Universidad de las Ciencias Informáticas

Facultad # 6: Bioinformática

Título: Módulo de la predicción de la relación Estructura-Actividad de compuestos orgánicos partiendo de Fragmentos ponderados por la refractividad atómica, utilizando Máquinas de Soporte Vectorial para la Plataforma GRATO.

.

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autor(es): Yaikiel Hernández Díaz
Andy Machín González

Tutor(es): Dr. Ramón Carrasco Velar
MSc. Aurelio Antelo Collado

Co-tutor: Lic. Maybel Anido Bada.

Consultante: Ing. Lourdes Escalona Peral.

Asesor: Dra. Isneri Talavera Bustamante
Ing. Noslen Hernández González

Julio ,2007.

¨Año 49 de la Revolución¨

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Yaikiel Hernández Díaz

_____

Firma del Autor

MSc. Aurelio Antelo Collado

_____

Firma del Tutor

Andy Machín González

_____

Firma del Autor

Dr. Ramón Carrasco Velar

_____

Firma del Tutor

El misterio es la cosa más bonita que podemos experimentar.
Es la fuente de todo arte y ciencia verdaderos.

Albert Einstein

AGRADECIMIENTOS

Agradecemos a nuestros padres por el apoyo insustituible y mantenernos motivados para terminar el camino.

A nuestros tutores Ramón Carrasco y Aurelio Antelo Collado por su paciencia y ayuda.

A nuestra co-tutora Maybel por haber sido tan incondicional.

A nuestra amiga Lourdes Escalona por todo el tiempo que nos dedicó y por habernos sacado del ¨bache¨.

A Isneri y a Noslen por caernos del cielo cuando pensábamos que todo estaba perdido.

A nuestro Comandante en Jefe Fidel y a la Revolución Cubana por enseñarnos que un mundo mejor es posible.

A nuestros amigos por todo su cariño.

DEDICATORIA

De Yaikiel.

Siempre los sueños se hacen realidad cuando el amor los apoya y la vida les permite tener a personas muy especiales junto a él. Por eso quiero dedicar esta tesis a las personas más especiales que tengo:

Primero que todo a la persona que es mi alma gemela y que llevo conmigo su foto para darle un beso todos los días, la amo eternamente y mi vida no tiene sentido sin ella, mi mami Titica.

A quien me enseñó todo para ser un hombre de bien, alguien muy especial que tengo el honor de tenerlo como padre y que amo con todo mi corazón, a Nelson.

A mis segundos grandes amores, las personas más dulces y amorosas que tengo, mis abuelos Pino y Moisés.

A mi otro yo, a quien mecí por horas, corrimos juntos y tenemos un amor muy especial, mi hermanita Lili.

A mis tíos Coco, Deisy y Moraima por ser tan especiales siempre que los necesité y por su amor tan incondicional.

A todos mis amigos que siempre llevaré conmigo y en especial a mi hermano y compañero de tesis Andy.

A mi pequeño pony, mi bebé y una de las personas que más quiero, quien a compartido mis desvelos y los mejores momentos, mi novia Elsy.

A mi gente de barrio, mi otra familia que no deja de ser especial y que siempre tendrán mi respeto, amor y cariño cuando lo necesiten, Juan y María.

A todos gracias por existir.

De Andy.

Dedico esta tesis, ya que marca el fin y el comienzo de etapas trascendentales en mi vida, primeramente, a las dos mujeres que más yo quiero, mi madre querida y a mi abuela, que como no tengo palabras de agradecimiento y de cariño que pueda expresar como hombre, adolescente, niño… solamente quiero decirle “Gracias, por haber confiado en mi”.

A mi brother Amed, por estar siempre tan preocupado por su hermano mayor!..Te quiero man...al igual que a Alien para que lo motive a prepararse, a estudiar y a graduarse !!!...

A mi padre Carlos Alberto Machin y a mi abuelo que siempre han estado ahí, dando el ejemplo, gracias a los dos. A mi abuela Umbelina y a mi tía Bárbara por ser tan atentas conmigo.

También dedico todo este trabajo a mi compañero de tesis Yaikiel, por ser tan persistente, y por haber confiado en mí como un hermano menor, jajaja...te quiero man ...y no llores !...jaja.

A todos mis amigos y compañeros que han compartido junto conmigo estos 5 años que han hecho de verdad mi vida más interesante... para todos ellos muchas gracias...

RESUMEN

El presente trabajo está enmarcado dentro del proyecto de investigación conjunta Plataforma para la Predicción de Actividad Biológica de Compuestos Orgánicos entre el Centro de Química Farmacéutica y la Facultad 6 de la Universidad de las Ciencias Informáticas. Es un proyecto concebido modularmente, uno de cuyos módulos incluye técnicas de inteligencia artificial para la predicción de la actividad. En el presente trabajo se muestran los resultados del análisis, diseño, implementación y validación de los modelos de las máquinas de soporte vectorial para el desarrollo de modelos de regresión y clasificación. Como descriptores de las moléculas se emplean los fragmentos moleculares ponderados por el Índice del Estado Refractotopológico Total. Con el modelo de clasificación se obtiene un 73% de acierto en la predicción y con el modelo de regresión se determina un coeficiente de correlación de 0.73. Se determinó que el error relativo medio es 69%. Para la creación de los modelos se emplearon variaciones en la variable gamma, así como sus respectivos parámetros y el kernel FBR, realizando comparaciones entre los modelos para determinar el de mejor calidad en la predicción.

Palabras Claves:
Fragmentos.
Índice del Estado Refractotopológico Total.
Actividad Biológica.
Funciones *Kernels*.

Índice.

INTRODUCCIÓN

El presente trabajo está enmarcado dentro del proyecto de investigación conjunta entre el Centro de Química Farmacéutica y la Facultad 6 de la Universidad de las Ciencias Informáticas denominado una Plataforma para la Predicción de Actividad Biológica de Compuestos Orgánicos.

La predicción de la actividad biológica de compuestos químicos es hoy día un objetivo principal dentro de la Industria Médico Farmacéutica Mundial. El alto costo del proceso de investigación - desarrollo de nuevos fármacos, ha obligado a este sector económico a adoptar la estrategia del uso de técnicas de la computación y la informática para acelerar el proceso y disminuir los costos. Esta necesidad, junto con los avances en el sector de la bioorgánica, los extraordinarios progresos de la fisiología, la bioquímica, la medicina y las técnicas de computación han promovido una revolución en el ámbito del diseño y producción de fármacos. En los últimos años, la industria farmacéutica ha reorientado sus investigaciones y prestado más atención a aquellos métodos que permitan una selección racional o el diseño de nuevos compuestos con propiedades deseadas[1]. En la literatura se han reportado varios enfoques para el diseño molecular asistido por computadora, muchos de los cuales están basados en la correlación entre estructura química y diferentes propiedades de las moléculas. La efectividad de estos métodos depende en gran medida de la forma de describir la estructura química, así como de la técnica de procesamiento de los datos. Un enfoque ampliamente utilizado por los químicos consiste en dividir la molécula en partes de ella, para poder determinar las regiones responsables o no de determinada respuesta.

A partir del estudio realizado en las estadísticas mundiales en el año 2005 murieron de cáncer 7,6 millones de personas con tendencia a aumentar a 9 millones en los venideros años, y es a comienzos de este siglo la principal causa de muerte y una de las enfermedades con más estudios realizados en este treintenio. Cuba es uno de los países con mayor cantidad de programas para la atención y prevención de esta enfermedad, con una mortalidad de 102 personas por cada 100 000 habitantes. El cáncer tiene su mayor incidencia en el de pulmón y el de mamas[2]. Esto hace muy necesaria la producción de nuevos y más eficientes medicamentos para su tratamiento. Para el desarrollo de los mismos juega un papel fundamental el diseño racional de fármacos, pues la búsqueda de nuevas y mejores entidades biológicamente activas candidatas a medicamentos es un proceso complejo que suele llevar de 10 a 12 años de investigación y desarrollo con un gasto medio de 800 millones de €, tras lo cual tan solo un 5-10% de las moléculas que llegan a fase de ensayo clínico terminan siendo comercializadas[3]. Esta situación

crítica conduce al surgimiento del proyecto mencionado, herramienta que en su primera versión permitirá el tamizaje virtual de moléculas orgánicas potencialmente activas contra el cáncer. En la Plataforma se aplicará, como forma de descripción de la estructura química, la ponderación de los vértices del grafo químico por un descriptor atómico recientemente desarrollado en Cuba, el Índice del Estado Refractotopológico. Al aplicarlo a fragmentos moleculares se realiza la suma de cada uno de los átomos pesados de dicho fragmento y se le denominará como Índice del Estado Refractotopológico Total de ese fragmento. Al estar asociados esos fragmentos con su valor de índice a una actividad biológica medida, será posible establecer correlaciones cuantitativas o cualitativas, según sea el caso entre la estructura química y la actividad biológica.

Es entonces una necesidad de esta plataforma conocer cuáles son los resultados de predicción de la actividad biológica anticancerígena y para el procesamiento de los datos se empleará una técnica de Inteligencia Artificial (IA) a partir de la descripción de la molécula por diferentes vías: fragmentos o descriptores. Lo anteriormente planteado nos conduce a la formulación del siguiente problema: ¿Cómo predecir actividad biológica anticancerígena en la Plataforma?

Existen diferentes técnicas de IA que han sido aplicadas con éxito en la modelación de moléculas con posible actividad biológica. Las más importantes son las redes neuronales, lógica difusa, algoritmos genéticos y sus diferentes variantes. Por lo que se define como **objeto de estudio**:
La inteligencia artificial aplicada a la predicción de actividad biológica.

Una técnica recientemente introducida son las Máquinas de Soporte Vectorial (MSV). Las MSV son un nuevo sistema de aprendizaje el cual ha tenido un gran desarrollo en los últimos años tanto en la generación de nuevos algoritmos como en las estrategias para su implementación. MSV es un sistema de aprendizaje basado en el uso de un espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización el cual utiliza elementos de la teoría de generalización. Además MSV es un sistema para entrenar máquinas de aprendizaje lineal eficientemente, y tanto que para la clasificación como para la regresión se han encontrado muchas aplicaciones como clasificación de imágenes, reconocimiento de caracteres, detección de proteínas, clasificación de patrones, identificación de funciones, etc. Las MSV están basadas en el principio de minimización del riesgo estructural, principio originado de la teoría de

aprendizaje estadístico desarrollada por Vapnik, el cual ha demostrado ser superior al principio de minimización del riesgo empírico, utilizado por las redes neuronales convencionales. Algunas de las razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo solo depende de los datos con más información llamados vectores de soporte. Las ventajas que este método posee son:

1) Una excelente capacidad de generalización debido a la minimización del riesgo estructurado.
2) Existen pocos parámetros a ajustar; el modelo solo depende de los datos con mayor información.
3) La estimación de los parámetros se realiza a través de la optimización de una función de costo convexa, lo cual evita la existencia de un mínimo local.
4) La solución de MSV es escasa (*sparse*), esto es, que la mayoría de las variables son cero en la solución de MSV, esto quiere decir que el modelo final puede ser escrito como una combinación de un número muy pequeño de vectores de entrada, llamados vectores de soporte.[4]

Por lo que se formula como **Campo de Acción**:
Las MSV aplicada al estudio de la relación estructura-actividad anticancerígena de compuestos orgánicos utilizando fragmentos ponderados por el Índice del Estado Refractotopológico Total.

Teniendo en cuenta lo planteado en los párrafos precedentes se puede formular, como **hipótesis** del trabajo que: *Si se desarrolla un módulo basado en MSV, a partir de fragmentos estructurales ponderados por el Índice del Estado Refractotopológico Total, entonces la Plataforma podrá predecir actividad anticancerígena de compuestos orgánicos.*

Para poder cumplir con la hipótesis de trabajo se trazó como **Objetivo general:** Desarrollar un módulo para la Plataforma capaz de predecir actividad anticancerígena de compuestos orgánicos a partir de fragmentos ponderados por el Índice del Estado Refractotopológico Total, utilizando MSV como técnica de inteligencia artificial.

Para dar cumplimiento al objetivo general se definieron como **objetivos específicos**:

- Analizar un módulo basado en MSV que permita predecir estructura-actividad.
- Diseñar el módulo analizado.

- Implementar el módulo diseñado.
- Validar el modelo desarrollado.

Se definieron las siguientes **tareas**:

- Revisión del estado del arte acerca de sistemas de predicción existentes en el mundo.
- Determinación de la función kernel.
- Determinación del modelo de dominio.
- Determinación de los requisitos funcionales y no funcionales.
- Realización del Diagrama de Casos de Uso del Sistema.
- Realización de la descripción de dichos casos de uso.
- Realización de los Diagramas de Interacción.
- Realización de los Diagramas de Clases del Diseño.
- Realización de los Diagramas de Componentes.
- Realización de las pruebas de validación del modelo.

Como **aporte práctico** se espera:
Brindar una herramienta computacional capaz de predecir actividad anticancerígena de compuestos orgánicos a partir de fragmentos ponderados por el índice del estado refractotopológico total.
El presente trabajo está compuesto por Resumen, Introducción, 4 capítulos que dan cuerpo a la tesis, Conclusiones, Recomendaciones Bibliografía y Referencias bibliográficas. Los capítulos principales están estructurados según se describe:

Capitulo 1. Fundamentación Teórica. Se brinda una panorámica de la historia de la farmacología, las relaciones estructura actividad cuantitativa (QSAR) aplicadas al diseño racional de fármacos. Se brinda un estado del arte de las redes neuronales artificiales y las máquinas de soporte vectorial como forma de aprendizaje supervisado tanto para la clasificación como para la regresión, además de las últimas investigaciones realizadas en esta rama aplicando esta técnica de inteligencia artificial. Se explican las tecnologías empleadas para el desarrollo de ésta herramienta.

Capítulo 2. Requisitos. En este capítulo se comienza con la fase de inicio del proceso unificado de desarrollo (RUP) en el flujo de trabajo Modelamiento del negocio, donde se muestra el modelo de dominio y se definen los requisitos funcionales y no funcionales del sistema y se construyen los artefactos correspondientes a este flujo de trabajo.

Capitulo 3. Descripción de la Solución Propuesta. Se muestran las clases, tanto del análisis como del diseño, además de los diagramas de interacción pertenecientes a este capítulo. Se muestra además la arquitectura empleada para el desarrollo, así como una descripción de todas las clases que intervienen en este.

Capitulo 4. Implementación del Módulo y Validación del Modelo. En este capítulo, a partir de una muestra de datos reales se desarrollan modelos de la actividad y se realiza la predicción, tanto cuantitativa como cualitativa. Los datos están basados en la descripción de las moléculas a partir de fragmentos moleculares ponderados por el Índice del Estado Refractotopológico Total. Se analizan muestras de diferente composición y se comprueba la calidad en la predicción. Los resultados dependen de la naturaleza de la muestra.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA.

