIP--KSTAR—SMO	0.92424242	0.909	0.717
FDLA--JRIP--KSTAR--NAIVEBAYES—SMO	***0.96969697***	***0.936***	***0.871***
EXTRATREE--FDLA--JRIP--NAIVEBAYES--SMO	0.86363636	0.874	0.563

TABLA 16: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN CUATRO DE LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR STACKING

Como se puede apreciar en la Tabla 15 ninguna de las combinaciones posee un desempeño superior al mejor clasificador individual por tal razón no serán tenidas en cuenta para la selección del mejor multiclasificador. En este caso el mejor resultado iguala al desempeño del mejor clasificador individual. Sin embargo en la Tabla 16 se encontró una combinación que supera en desempeño al mejor clasificador individual. Esta combinación de clasificadores permite identificar una instancia más que el mejor clasificador individual, por lo que estamos en presencia del segundo modelo multiclasificador a utilizar como solución al problema.

Estos resultados permiten demostrar que la forma de determinar diversidad reportada en la literatura no está correlacionada con una mejora en el desempeño del multiclasificador utilizado. En este caso se identificó una combinación de clasificadores que, según las medidas de diversidad, no era muy diversa, sin embargo garantizó con su utilización, una mejora en el desempeño del multiclasificador evaluado. Facilitando obtener un desempeño superior al mejor alcanzado por el clasificador individual utilizado para la evaluación de los multiclasificadores.

3.3.5 Resultados de la experimentación con el multiclasificador Vote.

En el epígrafe 2.4.5 se especifica que este algoritmo, al igual que el Stacking, no garantiza la diversidad, razón por la cual es necesario garantizar clasificadores individuales diversos para maximizar el espacio de búsqueda.

Para la realización de la experimentación con el algoritmo Vote, se utilizan las mismas combinaciones utilizadas para realizar la experimentación con el multiclasificador Stacking.

Vote posee la particularidad que utiliza varios métodos para la combinación de las salidas de los clasificadores utilizados como base, los cuales están descritos en el epígrafe 1.5.4.1. En este caso solo se experimentó utilizando como método de combinación “Voto Mayoritario” y “Promedio de las probabilidades”.

A continuación se muestran los resultados de la experimentación realizada con el multiclasificador Vote para los dos grupos de clasificadores diversos registrados en las Tabla 13 y Tabla 14. En ambos casos, fueron utilizados los métodos de combinación especificados con anterioridad.

Método de combinación	**Promedio de probabilidades**		
Subconjunto de Clasificadores	**ACC**	**AUC**	**KAPPA**
JRIP—MULTILAYERPERCEPTRON	0.92424242	0.975	0.764
EXTRATREE—SMO	0.93939394	0.980	0.742
FDLA—KSTAR	0.92424242	0.965	0.692
EXTRATREE—FDLA	0.95238095	0.969	0.797
EXTRATREE—NAIVEBAYES	0.95454545	0.980	0.797
JRIP—SMO	0.92424242	0.928	0.717
FDLA—JRIP	0.93939394	0.969	0.716
JRIP—NAIVEBAYES	0.92424242	0.927	0.717
KSTAR—NAIVEBAYES	0.92424242	0.975	0.717
KSTAR—SMO	0.92424242	0.975	0.717

EXTRATREE--JRIP—SMO	0.95454545	0.978	0.797
EXTRATREE--JRIP—NAIVEBAYES	0.95454545	0.978	0.797
EXTRATREE--FDLA—JRIP	0.95454545	0.967	0.797
FDLA--JRIP—KSTAR	0.93939394	0.965	0.716
JRIP--KSTAR—SMO	0.92424242	0.975	0.692
JRIP--KSTAR—NAIVEBAYES	0.92424242	0.975	0.692
EXTRATREE--FDLA--JRIP—NAIVEBAYES	0.95454545	0.977	0.797
FDLA--JRIP--KSTAR—NAIVEBAYES	0.92424242	0.975	0.692
EXTRATREE--JRIP--NAIVEBAYES—SMO	0.93939394	0.978	0.764
JRIP--KSTAR--NAIVEBAYES—SMO	0.93939394	0.975	0.764

TABLA 17: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR VOTE POR PROMEDIO DE PROBABILIDADES

Método de combinación	**Voto mayoritario**		
Subconjunto de Clasificadores	**ACC**	**AUC**	**KAPPA**
JRIP—MULTILAYERPERCEPTRON	0.92424242	0.816	0.662
EXTRATREE—SMO	0.92424242	0.863	0.692
FDLA—KSTAR	0.92424242	0.863	0.692
EXTRATREE—FDLA	0.92424242	0.863	0.692
EXTRATREE—NAIVEBAYES	0.92424242	0.980	0.797
JRIP—SMO	0.90769231	0.807	0.614
FDLA—JRIP	0.90909091	0.807	0.614
JRIP—NAIVEBAYES	0.90909091	0.807	0.614
KSTAR—NAIVEBAYES	0.92424242	0.863	0.692
KSTAR—SMO	0.92424242	0.863	0.692
EXTRATREE--JRIP—SMO	0.95454545	0.880	0.797
EXTRATREE--JRIP—NAIVEBAYES	0.95454545	0.880	0.797
EXTRATREE--FDLA—JRIP	0.95454545	0.880	0.797
FDLA--JRIP—KSTAR	0.95454545	0.880	0.797
JRIP--KSTAR—SMO	0.95454545	0.880	0.797
JRIP--KSTAR—NAIVEBAYES	0.95454545	0.880	0.797
EXTRATREE--FDLA--JRIP—NAIVEBAYES	0.93939394	0.871	0.742
FDLA--JRIP--KSTAR—NAIVEBAYES	0.93939394	0.871	0.742
EXTRATREE--JRIP--NAIVEBAYES—SMO	0.93939394	0.871	0.742
JRIP--KSTAR--NAIVEBAYES—SMO	0.93939394	0.871	0.742

TABLA 18: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR VOTE POR VOTO MAYORITARIO

Método de combinación	**Promedio de las probabilidades**		
Subconjunto de Clasificadores	**ACC**	**AUC**	**KAPPA**
JRIP--NAIVEBAYES—SMO	0.93939394	0.928	0.764
FDLA--JRIP—NAIVEBAYES	0.93939394	0.971	0.764
JRIP--LOGISTIC—NAIVEBAYES	0.92424242	0.973	0.692
FDLA--JRIP—MULTILAYERPERCEPTRON	0.93939394	0.975	0.764
FDLA--JRIP—SMO	0.92424242	0.969	0.717
JRIP--MULTILAYERPERCEPTRON—SMO	0.93939394	0.975	0.764
JRIP--MULTILAYERPERCEPTRON—NAIVEBAYES	0.93939394	0.973	0.764

EXTRATREE--FDLA--JRIP—SMO	0.95454545	0.977	0.981
FDLA--JRIP--KSTAR—SMO	0.92424242	0.975	0.692
FDLA--JRIP--KSTAR--NAIVEBAYES—SMO	0.93939394	0.975	0.764
EXTRATREE--FDLA--JRIP--NAIVEBAYES--SMO	0.93939394	0.977	0.764

TABLA 19: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN CUATRO DE LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR VOTE POR PROMEDIO DE PROBABILIDADES

Método de combinación	**Voto mayoritario**		
Subconjunto de Clasificadores	**ACC**	**AUC**	**KAPPA**
JRIP--NAIVEBAYES—SMO	0.93939394	0.918	0.764
FDLA--JRIP—NAIVEBAYES	0.93939394	0.918	0.764
JRIP--LOGISTIC—NAIVEBAYES	0.92424242	0.963	0.692
FDLA--JRIP—MULTILAYERPERCEPTRON	0.93939394	0.918	0.764
FDLA--JRIP—SMO	0.92424242	0.909	0.717
JRIP--MULTILAYERPERCEPTRON—SMO	0.93939394	0.918	0.764
JRIP--MULTILAYERPERCEPTRON—NAIVEBAYES	0.93939394	0.918	0.764
EXTRATREE--FDLA--JRIP—SMO	0.92424242	0.863	0.692
FDLA--JRIP--KSTAR—SMO	0.93939394	0.863	0.692
FDLA--JRIP--KSTAR--NAIVEBAYES—SMO	0.93939394	0.918	0.764
EXTRATREE--FDLA--JRIP--NAIVEBAYES--SMO	0.93939394	0.7647	0.918

TABLA 20: RESULTADOS DEL DESEMPEÑO DE LAS COMBINACIONES DE CLASIFICADORES EN CUATRO DE LAS CINCO MEDIDAS DE DIVERSIDAD CON EL MULTICLASIFICADOR VOTE POR VOTO MAYORITARIO

Como resultado del análisis de la información registrada en las Tablas 17-20 no se identificó ninguna combinación que superara al desempeño del mejor clasificador individual, por tal razón no serán tenidas en cuenta para la selección del mejor multiclasificador.

3.4 Selección del multiclasificador para la solución del problema.

A partir de análisis realizado, sobre los resultados obtenidos con la utilización de los multiclasificadores propuestos, se identificaron dos modelos multiclasificadores que mostraron un desempeño superior al mejor desempeño logrado por el mejor clasificador individual. Los resultados se resumen en la Tabla 21.

