UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERÍA

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

"APLICACIÓN DE MODELOS DE APRENDIZAJE AUTOMATIZADO PARA EL MODELADO DE PARÁMETROS DE UN BIOSENSOR ELECTROQUÍMICO"

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS

PRESENTA

Edwin R. García Curiel

DIRECTOR

Dra. Larysa Burtseva

CODIRECTOR

Dra. Margarita Stoytcheva

Mexicali, B. C. Noviembre, 2012

TESIS DEFENDIDA POR

EDWIN R. GARCÍA CURIEL

Y APROBADA POR EL SIGUIENTE COMITÉ

Dra. Larysa Burtseva
Director de Tesis

Dra. Margarita Stoytcheva
Co-Director de Tesis

Dr. Félix Fernando Gonzales Navarro
Miembro del Comité

M. C. Jorge Eduardo Ibarra Esquer
Miembro del Comité

Dr. Néstor Santillán Soto
Coordinador General del programa de Maestría y Doctorado en ciencias e Ingeniería MYDCI

RESUMEN de la Tesis de EDWIN R. GARCÍA CURIEL, presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en CIENCIAS DE LA COMPUTACIÓN. Mexicali, Baja California, México. Noviembre de 2012.

APLICACIÓN DE MODELOS DE APRENDIZAJE AUTOMATIZADO PARA EL MODELADO DE PARÁMETROS DE UN BIOSENSOR ELECTROQUÍMICO

Resumen aprobado por:

_____	_____
Larysa Burtseva	Margarita Stoytcheva
Directora de Tesis	CoDirectora de Tesis

Con el objetivo de mejorar las características analíticas de un biosensor electroquímico de acetilcolinesterasa de segunda generación se evalúa la corriente en estado estacionario producida por la interacción del pH, la temperatura y la concentración del substrato, empleando modelos de aprendizaje computacional. Las redes neuronales y las máquinas de soporte vectorial han demostrado excelentes resultados simulando la corriente, sin importar que la cantidad de muestras sea limitada. Las predicciones provistas por ambos modelos fueron evaluadas y comparadas para determinar cuál de ellas representa la respuesta generada por el sensor de una manera más exacta. Se ha localizado el área del óptimo de la corriente.

Palabras clave: Acetilcolinesterasa, Biosensor, Aprendizaje Supervisado, MSV, Red Neuronal.

ABSTRACT of the thesis, presented by EDWIN R. GARCIA CURIEL, in order to obtain the MASTER IN SCIENCE DEGREE in COMPUTER SCIENCE. Mexicali, Baja California, México. November, 2012.

APPLYING MACHINE LEARNING MODELS FOR MODELING AN ELECTROCHEMICAL BIOSENSOR PARAMETERS

Approved by:

Larysa Burtseva Thesis Advisor	Margarita Stoytcheva Thesis Co-Advisor

The steady-state current response of an acetylcholinesterase electrochemical sensor of second generation resulting from the interaction of pH, temperature and substrate concentration were evaluated to improve biosensor's analytical characteristics using computational learning models. Neural networks and support vector machines demonstrated excellent results, despite of the limited number of samples. The predictions provided by both models were compared in order to determine which generalizes better the response of the acetylcholinesterase sensor signal. It has been located the current's optimum area.

Keywords: Acethilcolinesterase, Biosensor, Machine Learning, SVM, Neural Network.

ÍNDICE

1. INTRODUCCIÓN 11
1.1. Planteamiento del problema 11
1.2. Biosensor Electroquímico 13
1.2.2. Enzima Acetilcolinesterasa 16
1.2.3. Diseño de biosensor 16
1.3. Objetivos 18
1.4. Metodología 19
1.5. Esquema general de la tesis 19
2. MODELOS DE APRENDIZAJE AUTOMATIZADO 20
2.1. Aprendizaje Automático 20
2.1.1. Modelos de Aprendizaje Automático 20
2.1.2. Aprendizaje Supervisado 22
2.2. Redes Neuronales Artificiales 23
2.2.1. Fundamentos de las Redes Neuronales 23
2.2.2. Modelo de Regresión 25
2.3. Máquinas de Soporte Vectorial 26
2.3.1. Fundamentos 26
2.3.2. Modelo de Regresión 27
2.4. Pre-procesamiento de datos 30
2.5. Evaluación y selección de modelos 31
2.5.1. Validación simple 33
2.5.2. Validación cruzada 33
2.6. Conclusiones del capitulo 35
3. ANÁLISIS DEL EXPERIMENTO ELECTROQUÍMICO 36
3.1. Experimento Electroquímico 36
3.2. Influencia de los parámetros 37
3.2.1. Concentración del Substrato 37
3.2.2. pH 38
3.2.3. Temperatura 39
3.3. Conclusiones del capitulo 41
4. MODELO PARA LA PREDICCIÓN DE PARÁMETROS 42

4.1. Proceso General 42
4.2. Elección del modelo 43
4.3. Pre-procesamiento y división de datos 44
4.4. Desarrollo del modelo de regresión 45
4.4.1. RNA: Entrenamiento y validación 45
4.4.2. MSV: Entrenamiento y validación 46
4.5. Evaluación del modelo 48
4.5.1. Pruebas: ANN 48
4.5.2. Pruebas: SVM 50
4.5.3. Análisis Comparativo 52
4.6. Conclusiones del capítulo 54
5. SIMULACIÓN Y RESULTADOS 55
5.1. Modelo Obtenido 55
5.2. Simulación 56
5.3. Resultados 57
5.3.1. Generales 57
5.3.1. Optimo 59
5.4. Conclusiones del capitulo 62
6. CONCLUSIONES 63
7. PRODUCTOS DE INVESTIGACIÓN 65
8. TRABAJO FUTURO 66
REFERENCIAS 67
ANEXOS 70

Índice de Figuras

Figura **Página**

Figura 1.1 Esquema de un Biosensor de Acetilcolinesterasa 17
Figura 2.1 Red Neuronal Biológica..... 23
Figura 2.2 Red Neuronal Artificial. 24
Figura 2.3 Ejemplo de una red MLP-BP. 25
Figura 2.4 Problema bi-clase de clasificación de datos linealmente separables..... 27
Figura 2.5 Esquema de la transformación de un conjunto de datos mediante una función núcleo de base radial..... 29
Figura 2.6 Ejemplos de diferentes ajustes en una caso de clasificacion bi-clase: 32
Figura 2.7 Funcionamiento de un K-Fold CV con K=3..... 34
Figura 3.1 El resultado máximo de la corriente resultante (z) en función del pH predefinido a 7, temperatura T (x) 25-70 ˚C, y concentración de la sustancia Cs (y) 0.2-1.0 µmoles L^{-1}..... 37
Figura 3.2 Grafica que presenta el resultado máximo de la corriente resultante (z) en función de la Temperatura T predefinida a 60 ˚C, pH (x) 5-9, y concentración de ACh (y) 0.2-1.0 µmoles L^{-1}..... 39
Figura 3.3 Grafica que presenta el resultado máximo de la corriente resultante (z) en función del pH predefinida a 7, temperatura T (x) 25-70 ˚C y concentración de ACh (y) 0.2-1.0 µmoles L^{-1}..... 40
Figura 4.1 Proceso para determinar el modelo de aproximación..... 43
Figura 4.2 Graficas de la corriente resultante de los datos de prueba contra los datos predichos por el modelo ANN-MLP. 48
Figura 4.3 Graficas comparativas de la simulación de datos realizada por el modelo ANN-MLP y de los datos experimentales: (a) Grafica de los datos experimentales; (b) Grafica de los datos predichos por la ANN-MLP. 49
Figura 4.4 Datos de la corriente resultante del conjunto de prueba contra los datos predichos por el modelo SVR. 51