La farmacología es la ciencia que estudia el origen, las acciones y las propiedades de las sustancias químicas en los organismos vivos. En un sentido más estricto se considera a la farmacología como el estudio de los fármacos, es decir, aquellas sustancias utilizadas en el tratamiento, cura, prevención o diagnóstico de una enfermedad. Su historia se limita al corto período de ciento cincuenta años que se inicia a mediados del siglo XIX. Uno de los propósitos más ambiciosos de la química moderna es encontrar la relación entre la estructura molecular y la función biológica que cumplen. Estas investigaciones comenzaron a ser realizadas por los químicos a partir de la década del veinte, para lograr mayores pasos en la producción de fármacos.[1] El mundo de los medicamentos es extraordinariamente complejo y multidisciplinario en la actualidad. Se ha avanzado mucho en el entendimiento de las interacciones entre las moléculas, en gran parte debido a que la automatización y robotización de las técnicas de rayos X y la resonancia magnética nuclear están acelerando sustancialmente el descubrimiento de la estructura tridimensional de muchas proteínas. Las relaciones estructura actividad cuantitativas (QSAR Quantitative Structure Activity Relationships) fueron introducidos por Hansch y Fujita en la década de 1960, las cuales requieren disponer de datos biológicos de calidad, definir descriptores químicos relevantes y elegir un modelo adecuado para establecer dichas relaciones.

El diseño racional de fármacos depende en gran medida del empleo de las técnicas QSAR. A partir del desarrollo computacional existente se han creado vínculos entre el análisis multivariado y la inteligencia artificial aplicados al desarrollo de modelos empleando descriptores atómicos y moleculares. Es tal la variedad de ciencias y tecnologías utilizadas para el estudio de la interacción entre la química y la vida que Corwin Hansch, la ha calificado como la Ciencia sin Nombre.

*1.1 Estructuras Moleculares***.**

Dos o más elementos combinados en una sustancia forman lo que se conoce como un compuesto químico, el cual consiste en átomos unidos mediante enlaces químicos formando moléculas. Se puede formar un enorme número de compuestos químicos por la combinación de unos 120 elementos; hasta la fecha, alrededor de treinta millones han sido caracterizados e identificados. Los compuestos basados en los átomos de carbono e hidrógeno se denominan compuestos orgánicos, y aquellos basados en otros elementos se denominan compuestos inorgánicos.

Cada molécula puede tener asociada una actividad biológica determinada, lo cual da la magnitud de la importancia de poder predecir la actividad biológica de cualquier compuesto químico. Dicha predicción puede realizarse utilizando descriptores estructurales asociados a propiedades químico-físicas de las moléculas. Estos están basados en cálculos de química teórica o por métodos de grafo-teóricos, entre otros. También pueden ser utilizados los fragmentos de las moléculas, como forma para determinar teóricamente, que parte ejerce la actividad biológica. Dentro de este proyecto se aplicarán las diferentes técnicas para compuestos orgánicos anticancerígenos, aunque los principios y procedimientos aplicados se pueden emplear en cualquier otra actividad de la que se posea información estructural y biológica.

Las relaciones cuantitativas estructura-actividad se expresan mediante modelos matemáticos que permiten la predicción de propiedades fisicoquímicas y/o biológicas a partir de parámetros estructurales, electrónicos, topológicos, etc. El Índice del Estado Refractotopológico, el cual se obtiene de la Teoría de Grafo Químico y de la partición de la refractividad molecular definida por Ghose y Crippen, se basa en la influencia de las fuerzas dispersivas de cada átomo en los otros átomos en la molécula, modificados por la topología molecular.

*1.2 Redes Neuronales***.**

Sin remontarnos a épocas históricas, donde los intentos resultaban prematuros en relación a la tecnología disponible, podemos considerar que el camino hacia la construcción de máquinas inteligentes comienza en la Segunda Guerra Mundial, con el diseño de ordenadores analógicos ideados para controlar cañones antiaéreos o para la navegación. A partir de 1937 comienza el desarrollo de las primeras computadoras como la Máquina de Turing hasta llegar a 1957 donde A. Newell, H. Simon y J. Shaw presentaron el primer programa capaz de razonar sobre temas arbitrarios. Hacia 1960 John McCarthy, acuña el término de inteligencia artificial, para definir los métodos algorítmicos capaces de simular el pensamiento humano en los ordenadores. Entre los métodos teóricos más utilizados están las Redes Neuronales Artificiales (RNA). Las RNA se han integrado dentro de los métodos ya clásicos del análisis de las relaciones cuantitativas entre la estructura y la actividad biológica u otras propiedades. Las mismas son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el cerebro del hombre. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida empleando métodos matemáticos, teniendo en cada neurona que la compone, funciones de propagación, activación y transferencia. Las RNA o simplemente redes neuronales,

constituyen una de las áreas de la inteligencia artificial que ha despertado mayor interés en los últimos años. La razón principal es que las redes neuronales potencialmente son capaces de resolver problemas cuya solución por otros métodos convencionales resulta extremadamente difícil dada su capacidad de aprender. Estos modelos de aprendizaje se clasifican en: Híbridos, Supervisados, No Supervisados y Reforzados. Dentro de los Supervisados Unidireccionales se encuentran las técnicas: Perceptrón, Perceptrón Multicapa, Adalina, Máquinas de Soporte Vectorial (MSV) entre otras.[5]

En el presente trabajo se utiliza el sistema de aprendizaje de Máquinas de Soporte Vectorial y Kernels, como herramienta de clasificación de patrones para el problema de predicción estructura actividad de fragmentos ponderados por el Índice del Estado Refractotopológico Total (Rstate (t)) aplicado a un banco de datos de moléculas anticancerígenas.

*1.3 Máquinas de Soporte Vectorial***.**

El algoritmo Vector de Soporte (VS) es una generalización no-lineal del algoritmo Semblanza Generalizada, desarrollado en la Rusia en los años sesenta. Como tal, este está firmemente enlazado con la Teoría del Aprendizaje Estadístico, el cual se desarrolló a finales de las últimas tres décadas por Vapnik [1982, 1995] y Chervonenkis [1974]. Por otra parte, esta Teoría de Aprendizaje Estadístico caracteriza las propiedades de aprendizaje, habilitándole el poder de generalización de datos desconocidos.

El desarrollo de los VS trae consigo el surgimiento de las Máquinas de Soporte Vectorial (MSV). Estas son sistemas de aprendizaje que usan un espacio de hipótesis de funciones lineales en un espacio de rasgos de mayor dimensión, entrenadas por un algoritmo proveniente de la teoría de optimización. Este algoritmo nos da las propiedades que debe cumplir la solución del problema de optimización las cuales nos llevan a su vez a la implementación de algoritmos de aprendizaje derivados de la Teoría del Aprendizaje Estadístico.

A partir de la Teoría de la Generalización se considera un conjunto de entrenamiento dado $(Xi,Yi)=\{(x1,y1),(x2,y2),\ldots(xN,yN)\}$,con la entrada $xi \epsilon Rn$ y la salida $yi \epsilon R$. El problema consiste en encontrar una función f (x) que aproxime de manera óptima a una salida deseada y=f(x) y que dicha función sea capaz de clasificar correctamente los elementos de las muestras. Esta función se puede encontrar optimizando una medida de desempeño del modelo entrenado y esta medida es el Riesgo

Esperado. Evaluar esta función de riesgo es un proceso complicado, pues usualmente no se conoce la función de distribución, de ahí la necesidad de emplear el riesgo sobre el conjunto de entrenamiento el cual se conoce como Riesgo Empírico. La minimización del riesgo empírico y la dimensión de Vapnik-Chervonenkis son fundamentales en las MSV.

*1.3.1 Minimización del Riesgo Empírico (MRE)***.**

Desde un punto de vista más simple, la MRE consiste en encontrar la función f(x) que minimice el riesgo promedio del conjunto de entrenamiento. En la medida en que tengamos mayor cantidad de vectores, se puede asegurar que el riesgo empírico converge asintóticamente al riesgo esperado cuando el número de valores tiende a infinito (n->∞).

*1.3.2 Dimensión de Vapnik-Chevonenkis***.**

La Dimensión de Vapnik Chevonenkis es el número máximo h de vectores z1...zn que se pueden separar de todas las maneras posibles 2^h por los hiperplanos, a partir de funciones que miden el mayor número de muestras que pueden ser explicadas por el sistema. Donde para un conjunto de datos R, la familia de hiperplanos que se genera está dada por el término R^D en donde D es al menos D+1 para valores de D mayor que cero. Teniendo como propiedades que:

i) La dimensión de VC h_k de cada conjunto S_k de funciones es finita. Por lo tanto, $h_1 \leq h_2 \ldots \leq h_n \ldots$.

ii) Cualquier elemento S_k de la estructura contiene un conjunto de funciones.

La relación existente entre los conceptos anteriormente explicados viene dada por:

$$R(f) \leq R_{emp}(f) + \underbrace{\sqrt{\frac{h(\ln(2N/h)+1) - \ln(\eta/4)}{N}}}_{\text{VC confidence}}$$

Donde:

N = Número de muestras entrenadas.

R (f)= Riesgo Esperado.

Remp (f)= Riesgo Empírico.

h= Dimensión de Vapnik-Chevonenkis.

Si se realiza un análisis de la misma se puede asegurar que: el modelo más adecuado se encuentra cuando se logra un balance entre el Riesgo Empírico (Remp(f)) y la dimensión de VC, este problema se conoce como la minimización del riesgo estructural. Además, a medida que la relación N/h se hace mayor, la confianza VC se hace menor y el Riesgo Esperado se acerca al Riesgo Empírico.

Una MSV mapea los puntos de entrada en un espacio de características de una dimensión mayor y encuentra el hiperplano óptimo que los separe y maximice el margen **m** entre las clases, en este espacio .**Figura1**.

.

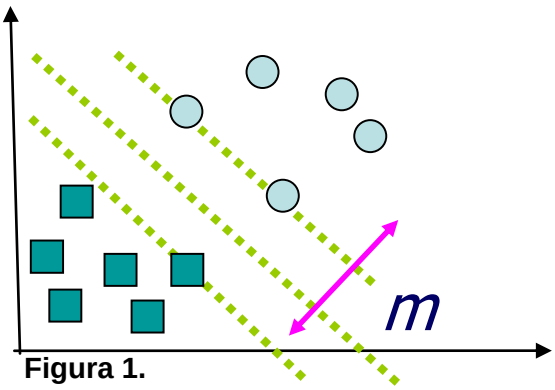


Figura 1.

A partir del conjunto de entrenamiento, donde el dominio de las imágenes se encuentran en el intervalo [-1; 1], las ecuaciones de los hiperplanos H1 y H2 viene dada por:
H1= wxi+b=1, H2= wxi+b=-1 y el hiperplano óptimo por wxi+b=0, partiendo de que ||w|| es la norma del vector normal al hiperplano para las clases que se representan. Por lo tanto el margen m se calcula por la expresión m=2/||w||. **Ver Figura 2**.[6]

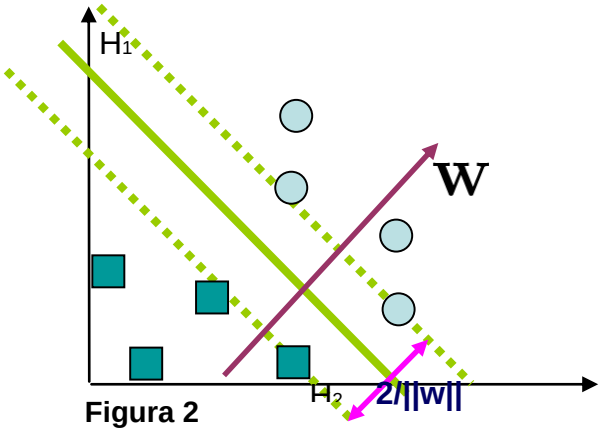


Figura 2

Maximizar dicho margen es un problema de programación cuadrática. Uno de los métodos para resolver el problema es a través de la teoría de Lagrange el cual fue extendido al problema de optimización con restricciones. Los conceptos principales de esta teoría son el concepto de multiplicadores de Lagrange (α) y la función del Lagrangiano (Lp (w, b, α)) relacionadas en la ecuación:

$$L_P(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(w^T x_i + b) - 1]$$

Dicha teoría es una generalización del resultado de Fermat en 1629, el cual plantea el problema de optimización sin restricciones.

En 1951 Kuhn y Tucker generalizaron la teoría de Lagrange permitiéndose entonces la introducción de restricciones de desigualdad en el problema de optimización. Solo los puntos que estén situados en unos de los hiperplanos cumple que α>0 y se les denomina vectores de soporte, que son los que ayudan a definir el hiperplano óptimo, según la expresión:

$$\frac{\partial J(w,b,\alpha)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

En el que se utiliza el producto del punto, con funciones en el espacio de características que son llamadas Kernels. La minimización del riesgo empírico y la dimensión de Vapnik-Chervonenkis son fundamentales para lograr la búsqueda de la solución óptima mediante funciones que en un espacio numérico determinado, logre el hiperplano separador. En la representación de estos conjuntos de entrenamiento, los mismos pueden estar linealmente separados. De no ser así, se utilizan las funciones Kernels para llevar estas muestras a un plano de mayor dimensión donde puedan ser linealmente separables. Dentro de los más utilizados se encuentran: el Kernel Polinomial y el de Base Radial. Las SVM pueden trabajar con dos o más clases en dependencia del problema al que se apliquen.[7]

*1.4 Máquinas de Soporte Vectorial para Regresión***.**

Las MSV se desarrollaron inicialmente para solucionar problemas de clasificación, pero se han ampliado para problemas de regresión. Los resultados finales a los que se puede arribar luego del empleo de las SVM pueden ser cualitativos o cuantitativos, para en el análisis cuantitativo se emplean MSV para regresión. Dicho método es una extensión del anteriormente explicado. Se conoce que la regresión crea una regla para predecir rasgos numéricos desconocidos desde rasgos numéricos conocidos y entre los cuales la relación que existe es la naturaleza del vector, por lo que esta es una forma de aprendizaje

supervisado. Los métodos para estimar dependencias funcionales basadas en los datos se introdujeron por los matemáticos Gauss (1777-1855) y Laplace (1749-1827) los cuales basaron su teoría en el cálculo de los mínimos cuadrados y el módulo de los mismos. Es necesario decir que aunque estos procedimientos matemáticos, que la fundamentan son antiguos, la Regresión Soportada por Vectores (RSV) fue creada en el año 1996.

El empleo de estimadores robustos para determinar los valores que tienen ruido dentro de la predicción se estiman a través de funciones de pérdida, donde los primeros pasos en este sentido se dieron por Tuckey quien demostró que en situaciones reales, se desconoce el modelo del ruido y dista de las distribuciones supuestas. A raíz de esto, Huber crea el concepto de estimadores robustos los cuales están determinados por funciones de pérdida. En la actualidad, las más utilizadas son: las funciones de pérdida cuadrática y lineal, y la de Huber, entre otras.

La función de pérdida conocida como ε *-insensitive*, le aporta una propiedad nueva a las soluciones que dan las máquinas de soporte vectorial, la llamada *esparcidad* de la solución. Ella tiene la ventaja de que no se necesitan todos los patrones de entrada para describir el vector de regresión w, posibilitando una vía más eficiente, desde el punto de vista computacional, para su implementación.

Las MSV para regresión utilizada están basadas en la función de costo ε *-insensitive*. Esta función tiene la estructura que combina dos funciones, una de las cuales es $f(u) = |u|$ y la función constante $f(u) = \text{const}$. Definida por y; -(w .x;)-b ≤ ϵ.