		ACC	**AUC**	**KAPPA**
Clasificador de control	**Mejor clasificador individual**			
	ExtraTree	0.95454545	0.928	0.7777
Multiclasificadores candidatos	**Multiclasificador AdaBoostM1**			
	MultilayerPerceptron	0.96969697	0.991	0.871
	Multiclasificador Stacking			
	FDLA--JRIP--KSTAR--NAIVEBAYES—SMO	0.96969697	0.936	0.871

TABLA 21: RESULTADOS DE LOS MEJORES MULTICLASIFICADORES Y EL MEJOR CLASIFICADOR INDIVIDUAL

Centrando el análisis en los modelos multiclasificadores identificados, se pudo determinar que los algoritmos no presentan diferencias significativas en su desempeño ya que poseen igual valor tanto en la efectividad como en el estadístico Kappa. Específicamente presentan una diferencia no significativa

en el valor del área bajo la curva ROC que no es suficiente como para indicar si alguno de ellos es superior respecto al otro.

Por tal razón ambos multiclasificadores pueden ser utilizados en la solución del problema. Constituyendo de esta manera, las propuestas a utilizar para la clasificación de los LI capaces de inhibir o no el funcionamiento de la enzima AChE. Ambos multiclasificadores maximizan tanto el TP_{rate} como el TN_{rate} y por tal razón logran clasificar el 96,96% de las instancias presentadas de forma correcta.

3.5 Conclusiones parciales.

En el capítulo fueron recopilados los resultados de la experimentación propuesta. Se presentaron los resultados de aplicar el preprocesamiento al conjunto de datos de los LI. Se evaluaron los clasificadores de la herramienta Weka para determinar aquellos que tenían un desempeño aceptable según el criterio de los especialistas. Fue evaluado el desempeño de estos clasificadores y se procedió a evaluar la utilización de los multiclasificadores. Se evaluaron los multiclasificadores Bagging y Boosting con la utilización de los clasificadores individuales elegidos según el criterio expuesto. A continuación se evaluaron los multiclasificadores Stacking y Vote para los cuales fue necesario calcular la diversidad de las combinaciones de los clasificadores bases a utilizar, debido a que estos algoritmos no garantizan la diversidad. Con esta información se realizó la comparación de los resultados obtenidos por los multiclasificadores contra el clasificador individual de mejor desempeño. El análisis realizado permitió determinar dos modelos multiclasificadores que podrán ser utilizados en la clasificación de los LI con capacidad de inhibir o no el funcionamiento de la enzima AChE.

Con los resultados obtenidos se puedo demostrar la veracidad de la hipótesis planteada y se logró darle cumplimiento al objetivo propuesto en la investigación.

CONCLUSIONES

Los líquidos iónicos son compuestos que han ganado la atención de los investigadores en la industria química. Las características físico-químicas que poseen y sus amplias posibilidades estructurales, permiten que estos compuestos puedan ser sintetizados para satisfacer propósitos específicos. La aplicación de estos compuestos, en la sustitución de solventes moleculares tóxicos para el medio ambiente, se justifica debido a la poca evaporación que poseen a temperatura ambiente, el reducido riesgo de oxidación exotérmica que los caracterizan y que son relativamente no tóxicos. Otras investigaciones han demostrado que algunos de estos compuestos también constituyen un riesgo tóxico para el medio ambiente. En la caracterización del perfil neurotóxico de los LI han sido realizadas diferentes baterías de pruebas biológicas para determinar el perfil físico-químico y su actividad biológica, permitiendo determinar el perfil neurotóxico de los LI.

La enzima AChE juega un papel imprescindible en el sistema nervioso central de los organismos, pues posee un centro activo conservador. El efecto que pueden provocar los LI sobre ella, puede ser extrapolado a diferentes especies que posean un sistema nervioso. Durante la década de los años 30 en países europeos se produjeron compuestos químicos que inhiben esta enzima y que provocan serios daños al organismo afectado. De ahí que la determinación de la capacidad inhibitoria de los LI sobre la AChE sea el indicador neurotóxico utilizado para la determinación del perfil neurotóxico de estos compuestos. La importancia de determinar las características que poseen los LI que son denominados tóxicos, por el efecto inhibitorio que tienen sobre la enzima AChE, juega un papel determinante en la utilización de los LI para la producción de compuestos químicos que respeten los principios de la química verde.

En el trabajo se identificaron multiclasificadores que tuvieron un buen desempeño ante la clasificación de nuevos LI. Para lograr este objetivo se obtuvo una base de datos con las características de los LI y su actividad ante la AChE. Luego se preprocesó esta base de casos comenzando con la realización de una limpieza manual siguiendo una metodología ya establecida en la literatura. Se cuantificó la estructura molecular de los LI mediante la utilización de softwares especializados. Se redujo la dimensionalidad de la base de casos, para evitar el sobreajuste, mediante la utilización del algoritmo mRMR y se eliminó el desbalance en la distribución de las clases con la utilización del algoritmo SMOTE. Ambos algoritmos se encuentran reportados en la literatura. A continuación se evaluaron los clasificadores presentes en la herramienta Weka y se seleccionaron aquellos que por su desempeño, podían ser utilizados en la conformación de los multiclasificadores según el criterio establecido en el trabajo.

Seguidamente se evaluaron los modelos multiclasificadores que podían generarse a partir de los clasificadores individuales seleccionados. Tomando como base el desempeño mostrado ante el problema, se identificaron dos modelos multiclasificadores cuyos resultados superaron al mejor de los clasificadores individuales utilizados para conformar los multiclasificadores. Lográndose de esta manera darle cumplimiento al objetivo de la investigación y demostrándose la veracidad de la hipótesis planteada.

RECOMENDACIONES

Se recomienda aplicar otros algoritmos de selección de atributos que, al introducir otras heurísticas en la identificación de los atributos relevantes, permitan verificar si es posible elevar el desempeño de los clasificadores utilizados.

Se recomienda aplicar otros algoritmos para modificar la distribución de las clases, con el objetivo de verificar si es posible garantizar una mejor discriminación entre las clases y contribuir en la mejora del desempeño de los clasificadores utilizados.

Se recomienda seguir investigando en el tema relacionado con el cálculo de las medidas de diversidad. Con el objetivo de identificar una medida que garantice la identificación de combinaciones de clasificadores diversos que propicien un aumento en el desempeño en la clasificación, debido a la variabilidad del desempeño encontrado en la clasificación realizada mediante la utilización de las combinaciones diversas encontradas.

Se recomienda la elaboración de un software que automatice los modelos propuestos como resultado de la experimentación realizada y que proporcione una herramienta de fácil uso para los especialistas interesados.

BIBLIOGRAFÍA

Aguiar, V., M. Gestal, M. Cruz-Monteagudo, J. R. Rabuñal, J. Dorado and C. R. Munteanu (2013). "Evolutionary Computation and QSAR Research." Current Computer-Aided Drug Design **9**(2).

Anastas, P. A. and J. C. Warner (1998). Green Chemistry: Theory and Practice. Oxford, Oxford University Press.

Arning, J., S. Stole and A. Boshen (2008). "Qualitative and quantitative structure activity relationship for the inhibitory effects of cationic head groups, functionalised side chains and anions of ionic liquids on acetylcholinesterase." Green Chemistry(10).

Arning, J., S. Stolte, A. Boschen, F. Stock, W. Pitner, U. Welz, B. Jastorff and J. Ranke (2008). "Qualitative and quantitative structure activity relationships for the inhibitory effects of cationic head groups, functionalised side chains and anions of ionic liquids on acetylcholinesterase." Green Chem **10**: 11.

Billsus, D. and M. J. Pazzani (1998). "Learning Collaborative Information Filters."

Bonet, I. and J. Arencibia (2012). "Predicción de resistencia a fármacos del VIH utilizando multiclasificadores." Revista Cubana de Ciencias Informáticas.

Bonet, I., P. Franco-Montero, V. Rivero, M. Tejeira, F. Borges, E. Uriarte and A. Morales (2013). "Classifier ensemble based on feature selection and diversity measures for predicting de affinity of A2B Adenosine Receptor Antagonist." Chemical Information and Modeling **53**: 16.

Bonet, I., A. Rodríguez, M. M. García and R. Grau (2012). "Combinación de clasificadores para bioinformática." Computación y Sistemas **16**(2): 10.

Bort, J. (2001). Química teórica y computacional, España.

Breiman, L. (1996). "Bagging predictors." Machine Learning **24**: 17.

Brown, G., J. Wyatts, R. Harris and X. Yao (2005). "Diversity creation methods: A survey and categorization." Information fusion **6**.

Cabrera, L. and G. Casas (2013). Estudios de medidas de diversidad en sistemas multiclasificadores. MsC, Universidad de las Villas. Martha Abreu.

Casas, G., J. L. Morales, L. Denoda, L. Cabrera, L. Arguelles and E. Gonzalez (2013). High cardiovascular risk models for a city in the central region of Cuba. Fourth International workshop on knowledge discover, knowledge management and decision making, Eureka Workshop, Atlantis Press.