Figura 4.5 Graficas comparativas de la simulación de datos realizada por el modelo SVR y de los datos experimentales: (a) Grafica de los datos experimentales; (b) Grafica de los datos predichos por la SVR. 52
Figura 4.6 Comparación de los datos experimentales de prueba contra los datos predichos por los modelos ANN-MLP y SVR 53
Figura 5.1 Simulación utilizando la escala mínima de factores. 58
Figura 5.2 Simulación perspectiva pH. 58
Figura 5.3 Comparación perspectiva Temperatura: Simulación vs datos originales 59
Figura 5.4 Simulación zonas de valores resultantes. 60
Figura 5.5 Localización del óptimo. 62

Índice de Tablas

Tabla **Página**

Tabla 1.1 Criterios de clasificación de los biosensores 14
Tabla 1.2 Biosensores con aplicaciones ambientales 15
Tabla 2.1 Modelos de Aprendizaje Automático 21
Tabla 2.2 Descripción de los métodos de normalización más utilizados. 31
Tabla 3.1 Parámetros Experimentales 36
Tabla 3.2 Resultados óptimos en función de la concentración de ACh 38
Tabla 3.3 Resultados óptimos en función de la temperatura 38
Tabla 3.4 Resultados óptimos en función del pH 40
Tabla 4.1 Datos antes y después de normalizar. 44
Tabla 4.2 Mejores resultados del algoritmo SVR con diferentes kernel 47
Tabla 4.3 Mejores resultados del algoritmo SVR 50
Tabla 4.4 Comparación Cuantitativa de los modelos 54
Tabla 5.1 Escala de factores para simulación 56
Tabla 5.2 Intervalos de localización del optimo. 61

Índice de Anexos

Anexo	**Página**
Anexo A. Estructura de Archivos Anexados	70
Anexo B. Conjunto Original de Datos	70
Anexo C. Conjunto de Datos de Entrenamiento/Validación.	70
Anexo D. Conjunto de Datos de Prueba	70
Anexo E. Predicción de Conjunto de Pruebas: Red Neuronal	70
Anexo F. Predicción de Conjunto de Pruebas: SVM	70
Anexo G. Simulador de las funciones de regresión: SVM y Red Neuronal (Matlab)	70

1. INTRODUCCIÓN

1.1. Planteamiento del problema

Dentro de los químicos que dañan al medio ambiente, se encuentran los inhibidores de la enzima acetilcolinesterasa (AChE), tales como los organofosfatos y carbamatos, que son insecticidas predominantes en la agricultura por su gran utilidad, y al mismo tiempo representan graves riegos para la salud y el medio ambiente.

La forma de detección más utilizada con los inhibidores de AChE, en especial aquellos basados en carbamatos y organofosfatos, es el uso de biosensores. Un *biosensor* es un dispositivo capaz de realizar un análisis cuantitativo de la señal resultante de una reacción química entre un compuesto biológico y alguna otra sustancia.

Los biosensores electroquímicos miden los cambios eléctricos que ocurren cuando el elemento biológico reacciona con alguna sustancia química, en este caso la reacción producida entre la enzima AChE y el neurotransmisor acetilcolina (ACh). Este tipo de reacción química depende mucho de factores externos, tales como el pH y la temperatura, y así como afectan a la reacción, de igual manera afectan a la medición realizada con el biosensor, tanto en la exactitud, como en la potencia eléctrica observada. El pH toma diferentes valores, uno de los cuales muestra la mayor potencia eléctrica y es considerado como el valor óptimo. En el caso de la temperatura, esta hace que la corriente eléctrica tenga un comportamiento exponencial, sin embargo, llega a un punto máximo (alrededor de 60 °C) donde la enzima sufre un proceso de desnaturalización, y la corriente eléctrica baja hasta casi desaparecer.

Teniendo en cuenta que la corriente producida por la reacción química depende de la interacción de todas las variables, y que no se explica mediante simple observación, se requiere el uso de enfoques matemáticos y estadísticos, para modelar, definir y optimizar el efecto que tiene cada una de las variables en la reacción química.

Se han hecho diferentes esfuerzos de modelar la corriente resultante de diversos tipos de biosensores, de donde destacan:

- El modelado digital de la respuesta de la corriente en estado estacionario obtenida por un biosensor de tipo amperométrico en un sistema de glucosa (Mell & Malloy, 1975), mantiene una sólida base matemática, pero el desarrollo del modelo está comprometido a simular bajas mediciones eléctricas, de lo contrario su margen de error aumenta de gran manera.
- El modelado de la corriente resultante de un biosensor electroquímico de oxigeno (Rangelova, et al., 2010), sin embargo solo muestra los modelos utilizados y no la manera en la que fueron implementados, ni cómo se realizaron las pruebas.
- La implementación de un chip utilizando una red neuronal para modelar la corriente de un biosensor enzimático, muestra el proceso utilizado, pero no habla de la cantidad de muestras utilizadas para el entrenamiento del modelo, ni de cómo fueron realizadas las pruebas (Alonso, et al., 2010).

A diferencia de las investigaciones anteriores, se plantea en esta tesis que el modelo de regresión computacional que se propone, determinará la señal eléctrica producida por la interacción entre la enzima AChE y el substrato ACh en una sustancia con una temperatura y pH determinados; el modelo será seleccionado de un análisis comparativo entre diferentes algoritmos de regresión, a través del margen de error de cada algoritmo. Además, todo el desarrollo del modelo contará con una sólida base matemática y estadística, a través de un proceso bien definido:

- Análisis de datos,
- Pre-procesamiento,
- Configuración del algoritmo,
- Entrenamiento y
- Validación,

y su evaluación:

- Pruebas,
- Análisis comparativo,

lo cual permitirá conocer y analizar la importancia que tiene cada uno de los factores en la reacción.