Este tipo de funciones pueden crearse además, en la medida en que se desarrolle una estimación determinada, dando paso así a nuevas definiciones y modificaciones en los datos de entrada: (x_1, y_1),..., $(x_l, y_l) \in \chi \times \Re$ donde χ denota el espacio de entrada de los patrones.

En el método de regresión por vectores de soporte, usando la función de costo ε *-insensitive*, el objetivo es encontrar una función $f(x)$ que tenga como máximo una desviación ε de los valores de salida y_i

(Figura 3) para todos los datos de entrenamiento y al mismo tiempo sea tan lineal como sea posible. Se describe la función como: $f(x) = \langle w, x \rangle + b \quad \text{con} \quad w \in \chi, \quad b \in \Re.$

La solución de esta ecuación se convierte en un problema de optimización convexo, tal como se representa a continuación.

$$\text{minimizar} \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$\text{sujeto a} \begin{Bmatrix} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{Bmatrix}$$

Donde ξ_i, ξ_i^* son variables de holgura introducidas para las situaciones en las que es necesario permitir errores y la constante $C > 0$ determina el compromiso entre el aplanamiento de f y la cantidad hasta la cual las desviaciones mayores que ε, se toleran, siendo un parámetro de regularización y libre de ser ajustado a la maximización del margen m. La formulación anterior es la correspondiente a trabajar con la función de pérdida ε *-insensitive* que se describe como:

$$|\xi|_\varepsilon = \begin{Bmatrix} 0 & \text{si } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{en otro caso} \end{Bmatrix}$$

Solamente los puntos que se encuentren fuera de esta condición son los que contribuyen con las funciones de pérdida. Ver **Figura 3**.

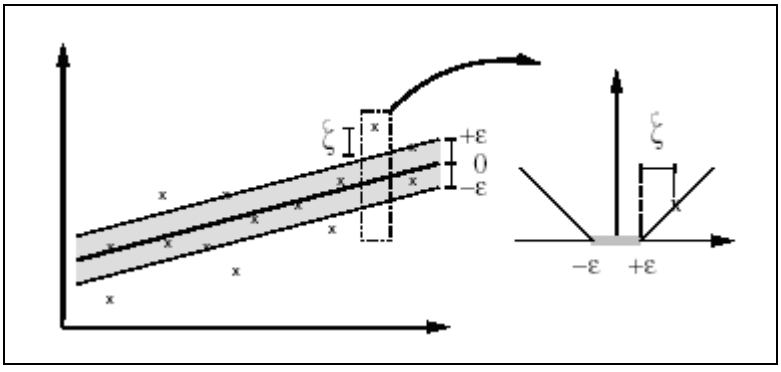


Figura 3.

La solución de este problema de optimización se realiza, al igual que con las MSV, mediante el empleo de los multiplicadores de Lagrange y la función del Lagrangiano.

Por lo tanto, se puede afirmar que en este tipo de técnicas, su estructura se determina sobre la base del conjunto de entrenamiento necesitándose pocos parámetros para el mismo. Dicho entrenamiento se reduce a la solución de un problema de optimización que se reduce a un problema de programación cuadrática. Al mismo tiempo, el uso de las funciones Kernels muestra una gran eficiencia en el resultado de la predicción.

*1.5 Estudios realizados en los últimos años***.**

Las MSV han tenido gran aplicación en la bioinformática, principalmente en la predicción de actividad biológica de compuestos, por ejemplo:

Año 2001

Se emplearon en el diseño de medicamentos para análisis de datos farmacéuticos en la Universidad de London.[8]

Año 2003

Fueron utilizadas para el diseño de fármacos y el estudio de similitud de los Medicamentos y predicción de la inhibición de enzimas. En el Departamento de Química de la Universidad de Moscú.[9]

Año 2004

Las MSV fueron utilizadas en la Universidad de París para desarrollar los modelos de QSAR que correlacionan las estructuras moleculares a su toxicidad y bioactividades. El funcionamiento y la capacidad predictiva de MSV, fueron descritas usando los parámetros fisicoquímicos o los descriptores moleculares. En ambos casos estudiados, la capacidad predictiva del modelo de MSV es comparable o superior. Los resultados indican que MSV se puede utilizar como herramienta para modelar los estudios QSAR.[10]

Se empleó en la Universidad de París en el estudio cuantitativo de la relación estructura-movilidad de los ácidos carboxílicos en la electroforesis de vasos capilares.[11]

Las MSV fueron utilizadas en el Departamento de Química de la Universidad de Lanzhou para desarrollar un modelo cuantitativo de la relación de la Estructura-Característica (QSPR) de la energía en enlace de la disociación del Oxígeno con el hidrógeno(O-H) de 78 fenoles substituidos.[12]
Las MSV, como tipo novedoso de máquina que aprendía, fueron utilizadas para desarrollar un modelo cuantitativo de la relación de la estructura-movilidad (QSMR) de 58 alifáticos y de ácidos carboxílicos aromáticos basados en los descriptores moleculares.[13]

Año 2005

La Universidad de Lanzhou utilizó MSV en el estudio QSAR de los compuestos de interrupción de la endocrina natural, sintética y ambiental para atar al receptor del andrógeno. Donde a partir de cinco descriptores el método mostró excelentes resultado.[14]

Año 2006

Fue empleada para determinar la predicción de un modelo cuantitativo de la relación de la estructura-característica (QSPR) se ha desarrollado para la degradación electroquímica de fenoles substituidos. Comparado con los modelos desarrollados con los cuadrados (PLS) y la regresión linear múltiple (MLR), donde estaban 0.804 y 0.799 Q2 respectivamente, MSV demostraron rendimientos más altos.[15]
Se emplearon para predecir relaciones cuantitativas de la estructura característica para los pesticidas en la universidad de Lanzhou, donde se probaron con: método heurístico (HM) y MSV y los modelos lineares y no lineales, generados por ambos se compararon los resultados de estos dos métodos, resultando las MSV el mejor método.[16]

Se emplearon en relación cuantitativa de la estructura-actividad de modelos para la predicción de los irritantes sensoriales de productos químicos orgánicos volátiles. De ellos se han desarrollado: modelos de clasificación y de regresión para la predicción de los irritantes sensoriales (LOGRD50) de los productos químicos orgánicos volátiles (VOCs). Cada compuesto fue representado por los descriptores estructurales.[17]

*1.6 Tendencias y Tecnologías actuales***.**

Metodología de Desarrollo.

La metodología utilizada para mantener una compatibilidad con las normas de calidad del centro (UCI) es el Proceso Unificado de Desarrollo (RUP), este es un proceso de desarrollo creado por la Corporación Rational Software, ahora una división de IBM, como una plataforma adaptable de procesos para describir cómo crear productos efectivos a través de técnicas de alta fidelidad. Aunque RUP abarca un determinado número de actividades diferentes, está diseñado para poder ajustarse en la selección de procesos específicos destinados a un proyecto u organización de desarrollo en particular y es reconocida en medio de grandes equipos de trabajo que llevan a cabo el manejo de complicadas aplicaciones de software.

RUP se basa en casos de uso para describir lo que se espera del software y esta muy orientado a la Arquitectura del sistema, documentándose lo mejor posible, basándose en UML (Unified Modeling Language) como herramienta principal. Unifica los mejores elementos de metodologías anteriores.

Herramienta Case.

La herramienta *CASE* utilizada para modelar el programa es el *Visual Paradigm* que es un producto de *Visual Paradigm UML Community*. Tiene disponible distintas versiones: *Enterprise*, *Professional*, *Standard*, *Modeler*, *Personal* y *Community* la cual es gratuita. Unas de las principales diferencias entre el *Visual Paradigm* y otros sistemas basados en *Windows,* está en la fuerza de la interfaz de la aplicación presentado a los usuarios y la base de todas las aplicaciones de base de datos. *Visual Paradigm* provee a los usuarios de una presentación con *Multiple Document Interface* (MDI). Todas las vistas tienen menús *pull-down*, barra de botones, de teclado y de navegación de los datos. Las vistas proporcionan menús *pop-up* del contexto, ayuda del estado en todos los campos, listas de la selección y las tablas de la validación, así como confirmaciones y advertencias en todas las operaciones de los datos.
Visual Paradigm es una herramienta que sirve para realizar modelado UML siguiendo el estándar UML. Esta herramienta tiene unas características graficas muy cómodas que facilitan la realización de los diagramas de modelado. Entre otras características importantes que tiene es la integración con algunos IDE de programación como Eclipse desarrollado por IBM, *Netbeans de Sun* o *JBuilder de Borland*.

Plataforma de Desarrollo.

En cuanto a plataforma escogida se optó por jdk version 1.5.0_10 y como lenguaje de programación Java debido a que es un lenguaje de propósito general y al hecho de ser independiente de la plataforma, buscando la portabilidad en los diferentes sistemas operativos y plataformas de hardware.

Entorno de Desarrollo.

Se utilizó Eclipse como entorno de desarrollo, es una plataforma de software de código abierto, extensible. Esta plataforma, típicamente ha sido usada para desarrollar entornos integrados de desarrollo, como el IDE de Java llamado *Java Development Toolkit* (JDT) y el compilador (ECJ) que se embarca como parte de Eclipse y que son usados también para desarrollar este entorno. Sin embargo, también se puede usar para otros tipos de aplicaciones cliente además de ser un entorno de desarrollo integrado que ofrece el control del editor de códigos, del compilador y del depurador desde una única interfaz de usuario. Este entorno de desarrollo integrado ofrece, el control del editor de código, del compilador y del depurador desde una única interfaz de usuario. Además Eclipse es una plataforma universal para integrar herramientas de desarrollo, basada en plug-ins.

Gestor de Bases de Datos.

Finalmente como sistema gestor de bases de datos se ha utilizado MySQL en su versión 5.0 debido a ser software libre, es uno de los gestores más rápidos que se encuentran en el mercado, presenta versiones en varios sistemas operativos y compatibilidad entre sus versiones y es muy fácil de usar.

MySQL es un sistema de administración de bases de datos relacional (RDBMS). Se trata de un programa capaz de almacenar gran cantidad de datos y distribuirlos para cubrir las necesidades de cualquier tipo de organización, desde pequeños establecimientos comerciales a grandes empresas y organismos administrativos. Es una aplicación de código abierto que permite redistribuir una aplicación que la contenga y modificar su código para mejorarla o adaptarla a nuestras necesidades. Además, existe la seguridad de contar con una importante cuota de mercado y de saber que es una solución estable, mantenida por un buen equipo de desarrolladores y e incluso con soporte de pago. Compite con sistemas RDBMS como Oracle, *SQL Server* entre otros. Incluye todos los elementos necesarios para instalar el programa, preparar diferentes niveles de acceso de usuarios, administrar el sistema y proteger los datos. Utiliza el lenguaje de consulta estructurado.

Facilidades.

MySQL es un sistema fácil de instalar y configurar en servidores Windows, Linux entre otros. Además ofrece facilidades para la realización de consultas para el manejo de los datos. Permite centralizar la base de datos y brinda mayor seguridad para estas.

Funcionalidad.

MySQL dispone de muchas funciones vitales para el desarrollo profesional cómo puede ser el volcado online, la duplicación etc.

Portabilidad.

MySQL puede correr en la inmensa mayoría de sistemas operativos, por lo que junto a otro lenguaje de programación de lado de servidor de alta portabilidad como Java y PHP nos permite el desarrollo de aplicaciones, el acceso y copia de los datos desde cualquier Sistema Operativo.

CAPÍTULO 2: REQUISITOS.

La inteligencia artificial aplicada a la predicción de actividad biológica, es el campo de acción de esta investigación, donde el proceso de negocio definido es muy simple lo que trae consigo la realización de un modelo de dominio para identificar los procesos de automatización: realizar predicción y realizar entrenamiento. Teniendo como actores principales de nuestro análisis al Especialista Químico y el Administrador para lograr una predicción de actividad biológica anticancerígena a partir de fragmentos moleculares, empleando MSV como técnica de inteligencia artificial.

2.1 Modelo de Dominio

El Flujo de Trabajo de Modelamiento del Negocio describe los procesos del negocio, identificando quienes participan y las actividades que requieren automatización. Teniendo como objetivo:

- Comprender la estructura y la dinámica de la organización en la cual se va a implementar un sistema.
- Comprender los problemas actuales de la organización e identificar las mejoras potenciales.
- Asegurar que los consumidores, usuarios finales y desarrolladores tengan un entendimiento común de la organización.
- Derivar los requerimientos del sistema que va a soportar la organización.

Como parte del Modelo de Negocio se encuentra el Modelo de Dominio el cual captura los tipos más importantes de objetos en el contexto del sistema. Los objetos del dominio representan los eventos que suceden en el entorno en que trabaja el sistema. Los conceptos definidos para dicho modelo son:

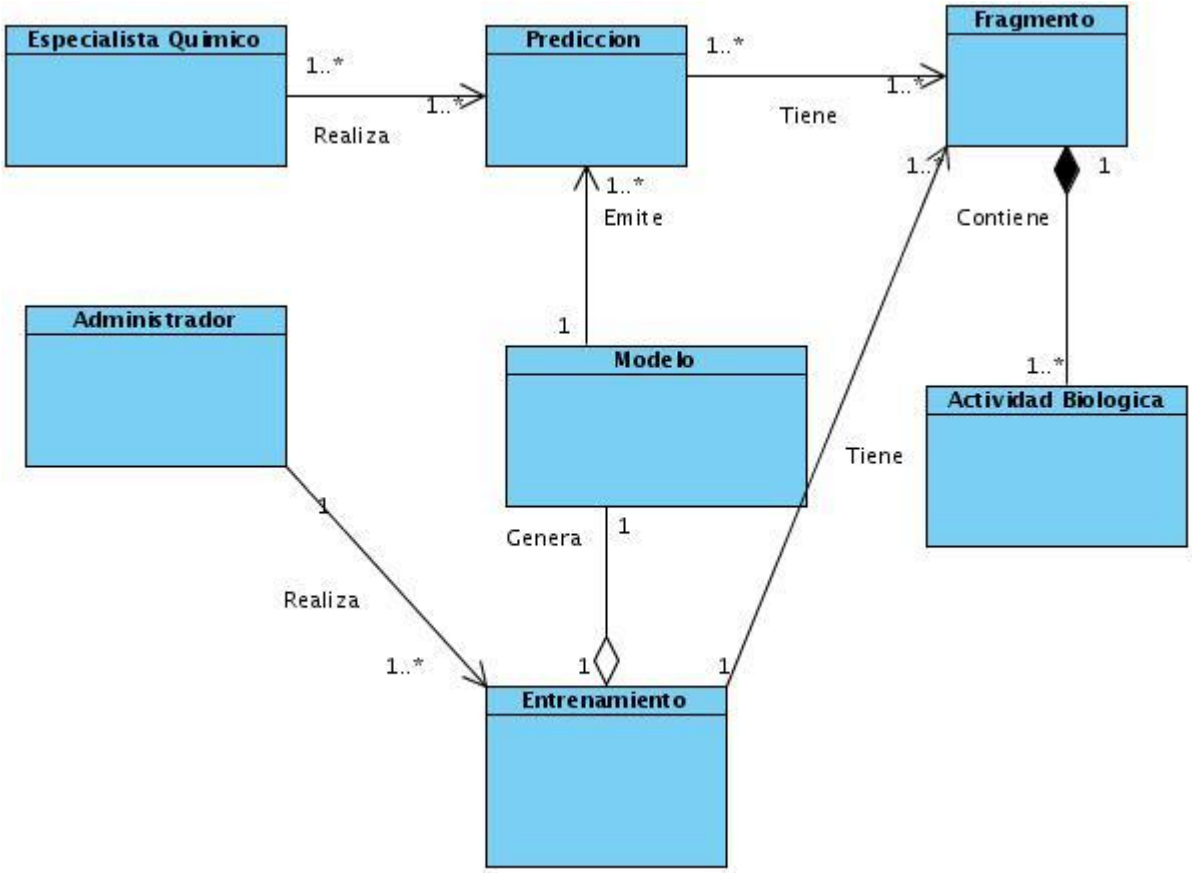
- Especialista Químico: Cliente que utilizará en sistema.
- Administrador: Persona que se encarga de generar los modelos de predicción a partir del entrenamiento de las máquinas de soporte vectorial.
- Entrenamiento: Acción que se realiza para el aprendizaje de la máquina de soporte vectorial.
- Modelo: Modelo de máquina de soporte vectorial que se crea para realizar una predicción.
- Predicción: Acción que el sistema realiza para emitir un resultado.
- Fragmento: Porción de una molécula que tiene asociada una actividad biológica.