Couling, D., R. J. Bernot, K. M. Docherty, J. K. Dixon and E. J. Maginn (2006). "Assessing the factors responsible for ionic liquid toxicity to aquatic organisms via quantitative structure–property relationship modeling." Green Chem **8**.

Cruz-Monteagudo, M., E. Ancede, M. J. M and N. Cordeiro (2013). "Chemoinformatics profiling of ionic liquids – Automatic and chemically interpretable cytotoxicity profiling, virtual screening and cytotoxicophore identification." Green Chemistry.

Chavéz, M. d. C. (2008). Modelos de redes bayesianas en estudios de secuencias genómicas y otros problemas biomédicos. PhD, Universidad de Las Villas Martha Abreu.

Chawla, N., K. Bowyer, L. Hall and W. P. Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Artificial intelligence research **16**.

ChemAxon (2012). Standarizer.

D. Zhao, Y. L., Z. Zhang (2007). "Toxicity of Ionic Liquids." Clean Prod. Proces. **35**: 6.

Davis, J. and M. Goadrish (2006). The relationship between Precision-Recall and ROC curve. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh.

Dehmer, M., K. Varmuza, D. Bonchev and F. Emmert-Streib (2012). Statistical Modelling of Molecular Descriptors in QSAR/QSPR Quantitative and Network Biology.

Ding, C. and H. Peng (2005). "Minimun redundancy feature selection from microarray gene expression data." Bioinformatics and Computational Biology **3**(2): 20.

Dragon (2006). DRAGON for windows(For molecular descriptor calculations).

Freund, Y. and R. Schapire (1999). "A Short Introduction to Boosting." Journal of Japanese Society for Artificial Intelligence **14**(5): 9.

Freund, Y. and R. E. Schapire (1996). Experiments with a New Boosting Algorithm.

García, A., E. Tojo, J. Tojo, M. R. Teijeira, F. J. , M. P. González and V. S. Martínez (2008). "Cytotoxicity of selected imidazolium-derived ionic liquids in the human Caco-2 cell line. Sub-structural toxicological interpretation through a QSAR study." Green Chem **10**.

García, S., A. Fernández, L. Luengo and F. Herrera (2008). "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability." Soft Comput **13**.

Guo, H. and H. L. Viktor "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach." Sigkdd Explorations. **6**: 30.

Hall, M., H. Frank, H. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009). "The WEKA Data Mining Software: An Update." Sigkdd Explorations. **11**(1).

Hansh, C. (1964). "Method for correlation of biological activity of chemical." Journal of American Chemical Society.

Hossain, M., B. Belhaouari, M. El-Harbawi, A. Nusaibah, M. I. Abdul, G. Hefter and C.-Y. Yin (2011). "Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities." Chemosphere **85**: 4.

Inc, S. (2006). Cap 23.Análisis Discriminante: El procedimiento Discriminante. Introducción al SPSS.

ISIDA (2013). ISIDA.

Ivanciuc, O. I., T.; Klein, D.J. (2001). " Quantitative structure-property relationships generated with optimizable even/odd Wiener polynomial descriptors. ." SAR QSAR Environ. Res. **12**: 16.

Jchem (2013). Jchem for Excel.

Kittler, J., M. Hatted, R. Duin and J. Matas (1998). "On combining classifiers." IEEE Transaction in pattern recognizicion and machine intelligent **20**.

Kohavi, R. and G. John (1997). "Wrapper for feature subset selection." Artificial intelligence.

Koolman, J. and K. H. Roehm (2005). Color Atlas of Biochemistry. Estados Unidos, Thieme.

Kuncheva, L. and C. Whitaker (2003). "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy." Machine Learning **51**(2): 181-207.

Kuncheva, L. I. (2004). "Combining Pattern Classifiers: Methods and Algorithms." Neural Networks, IEEE Transactions on **18**(3): 964-964.

Kuncheva, L. I. (2007). "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, L.I.; 2004) [book review]." Neural Networks, IEEE Transactions on **18**(3): 964-964.

Lacrama, A., M. Putz and V. Ostafe (2007). "A Spectral-SAR Model for the Anionic-Cationic Interaction in Ionic Liquids: Application to Vibrio fischeri Ecotoxicity." Int. J. Mol. Sci **8**: 21.

Larose, D. T. (2005). Discovering knowledge in data. New York, John Wiley & Sons, Inc.

Lehninger (2005). Principios de Bioquímica.

Luis, P., A. Garea and A. Irabien (2010). "Quantitative structure – activity relationships (QSARs) to estimate ionic liquids ecotoxicity EC50 (Vibrio fi scheri)." Molecular Liquids **152**: 5.

Luis, P., I. Ortiz, R. Aldaco and A. Irabien (2007). "A novel group contribution method in the development of a QSAR for predicting the toxicity (Vibrio fischeri EC50) of ionic liquids." Ecotoxicol. Environ. Saf **67**: 6.

Martin, B. and A. Sanz (2001). Redes neuronales y sistemas difusos. Saragoza, España.

Mathews., Holde. and Ahern. (2005). BioChemistry.

Metzler, D. E. (2003). Biochemistry. The Chemical Reactions of living cells., ELSEVIER. Academic Press.

Mitchel, T. (2007). Machine Learning, McGraw-Hill Science.

Mitchell, T. M. (1997). "*Machine Learning*." McGraw-Hill Science/Engineering/Math: .

Mohammad, H., A. Fatemi and P. Izadiyan (2011). "Cytotoxicity estimation of ionic liquids based on their effective structural features." Chemosphere **84**: 10.

Mohammad, H., A. Fatemi and P. Izadiyan (2011). "In silico cytotoxicity estimation of ionic liquids based on their two- and three-dimensional structural descriptors." Monatsh Chem **142**: 8.

Molina, Y. (2010). Desarrollo de modelos QSAR utilizando Programación Genética y Árboles de Regresión. Master, Universidad de las Ciencias Informáticas.

Peng, H., F. Long and C. Ding (2005). "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy." IEEE Transaction in pattern recognizicion and machine intelligent **27**(8).

Putz, M., A. Lacrama and V. Ostafe (2007). Res. Lett. Ecology **7**: 4.

Quinlan, R. (1986). "ID3 System."

Quinlan, R. (1993). "C4.5: “Programs for Machine Learning”."

R. F. Frade, C. A. A. (2010). Hum. Exp. Toxicol. **29**: 14.

Ramentol, E., Y. Caballero, R. Bello and F. Herrera (2011). "SMOTE-RS B∗: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory." Knowl Inf Syst.

Ramentol, E., F. Herrera, R. Bello, Y. Caballero and Y. Sánchez (2009). Edición de Conjuntos de Entrenamiento no Balanceados, haciendo uso de Operadores Genéticos y la Teoría de los Conjuntos Aproximados. VI Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB'09), España.

Ramentol, E., E. VERBIEST, R. Bello, Y. Caballero, C. CORNELIS and F. Herrera (2012). SMOTE-FRST: A NEW RESAMPLING METHOD USING FUZZY ROUGH SET THEORY. Flins12.

Randi, M. (1991). "Orthogonal molecular descriptors." New J. Chem. **15**: 7.

Randi, M. (1995). "Molecular profiles. Novel geometry-dependent molecular descriptors. ." New J. Chem. **19**: 10.

Randi, M. (1998). "On characterization of molecular attributes." Acta Chim **45**: 13.

Richert, W. and L. P. Coelho (2013). Building machine learning systems with Python., Packt Publishing Ltd.

Ruiz, R. (2005). Selección de Atributos mediante proyecciones. DrC, Universidad de Sevilla.España.

Schapire, R. E. (1990). "The strength of weak learnability." Machine Learning **5**: 10.

Segrera, S. and M. N. Moreno (2006). Multiclasificadores: Métodos y Arquitecturas. D. d. I. y. A. U. d. Salamanca.

Sierra, J., A. Garau, E. Martí and R. Cruañas (2006). QUÍMICA VERDE: TOXICIDAD Y BIODEGRADABILIDAD DE LÍQUIDOS IÓNICOS. Departament de Productes Naturals, Biologia Vegetal i Edafologia, Unitat d’Edafologia, Universitat de Barcelona**:** 34.

Sierra, J., A. Garau, E. Martí and R. Cruañas (2006). Usos y aplicaciones de los líquidos iónicos

SPSS (2010). SPSS.

Stock, F., J. Hoffmann, J. Ranke, R. Stormann, B. Ondruschka and B. Jastorff (2004). "Effects of ionic liquids on the acetylcholinesterase -a structure-activity relationship consideration." Green Chem **6**: 5.

Störmann, R., J. Arning, S. Stolte, M. Matzke, B. Jastorff and J. Thöming. (2011). "UFT Merck Ionic Liquids Biological Effects Database." 2014, from http://www.il-eco.uft.uni-bremen.de.

Tax, D., M. Breukelen, R. Duin and J. Kittler (2003). "Combining multiples classifiers by averaging or multiplying? ." Pattern Recognition **33**.

Tejedor, C. (2005). "Bioquímica Ambiental.", 2014, from http://www.uah.es/tejedor_bio/bioquimica_ambiental/tema12/tema%2012-indice.htm.