1.2. Biosensor Electroquímico

1.2.1 Clasificación de biosensores

De acuerdo a la definición de IUPAC (*International Union of Pure and Applied Chemistry*), un biosensor es un dispositivo analítico que combina un elemento biológico para reconocimiento molecular con un dispositivo de procesamiento de señal llamado transductor, el cual permite detectar y medir la señal producida por la interacción del elemento biológico y la sustancia de interés, de manera rápida y precisa. El transductor, que normalmente asegura una alta eficiencia del sensor, puede ser térmico, óptico, magnético, nanomecánico, piezoeléctrico o electroquímico. Por otro lado, la selectividad de la detección es asegurada por el elemento biológico de reconocimiento (*biological recognition element*), el cual está basado en un bioligando (ácido desoxirribonucleico DNA, ácido ribonucleico RNA, anticuerpos, etc.) o en un biocatalizador (algunas proteínas redoxantes, enzimas individuales y sistemas enzimáticos, como células, membranas, micro-organismos completos, telas) (Thévenot, 1999). En la Tabla 1.1 se presentan varios criterios de clasificación de los biosensores que consideran:

- Tipo de interacción entre los componentes de la reacción a detectar;
- Forma de detección de la interacción;
- Elemento biológico a reconocer ;
- Tipo de sistema de transducción que posee el dispositivo.

Tabla 1.1 Criterios de clasificación de los biosensores

Tipo de Interacción	**Detección de la Interacción**
∗ Biocatalítica ∗ Bioafinidad	∗ Directa ∗ Indirecta
Elemento de Reconocimiento	**Sistema de Transducción**
∗ Enzima ∗ Orgánulo, tejido o célula completa ∗ Receptor biológico ∗ Anticuerpo ∗ Acido nucleicos ∗ PIM, PNA, aptámero	∗ Electroquímico ∗ Óptico ∗ Piezoeléctrico ∗ Termométrico ∗ Nanomecánico

Los biosensores con transducción electroquímica se dividen en dos tipos:

- Potenciométricos;
- Amperométricos.

Los biosensores que solo miden un cambio de potencial en la interface sensor-analito con respecto al electrodo de referencia se conocen como sensores o biosensores potenciométricos. Los sensores en los que se impone un potencial externo para efectuar la transformación electroquímica son conocidos como biosensores amperométricos. La inhibición de la enzima en estos biosensores es monitoreada por el cambio de la corriente de oxidación o reducción a detectar, a un cierto potencial. En la Tabla 1.2 se describen diversos biosensores ambientales de acuerdo al tipo de transductor, elemento biológico utilizado, compuesto a detectar y área donde se busca el analito.

Tabla 1.2 Biosensores con aplicaciones ambientales

Tipo de Transductor	Elemento Biológico	Analito	Área
Electroquímico	Anticuerpos	Atrazina	*
Óptico	Anticuerpos	Simazina	Suelo, superficie del agua
Óptico	Anticuerpos	Pesticidas	Agua de ríos
Electroquímico	Anticuerpos	Tensoactivos (alquilfenoles y sus etoxilatos)	*
Electroquímico	Anticuerpos	Estradiol	*
Electroquímico	Anticuerpos	Eschenchacoll	Agua potable
Óptico	Anticuerpos	Salmonella enteriditidLymeria monocytogenes	*
Acústico	Anticuerpos	Salmonella typhymurium	*
Óptico	Enzima (AChE)	Compuestos organofosforados	Agua
Electroquímico	Enzima (AChE)	Paraoxon y carbofuran (pesticidas)	Aguas negras
Electroquímico	Enzima (tirosinasa)	Fenoles	Suelos, lodos, aguas (después de extracción)

1.2.2. Enzima Acetilcolinesterasa

La AChE es una parte clave en la trasmisión colinérgica ya que cataliza la hidrólisis del neurotransmisor acetilcolina en un acetato y colina. Los agentes inhibidores de esta enzima, tales como los carbamatos y organofosfatos interfieren con este proceso, lo cual conduce a un deterioro grave de las funciones nerviosas. Sin embargo, los inhibidores de AChE son los insecticidas que predominan en la agricultura. Estos insecticidas son muy útiles y al mismo tiempo son muy dañinos para la salud.

La inhibición de AChE provoca las siguientes reacciones:

- Disminución de presión intraocular;
- Bradicardia;
- Hipotensión;
- Hipersecreción;
- Bronco constricción;
- Contracción muscular prolongada;
- Casos mortales.

1.2.3. Diseño de biosensor

Las mediciones fueron realizadas utilizando un biosensor electroquímico, formado por un par de electrodos, un contra-electrodo (electrodo auxiliar) y un electrodo referencial, en conjunción con un electrodo aislado (electrodo de trabajo), el cual en su parte inferior, tiene colocada una porción (mg/cm^3) de la enzima AChE (Ringsdorf Werke, Germany). En la Figura 1.1 se muestra el esquema del biosensor de AChE utilizado en el experimento.

El biosensor (electrodo de trabajo) utilizado en el experimento se elaboró inmovilizando la enzima acetilcolinesterasa por enlaces químicos sobre la superficie de un electrodo de grafito. El electrodo auxiliar era de carbón vítreo y el electrodo de referencia – de calomelanos saturado.