- Actividad Biológica: Actividad asociada a un fragmento molecular.

*2.1.1 Descripción del modelo de dominio***.**

El Especialista Químico es quien realizará la predicción de los fragmentos que tendrán asociado una actividad biológica anticancerígena, dicha predicción se emite a partir de un modelo de datos que no son más que fragmentos que han sido entrenados y que tienen asociado también una actividad biológica, este entrenamiento es realizado por el administrador quien es el encargado de generar estos modelos.

Modelo de Dominio



*2.2 Requerimientos del sistema***.**

*2.2.1 Requisitos Funcionales***.**

Los requerimientos funcionales son capacidades o condiciones que el sistema debe cumplir. Para el sistema propuesto se definen:
R1: Realizar predicción.
R2: Realizar entrenamiento.
R3: Crear modelos.
R4: Cargar fichero.

2.2.2 Requerimientos No Funcionales.

Los requerimientos no funcionales son propiedades o cualidades que el producto debe tener. Representan las características del producto.

Apariencia o interfaz externa.

- La aplicación debe estar diseñada con una interfaz de fácil uso, de forma tal que el usuario navegue sin dificultad alguna, ajustándose a los estándares establecidos para el desarrollo de un buen diseño.

Usabilidad.

- El sistema podrá ser usado por cualquier tipo de persona que posea conocimientos básicos en el manejo de la computadora. En este caso el Especialista Químico y el Administrador.
- Se necesita que el usuario sea un especialista en química para entender los resultados dados por la aplicación.

Rendimiento.

- Al estar concebida para un ambiente cliente/servidor, se trata de garantizar la rapidez de respuesta del sistema ante las solicitudes de los usuarios, al igual que la velocidad de procesamiento de la información. Para lo cual se realiza la validación de los datos y la manipulación de eventos en el cliente y en el servidor aquellas que por cuestiones de seguridad, o de acceso a los datos lo requieran. Lográndose así un tiempo de respuesta más rápido, una mayor velocidad de procesamiento, y un mayor aprovechamiento de los recursos.

Soporte.

- El sistema debe propiciar su mejoramiento y la incorporación de otras opciones en un futuro.

Portabilidad.

- El sistema puede ser ejecutado sobre los sistemas operativos Linux y Windows, por su característica de ser multiplataforma.

Seguridad.

- El especialista solo tiene control a predecir y el administrador a la realización del modelo a través del entrenamiento.

Confiabilidad.

- El sistema debe ser confiable y preciso en la información que le suministra al usuario para evitar cualquier tipo de error.

Ayuda.

- La aplicación posee ayuda, en la que se explica de forma clara el uso de las opciones del sistema, garantizando así el buen desempeño de los usuarios a la hora de interactuar con el mismo.

Software.

- Se debe disponer de sistemas operativos *Linux, Windows* 95 o superior para la instalación de la aplicación.
- Debe tenerse instalado el *Java Runtime Environment* (JRE) versión 1.5 o superior.

Hardware.

Para el desarrollo y puesta en práctica del proyecto se requieren máquinas con los siguientes requisitos:

- Procesador *Pentium* 3 o superior.
- 256 Mb de RAM.
- 50 Mb de capacidad del disco duro.

*2.3 Definición del Sistema***.**

*2.3.1 Actores***.**

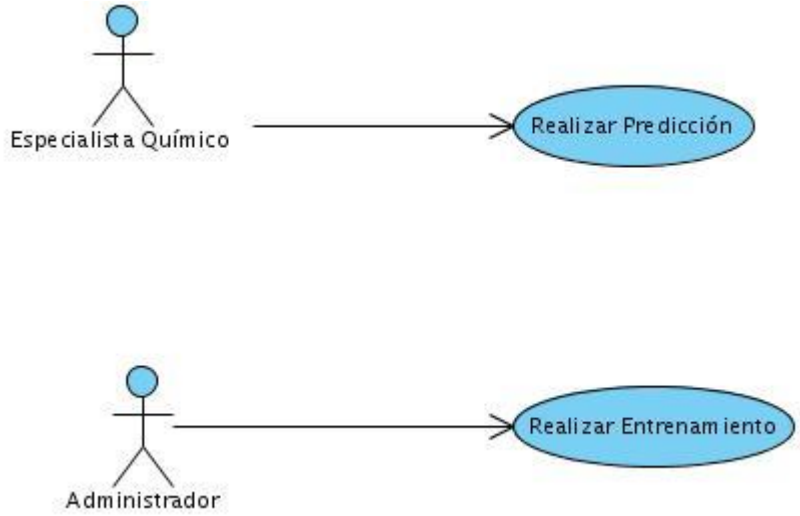
Los actores del sistema son los trabajadores del negocio. Es decir el personal que trabaja directamente con la aplicación. El especialista químico y el administrador son los actores definidos en este contexto.

Actores	Descripción
Especialista Químico	Usuario que va ha hacer uso del sistema, para obtener la predicción de algún compuesto químico.
Administrador	Es el encargado de generar nuevos modelos, así como el entrenamiento de los mismos.

*2.3.2 Casos de Uso del Sistema***.**

Los casos de uso del sistema son la funcionalidad que el mismo tendrá, estarán estrechamente relacionados con los requisitos funcionales planteados anteriormente.

*2.3.3 Diagrama de Casos de Uso del Sistema***.**



2.3.4 Descripción de los Casos de Uso del Sistema.

Caso de Uso:	Realizar Predicción
Actores:	Especialista Químico
Propósito:	Realizar la predicción de los fragmentos moleculares y devolver dicha predicción
Resumen:	El especialista químico inicia el caso de uso al cargar un fichero de prueba con las/la molécula de la cual desea saber su predicción, escoge un modelo que corresponda al entrenamiento realizado a ese tipo de datos, y asigna la dirección de un archivo vacío para la escritura de los resultados.
Referencia:	R1, R4.
Precondiciones:	Debe existir un fichero con los modelos entrenamiento, otro con los datos de prueba y por ultimo uno para escribir los resultados.
Poscondiciones:	Se escribe un fichero con los datos de la predicción y se muestra un resultado.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El especialista químico selecciona el botón cargar molécula.	1.1 El sistema carga el fichero con lo datos de prueba.
2. El especialista selecciona el botón para cargar el fichero modelo.	2.1 El sistema carga el fichero modelo.
3. El especialista selecciona el directorio en el cual esta el fichero vacío.	3.1 El sistema guarda el directorio donde se encuentra el camino al fichero vacío.
4. El especialista selecciona el botón para realizar la predicción.	4.1 El sistema realiza la predicción con el fichero de prueba y el modelo de entrenamiento. 4.2 El sistema escribe el fichero vacío con los resultados. 4.3 El sistema muestra al especialista valores resultantes de la Predicción.

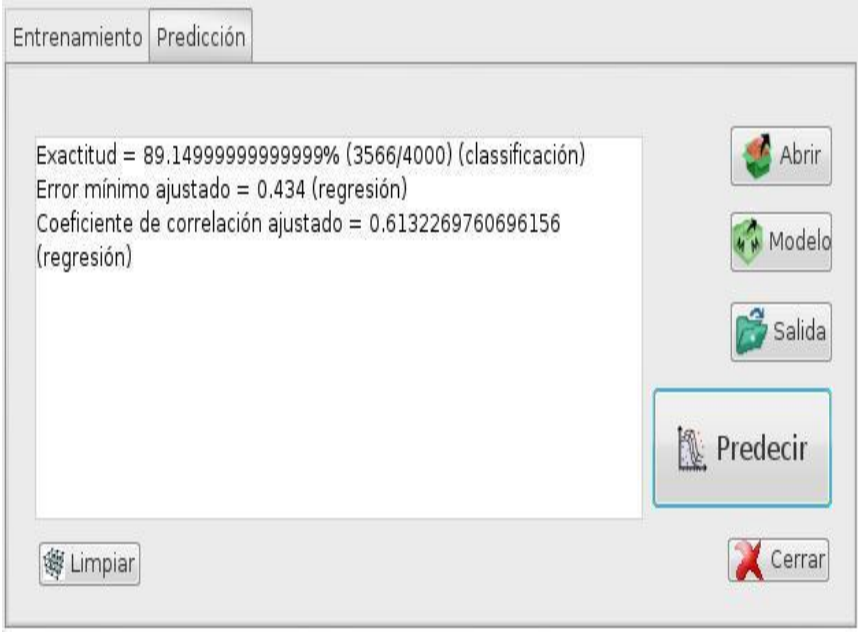
Prototipo de Interfaz	 Formulario para predecir los fragmentos.
Prioridad:	Crítico.

Tabla 1.1 Descripción Caso de Uso Realizar Predicción

Caso de Uso:	Realizar Entrenamiento.
Actores:	Administrador
Propósito:	Crear modelos para la predicción a partir del entrenamiento de los datos.
Resumen:	El administrador inicia el caso de uso cuando selecciona las opciones para realizar el entrenamiento y carga el fichero con los datos de los fragmentos que presentan las actividades biológicas asociadas y el índice del estado refractotopológico total, el sistema realiza el entrenamiento y devuelve el modelo.
Referencia:	R2, R3, R4.
Precondiciones:	Deben existir el fichero con los datos para el entrenamiento.
Poscondiciones:	Genera un modelo.
Flujo Normal de Eventos	

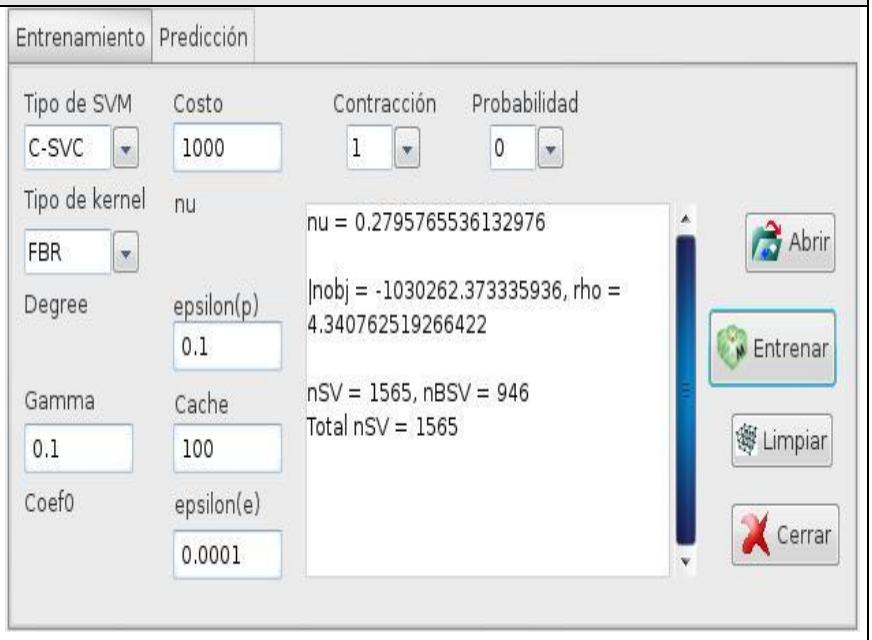
Acción del Actor	Respuesta del Sistema
1. El administrador selecciona el botón buscar fragmentos y busca el fichero con los datos de fragmentos para realizar el entrenamiento.	1.1 El sistema carga los datos.
2. El administrador selecciona los parámetros para el entrenamiento.	2.1 El sistema realiza el entrenamiento de los fragmentos. 2.2 El sistema crea el modelo.
Prototipo de Interfaz	 Formulario para el entrenamiento de los fragmentos.
Prioridad:	Crítico.

Tabla 1.2 Descripción del Caso de Uso Realizar Entrenamiento.

Como resultado del flujo de trabajo de requisitos se obtuvo una vista externa del sistema, representado a través del Diagrama de casos de uso propuesto en el capítulo anterior. A partir de aquí se debe profundizar en los casos de usos, detallándolos de manera que permitan reflejar una vista interna del sistema descrita con el lenguaje de los desarrolladores. En esta vista interna se especifican mejor los casos de uso y se determinan las clases necesarias para llevar a cabo las funcionalidades en ellos contenidos.

3.1 Modelo de Análisis.

Este proceso se desarrolla fundamentalmente dentro de la fase de elaboración y se corresponde principalmente con el flujo de trabajo de análisis y diseño según RUP. Definiendo los artefactos, actividades y trabajadores que participan en el análisis.

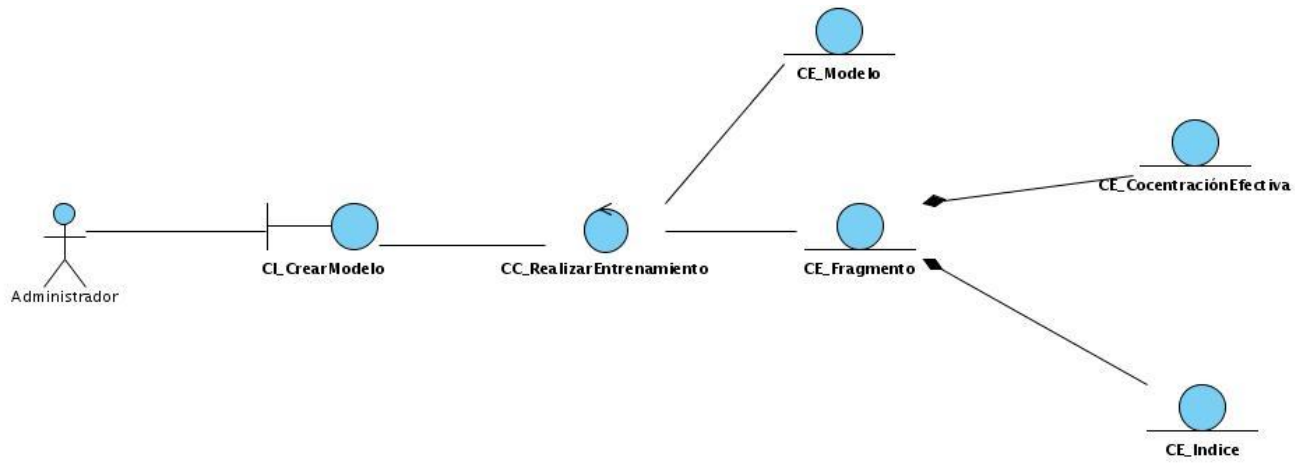
A pesar de que el modelo del análisis hay un refinamiento de los requisitos, no se toman en cuenta el lenguaje de programación a usar en la construcción, la plataforma en la que se ejecutará la aplicación, los componentes prefabricados o reutilizables de otras aplicaciones, entre otras características que afectan al sistema, ya que el objetivo del análisis es comprender perfectamente los requisitos del software y no precisar cómo se implementará la solución.