Todeschini, R. C., V. (2000.). Handbook of Molecular Descriptors., Wiley-VCH Publishers: Weinheim.

Torrecilla, J., J. Palomar, J. Lemus and F. Rodriguez (2010). "A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids." Green Chemistry **12**(1): 123-134.

Torrecilla, J., J. Palomar, J. Lemus and F. Rodríguez (2010). "A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids." Green Chem **12**: 11.

Tropsha, A. (2010). "Best practices for QSAR Model Development, Validation and Exploration." Molecular Informatic **29**: 12.

Tropsha, A., D. Fourche and E. Muratov (2010). "Trust, but verify: On importance of Chemical Structure Curation in cheminformatic and QSAR Modeling Research." J. Chem. Inf. Model **50**(7).

Vaskin, I. and A. Varnek (2008). "Building a Chemical Space based on fragments descriptors." Combinatorial Chemistry & High Throughput Screening **11**.

Witten, I. H. and E. Frank (2005). Data Mining. A practical machine learning tools and techniques.

Wolper, D. H. (1992). "Stacked Generalization." Neural Networks **5**: 18.

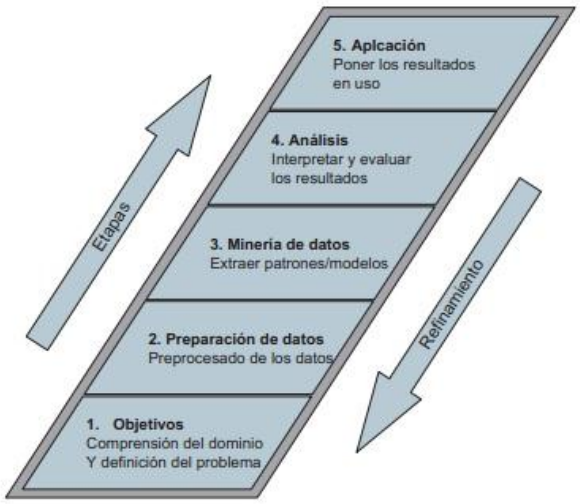
Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand and D. Steinb (2008). "Top 10 algorithms in data mining." Knowl Inf Syst **14**.

Yao, X. and Y. Liu (2005). Machine Learning. Search methodologies. Introductory Tutorial in optimization and decision support techniques, Springer.

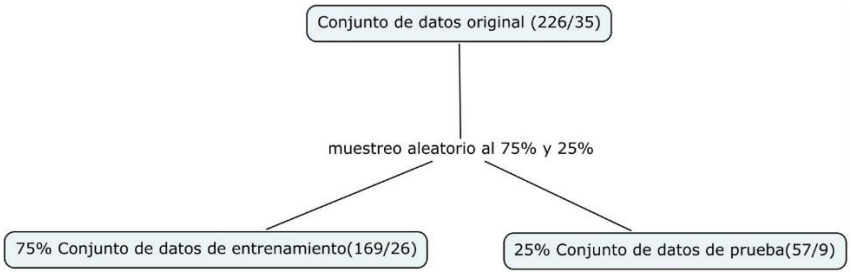
ANEXOS

ANEXO 1: APLICACIÓN DE LOS LÍQUIDOS IÓNICOS

ANEXO 2: ESQUEMA GENERAL DEL KDD



ANEXO 3: PARTICIÓN DEL CONJUNTO DE DATOS.



ANEXO 4: DESCRIPTORES MOLECULARES

[A-][Br-]	[A-]O-S-O-C-C-O	[C+]C*C*C*C*N	[C+]C-C-C-C-C-C-C-C-C-C-N
[A-][Cl-]	[A-]C-O-C-C-O-S-O	[C+]C-N	[C+]C-C-N-C
[A-][I-]	[A-]C-C-O-C-C-O-S-O	[C+]C-C-N	[C+]C-N-C-C*C
[A-]B-F	[A-]O-S-O-C-C-O-C-C-O	[C+]C-C-C-N	[C+]C-N-C-C*C*C
[A-]F-B-F	[A-]C-O-C-C-O-C-C-O-S-O	[C+]C-C-C-C-N	[C+]C-N-C-C*C*C*C
[A-]S-N	[A-]S-O-C-C-O	[C+]C-C-C-C-C-N	[C+]C-C-C-N-C
[A-]N-S=O	[A-]S-O-C-C-O-C	[C+]C-C-C-C-C-C-N	[C+]C-C-C-C-N-C
[A-]C-S-N	[A-]S-O-C-C-O-C-C	[C+]C-C-C-C-C-C-C-N	[C+]C-C-C-C-C-N-C
[A-]N-S-C-F	[A-]S-O-C-C-O-C-C-O	[C+]C-C-C-C-C-C-C-C-N	[C+]C-C-C-C-C-C-N-C
[A-]S=O	[A-]S-O-C-C-O-C-C-O-C	[C+]C*C	[C+]C-C-C-C-C-C-C-N-C
[A-]S-C	[A-]O-C-C-O-S=O	[C+]C*C*C	[C+]C-C-C-C-C-C-C-C-N-C
[A-]S-C-F	[A-]C-O-C-C-O-S=O	[C+]C*C*N*C	[C+]C-C-C-C-C-C-C-C-C-N-C
[A-]S-N-S	[A-]C-C-O-C-C-O-S=O	[C+]C*C*C*N*C	[C+]C-C-C-C-C-C-C-C-C-C-N-C
[A-]S-N-S=O	[A-]O-C-C-O-C-C-O-S=O	[C+]C*C*C*C*N*C	[C+]C-C*C*C*C
[A-]S-N-S-C	[A-]C-O-C-C-O-C-C-O-S=O	[C+]C*N*C	[C+]C-C-N-C-C
[A-]S-N-S-C-F	[A-]O-C-C-O	[C+]C-N*C	[C+]C-C-C-N-C-C
[A-]O=S=O	[A-]C-O-C-C-O	[C+]C-C-N*C	[C+]C-C-C-C-N-C-C
[A-]C-S=O	[A-]C-C-O-C-C-O	[C+]C-C-C-N*C	[C+]C-C-C-C-C-N-C-C
[A-]O=S-C-F	[A-]O-C-C-O-C-C-O	[C+]C-C-C-C-N*C	[C+]C-C-C-C-C-C-N-C-C
[A-]O=S-N-S=O	[A-]C-O-C-C-O-C-C-O	[C+]C-C-C-C-C-N*C	[C+]C-C-C-C-C-C-C-N-C-C
[A-]C-S-N-S=O	[A-]C-C-O-C	[C+]C-C-C-C-C-C-N*C	[C+]C-C-C-C-C-C-C-C-N-C-C
[A-]O=S-N-S-C-F	[A-]C-C-O-C-C	[C+]C-C-C-C-C-C-C-N*C	[C+]C-C-C-C-C-C-C-C-C-N-C-C

[A-]C-F	[A-]C-C-O-C-C-O-C	[C+]C-C-C-C-C-C-C-C-N*C	[C+]C-C-C-C-C-C-C-C-C-C-N-C-C
[A-]C-S-N-S-C	[A-]C-O-C	[C+]C*C*C*C	[C+]C-C-N-C-C*C
[A-]C-S-N-S-C-F	[A-]C-O-C-C-O-C	[C+]C*C*C*C*C	[C+]C-C-C-N-C-C*C
[A-]F-C-F	[A-]B-O	[C+]C*C*C*N*C*C	[C+]C-C-C-C-N-C-C*C
[A-]F-C-S-N-S-C-F	[A-]B-O-C	[C+]C*C*N*C*C	[C+]C-C-C-C-C-N-C-C*C
[A-]C-N	[A-]B-O-C*C	[C+]C-N*C*C	[C+]C-C-C-C-C-C-N-C-C*C
[A-]N-C+N	[A-]B-O-C*C*C	[C+]C-C-N*C*C	[C+]C-C-C-C-C-C-C-N-C-C*C
[A-]C+N	[A-]C*C-O	[C+]C-C-C-N*C*C	[C+]C-C-C-C-C-C-C-C-N-C-C*C
[A-]C-N-C	[A-]C*C*C-O	[C+]C-C-C-C-N*C*C	[C+]C-C-C-C-C-C-C-C-C-N-C-C*C
[A-]C-N-C+N	[A-]C*C*C*C-O	[C+]C-C-C-C-C-N*C*C	[C+]C-C-C-C-C-C-C-C-C-C-N-C-C*C
[A-]N+C-N-C+N	[A-]O-B-O	[C+]C-C-C-C-C-C-N*C*C	[C+]C-C-N-C-C*C*C
[A-]P-F	[A-]C-O-B-O	[C+]C-C-C-C-C-C-C-N*C*C	[C+]C-C-C-N-C-C*C*C
[A-]F-P-F	[A-]C*C-O-B-O	[C+]C-C-C-C-C-C-C-C-N*C*C	[C+]C-C-C-C-N-C-C*C*C
[A-]B-C	[A-]C*C*C-O-B-O	[C+]C-N*C*C*C	[C+]C-C-C-C-C-N-C-C*C*C
[A-]B-C+N	[A-]C-O-B-O-C	[C+]C-C-N*C*C*C	[C+]C-C-C-C-C-C-N-C-C*C*C
[A-]C-B-C	[A-]C-O-B-O-C*C	[C+]C-C-C-N*C*C*C	[C+]C-C-C-C-C-C-C-N-C-C*C*C
[A-]C-B-C+N	[A-]C-O-B-O-C*C*C	[C+]C-C-C-C-N*C*C*C	[C+]C-C-C-C-C-C-C-C-N-C-C*C*C
[A-]N+C-B-C+N	[A-]C*C-O-B-O-C*C	[C+]C-C-C-C-C-N*C*C*C	[C+]C-C-C-C-C-C-C-C-C-N-C-C*C*C
[A-]N-C-F	[A-]C*C-O-B-O-C*C*C	[C+]C-C-C-C-C-C-N*C*C*C	[C+]C-C-C-C-C-C-C-C-C-C-N-C-C*C*C
[A-]C-N-C-F	[A-]C*C*C-O-B-O-C*C*C	[C+]C-C-C-C-C-C-C-N*C*C*C	[C+]C-C-N-C-C*C*C*C
[A-]F-C-N-C-F	[A-]S-C+N	[C+]C-C-C-C-C-C-C-C-N*C*C*C	[C+]C-C-C-N-C-C*C*C*C
[A-]P-C	[A-]C-C-C-O	[C+]C-N*C*C*C*C	[C+]C-C-C-C-N-C-C*C*C*C
[A-]P-C-F	[A-]C-C-C	[C+]C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-N-C-C*C*C*C