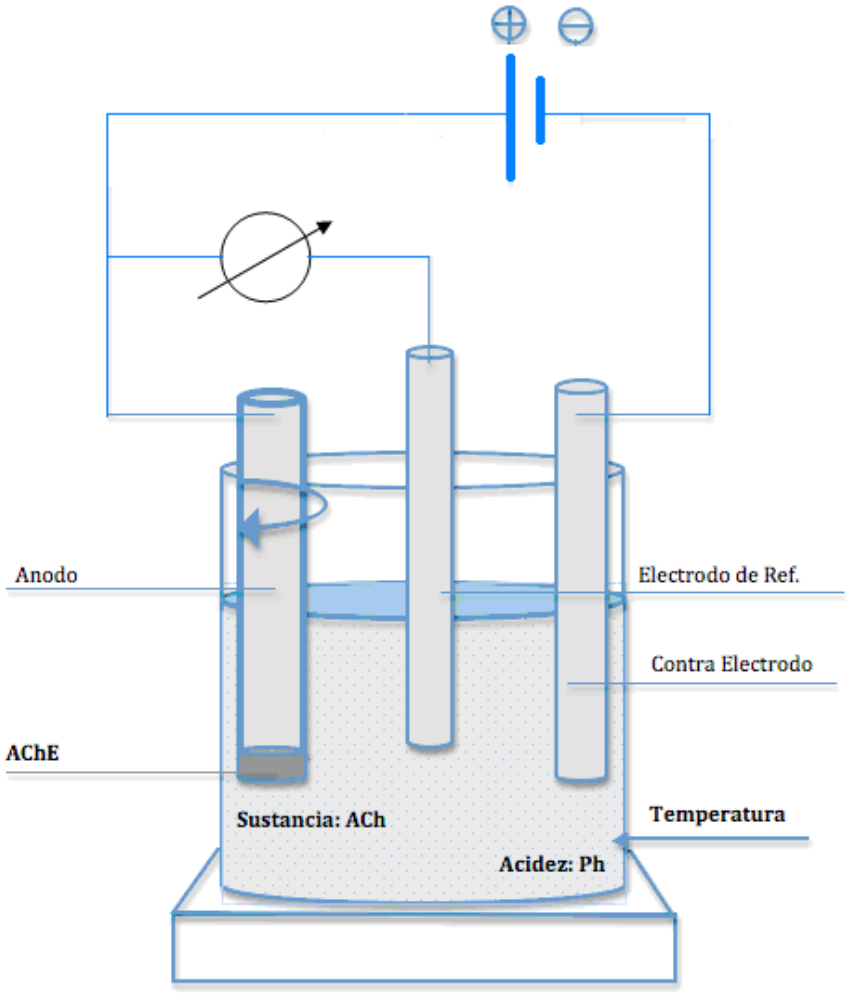


Figura 1.1 Esquema de un Biosensor de Acetilcolinesterasa

Posteriormente se sumergen los electrodos en una celda electroquímica convencional, la cual contiene una solución tampon con una determinada concentración (mMoles/L) del sustrato ACh. La solución recibió un ajuste determinado en su acidez (pH) y su temperatura (˚C). Durante el estudio el electrodo mantuvo una rotación de 1000 rpm (rotaciones por minuto).

Las ecuaciones de las reacciones químicas, donde interviene la AChE, se muestran a continuación:

$$\text{ACh} + H_20 \xrightarrow{\text{AChE}} \text{Ch} + \text{Ácido Acético} \quad (1.1)$$

$$2\text{Ch} = \text{Ch-Ch} + 2e^- + 2H^+ \quad (1.2)$$

Durante el estudio se midió la corriente producida por la reacción de oxidación del producto Ch (Ec. 1.2) de la reacción enzimática (Ec. 1.1).

1.3. Objetivos

Objetivo General

A partir del análisis numérico y modelación matemática, encontrar la combinación de niveles de concentración de ACh, pH y temperatura, que maximicen la sensibilidad de la determinación de la ACh.

Objetivos Específicos

- Diseñar un modelo de regresión basado en aprendizaje automatizado, que mejor se ajuste al comportamiento de la variable respuesta del biosensor.
- Encontrar el valor máximo de la corriente, utilizando el modelo de regresión desarrollado.

1.4. Metodología

1. Revisión de métodos y factores para modelar la corriente resultante en biosensores enzimatico-amperometricos.
2. Estudio e interpretación de la interacción entre los parámetros a optimizar.
3. Elección de métodos para la modelación de la variable respuesta.
4. Diseño y programación de modelos de regresión.
5. Simulación y evaluación del modelo de regresión.
6. Análisis e interpretación de resultados.

1.5. Esquema general de la tesis

En el capítulo 2 se presenta un breve estado del arte del aprendizaje automático, así como el de las Redes Neuronales y de las Maquinas de Soporte Vectorial como algoritmos de aprendizaje orientados a problemas de regresión. De igual manera se expone la forma en la que se evaluarán los diferentes modelos de aprendizaje a utilizar.

El capítulo 3 se enfoca a explicar el conjunto de datos utilizado, la manera en la que se obtuvo, sus diferentes variables y la relación que se presentan entre estas.

En el capítulo 4 se describe el proceso de desarrollo de los diferentes modelos, el pre procesamiento del conjunto de datos, la búsqueda de la configuración de los algoritmos, así como la evaluación y selección del mejor modelo. Se presentan comparaciones cualitativas y cuantitativas.

En el capítulo 5 se muestra el análisis de resultados provistos por una simulación de datos utilizando el modelo de regresión desarrollado en el capítulo anterior, con el mínimo incremento perceptible de los factores C_s, Temperatura y pH.

En el capítulo 6 y 7 se presentan las conclusiones de la investigación y los productos que de ella se derivan, respectivamente, finalizando la tesis.

2. MODELOS DE APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es una conocida rama de la inteligencia artificial que permite desarrollar algoritmos computacionales que "aprendan" el comportamiento de fenómenos de diferentes campos de investigación. Con el objetivo de entender y recrear el funcionamiento del biosensor de AChE, a continuación se describen algunos modelos de aprendizaje automático, de la misma manera se presentan algunos otros métodos que intervienen en el proceso de aprendizaje de estos modelos.

2.1. Aprendizaje Automático

2.1.1. Modelos de Aprendizaje Automático

Se le llama Aprendizaje Automático al estudio de algoritmos computacionales, los cuales "aprenden" automáticamente a través de la experiencia; son considerados modelos de inducción de conocimiento (Mitchell, 1997). Este tipo de aprendizaje se considera importante en diversas áreas de investigación (ver Tabla 2.1).

Por el tipo de problema que resuelve, el aprendizaje automático se clasifica de la siguiente manera:

- **Aprendizaje supervisado** entrena un modelo que tiene una correspondencia entre los valores de entrada (conjunto de entrenamiento) y valores de salida (etiquetas);
- **Aprendizaje no supervisado** (*clustering*) busca la agrupación natural de los datos según su similitud, encontrando patrones distintivos de cada muestra. A diferencia de otros métodos de aprendizaje, el no supervisado no cuenta con las etiquetas distintivas de cada muestra (clases).
- **Aprendizaje semi-supervisado** es la mezcla del aprendizaje supervisado y no supervisado, con el objetivo de predecir de manera adecuada las etiquetas de clase;

Tabla 2.1 Modelos de Aprendizaje Automático

Modelo	Forma de Aprendizaje	Campos de aplicación	Referencias
Clasificadores Bayesianos	Clasificación, Regresión	Minería de datos	Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009.
Redes Neuronales	Clasificación, Regresión, Clustering	Visión por computador, reconocimiento de patrones	Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009; Mitchell, 1997
Máquinas de Soporte Vectorial	Clasificación, Regresión	Bioinformática, minería de texto, procesamiento de imágenes, quimiometría.	(Guide, 2003; Hammel, 2009; Pedroza, 2007; Smola, 2012)
Arboles de decisión	Clasificación, Regresión, Clustering	Minería de datos	Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009; Mitchell, 1997
K- NN	Clasificación, Clustering	Minería de datos	Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009.
K-Means	Clasificación, Clustering	Minería de datos.	(Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009.

- **Aprendizaje por refuerzo** incluye algoritmos, los cuales se entrenan bajo el modelo de "fallo o acierto" (premio o castigo) desarrollado según la decisión que este haya obtenido en base a una percepción tomada de su entorno.