*3.1.1 Diagrama de Clases del análisis***.**

DIAGRAMA DE CLASES DEL ANALISIS

Caso de Uso Realizar Predicción.

DIAGRAMA DE CLASES DEL ANÁLISIS



Caso de Uso Realizar Entrenamiento.

3.2 Modelo de Diseño.

El flujo de trabajo de diseño tiene el propósito de formular los modelos que se centran en los requisitos no funcionales, en el dominio de la solución y que prepara para la implementación y prueba del sistema creando un plano del modelo de implementación.

Durante esta fase se analiza si es posible dar una solución que satisfaga a los requerimientos significativos de la arquitectura la cual es una descripción de los subsistemas y los componentes de un sistema informático y las relaciones entre ellos.

*3.2.1 Estilo de Arquitectura***.**

La arquitectura se centra en un modelo de tres capas:

Capa de Presentación.

En esta capa se encuentran las interfaces con las que interactúa el usuario.

Capa de Lógica de Negocio.

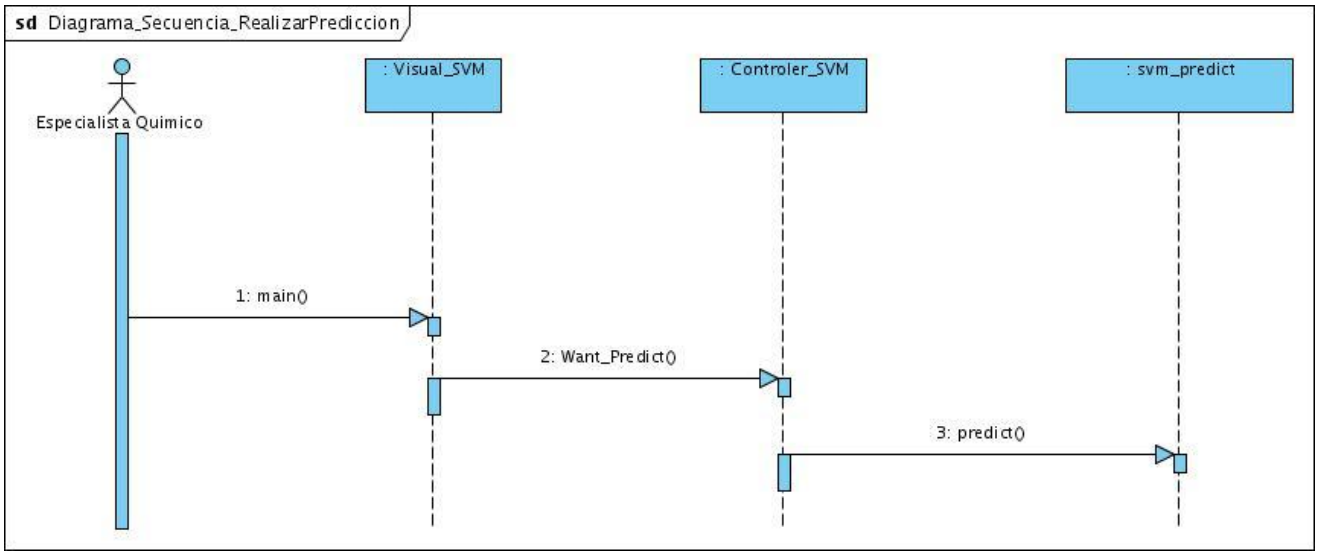
Esta capa está compuesta por las clases controladoras y la biblioteca (LibSVM).

Capa de Acceso a Datos.

Está compuesta por los ficheros de datos manejados en la capa de negocio.

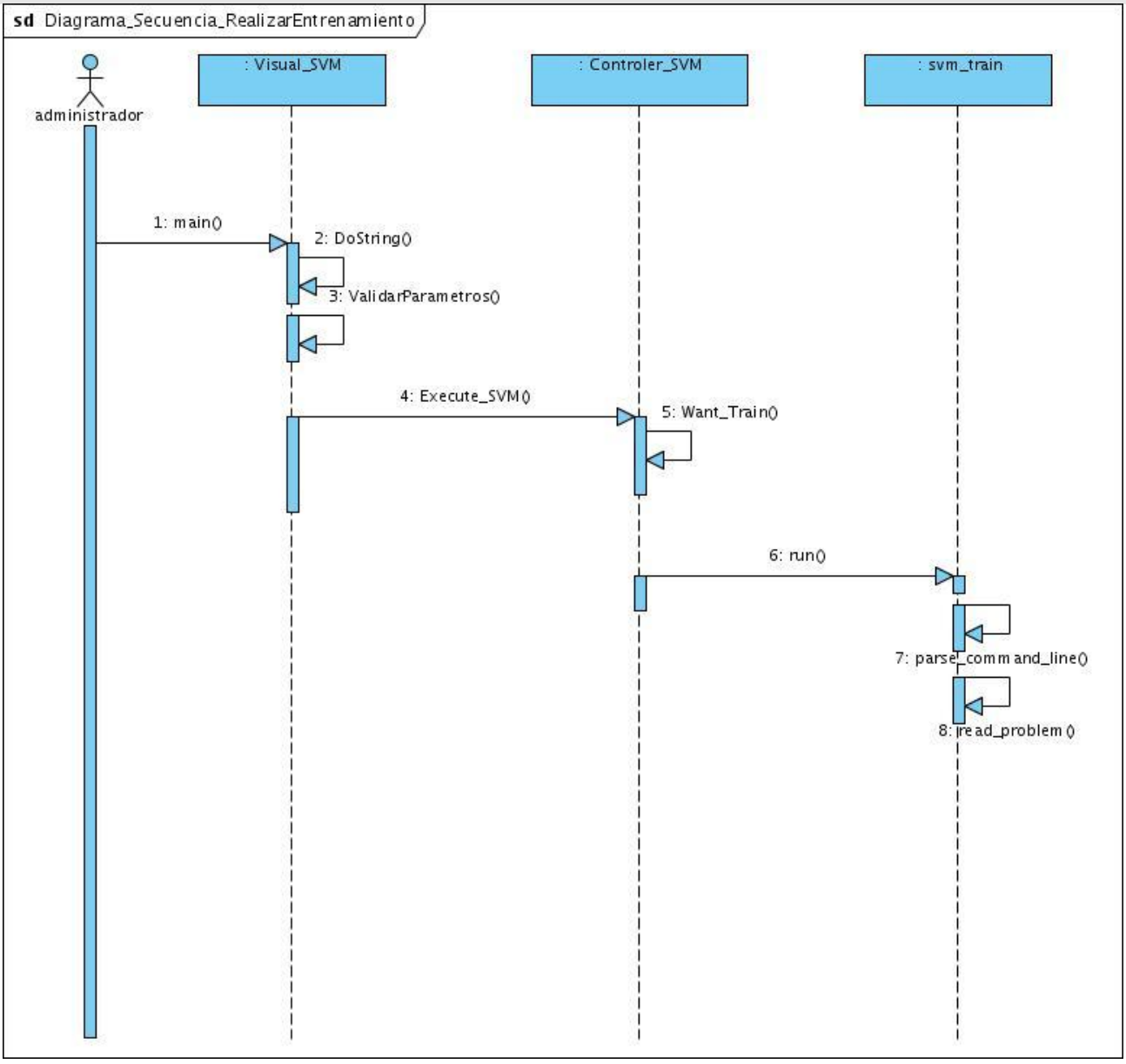
*3.2.2 Diagramas de Secuencia***.**

DIAGRAMA DE INTERACCION



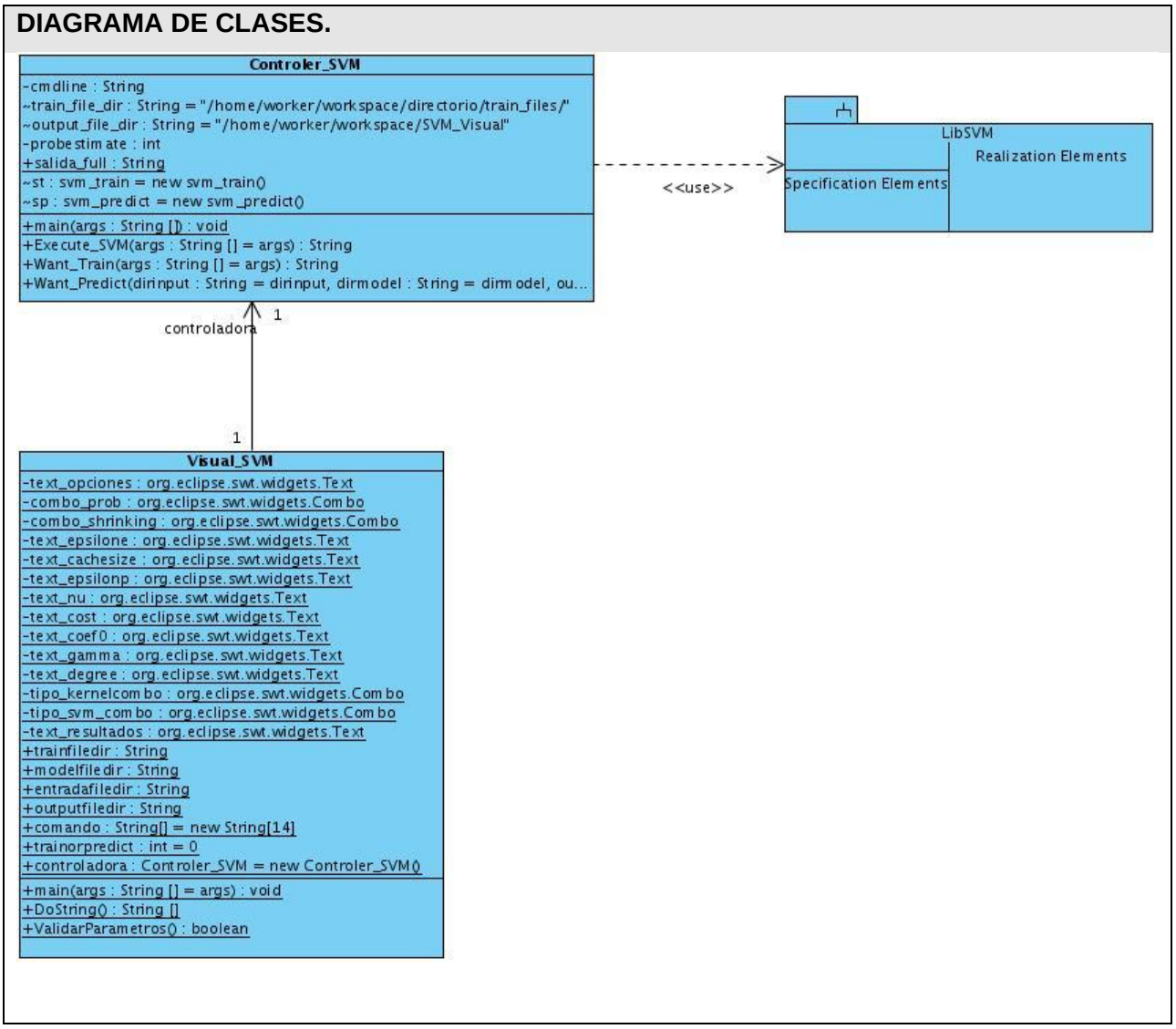
Caso de Uso Realizar Predicción

DIAGRAMA DE INTERACCION



Caso de Uso Realizar Entrenamiento.

*3.2.3 Diagrama de Clases del Diseño***.**



*3.2.4 Descripción de las clases***.**

Nombre: Visual_SVM	
Tipo de clase: Vista	
Modificador: -	
Padre: -	
Implementa: Serializable	
Principales Responsabilidad	
Nombre	**Tipo**
main	public static void
DoString	public static String
ValidarParametros	public static boolean
Descripción:	
Visual_SVM es la clase de interfaz visual, de donde se van a entrar los parámetros para la realización del entrenamiento y se carga un fichero con los datos para dicha acción. También soporta esta interfaz la entrada de ficheros y modelos para la realización de la predicción.	

Nombre: Controler_SVM	
Tipo de clase: Entidad	
Modificador: -	
Padre:-	
Implementa: Serializable	
Principales Responsabilidad	
Nombre	**Tipo**
Execute_SVM	public void
Want_Train	public String
want_Predict	public String
Descripción:	
La clase controladora Controler_SVM es la que manda a ejecutar las acciones para la realización del entrenamiento y de la predicción.	

CAPÍTULO 4: IMPLEMENTACIÓN DEL MÓDULO Y VALIDACIÓN DEL MODELO.

La peculiaridad del diseño es que al modelar el sistema, encuentra su arquitectura para que soporte todos los requisitos. Para la implementación se parte del resultado del diseño y se implementa el sistema en términos de componentes, es decir, ficheros de código fuente, scripts, ficheros de código binario, ejecutables y similares. El Flujo de Implementación está determinado por el lenguaje de programación, en él se desarrolla la arquitectura y el sistema como un todo. Este flujo tiene como objetivos:

- Definir la organización del sistema en términos de Subsistemas de Implementación organizados en capas.
- Implementar los elementos del diseño.
- Probar los componentes desarrollados de forma independiente.
- Integrar los resultados producidos por los desarrolladores.

*4.1 Diagrama de Componentes***.**

Los diagramas de componentes se usan para estructurar el modelo de implementación en términos de subsistemas de implementación y mostrar las relaciones entre los elementos de implementación.

4.2 Validación del Modelo.

La creación de los ficheros de entrenamiento y prueba, a partir de la muestra original de nuestro sistema se realizó con criterio químico, considerando las agrupaciones de átomos, las cuales constituyen muchas de las funcionalidades en las moléculas orgánicas. A partir de las facilidades de cómputo existentes, se tomaron 8000 fragmentos para conformar los ficheros de entrenamiento y prueba en clasificación, asignando 4000 a cada muestra. Para la regresión se utilizaron 4000 y 1000 fragmentos para los ficheros de entrenamiento y prueba respectivamente.

En la biblioteca para las MSV (LibSVM) implementada en Java, están definidos los siguientes tipos de máquinas: C-MSV y nu-MSV para clasificación, mientras que para regresión están: epsilon-MSV y nu-MSV respectivamente.

El formato para los ficheros de entrenamiento y prueba es el siguiente:

<etiqueta> <indice1>:<valor1> <indice2>:<valor2> . . .

Para clasificación <etiqueta> es un número entero que indica la etiqueta de la clase que determina si el fragmento es activo o inactivo. En regresión, <etiqueta> es un valor único que define al fragmento y esta asociado a su actividad biológica, el cual puede ser cualquier número. El formato <índice>:<valor> muestra un valor de un campo (atributo), definido a partir de: el <índice>, que es un número entero a partir de 1 ordenado ascendentemente y el <valor> es un número real. Las etiquetas en el archivo de prueba se utilizan solamente para calcular exactitud o errores. Si estos son valores desconocidos, se completa la primera columna con números aleatorios.