[A-]P-C-C	[A-]C-C-C=O	[C+]C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-N-C-C*C*C*C
[A-]P-C-C-F	[A-]O-C-C=O	[C+]C-C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-C-N-C-C*C*C*C
[A-]C-P-F	[A-]N-S-O	[C+]C-C-C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-C-C-N-C-C*C*C*C
[A-]F-P-C-F	[A-]P-O	[C+]C-C-C-C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-C-C-C-N-C-C*C*C*C
[A-]C-C-P-F	[A-]O-P=O	[C+]C-C-C-C-C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-C-C-C
[A-]F-P-C-C-F	[A-]C-P-O	[C+]C-C-C-C-C-C-C-C-N*C*C*C*C	[C+]C-C-C-C-C-C-C-C-C-C
[A-]C-C	[A-]C-C-P-O	[C+]C-C	[C+]N*C*N
[A-]C-C-F	[A-]C-C-C-P-O	[C+]C-C-C	[C+]C-N*C*N
[A-]C-P-C	[A-]C-C-C-C-P-O	[C+]C-C-C-C	[C+]C-C-N*C*N
[A-]C-P-C-F	[A-]C-C-C-C-C-P-O	[C+]C-C-C-C-C	[C+]C-C-C-N*C*N
[A-]C-P-C-C	[A-]P=O	[C+]C-C-C-C-C-C	[C+]C-C-C-C-N*C*N
[A-]C-P-C-C-F	[A-]P-C-C-C	[C+]C-C-C-C-C-C-C	[C+]C-C-C-C-C-N*C*N
[A-]F-C-C-F	[A-]P-C-C-C-C	[C+]C-C-C-C-C-C-C-C	[C+]C-C-C-C-C-C-N*C*N
[A-]F-C-P-C-F	[A-]P-C-C-C-C-C	[C+]N-C*C*C*N	[C+]C-C-C-C-C-C-C-N*C*N
[A-]C-C-P-C-F	[A-]C-P=O	[C+]C-N-C*C*C*N	[C+]C-C-C-C-C-C-C-C-N*C*N
[A-]F-C-P-C-C-F	[A-]C-C-P=O	[C+]C*C*C-N	[C+]C-C-C-C-C-C-C-C-C-N*C*N
[A-]C-C-P-C-C	[A-]C-C-C-P=O	[C+]C-N-C*C*C	[C+]C-C-C-C-C-C-C-C-C-C-N*C*N
[A-]C-C-P-C-C-F	[A-]C-C-C-C-P=O	[C+]C*C-N	[C+]C-N*C*N-C
[A-]F-C-C-P-C-C-F	[A-]C-C-C-C-C-P=O	[C+]C-N-C*C	[C+]C-C-N*C*N-C
[A-]S-O	[A-]C-C-C-C	[C+]C-N-C	[C+]C-C-C-N*C*N-C
[A-]O-S=O	[A-]C-C-C-C-C	[C+]C-N*C*C*C-N	[C+]C-C-C-C-N*C*N-C
[A-]O-S-O	[A-]C-P-C-C-C	[C+]C-C-N*C*C*C-N	[C+]C-C-C-C-C-N*C*N-C
[A-]C-O-S-O	[A-]C-P-C-C-C-C	[C+]C-C-C-N*C*C*C-N	[C+]C-C-C-C-C-C-N*C*N-C

[A-]S-O-C	[A-]C-P-C-C-C-C-C	[C+]C-C-C-C-N*C*C*C-N	[C+]C-C-C-C-C-C-C-N*C*N-C
[A-]C-O-S=O	[A-]C-C-P-C-C-C	[C+]C-C-C-C-C-N*C*C*C-N	[C+]C-C-C-C-C-C-C-C-N*C*N-C
[A-]C-O	[A-]C-C-P-C-C-C-C	[C+]C-C-C-C-C-C-N*C*C*C-N	[C+]C-C-C-C-C-C-C-C-C-N*C*N-C
[A-]C-S-O	[A-]C-C-P-C-C-C-C-C	[C+]C-N-C*C*C*N-C	[C+]C-C-C-C-C-C-C-C-C-C-N*C*N-C
[A-]O-S-C-F	[A-]C-C-C-P-C-C-C	[C+]C-C-N*C*C*C-N-C	[C+]C-C-N*C*N-C-C
[A-]C*C-S-O	[A-]C-C-C-P-C-C-C-C	[C+]C-C-C-N*C*C*C-N-C	[C+]C-C-C-N*C*N-C-C
[A-]C*C*C-S-O	[A-]C-C-C-P-C-C-C-C-C	[C+]C-C-C-C-N*C*C*C-N-C	[C+]C-C-C-C-N*C*N-C-C
[A-]C*C*C*C-S-O	[A-]C-C-C-C-P-C-C-C-C	[C+]C-C-C-C-C-N*C*C*C-N-C	[C+]C-C-C-C-C-N*C*N-C-C
[A-]C-C*C*C*C-S-O	[A-]C-C-C-C-P-C-C-C-C-C	[C+]C-C-C-C-C-C-N*C*C*C-N-C	[C+]C-C-C-C-C-C-N*C*N-C-C
[A-]S-C*C	[A-]C-C-C-C-C-P-C-C-C-C-C	[C+]C-C*C*N	[C+]C-C-C-C-C-C-C-N*C*N-C-C
[A-]S-C*C*C	[A-]C-C-C-C-O	[C+]C-C*C	[C+]C-C-C-C-C-C-C-C-N*C*N-C-C
[A-]S-C*C*C*C	[A-]C-C-C-C-C-O	[C+]C-C*C*C	[C+]C-C-C-C-C-C-C-C-C-N*C*N-C-C
[A-]S-C*C*C*C-C	[A-]C-C-C-C-C-C-O	[C+]C-C*C*N*C	[C+]C-C-C-C-C-C-C-C-C-C-N*C*N-C-C
[A-]C*C-S=O	[A-]C-C-C-C-C-C-C-O	[C+]C-C*C*N-C	[C+]C-C-C-C-C-C-C-C-C-N*C
[A-]C*C*C-S=O	[A-]C-C-C-C-C-C-C-C-O	[C+]C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-N*C
[A-]C*C*C*C-S=O	[A-]C-C-C-C-C-C-C-C-C-O	[C+]C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-N*C*C
[A-]C-C*C*C*C-S=O	[A-]C-C-C-C-C-C-C-C-C-C-O	[C+]C-C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-N*C*C
[A-]C*C	[A-]C-C-C-C-C-C	[C+]C-C-C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-C-N*C*N
[A-]C*C*C	[A-]C-C-C-C-C-C-C	[C+]C-C-C-C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-C-C-N*C*N
[A-]C*C*C*C	[A-]C-C-C-C-C-C-C-C	[C+]C-C-C-C-C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-C-C-N*C*N-C
[A-]C-C*C*C*C	[A-]C-C-C-C-C-C-C-C-C	[C+]C-C-C-C-C-C-C-C-N*C*C-C	[C+]C-C-C-C-C-C-C-C-C-C-C-N*C
[A-]C-C*C*C	[A-]C-C-C-C-C-C-C-C-C-C	[C+]C-C*N	[C+]C-C-C-C-C-C-C-C-C-C-C-N*C
[A-]C-C*C	[A-]C-C-C-C=O	[C+]C-C*N*C*C	[C+]C-C-C-C-C-C-C-C-C-C-C-C-N*C