En la presente investigación se aplican técnicas de aprendizaje supervisado, por la necesidad de modelar la función producida por la corriente de respuesta *I* como una función

de *ACh*, *pH* y *t* en un biosensor de AChE, con el objetivo de optimizar la exactitud de medición bajo diferentes condiciones. Esta corriente no se interpreta por expresiones lineales, sino por métodos basados en análisis de regresión que aproximen procesos multifactoriales.

2.1.2. Aprendizaje Supervisado

La idea principal del Aprendizaje Supervisado es realizar el entrenamiento de diferentes modelos computacionales de predicción utilizando un conjunto de datos etiquetados (categorizados), con el objetivo de que estos generen las etiquetas de clase para futuros datos desconocidos. Estos modelos producen una función que establece una correspondencia entre las entradas y las salidas deseadas de un modelo.

Un modelo de predicción representa un conjunto de datos $\mathbf{D} = \{x_i, y_i\}$, $i = 1, 2,\ldots, n$, donde $\mathbf{x} \in \mathbf{R}^d$ siendo el conjunto de datos de entrada, y el vector $\boldsymbol{y}$ las etiquetas de clase, que pueden ser continuas ($\mathbf{y} \in \mathbf{R}$) o discretas ($\mathbf{y} \in \{0,1\}$), dependiendo del tipo de predicción que se utilice. Existen dos formas de predicción conocidas como clasificación y regresión.

Clasificación: Cuando las etiquetas y_i son finitas o por lo menos discretas, su forma más conocida es la llamada clasificación binaria o bi-clase, donde $\mathbf{y} = \{0,1\}$ o $\mathbf{y} = \{-1,1\}$ y $\mathbf{x} \in \mathbf{R}^d$. Su función es predecir la etiqueta de clase de cada muestra nueva, ajena al modelo original. En este caso siempre se obtendrán valores idénticos a las etiquetas originales.

Regresión: Cuando las etiquetas y_i son continúas, de donde se deriva que $\mathbf{y} \in \mathbf{R}$, y el $\mathbf{x}$ (conjunto de datos de entrada) es un subconjunto de $\mathbf{R}^d$, su función es predecir la etiqueta de clase para cada muestra que no pertenece al conjunto original, sin embargo no se obtienen los mismos valores de **y**, en cambio se genera una **y'** que tiende a valores iguales o semejantes a **y.**

2.2. Redes Neuronales Artificiales

2.2.1. Fundamentos de las Redes Neuronales

Las redes neuronales son modelos computacionales basados en la estructura de conexiones de las células nerviosas encontradas en el cerebro de los seres vivos, y de igual manera intentan imitar su funcionamiento (Russell & Norvig, 2010; Wolfgang, 2011; Luger, 2009; Mitchell, 1997). En la Figura 2.1 se presenta el esquema de una red neuronal biológica.

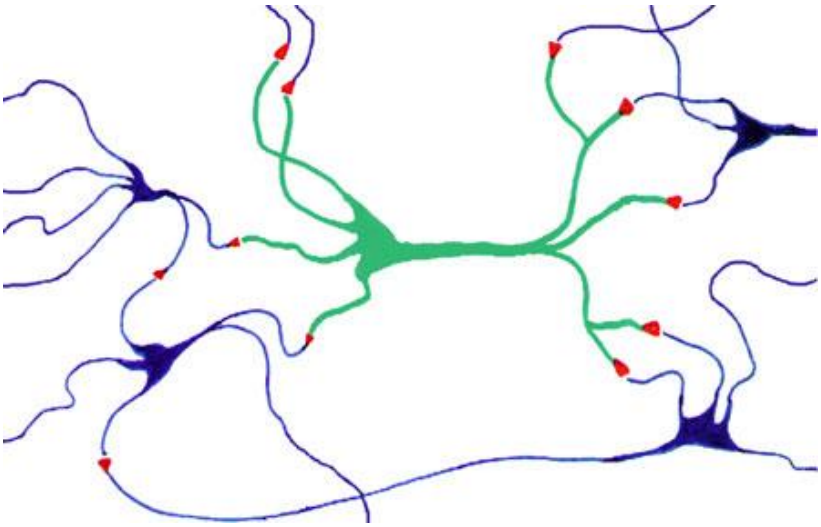


Figura 2.1 Red Neuronal Biológica

Una red neuronal está conformada por una o más unidades (neuronas), las cuales poseen una serie de conexiones, que sirven como entradas (dendritas), salidas (axiomas), o para comunicar una unidad con otras (sinapsis). Cada conexión tiene un determinado peso numérico asociado, con excepción de los axiomas, ya que estos no están conectados con otras neuronas. Una neurona procesa la información que recibe y obtiene un resultado, el cual se considera como una entrada a otra neurona o como un resultado final. En la Figura 2.2 se muestra el esquema de una red neuronal artificial obtenida a partir de la red neuronal biológica presentada en la Figura 2.1.

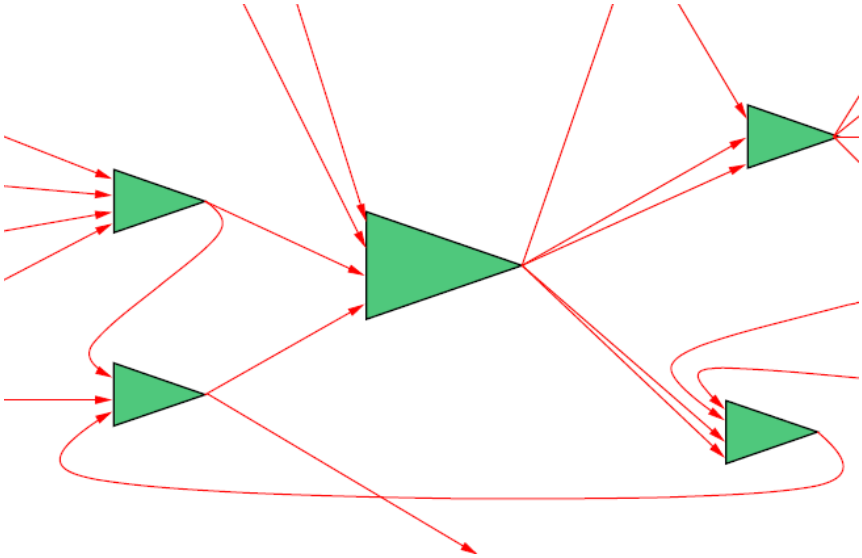


Figura 2.2 Red Neuronal Artificial.