Para realizar el entrenamiento se cuenta además con cuatro funciones kernels básicas:

- Lineal: K(u, v) = u' * v.
- Polinomial: K(u, v) = (gamma * u' * v + coef0)^degree.
- Función de Base Radial (FBR): K(u, v) = exp(-gamma * |u – v|^2).
- Sigmoidal: K(u, v) = tanh(gamma * u * v + coef0).

Los parámetros que se tienen cuenta para el cálculo de estas predicciones a los cuales se enfocó esta investigación son:

- degree: Para modificar el parámetro del grado de la función polinómica (por defecto: 3).
- gamma: Para modificar este parámetro en la función kernel (por defecto: 1/k, donde k es el número de atributos en los datos de entrada).
- coef0: Para modificar este parámetro en la función kernel (por defecto: 0).
- Costo (C): Parámetro de penalización (por defecto: 100).
- nu: Parámetro que controla el número de vectores de soporte y los errores de entrenamiento (v ϵ (0, 1], por defecto: 0.5).
- epsilon(p): modifica el valor de epsilon en la función de perdida (por defecto: 0.1).
- epsilon(e): Modifica la tolerancia en el criterio de parada (por defecto: 0.001).

El Kernel empleado para el entrenamiento de las muestras, tanto para clasificación como para regresión, es FBR, debido a que es una primera opción razonable, ya que mapea no-linealmente la muestra en un espacio de mayor dimensión, por lo que a diferencia del Kernel Lineal puede manejar la relación no-lineal entre etiqueta y atributos. Además, para ciertos valores de C, el Kernel Lineal tiene el mismo rendimiento que el FBR con algunos valores de C y gamma, así como el Kernel Sigmoidal se comporta como el FBR para ciertos parámetros.

Otra de las razones es que el número de hiperparámetros, que no son más que los valores evaluados en la función Kernel y llevados a un hiperplano de mayor dimensión, influyen en la complejidad de la selección del modelo. El FBR tiene menor cantidad que el Kernel Polinomial, por lo que tiene menos dificultades numéricas.[18]

4.2.1 Análisis de los resultados.

Se realizó una modelación y predicción de actividad biológica empleando datos reales obtenidos del ensayo NCI Yeast Anticancer Drug Screen. La muestra de trabajo empleada tiene una relación de 1x1 entre moléculas activas e inactivas. Se realizaron predicciones con los ficheros de prueba y entrenamiento, para conocer el error del modelo para el conjunto de parámetros empleados.

Se evaluó RBF con distintos valores de C, nu, gamma, epsilon(p), epsilon(e), degree y coef0 explicados anteriormente. Se compararon los resultados obtenidos entre los tipos de MSV: C-SVC y nu-SVC, para clarificación, y epsilon-SVR, nu-SVR para regresión.

El procedimiento a seguir para realizar las predicciones, es el siguiente:

Clasificación

Primero se selecciona el tipo de MSV (C-SVC o nu-SVC). En este caso se empezó por C-SVC, el tipo de kernel: FBR, con valores de gamma de 0.1, 0.3 y 0.5, C = 1000, p = 0.1, e = 0,0001. Se generaron los tres modelos para clasificar y se pudo valorar en estos casos, la influencia de los diferentes valores de gamma, en la efectividad entre modelos.

Se generan entonces las predicciones para nu-SVC, con nu = 0.5, valores de gamma 0.1, 0.3 y 0.5, e = 0.0001. (Anexos: 1 ,2 y 3)

A partir del análisis de los resultados, los % de buena clasificación en los modelos de C-SVC están entre el 89% y el 95% para la muestra de entrenamiento, y en nu-SVC se mantiene entre 81-82%. No obstante, puede afirmarse que sus predicciones están bastantes equilibradas para la muestra de prueba, manteniéndose en ambos casos entre un 70 y un 75 por ciento. Esto demuestra que, para la muestra limitada de trabajo, con el reducido número de descriptores empleados, las MSV brindan una solución aceptable para la clasificación de fragmentos ponderados con el Índice del Estado Refractotopológico Total (Rstate (t)). (Ver Tabla 4.1)

Regresión

En este tipo de predicción se efectúa el mismo procedimiento que se usó para la clasificación. Se seleccionó C = 1000, e = 0.00001, p = 0.1, gamma con valores de 0.1, 0.3 y 0.5, FBR y e-SVR como MSV. Los valores de predicción para nu-SVR son: FBR, C = 1000, e = 0.00001, nu = 0.5 y gamma con valores de 0.1, 0.3 y 0.5.

En las tablas 4.2 y 4.3 se muestran el por ciento de los resultados de la predicción para la regresión:

Gamma	C-SVC	Muestra de entrenamiento (C-SVC)	nu-SVC	Muestra de entrenamiento (nu-SVC)
0.1	73.575	89.15	74.0	81.875
0.3	71.0	93.50	73.95	81.475
0.5	70.15	95.70	73.8	81.4

Tabla 4.1 Resultados de clasificación.

Gamma	e-SVR	Muestra de entrenamiento (e-SVR)	nu-SVR	Muestra de entrenamiento (nu-SVR)
0.1	0.7362861	0.66183890	0.73406967	0.66231227
0.3	0.607990	0.66183297	0.61278325	0.66234425
0.5	0.5055281	0.6618329	0.51248035	0.662346035

Tabla 4.2 Resultados de los valores de coeficientes de correlación ajustado para la regresión.

Gamma	e-SVR	Muestra de entrenamiento (e-SVR)	nu-SVR	Muestra de entrenamiento (nu-SVR)
0.1	69.44296	52.932312	90.1346471	52.4634155
0.3	88.171425	52.932226	92.11092491	52.453223399
0.5	107.678377	52.932148	111.367694	52.452334517

Tabla 4.3 Resultados del error máximo ajustado para la regresión.

Según lo mostrado en las tablas anteriores, a medida que los valores de gamma aumentan, los coeficientes de correlación ajustados disminuyen y los errores máximos ajustados también aumentan, lo que se puede concluir que para las MSV (e-SVR y nu-SVR), el modelo generado con un valor de gamma de 0.1 y FBR, para fragmentos con Rstate (t), brinda un mejor resultado.

CONCLUSIONES.

Se analizó, diseñó e implementó una aplicación para la predicción de actividad biológica de compuestos orgánicos empleando Máquinas de Soporte Vectorial y Regresión Soportada por Vectores, con el uso de fragmentos ponderados por el Índice del Estado Refractotopológico Total como método de descripción de la estructura química.

Se determinó que para las Máquinas de Soporte Vectorial empleadas (e-SVR y nu-SVR), el modelo se debe generar con valores pequeños de gamma para alcanzar los mejores resultados en la regresión.

Se obtuvieron modelos de clasificación con un por ciento de acierto mayor del 70%, lo cual constituye un resultado alentador dentro de esta primera versión. De igual manera, los valores de coeficiente de correlación de 0.66 y 0.75 para las muestras de entrenamiento y prueba respectivamente, constituyen también un buen punto de partida para el perfeccionamiento del sistema.

Se proponen las máquinas C-SVC y e-SVR para clasificación y para regresión respectivamente, teniendo en cuenta los resultados mostrados con ambas en el modelado y la predicción.

Se implementó una aplicación como plug-in para su incorporación al visualizador del sistema, para la predicción de actividad biológica de compuestos orgánicos utilizando como lenguaje de programación Java.

RECOMENDACIONES.

- Investigar los posibles Kernels para realizar futuras comparaciones con fragmentos.
- Probar nuevos tipos de combinaciones posibles entre los parámetros, para lograr una mejor predicción.
- Utilizar mayores cantidades de datos de fragmentos, para generar un modelo mejor.
- Hacer un estudio más profundo del tema para futuras funcionalidades a otras ramas de la Bioinformática.

REFERENCIA BIBLIOGRÁFICA.

1. ESCALONA, J. C.; CARRASCO, R.*, et al. Introducción al diseño de Fármacos* [Consultado el: 20 de junio de 2007]. 17. Disponible en: http://www.fq.uh.cu/investig/lqct/imagenes2/diseno.pdf.

2. BALMASEDA, J. C. L. *El Cáncer de los Cubanos* [Consultado el: 22 de junio de 2007]. Disponible en: http://medicinacubana.blogspot.com/2007/04/el-cncer-de-los-cubanos.html.

3. GUTIÉRREZ, H. *Modelización molecular de los receptores de adenosina y sus ligandos en el marco de diseño de fármacos asistido por ordenador* Doctorado, Ciencias Experimentales. Universidad de Pompeu Fabra, 2004.

4. RESENDIZ, J. A. *Las Máquinas de Soporte Vectorial para identificación en Linea.* Maestria, Control Automático. Instituto Politecnico Nacional, 2006.

5. RÍO, B. M. D. y MOLINA, A. S. *Redes Neuronales y Sistemas Difusos.* Disponible en: http://bibliodoc.uci.cu/pdf/reg00050.pdf.

6. CHEN, N.; YANG, J.*, et al. Support Vector Machine in Chemistry.* World Scientific, 2004,

7. YANG, H. *Margin Variations in Support Vector Regression for the Stock Market Prediction.* Maestria, Ingenieria y Ciencias de la Computacion. Universidad de Hong Kong, 2003.

8. R, B.; M, T.*, et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis* [Consultado el: 23 de junio de 2007]. Disponible en: http://citeseer.ist.psu.edu/528480.html.

9. VV, Z.; KV, B.*, et al. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions* [Consultado el: 23 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2003/43/i06/abs/ci0340916.html.

10. HX, L.; CX, X.*, et al. Quantitative prediction of logk of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine*

[Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i06/abs/ci049891a.html.

11. CX, X.; RS, Z.*, et al. Study of the quantitative structure-mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines.* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i03/abs/ci034280o.html.

12. XUE, C. X.; ZHANG, R. S.*, et al. An Accurate QSPR Study of O-H Bond Dissociation Energy in Substituted Phenols Based on Support Vector Machines*
[Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i02/abs/ci034248u.html.

13. S, W.; C, X.*, et al. Study on the quantitative relationship between the structures and electrophoretic mobilities of flavonoids in micellar electrokinetic capillary chromatography* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.cababstractsplus.org/google/abstract.asp?AcNo=20043056516.

14. CY, Z.; RS, Z.*, et al. QSAR study of natural, synthetic and environmental endocrine disrupting compounds for binding to the androgen receptor* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.informaworld.com/smpp/content~content=a727214684~db=all.

15. S, Y.; M, X.*, et al. Quantitative structure-property relationship studies on electrochemical degradation of substituted phenols using a support vector machine.* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.informaworld.com/smpp/content~content=a757824679~db=all~jumptype=rss.

16. W, M.; F, L.*, et al. Quantitative structure-property relationships for pesticides in biopartitioning micellar chromatography* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.aapspharmaceutica.com/search/view.asp?ID=74403.

17. F, L.; W, M.*, et al. Quantitative structure-activity relationship models for prediction of sensory irritants (logRD50) of volatile organic chemicals.* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=16307788&dopt=Abstract.

18. HSU, C.-W.; CHANG, C.-C., *et al. A Practical Guide to Support Vector Classification* [Consultado el: 20 de junio de 2007]. 1-12. Disponible en: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

BIBLIOGRAFÍA.

1. XUE, C. X.; ZHANG, R. S., *et al. An Accurate QSPR Study of O-H Bond Dissociation Energy in Substituted Phenols Based on Support Vector Machines* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i02/abs/ci034248u.html.

2. CY, Z.; HX, Z., *et al. Application of support vector machine (SVM) for prediction toxic activity of different data sets.* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TCN-4H877RX-1&_user=2342189&_coverDate=01%2F16%2F2006&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000056883&_version=1&_urlVersion=0&_userid=2342189&md5=30f145309eb9dd80ed8adba10558bd03.

3. WEI, C. *BAYESIAN APPROACH TO SUPPORT VECTOR MACHINES.* Doctorado, Ingeniería Mecánica. Universidad Nacional de Singapore, 2003.

4. SALUD, O. M. D. L. *Cáncer* [Consultado el: 21 de junio del 2007 Disponible en: http://www.who.int/mediacentre/factsheets/fs297/es/print.html.

5. F, L.; R, Z., *et al. Classification of the carcinogenicity of N-nitroso compounds based on support vector machines and linear discriminant analysis.* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/crtoec/2005/18/i02/abs/tx049782q.html.

6. XJ, Y.; A, P., *et al. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression* Paris: [Consultado el: 22 de junio Disponible en: http://cat.inist.fr/?aModele=afficheN&cpsidt=15967815.

7. RICARDO, I. F. A. C.; LÓPEZ, I. N. S., *et al. Conferencia 3: Fase de Inicio. Levantamiento de requisitos.* La Habana: 2006, Disponible en: http://inter-nos/Teleclases/Teleclases.asp?id_as=12.

8. RICARDO, I. F. A. C. y LÓPEZ, I. N. S. *Conferencia 5: Fase de Elaboración.Análisis-Diseño.* La Habana: 2006, Disponible en: http://inter-nos/Teleclases/Teleclases.asp?id_as=12.

9. CANCER, S. A. D. *Datos y Estadísticas de Cáncer en Hispanos/Latinos 2003-2005* [Consultado el: 22 de junio de 2007]. 1-9. Disponible en: http://www.cancer.org/downloads/STT/862301.pdf.

10. R, B.; M, T., *et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis* [Consultado el: 23 de junio de 2007]. Disponible en: http://citeseer.ist.psu.edu/528480.html.

11. VV, Z.; KV, B., *et al. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions* [Consultado el: 23 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2003/43/i06/abs/ci0340916.html.

12. BALMASEDA, J. C. L. *El Cáncer de los Cubanos* [Consultado el: 22 de junio de 2007]. Disponible en: http://medicinacubana.blogspot.com/2007/04/el-cncer-de-los-cubanos.html.

13. LÓPEZ, C. A. *El futuro de los medicamentos* Disponible en: http://www.ranf.com/pdf/arti/futuro.pdf.

14. JACOBSON, I.; BOOCH, G., *et al. El Proceso Unificado de Desarrollo de Software.* La Habana: Felix Varela, 2004. vol. I, 105-179 p.

15. CANTO, D. A. M.; CUETO, D. C. E. S., *et al. EPIDEMIOLOGÍA DEL CÁNCER DE PULMÓN. ESTUDIO DE CINCO AÑOS* [Consultado el: 21 de junio de 2007]. 1-7. Disponible en: http://www.amc.sld.cu/amc/2004/v8n1/865.pdf. ISBN 1025-0255.

16. VALERO, A. T. *Extracción de Información con Algoritmos de Clasificación.* Instituto Nacional de Astrofísica, Óptica y Electrónica., 2005.

17. DESHPANDE, M.; KURAMOCHI, M., *et al. Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds* [Consultado el: 17 de junio de 2007]. 2-13. Disponible en: http://doi.ieeecomputersociety.org/10.1109/ICDM.2003.1250900.

18. MAK, G. *THE IMPLEMENTATION OF SUPPORT VECTOR MACHINES USING THE SEQUENTIAL MINIMAL OPTIMIZATION ALGORITHM.* Master en Ciencias, School of Computer Science McGill University, Montreal, Canada, 2000.