[A-]O-C=O	[A-]C-C-C-C-C=O	[C+]C-C*N*C	[C+]C-C-C-C-C-C-C-C-C-C-N
[A-]C-C-O	[A-]C-C-C-C-C-C=O	[C+]C-C*N-C	[C+]C-C-C-C-C-C-C-C-C-C-C-N
[A-]O-C-C-F	[A-]C-C-C-C-C-C-C=O	[C+]C-C-N*C-C	[C+]C-C-C-C-C-C-C-C-C-C-C-C-N
[A-]C=O	[A-]C-C-C-C-C-C-C-C=O	[C+]C-C-C-N*C-C	[C+]C-C-C-C-C-C-C-C-C-C-C-C-C-N
[A-]C-C=O	[A-]C-C-C-C-C-C-C-C-C=O	[C+]C-C-C-C-N*C-C	[C+]C-C-C-C-C-C-C-C-C-C-N*C*C
[A-]O=C-C-F	[A-]C-C-C-C-C-C-C-C-C-C=O	[C+]C*C-C-N	[C+]C-C-C-C-C-C-C-C-C-C-C-N*C*C
[A-]C-C-O-S-O	[C+]C*N	[C+]C*C*C-C-N	[C+]C-C-C-C-C-C-C-C-C-C
[A-]S-O-C-C	[C+]C*C*N	[C+]C*C*C*C-C-N	[C+]C-C-C-C-C-C-C-C-C-C
[A-]C-C-O-S=O	[C+]C*C*C*N	[C+]C-C-C-C-C-C-C-C-N	[C+]C-C-C-C-C-C-C-C-C-C-C
[C+]C-C-C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-S-C	[C+]C-C-C-P-C-C-C-C-C-C-C-C	
[C+]C-C-C-C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-S-C-C	[C+]C-C-C-P-C-C-C-C-C-C-C-C-C	
[C+]C-C*C*C*N	[C+]C-C-S-C-C	[C+]C-C-C-P-C-C-C-C-C-C-C-C-C-C	
[C+]C-C*C-C	[C+]C-C*C*C*C-C-N*C*N	[C+]C-C-C-C-P-C-C-C-C	
[C+]C-C*C*C*N-C	[C+]C-C*C*C*C-C-N*C*N-C	[C+]C-C-C-C-P-C-C-C-C-C	
[C+]C-C-N*C*C*C-C	[C+]C-C*C*C*C-C-N*C	[C+]C-C-C-C-P-C-C-C-C-C-C	
[C+]C-C-C-N*C*C*C-C	[C+]C-C*C*C*C-C-N	[C+]C-C-C-C-P-C-C-C-C-C-C-C	
[C+]C-C-C-C-N*C*C*C-C	[C+]C-C*C*C*C-C-N*C*C	[C+]C-C-C-C-P-C-C-C-C-C-C-C-C	
[C+]C-C-C-C-C-N*C*C*C-C	[C+]C-C*C*C*C-C	[C+]C-C-C-C-P-C-C-C-C-C-C-C-C	
[C+]C-C-C-C-C-C-N*C*C*C-C	[C+]N*C*N-C-C-O	[C+]C-C-C-C-P-C-C-C-C-C-C-C-C-C	
[C+]C-C-C-C-C-C-C-N*C*C*C-C	[C+]C-O-C-C-N*C*N	[C+]C-C-C-C-C-P-C-C-C-C-C	
[C+]C-C-C-C-C-C-C-C-N*C*C*C-C	[C+]C-C-O-C-C-N*C*N	[C+]C-C-C-C-C-P-C-C-C-C-C-C	
[C+]C-C-C-C-C-N*C-C	[C+]C-N*C*N-C-C-O	[C+]C-C-C-C-C-P-C-C-C-C-C-C-C	
[C+]C-C-C-C-C-C-N*C-C	[C+]C-N*C*N-C-C-O-C	[C+]C-C-C-C-C-P-C-C-C-C-C-C-C-C	

[C+]N*C*N-C-C-C-C-C-C-C-C-O	[C+]C-C-O-C-C-N*C*N-C	[C+]C-C-C-C-C-P-C-C-C-C-C-C-C-C-C	
[C+]C-N*C*N-C-C-C-C-C-C-C-C-O	[C+]C-C-C-N-C-O	[C+]C-C-C-C-C-C-P-C-C-C-C-C-C	
[C+]C*N-C-C-C-C-C-C-C-C-O	[C+]C-C-C-N-C-O-C	[C+]C-C-C-C-C-C-P-C-C-C-C-C-C	
[C+]N-C-C-C-C-C-C-C-C-O	[C+]C-C-C-N-C-O-C-C	[C+]C-C-C-C-C-C-P-C-C-C-C-C-C-C	
[C+]C*C*N-C-C-C-C-C-C-C-C-O	[C+]N-C-C-C-O	[C+]N*C*N-C-O	
[C+]C-C-C-C-C-C-C-C-O	[C+]C-O-C-C-C-N	[C+]C-O-C-N*C*N	
[C+]C-C-C-C-C-C-C-O	[C+]C*N-C-C-C-O	[C+]C-C-O-C-N*C*N	
[C+]C-C-C-C-C-C-O	[C+]C-O-C-C-C-N*C	[C+]C-N*C*N-C-O	
[C+]C-C-C-C-C-O	[C+]C*C*N-C-C-C-O	[C+]C-N*C*N-C-O-C	
[C+]C-C-C-C-O	[C+]C-O-C-C-C-N*C*C	[C+]C-C-O-C-N*C*N-C	
[C+]C-C-C-O	[C+]C*C*C*N-C-C-C-O	[C+]N*C*N-C-C-C-C-O	
[C+]C-C-O	[C+]C-O-C-C-C-N*C*C*C	[C+]C-N*C*N-C-C-C-C-O	
[C+]C-O	[C+]C-C-C-O-C	[C+]C*N-C-C-C-C-O	
[C+]N-C-C-O	[C+]C-N-C-C-C-O	[C+]N-C-C-C-C-O	
[C+]C-O-C-C-N	[C+]C-N-C-C-C-O-C	[C+]C*C*N-C-C-C-C-O	
[C+]C-C-O-C-C-N	[C+]C-C-N-C-C-C-O	[C+]C-C-C-C-N-C-C-O	
[C+]C*N-C-C-O	[C+]C-C-N-C-C-C-O-C	[C+]N*C*N-C-C-C=O	
[C+]C-O-C-C-N*C	[C+]C-C-C-N-C-C-C-O	[C+]C-N*C*N-C-C-C=O	
[C+]C-C-O-C-C-N*C	[C+]C-C-C-N-C-C-C-O-C	[C+]C*N-C-C-C=O	
[C+]C*C*N-C-C-O	[C+]C-C-C-C-N-C-C-C-C	[C+]N-C-C-C=O	
[C+]C-O-C-C-N*C*C	[C+]C=C-C-N*C*N	[C+]C*C*N-C-C-C=O	
[C+]C-C-O-C-C-N*C*C	[C+]C-N*C*N-C-C=C	[C+]C-C-C=O	
[C+]C*C*C*N-C-C-O	[C+]C=C-C-N*C	[C+]C-C=O	

[C+]C-O-C-C-N*C*C*C	[C+]C=C-C-N	[C+]C=O	
[C+]C-C-O-C-C-N*C*C*C	[C+]C=C-C-N*C*C	[C+]N+C-C-N*C*N	
[C+]C-C-O-C	[C+]C-C=C	[C+]C-N*C*N-C-C+N	
[C+]C-C-O-C-C	[C+]C=C	[C+]O-C-C-N-C-O	
[C+]C-O-C	[C+]N-C-C+N	[C+]C-O-C-N-C-C-O	
[C+]C*C-C-N*C*N	[C+]C-N-C-C+N	[C+]C-C-O-C-N-C-C-O	
[C+]C*C*C-C-N*C*N	[C+]C-C-N-C-C+N	[C+]O-C-C-N-C-C-O	
[C+]C*C*C*C-C-N*C*N	[C+]C-C-C-N-C-C+N	[C+]C-O-C-C-N-C-C-O	
[C+]C-N*C*N-C-C*C	[C+]C-C+N	[C+]C-C-O-C-C-N-C-C-O	
[C+]C-N*C*N-C-C*C*C	[C+]C+N	[C+]N*C*N-C-C-C-C-C-C-C-C=O	
[C+]C-N*C*N-C-C*C*C*C	[C+]C*N-C-C+N	[C+]C-N*C*N-C-C-C-C-C-C-C-C=O	
[C+]C*C-C-N*C	[C+]C*C*N-C-C+N	[C+]C*N-C-C-C-C-C-C-C-C=O	
[C+]C*C*C-C-N*C	[C+]C*C*C*N-C-C+N	[C+]N-C-C-C-C-C-C-C-C=O	
[C+]C*C*C*C-C-N*C	[C+]C-C-O-C-C-N-C	[C+]C*C*N-C-C-C-C-C-C-C-C=O	
[C+]C*C-C-N*C*C	[C+]C-C-N-C-C-O-C-C	[C+]C-C-C-C-C-C-C-C-C=O	
[C+]C*C*C-C-N*C*C	[C+]N*C*N-C-C-C-O	[C+]C-C-C-C-C-C-C-C=O	
[C+]C*C*C*C-C-N*C*C	[C+]C-O-C-C-C-N*C*N	[C+]C-C-C-C-C-C-C=O	
[C+]C-C-C-N-C-C-C	[C+]C-N*C*N-C-C-C-O	[C+]C-C-C-C-C=O	
[C+]C-C-C-C-N-C-C-C	[C+]C-N*C*N-C-C-C-O-C	[C+]C-C-C-C=O	
[C+]N-C-O	[C+]P-C	[C+]O-C=O	
[C+]C-O-C-N	[C+]P-C-C	[C+]O-C-C-C-N-C-C-O	
[C+]C-C-O-C-N	[C+]P-C-C-C	[C+]C-O-C-C-C-N-C-C-O	
[C+]C-N-C-O	[C+]P-C-C-C-C	[C+]N*C*N-C-C-C-C=O	