Una red neuronal funciona de la siguiente manera. La entrada representa un vector **x** = (x_1, …, x_n), donde $x_j \in \mathbf{R}$, $j = 1,\ldots, n$. A cada elemento de entrada x_j se le asocia un peso numérico w_{ji}, donde i es el índice de neurona. Posteriormente estos valores se suman en cada neurona y se les aplica una función de transferencia (activación) f, la cual determina el valor de salida y_i.

$$y_i = f\left(\sum_{j=1}^{n}\left(w_{ji}x_j\right)\right) \tag{2.1}$$

La función de activación se elige previamente por ser una función de paro, pero por simplicidad matemática se utilizan funciones tangentes hiperbólicas (2.2) o funciones sigmoides (2.3):

$$\tanh(x) = \frac{1-e^{-x}}{1+e^{-x}} \tag{2.2}$$

$$\mathrm{sig}(x) = \frac{1}{1+e^{-x}} \tag{2.3}$$

Existen diferentes arquitecturas de redes neuronales, desarrolladas para distintos tipos de problemas. Si el problema consiste en modelar (regresión) o predecir (clasificación) datos, se emplean arquitecturas como el perceptrón de una sola capa (*single layer perceptron*, SLP), para problemas linealmente separables, o el perceptrón multicapa (*multilayer perceptron*, MLP), para problemas de alta dimensión o que no son linealmente separables. Además, estas arquitecturas se basan en diferentes algoritmos de aprendizaje, como el algoritmo *Least-Mean-Square* (LMS), o el algoritmo de propagación hacia atrás del error (*error backpropagation algorithm*, BP), considerado como el algoritmo de aprendizaje más utilizado por las redes de tipo MLP.

2.2.2. Modelo de Regresión

Se considera a la red MLP con algoritmo de entrenamiento BP (MLP-BP) como una de las mejores arquitecturas de redes neuronales utilizadas para tareas de regresión, por su adaptabilidad a diferentes problemas. En este caso, el problema de regresión consiste en aproximar una posible función no lineal f (**x**) con una red neuronal Y(**x**) de arquitectura MLP-BP**,** donde $\mathbf{x} \in \mathbf{R}^n$. En la Figura 2.3 se presenta el esquema de una red MLP-BP.

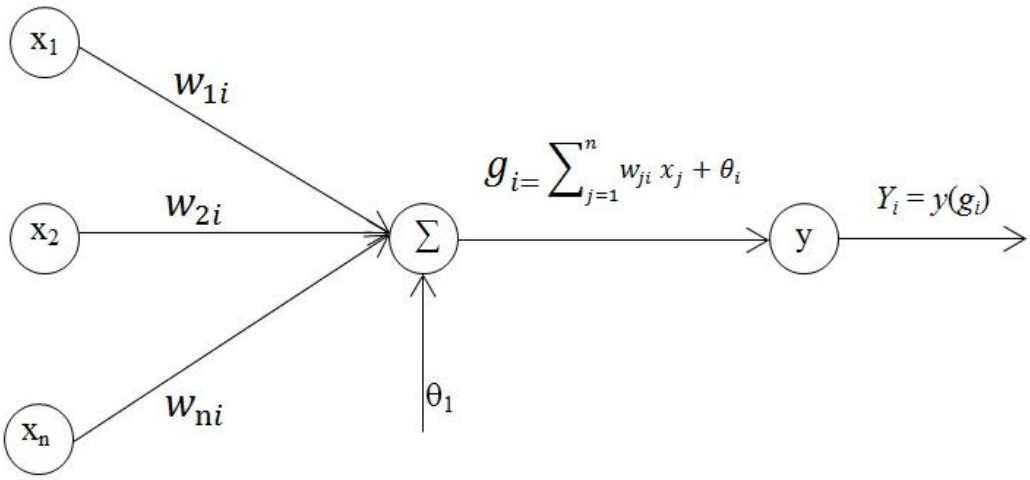


Figura 2.3 Ejemplo de una red MLP-BP.

Las entradas x_j, $j = 1,..., n$, a la neurona i se multiplican por los pesos w_{ji} y se suman al valor de sesgo θ_i. El resultado g_i es la entrada a la función de activación Y_i. El nodo de salida i se convierte en:

$$Y_i = y\left(\sum_{j=1}^{n} w_{ji}\, x_j + \theta_i\right) \tag{2.3}$$

2.3. Máquinas de Soporte Vectorial

2.3.1. Fundamentos

Las máquinas de soporte vectorial (MSV) son métodos de aprendizaje supervisado que generan una función de mapeo de un conjunto de datos de entrenamiento previamente etiquetado. Las funciones de mapeo pueden ser tanto una función de clasificación como una de regresión.

En el caso de clasificación, el mapeo se utiliza para transformar los datos de entrada, cuando estos no son linealmente separables, de modo que se elevan a una dimensión mayor, donde serán separables en contraste con la dimensión original. El modelo producido dependerá de solo un subconjunto de los datos originales de entrada cuya característica es que se crean los límites que separan una clase de datos, de las otras. Los datos que pertenecen a ese subconjunto reciben el nombre de vectores de soporte. De esta manera, se observa una distancia (margen) máxima entre las diferentes clases (Figura 2.1).

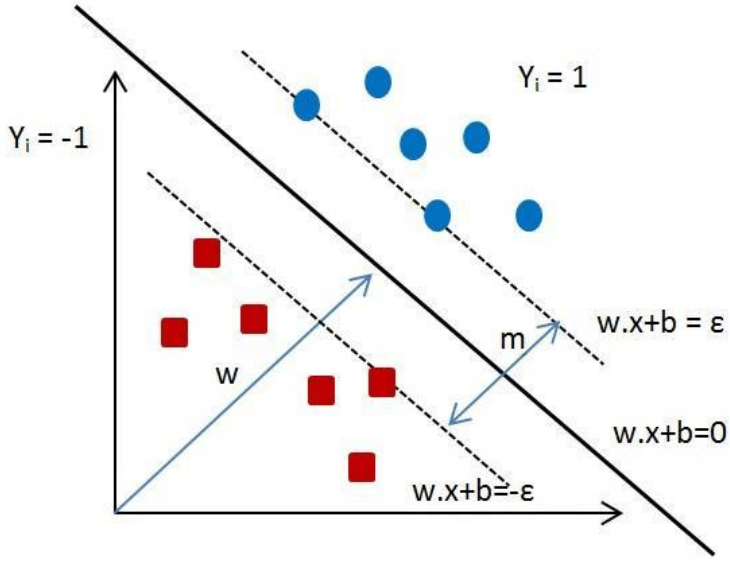


Figura 2.4 Problema bi-clase de clasificación de datos linealmente separables.

El objetivo de las MSV es crear un modelo computacional previamente entrenado, que prediga la etiqueta de clase a cada nueva muestra de datos.

Además de su sólida base matemática radicada en la teoría del aprendizaje estadístico, las MSV han demostrado un rendimiento altamente competitivo en numerosas aplicaciones, tales como bioinformática, minería de textos, reconocimiento de rostros, procesamiento de imágenes, las cuales han establecido a las MSV como una de las herramientas base del aprendizaje automático.