19. PRESMAN, R. S. *Ingenieria de Software. Un Enfoque Práctico.* . La Habana: Felix Varela, 2005. vol. 2, 165-323 p.

20. ESCALONA, J. C.; CARRASCO, R., *et al. Introducción al diseño de Fármacos* [Consultado el: 20 de junio de 2007]. 17. Disponible en: http://www.fq.uh.cu/investig/lqct/imagenes2/diseno.pdf.

21. BETANCOURT, G. A. *Las Máquinas de Soporte Vectorial* [Consultado el: 22 de junio de 2007]. 67-72. Disponible en: http://www.fet2005.cs.buap.mx/DM/AUDIOVISUAL-PRESENTACIONES/svm-bueno-util.pdf.

22. RESENDIZ, J. A. *Las Máquinas de Soporte Vectorial para identificación en Linea.* Maestria, Control Automático. Instituto Politecnico Nacional, 2006.

23. SMOLA, A. J. *Learning With Kernels.* 2002, ISBN 0-262-19475-9.

24. CHANG, C.-C. y LIN, C.-J. *LIBSVM: a Library for Support Vector Machines* [Consultado el: 22 de junio de 2007]. 1-22. Disponible en: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

25. YANG, H. *Margin Variations in Support Vector Regression for the Stock Market Prediction.* Maestria, Ingenieria y Ciencias de la Computacion. Universidad de Hong Kong, 2003.

26. GUTIÉRREZ, H. *Modelización molecular de los receptores de adenosina y sus ligandos en el marco de diseño de fármacos asistido por ordenador* Doctorado, Ciencias Experimentales. Universidad de Pompeu Fabra, 2004.

27. ESPINOSA, G. *Modelos QSPR/QSAR/QSTR basados en sistemas neuronales cognitivos* Doctorado, URV, 2002.

28. GARDNER, A. B. *A Novelty Detection Approach to Seizure Analysis from Intracranial EEG.* Doctorado, Instituto Tecnológico de Georgia, 2004.

29. HSU, C.-W.; CHANG, C.-C., *et al. A Practical Guide to Support Vector Classification* [Consultado el: 20 de junio de 2007]. 1-12. Disponible en: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

30. NEELANJAN MUKHERJEE, S. M. *Predicting Signal Peptides with Support Vector Machines* [Consultado el: 20 de junio de 2007]. 3-6. Disponible en: http://www.springerlink.com/index/MYCRNXQQFKFGVY1R.pdf.

31. F, L.; W, M., *et al. Prediction of pK(a) for neutral and basic drugs based on radial basis function Neural networks and the heuristic method* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=16132357&dopt=Abstract.

32. HX, L.; XJ, Y., *et al. Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=16317501&dopt=Citation.

33. DEVROYE, L.; GYORFI, L., *et al. A Probabilistic Theory of Pattarn Recognition.* Editado por: Springer. 1996, ISBN 0-387-94618-7.

34. CX, X.; RS, Z., *et al. QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i05/abs/ci049820b.html.

35. CY, Z.; RS, Z., *et al. QSAR study of natural, synthetic and environmental endocrine disrupting compounds for binding to the androgen receptor* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.informaworld.com/smpp/content~content=a727214684~db=all.

36. HX, L.; CX, X., *et al. Quantitative prediction of logk of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i06/abs/ci049891a.html.

37. F, L.; W, M., *et al. Quantitative structure-activity relationship models for prediction of sensory irritants (logRD50) of volatile organic chemicals.* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=16307788&dopt=Abstract.

38. S, Y.; M, X., *et al. Quantitative structure-property relationship studies on electrochemical degradation of substituted phenols using a support vector machine.* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.informaworld.com/smpp/content~content=a757824679~db=all~jumptype=rss.

39. ---. *Quantitative structure-property relationship studies on electrochemical degradation of substituted phenols using a support vector machine.* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.informaworld.com/smpp/content~content=a757824679~db=all~jumptype=rss.

40. W, M.; F, L., *et al. Quantitative structure-property relationships for pesticides in biopartitioning micellar chromatography* [Consultado el: 22 de junio de 2007]. Disponible en: http://lib.bioinfo.pl/pmid:16490199.

41. ---. *Quantitative structure-property relationships for pesticides in biopartitioning micellar chromatography* [Consultado el: 23 de junio de 2007]. Disponible en: http://www.aapspharmaceutica.com/search/view.asp?ID=74403.

42. RÍO, B. M. D. y MOLINA, A. S. *Redes Neuronales y Sistemas Difusos.* Disponible en: http://bibliodoc.uci.cu/pdf/reg00050.pdf.

43. SCHOLKOPF, B. *Statistical Learning and Kernel Methods* 29. Disponible en: http://www.iro.umontreal.ca/~pift6266/A06/refs/scholkopf-kernel-tutorial-2000.pdf.

44. CX, X.; RS, Z., *et al. Study of the quantitative structure-mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines.* [Consultado el: 22 de junio de 2007]. Disponible en: http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2004/44/i03/abs/ci034280o.html.

45. S, W.; C, X., *et al. Study on the quantitative relationship between the structures and electrophoretic mobilities of flavonoids in micellar electrokinetic capillary chromatography* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.cababstractsplus.org/google/abstract.asp?AcNo=20043056516.

46. CHEN, N.; YANG, J., *et al. Support Vector Machine in Chemistry.* World Scientific, 2004,

47. GUNN, S. R. ***Support Vector Machines for Classification and Regression***. 1998.

48. FANG, J. *Support Vector Machines in HTS Data Mining: Type I MetAPs Inhibition Study* [Consultado el: 22 de junio de 2007]. Disponible en: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=16418315&dopt=Abstract.

49. FUMERA, G. y ROLI, F. *Support Vector Machines with Embedded Reject Option.* 2002, 3-6 p. Disponible en: http://citeseer.ist.psu.edu/552636.html.

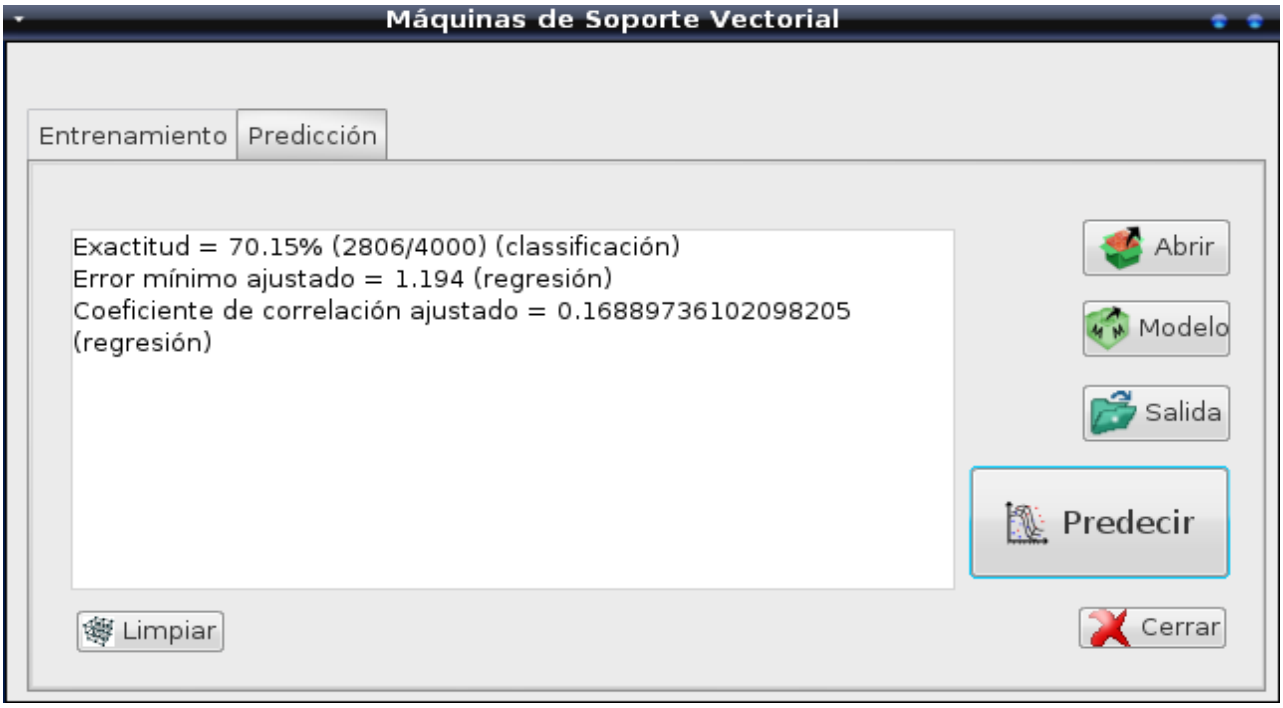
50. WELLING, M. *Support Vector Regression* [Consultado el: 20 de junio del 2007 1-3.

ANEXOS.

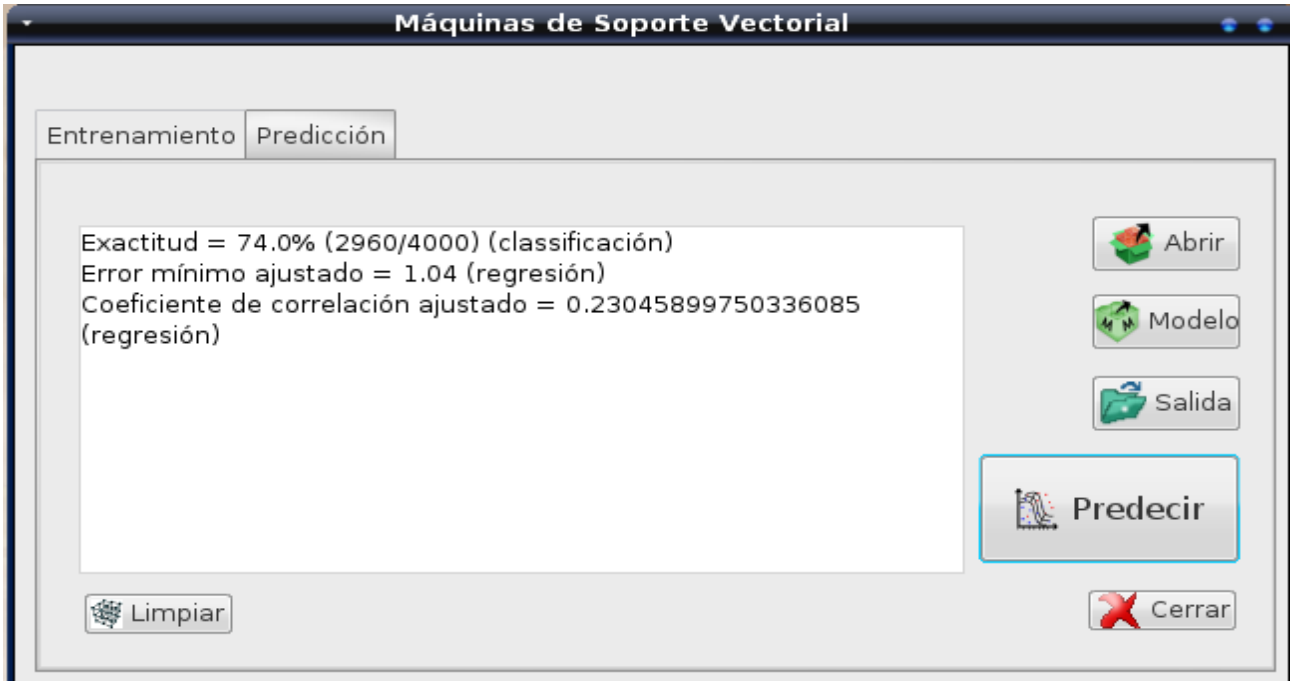
Anexo 1.

Anexo 2.

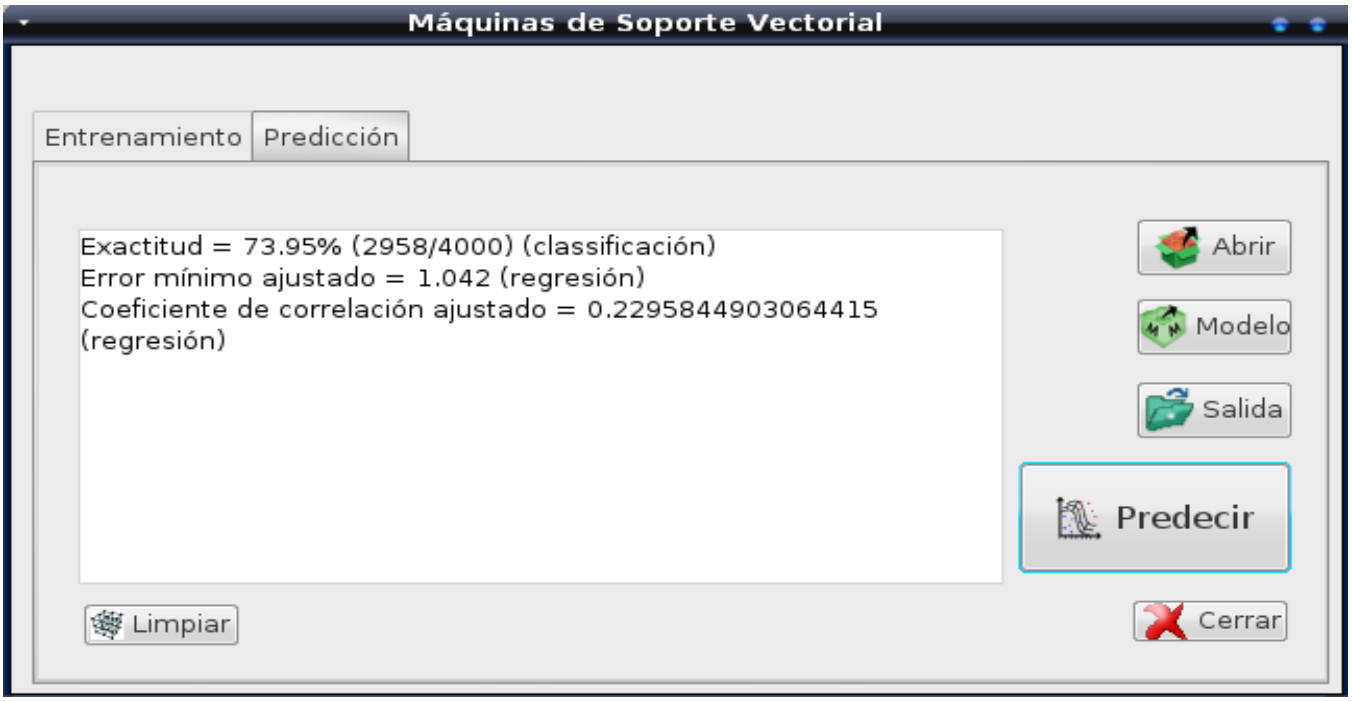
Anexo 3.