[C+]C-N-C-O-C	[C+]P-C-C-C-C-C	[C+]C-N*C*N-C-C-C-C=O	
[C+]C-C-O-C-N-C	[C+]P-C-C-C-C-C-C	[C+]C*N-C-C-C-C=O	
[C+]C-C-N-C-O	[C+]P-C-C-C-C-C-C-C	[C+]N-C-C-C-C=O	
[C+]C-C-N-C-O-C	[C+]P-C-C-C-C-C-C-C-C	[C+]C*C*N-C-C-C-C=O	
[C+]C-C-N-C-O-C-C	[C+]P-C-C-C-C-C-C-C-C-C	[C+]N-C-C=O	
[C+]C*C-C-C-N*C*N	[C+]P-C-C-C-C-C-C-C-C-C-C	[C+]C-N-C-C=O	
[C+]C*C*C-C-C-N*C*N	[C+]P-C-C-C-C-C-C-C-C-C-C-C	[C+]C-C-N-C-C=O	
[C+]C*C*C*C-C-C-N*C*N	[C+]P-C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-O-C=O	
[C+]C-N*C*N-C-C-C*C	[C+]P-C-C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-C-O-C=O	
[C+]C-N*C*N-C-C-C*C*C	[C+]P-C-C-C-C-C-C-C-C-C-C-C-C-C-C	[C+]S-C-C-C-N	
[C+]C-N*C*N-C-C-C*C*C*C	[C+]C-P-C	[C+]N-C-C-C-S=O	
[C+]C*C-C-C-N*C	[C+]C-P-C-C	[C+]N-C-C-C-S-O	
[C+]C*C*C-C-C-N*C	[C+]C-P-C-C-C	[C+]S-C-C-C-N*C	
[C+]C*C*C*C-C-C-N*C	[C+]C-P-C-C-C-C	[C+]C*N-C-C-C-S=O	
[C+]C*C-C-C-N	[C+]C-P-C-C-C-C-C	[C+]C*N-C-C-C-S-O	
[C+]C*C*C-C-C-N	[C+]C-P-C-C-C-C-C-C	[C+]S-C-C-C-N*C*C	
[C+]C*C*C*C-C-C-N	[C+]C-P-C-C-C-C-C-C-C	[C+]C*C*N-C-C-C-S=O	
[C+]C*C-C-C-N*C*C	[C+]C-P-C-C-C-C-C-C-C-C	[C+]C*C*N-C-C-C-S-O	
[C+]C*C*C-C-C-N*C*C	[C+]C-P-C-C-C-C-C-C-C-C-C	[C+]S-C-C-C-N*C*C*C	
[C+]C*C*C*C-C-C-N*C*C	[C+]C-P-C-C-C-C-C-C-C-C-C-C	[C+]C*C*C*N-C-C-C-S=O	
[C+]C-C-C*C	[C+]C-P-C-C-C-C-C-C-C-C-C-C-C	[C+]C*C*C*N-C-C-C-S-O	
[C+]C-C-C*C*C	[C+]C-P-C-C-C-C-C-C-C-C-C-C-C-C	[C+]S-C-C-C	
[C+]C-C-C*C*C*C	[C+]C-P-C-C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-C-C-S=O	

[C+]C-N-C-C-O	[C+]C-C-P-C-C	[C+]C-C-C-S-O	
[C+]C-N-C-C-O-C	[C+]C-C-P-C-C-C	[C+]C-C-S=O	
[C+]C-C-N-C-C-O	[C+]C-C-P-C-C-C-C	[C+]C-C-S-O	
[C+]C-C-N-C-C-O-C	[C+]C-C-P-C-C-C-C-C	[C+]C-S=O	
[C+]C-C-C-N-C-C-O	[C+]C-C-P-C-C-C-C-C-C	[C+]C-S-O	
[C+]C-C-C-N-C-C-O-C	[C+]C-C-P-C-C-C-C-C-C-C	[C+]S=O	
[C+]C*N-C-O	[C+]C-C-P-C-C-C-C-C-C-C-C	[C+]S-O	
[C+]C-O-C-N*C	[C+]C-C-P-C-C-C-C-C-C-C-C-C	[C+]O=S=O	
[C+]C-C-O-C-N*C	[C+]C-C-P-C-C-C-C-C-C-C-C-C-C	[C+]O-S=O	
[C+]C*C*N-C-O	[C+]C-C-P-C-C-C-C-C-C-C-C-C-C-C	[C+]C-O-C-C-N-C-C-O-C	
[C+]C-O-C-N*C*C	[C+]C-C-P-C-C-C-C-C-C-C-C-C-C-C-C		
[C+]C-C-O-C-N*C*C	[C+]C-C-C-P-C-C-C		
[C+]C*C*C*N-C-O	[C+]C-C-C-P-C-C-C-C		
[C+]C-O-C-N*C*C*C	[C+]C-C-C-P-C-C-C-C-C		
[C+]C-C-O-C-N*C*C*C	[C+]C-C-C-P-C-C-C-C-C-C		
[C+]S-C	[C+]C-C-C-P-C-C-C-C-C-C-C		
[C+]S-C-C	[C+]C-C-C-P-C-C-C-C-C-C-C-C		

ANEXO 5: DESCRIPTORES MOLECULARES MÁS RELEVANTES Y MENOS REDUNDANTES ENTRE ELLOS

[C+]C-N*C*C	[C+]C-C-C-C-N*C*C*C
[C+]O-C=O	[C+]N-C-C-C-S=O
[A-]C-P-C-C	[C+]C-C-C-C-C-C-C=O
[A-]C-C=O	[C+]N*C*N-C-C-C-C=O
[C+]N-C-C-C-S-O	[C+]C-C-N-C-C+N
[C+]C-C-O-C-C-N-C-C-O	[C+]C-C-C-C-P-C-C-C-C-C
[C+]O-C-C-C-N-C-C-O	[C+]C*N-C-C-C-O
[C+]C-C-C-C-C-C-C-C-C-C-C-C	[C+]C-N*C*N-C-C-O
[A-]C-C*C*C*C-S=O	[C+]C-N-C-C=O
[C+]C-C+N	[C+]C*C-C-C-N*C*N
[C+]C*C-C-N	[C+]C-O-C-C-N-C-C-O
[A-]C-C-C=O	[A-]C-C-C-C-C-C-C-O
[C+]C-C-C-C-C=O	[C+]C*N-C-C-C-S=O
[C+]C*N-C-C-C-C=O	[C+]C-C-N*C*N-C-C
[C+]C-C-O-C=O	[A-]C=O
[C+]N-C-O	[A-]C-C-C-C-C-P-C-C-C-C-C
[C+]C-O-C-C-C-N-C-C-O	[C+]C-C-C-C-C-C-C-C-C-N
[C+]S-C-C	[C+]C*C*N-C-C-C-C=O
[C+]C-C-P-C-C	[C+]C-C-C-C-C-C-O
[A-]O-C-C-O	[A-]C-C-P-F
[C+]C-C-C-C=O	[C+]N-C-C=O

[A-]O-C=O	[C+]C-C-P-C-C-C
[A-]C-C-C-C-P-C-C-C-C	[C+]S-C-C-C-N*C
[A-][I-]	[C+]C-C-C-N-C-C-C
[C+]C-O-C=O	[A-]O-C-C=O

ANEXO 6: DESEMPEÑO DE LOS CLASIFICADORES DE BUEN DESEMPEÑO

Grupo	Clasificadores	Acc	TPrate	TNrate	Kappa
Trees	ExtraTree	0.95454545	0.98245614	0.77777778	0.79
	HoeffdingTree	0.93939394	0.94736842	0.88888889	0.76
	RandomTree	0.95454545	0.98245614	0.77777778	0.79
	LMT	0.95454545	0.98245614	0.77777778	0.79
Bayes	BayesianLogisticRegression	0.92424242	0.92982456	0.88888889	0.71
	BayesNet	0.92424242	0.94736842	0.77777778	0.69
	ComplementeNaiveBayes	0.92424242	0.92982456	0.88888889	0.71
	NaiveBayes	0.92424242	0.92982456	0.88888889	0.71
	NaiveBayesMultinominal	0.92424242	0.92982456	0.88888889	0.71
Lazy	IBK estableciendo k=5	0.92424242	0.94736842	0.77777778	0.69
	Kstar	0.95454545	0.98245614	0.77777778	0.79
Rules	Jrip	0.93939394	0.98245614	0.66666667	0.71
	Nnge	0.87878788	0.87719298	0.88888889	0.59
Functions	FDLA	0.92424242	0.92982456	0.88888889	0.71
	KernelLogisticRegression	0.90909091	0.92982456	0.77777778	0.64
	Logistic	0.90909091	0.92982456	0.77777778	0.64
	MultilayerPerceptron,	0.93939394	0.94736842	0.88888889	0.76
	MultilayerPerceptronCS	0.93939394	0.94736842	0.88888889	0.76
	SimpleLogistic,	0.92424242	0.94736842	0.77777778	0.69
	SMO	0.92424242	0.92982456	0.88888889	0.71