2.3.2. Modelo de Regresión

Un problema de regresión se expresa en función de las MSV de la siguiente manera:

Esta dado un conjunto de datos de entrenamiento $\mathbf{D} = \{x_i, y_i\}$, $i = 1, \ldots, l$, de vectores de entrada $x_i \in \mathbf{R}^n$ y sus respectivas etiquetas $y_i \in \mathbf{R}^1$. A través de los parámetros $C > 0$ y $\varepsilon > 0$, la forma estándar de las MSV aplicadas a un problema de regresión es (Vapnik, 1998):

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^*$$

s. a

$$\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i \leq \varepsilon + \xi_i,$$

$$y_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b \leq \varepsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \dots, l.$$

Aquí los vectores de entrenamiento x_i son mapeados (de ser necesario) por la función $\phi(x_i)$ a un espacio de mayores dimensiones, donde la MSV encuentra un híper plano de separación lineal con la máxima distancia (margen); $C > 0$ es el parámetro de regularización (penalización) del error; ξ_i, ξ_i^* son variables primarias.

El método de dualización estándar (*standard dualization*) aplicando multiplicadores de Lagrange es:

$$min_{\alpha,\alpha^*} \quad \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T Q(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \qquad (2.4)$$

s. a

$$\boldsymbol{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0,$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \dots, l,$$

donde

$$Q_{ij} = K(x_i, x_m) \equiv \phi(x_i)^T\phi(x_m).$$

Después de resolver el problema (2.4), la función de aproximación es

$$g_{(x)} = \sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x) + b,$$

donde ε es una constante predefinida que controla la tolerancia al ruido. Con la función de perdida insensible ε, el objetivo es encontrar la función $g_{(x)}$ cuya desviación tiene a lo sumo el valor de la función de perdida ε a partir de las etiquetas obtenidas t_i para todos los datos de entrenamiento, y al mismo tiempo debe ser tan plana como sea posible. En otras palabras, al algoritmo de regresión no le importan los errores, siempre y cuando sean menores que ε, pero no aceptaran ninguna desviación más grande que ε.

A la función definida por:

$$K\ (x_i, x_m) \equiv \phi\ (x_i)^T (x_m)$$

se le conoce como función núcleo. Aunque los investigadores proponen continuamente nuevas funciones núcleo, se sugiere utilizar los siguientes (Pedroza, 2007), (Hammel, 2009), (Guide, 2003), (Chang, 2011):

- Lineal: $K(x_i, x_m) = x_i^T\ x_m$.
- Polinomial: $K(x_i, x_m) = (x_i^T x_m + r)^d$, d > 0.
- Función de base radial: $K(x_i, x_m) = exp(-\gamma \|x_i - x_m\|^2),\ \gamma > 0$.
- Sigmoidal: $K(x_i, x_m) = \tanh(\gamma {x_i}^T x_m + r)$.
- γ, r y d son parametros especificos de las diferentes funciónes nucleo.

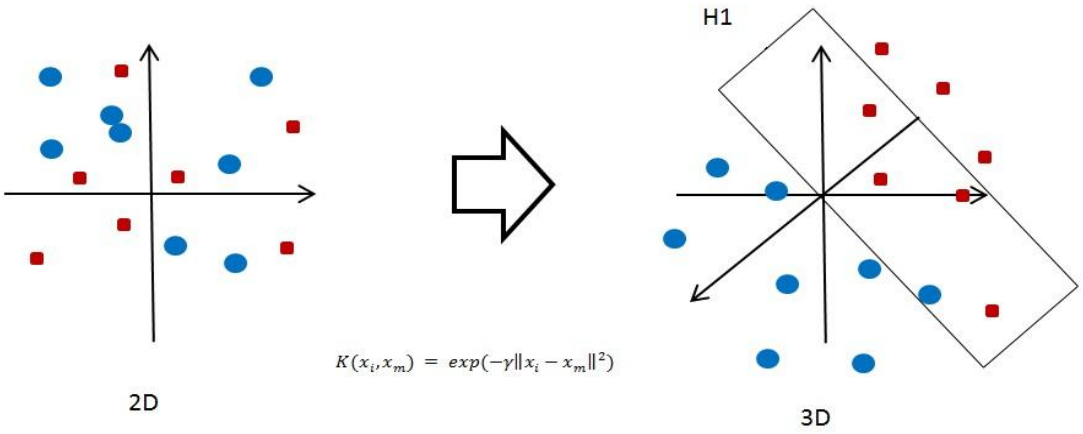


Figura 2.5 Esquema de la transformación de un conjunto de datos mediante una función núcleo de base radial

No es posible conocer de manera previa la función núcleo que mejor cumpla con los objetivos previamente descritos, por lo que el procedimiento utilizado para encontrarla no ha sido establecido. La forma de conocer la función ϕ a utilizar, está basada en la realización de diferentes ensayos con funciones ϕ conocidas. A este procedimiento se le conoce como **búsqueda de rejilla (***grid search***)**, además que algunas funciones requieren de parámetros específicos (polinomial, radial, sigmoidal, entre otras) los cuales determinan que tanto mejorará la separabilidad entre las clases o hasta que dimensión será elevada. El consumo de esfuerzo computacional dependerá de las dimensiones de la búsqueda de rejilla (cantidad de funciones a comparar y la cantidad de diferentes parámetros de cada función a utilizar), así como el tamaño del conjunto de datos.

Aun cuando existen diferentes funciones núcleo, es común el uso de la Función de Base Radial. Sin embargo puede optarse por el uso de otros núcleos dependiendo de los resultados obtenidos para un caso particular.

2.4. Pre-procesamiento de datos

El pre-procesamiento de datos se basa en la afirmación de que todos los datos tomados del mundo real (mediciones, imágenes, sonidos) no se encuentran en un estado óptimo. La mayoría de estos son corrompidos por valores externos, de los cuales no se tiene conocimiento de su existencia, o que aunque se conocen no se pueden controlar. Además de algunos otros factores que de igual manera afectan a los datos, tales como errores de medición, mal calibración del equipo y el error humano. Por lo tanto se requieren de algunos procesos o métodos que minimicen el efecto que tienen los valores externos o errores de medición en los datos que se desea utilizar.

La normalización es el método más utilizado entre los diferentes pre-procesamientos. Este método se basa en ajustar los datos en un rango determinado, con el propósito de minimizar las diferencias entre los parámetros de un conjunto de datos; en otras palabras, eliminar el peso excesivo que tiene un parámetro en comparación con otro.

Tabla 2.2 Descripción de los métodos de normalización más utilizados.