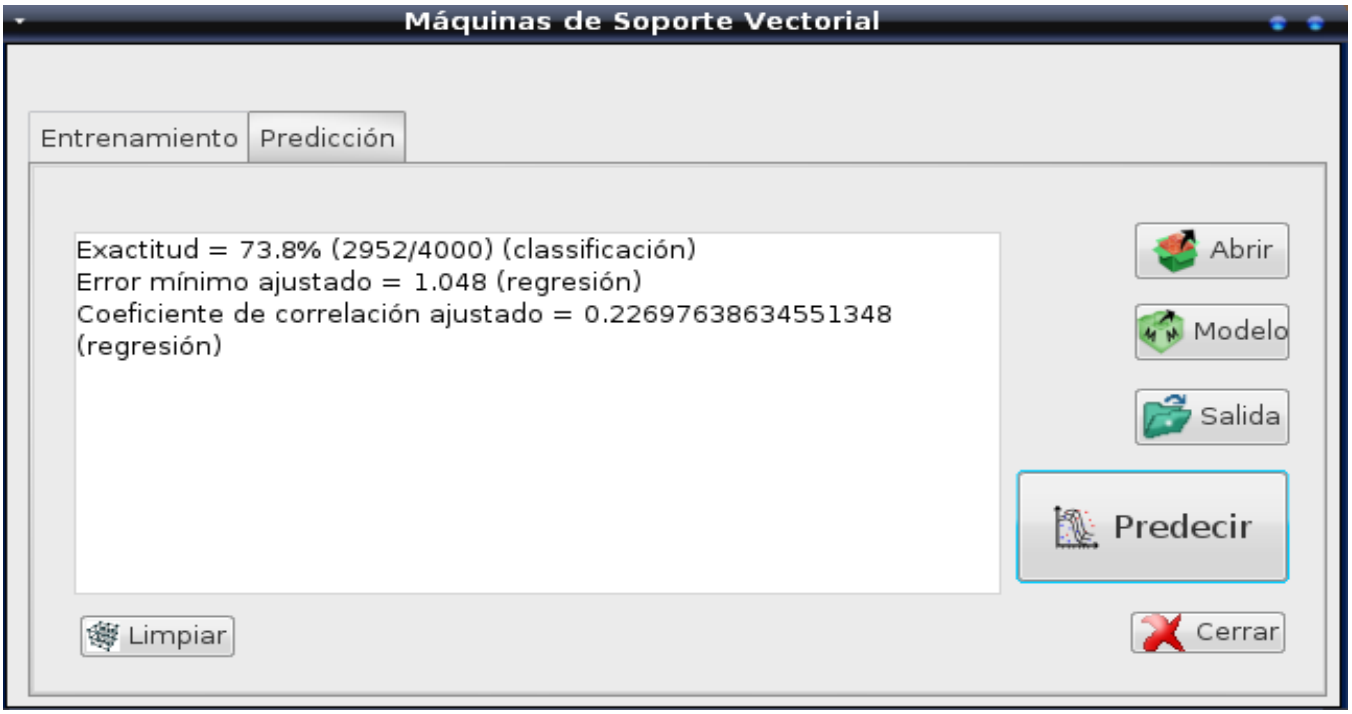
Anexo 4.



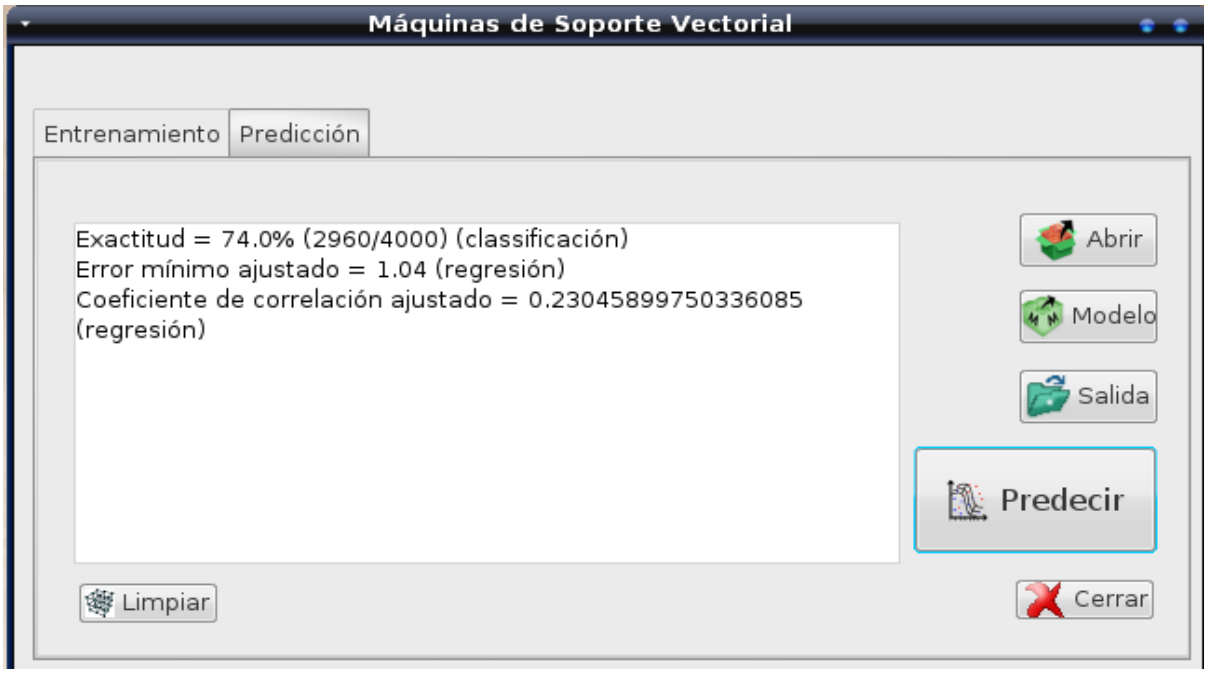
Anexo 5.



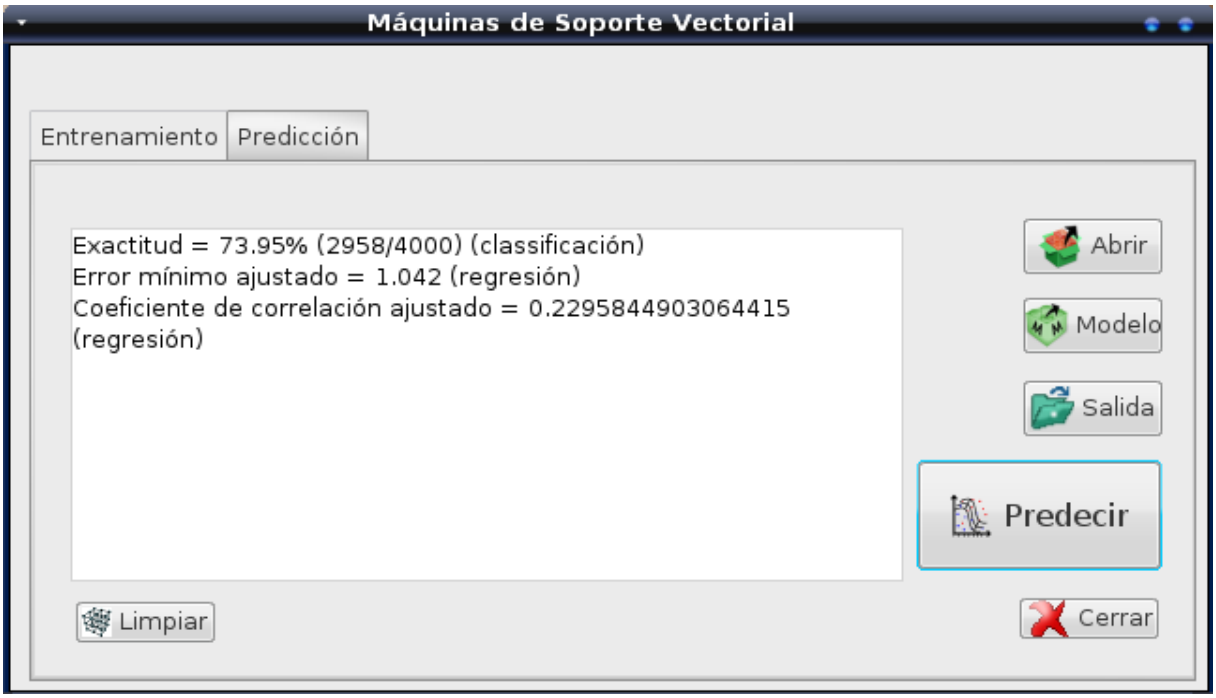
Anexo 6.



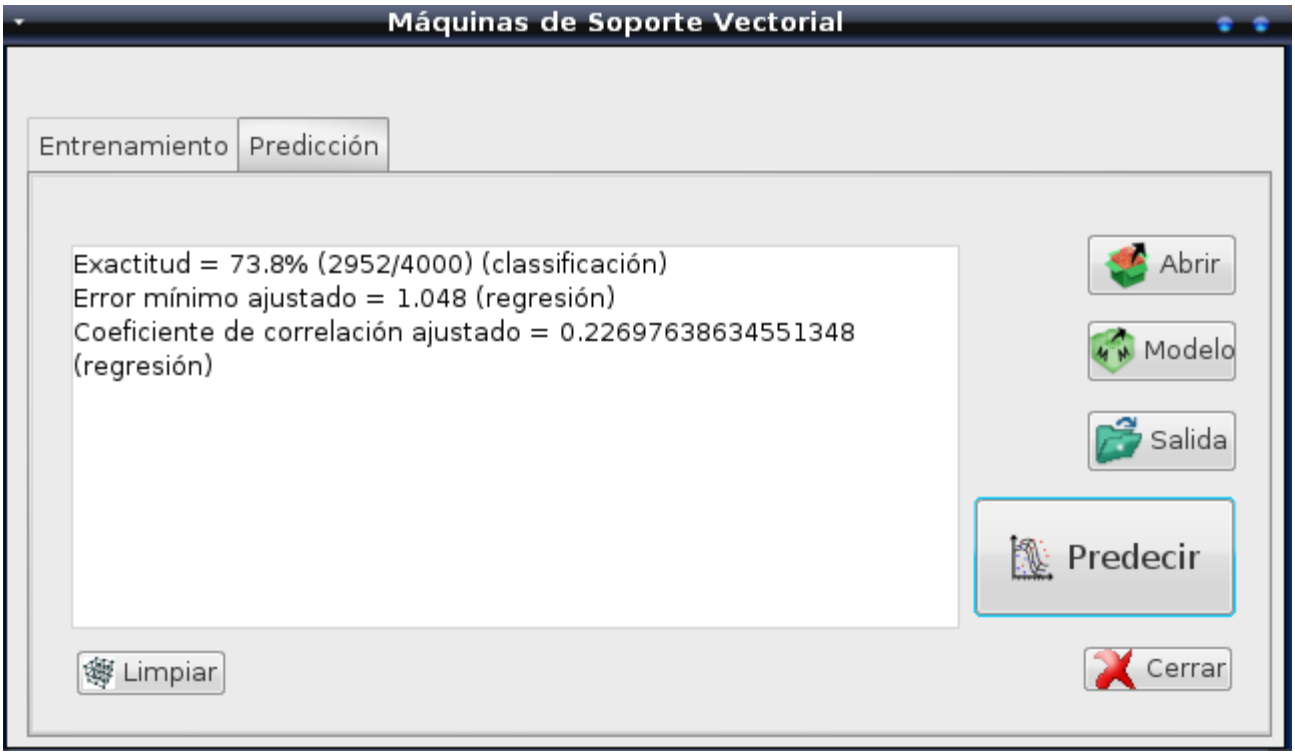
Anexo 7.



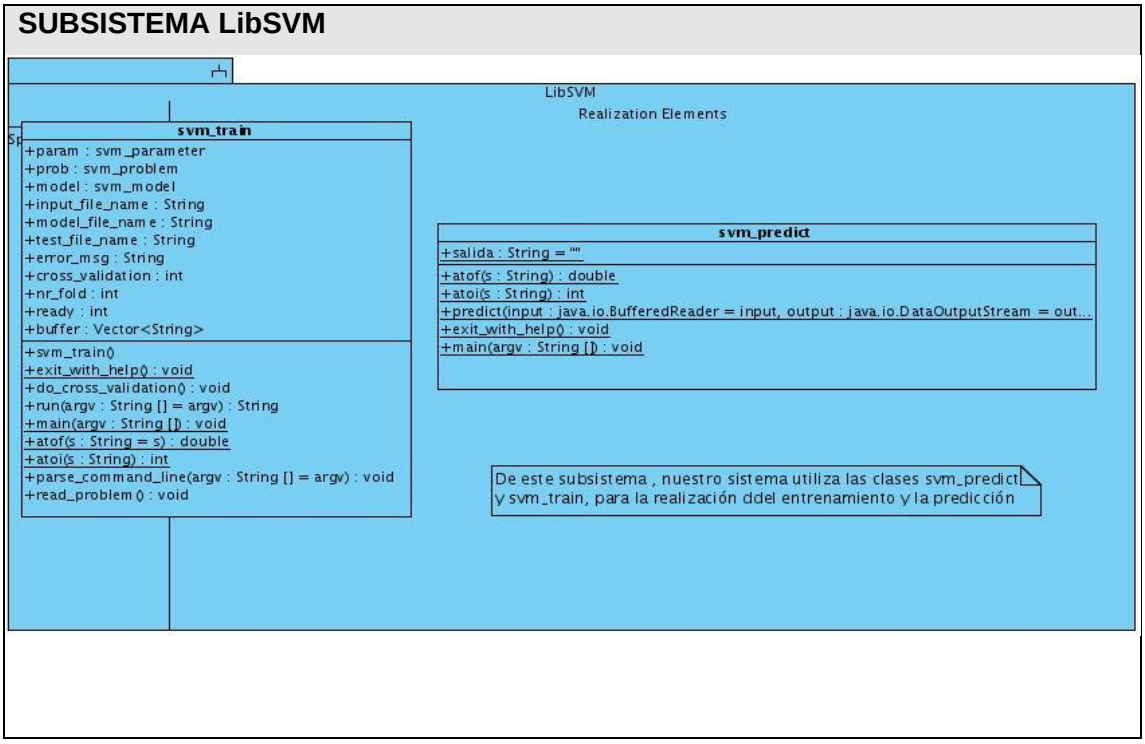
Anexo 8.



Anexo 9.



Anexo 10.



GLOSARIO DE TÉRMINOS

A

- ✓ **Actividad biológica:** Actividad que caracteriza el comportamiento biológico en compuestos químicos (Molécula o Fragmento).

B

- ✓ **Bioinformática:** Es la aplicación de los ordenadores y los métodos informáticos en el análisis de datos experimentales y simulación de los sistemas biológicos.

C

- ✓ **CQF:** Centro de Química Farmacéutica (CQF) es una institución para el desarrollo de investigaciones científico-tecnológicas dirigidas hacia la obtención de sustancias bioactivas para uso humano.
- ✓ **CASE:** Computer Aided Software Engineering (Herramientas de ingeniería de software asistida por computadora).
- ✓ **Compuestos Orgánicos:** Compuestos cuya composición fundamental es sobre la base del elemento químico carbono.

F

- ✓ **Fragmento:** Una pequeña parte de la molécula a la cual se llega a través de un método que es el encargado de fragmentarla.

M

- ✓ **Moléculas:** Una molécula es una partícula formada por un conjunto de átomos ligados por enlaces covalentes.

O

- ✓ **Open Source:** Cualidad de algunos software de incluir el código fuente en la distribución del programa. En general se usa para referirse al software libre.

P

- ✓ **Plug-ins:** aplicación informática que interactúa con otra aplicación para aportarle una función o utilidad específica.