ANEXO 7: RESULTADOS DE LAS MEDIDAS DE DIVERSIDAD PARA LAS 140 COMBINACIONES SELECCIONADAS DE LAS 240 COMBINACIONES POSIBLES A CREAR

p	D	Q	DF	Subconjunto de Clasificadores
0.47	0.06	0.94	0.03	JRIP—MULTILAYERPERCEPTRON
0.49	0.06	0.95	0.03	EXTRATREE—SMO
0.49	0.06	0.95	0.03	FDLA—KSTAR
0.49	0.06	0.95	0.03	EXTRATREE—FDLA
0.49	0.06	0.95	0.03	EXTRATREE—NAIVEBAYES
0.41	0.07	0.91	0.03	JRIP—SMO
0.41	0.07	0.91	0.03	FDLA—JRIP
0.41	0.07	0.91	0.03	JRIP—NAIVEBAYES
0.49	0.06	0.95	0.03	KSTAR—NAIVEBAYES
0.49	0.06	0.95	0.03	KSTAR—SMO
0.56	0.04	0.97	0.03	EXTRATREE—MULTILAYERPERCEPTRON
0.56	0.04	0.97	0.03	KSTAR—MULTILAYERPERCEPTRON
0.58	0.06	0.97	0.04	JRIP—LOGISTIC
0.59	0.05	0.95	0.03	EXTRATREE--JRIP—SMO
0.59	0.05	0.95	0.03	EXTRATREE--JRIP--NAIVEBAYES
0.59	0.05	0.95	0.03	EXTRATREE--FDLA--JRIP
0.59	0.05	0.95	0.03	FDLA--JRIP--KSTAR
0.53	0.06	0.94	0.04	JRIP--NAIVEBAYES--SMO
0.53	0.06	0.94	0.04	FDLA--JRIP--NAIVEBAYES
0.59	0.05	0.95	0.03	JRIP--KSTAR--SMO
0.57	0.06	0.95	0.04	JRIP--LOGISTIC--NAIVEBAYES

0.59	0.05	0.95	0.03	JRIP--KSTAR--NAIVEBAYE
0.59	0.05	0.97	0.04	EXTRATREE--NAIVEBAYES--SM
0.59	0.05	0.95	0.04	FDLA--JRIP--MULTILAYERPERCEPTRON
0.61	0.05	0.94	0.04	FDLA--JRIP--SMO
0.70	0.03	0.98	0.03	EXTRATREE--KSTAR--MULTILAYERPERCEPTRON
0.63	0.04	0.97	0.03	EXTRATREE--JRIP--MULTILAYERPERCEPTRON
0.63	0.04	0.97	0.03	JRIP--KSTAR--MULTILAYERPERCEPTRON
0.66	0.04	0.97	0.03	EXTRATREE--KSTAR--NAIVEBAYES
0.66	0.04	0.97	0.03	EXTRATREE--FDLA--KSTAR
0.66	0.04	0.97	0.03	EXTRATREE--KSTAR--SMO
0.59	0.05	0.97	0.04	EXTRATREE--FDLA--NAIVEBAYES
0.59	0.05	0.95	0.04	JRIP--MULTILAYERPERCEPTRON--SMO
0.59	0.05	0.95	0.04	JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES
0.63	0.05	0.98	0.04	EXTRATREE--LOGISTIC--NAIVEBAYES
0.57	0.05	0.95	0.04	EXTRATREE--FDLA--JRIP--NAIVEBAYES
0.57	0.05	0.95	0.04	FDLA--JRIP--KSTAR--NAIVEBAYES
0.61	0.05	0.95	0.04	EXTRATREE--FDLA--JRIP--SMO
0.57	0.05	0.95	0.04	EXTRATREE--JRIP--NAIVEBAYES--SMO
0.57	0.05	0.95	0.04	JRIP--KSTAR--NAIVEBAYES--SMO
0.61	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--MULTILAYERPERCEPTRON
0.61	0.05	0.95	0.04	FDLA--JRIP--KSTAR--SMO
0.61	0.04	0.96	0.04	EXTRATREE--JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES
0.61	0.04	0.96	0.04	EXTRATREE--JRIP--MULTILAYERPERCEPTRON--SMO

0.61	0.04	0.96	0.04	FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON
0.63	0.05	0.95	0.05	FDLA--JRIP--NAIVEBAYES--SMO
0.72	0.03	0.98	0.04	EXTRATREE--JRIP--KSTAR--MULTILAYERPERCEPTRON
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--KSTAR--MULTILAYERPERCEPTRON
0.68	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR
0.68	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--NAIVEBAYES
0.68	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--SMO
0.68	0.04	0.96	0.05	FDLA--JRIP--MULTILAYERPERCEPTRON--SMO
0.61	0.04	0.96	0.04	JRIP--KSTAR--MULTILAYERPERCEPTRON--SMO
0.62	0.04	0.97	0.04	EXTRATREE--FDLA--KSTAR--NAIVEBAYES
0.64	0.04	0.96	0.04	FDLA--JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES
0.64	0.04	0.96	0.04	JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.62	0.05	0.97	0.04	EXTRATREE--JRIP--LOGISTIC--NAIVEBAYES
0.62	0.05	0.97	0.04	JRIP--KSTAR--LOGISTIC--NAIVEBAYES
0.63	0.05	0.96	0.05	FDLA--JRIP--LOGISTIC--NAIVEBAYES
0.63	0.05	0.96	0.05	JRIP--LOGISTIC--NAIVEBAYES--SM
0.62	0.04	0.96	0.04	EXTRATREE--JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO-
0.62	0.04	0.96	0.04	JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.63	0.04	0.96	0.04	EXTRATREE--JRIP--KSTAR--NAIVEBAYES--SMO
0.62	0.04	0.96	0.04	FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES
0.62	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES
0.63	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--KSTAR--NAIVEBAYES
0.61	0.05	0.96	0.04	FDLA--JRIP--KSTAR--NAIVEBAYES--SMO

0.61	0.05	0.96	0.04	EXTRATREE--FDLA--JRIP--NAIVEBAYES--SMO
0.65	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--KSTAR--SMO
0.65	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--MULTILAYERPERCEPTRON--SMO
0.63	0.05	0.97	0.04	EXTRATREE--FDLA--JRIP--LOGISTIC--NAIVEBAYES
0.63	0.05	0.97	0.04	FDLA--JRIP--KSTAR--LOGISTIC--NAIVEBAYES
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON
0.66	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES
0.66	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--MULTILAYERPERCEPTRON--SMO
0.66	0.04	0.97	0.04	EXTRATREE--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.63	0.05	0.97	0.04	JRIP--KSTAR--LOGISTIC--NAIVEBAYES--SMO
0.63	0.05	0.97	0.04	EXTRATREE--JRIP--LOGISTIC--NAIVEBAYES--SMO
0.64	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--LOGISTIC--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--NAIVEBAYES
0.63	0.04	0.96	0.04	EXTRATREE--FDLA--JRIP--KSTAR--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON--SMO
0.64	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES
0.66	0.04	0.97	0.05	EXTRATREE--FDLA--JRIP--LOGISTIC--NAIVEBAYES--SMO
0.66	0.04	0.97	0.05	FDLA--JRIP--KSTAR--LOGISTIC--NAIVEBAYES--SMO
0.65	0.04	0.97	0.04	FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.65	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.68	0.04	0.98	0.04	EXTRATREE--FDLA--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.68	0.04	0.98	0.04	EXTRATREE--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--SMO

0.68	0.04	0.98	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON
0.67	0.04	0.98	0.04	EXTRATREE--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES
0.66	0.04	0.97	0.04	EXTRATREE--JRIP--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES
0.65	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.67	0.04	0.97	0.04	EXTRATREE--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.66	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--NAIVEBAYES--SMO
0.67	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES
0.69	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--SMO
0.69	0.04	0.97	0.05	EXTRATREE--FDLA--JRIP--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.69	0.04	0.97	0.05	FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO
0.68	0.04	0.97	0.04	EXTRATREE--FDLA--JRIP--KSTAR--LOGISTIC--MULTILAYERPERCEPTRON--NAIVEBAYES--SMO

ANEXO 8: CARACTERÍSTICAS DEL CONJUNTO DE DATOS DE ENTRENAMIENTO

Base de datos	Cantidad rasgos	Cantidad de clases	Distribución por clases	Total de casos
AChE_inhibitor	693	2	169-26	195

ANEXO 9: CARACTERÍSTICAS DEL CONJUNTO DE DATOS DE PRUEBA

Base de datos	Cantidad rasgos	Cantidad de clases	Distribución por clases	Total de casos
AChE_inhibitor	693	2	57-9	66