Normalización	Ecuación	Descripción
Min-Max	$v' = \frac{v - min_X}{max_X - min_X}(max'_X - min'_X) + \\ + min'_X$	Se hace una transferencia lineal de los datos, mediante la inserción de nuevos valores mínimos y máximos. Las escalas más utilizadas son: [0 a 1] y [-1 a 1]
Zscore o Zero Mean	$v' = (v - prom_X)/\sigma_X$	Los valores de los datos se transforman con respecto a su media y a su desviación estándar. Útil cuando el mínimo y el máximo de los datos son desconocidos.
Escala Decimal	$v' = {}^{v}/_{10^j}$	Donde *j* es el entero más pequeño tal que Max(|v'|) < 1

2.5. Evaluación y selección de modelos

Para resolver cualquier tipo de problema (regresión, clasificación) mediante algún método de aprendizaje, se requiere evaluar todos los posibles modelos, con el objetivo de seleccionar el que mejor se adapte a nuevos datos. La evaluación y selección se hace en base al desempeño que tiene cada modelo.

Una pobre o pequeña estimación de los datos que hace algún algoritmo, se le conoce como error de aproximación; cuando sobrestima los datos, a esto se le conoce como error de estimación. Si el error de aproximación es muy grande, el error de estimación tiende a ser pequeño, a esto se le conoce como infra-ajuste, y cuando el de estimación es muy grande, el de aproximación será pequeño, a esto se le conoce como sobre-ajuste. El mejor método es el que además de tener un buen desempeño al resolver el problema, tiene un error de aproximación y de estimación relativamente balanceados. En la Figura 2.6 se muestran casos de infra-ajuste y sobre-ajuste de un problema de clasificación bi-clase.

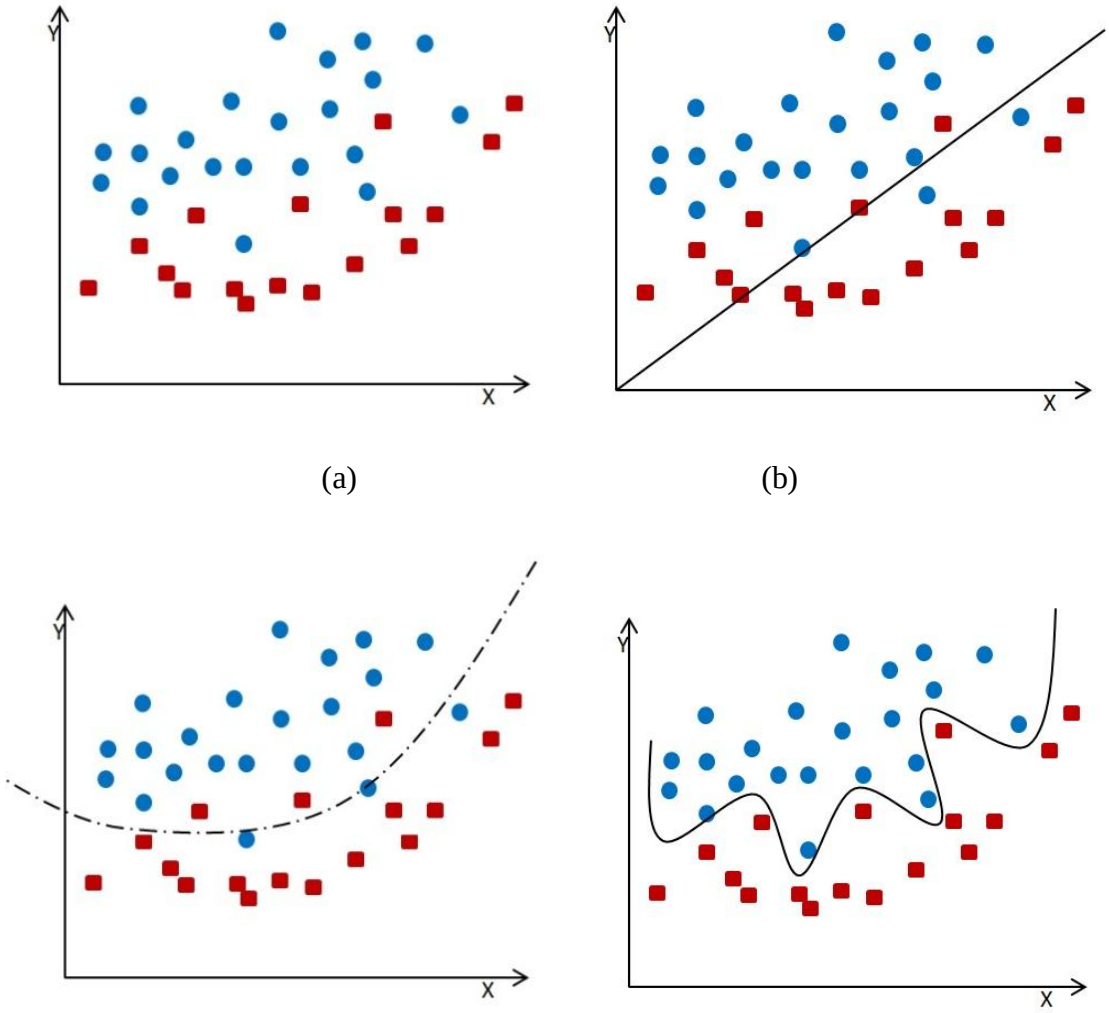


Figura 2.6 Ejemplos de diferentes ajustes en una caso de clasificacion bi-clase: a) Datos originales sin ajuste; b) Infra-ajuste del modelo; c) Datos ajustados; d) Sobre-Ajuste del modelo.

Los datos usados para el entrenamiento de un algoritmo son los mismos datos que se utilizan para evaluarlo, lo que implica el problema de la selección de muestras, el cual se resuelve a través de una división (partición) de los datos: una parte se usa para entrenamiento (conjunto de entrenamiento) y la otra parte se utiliza para validar el desempeño (conjunto de validación), ya que representa datos aparentemente "nuevos" siempre y cuando no hayan sido usados en el entrenamiento. Cuando los datos se dividen

en dos subgrupos, el entrenamiento y la validación, se considera que se está aplicando una validación *al modelo*, pero cuando se dividen de dos a *n*-1, donde *n* es la totalidad de los datos, se considera que se está aplicando una *validación cruzada* (*crossvalidation*) al modelo.

2.5.1. Validación simple

Una validación simple conocida como *Hold Out Validation* (Devroye and Wagner, 1979), es la forma más simple de validación, la cual consiste en separar un conjunto de datos $D_n = \{(x_i, y_i) | \ i = 1, 2, \ldots, n\}$ en dos subconjuntos, uno para entrenamiento D_{nt}, u otro, D_{nv} , para validación, donde n_t es el total de datos de entrenamiento y n_v es el de validación, por lo tanto $n = n_t + n_v$. En este caso se sabe que el tamaño de D_{nt} debe ser mayor que el de validación, es decir, $|D_{nt}| > |D_{nv}|$, o el modelo poseería una tendencia a infra-ajuste, pero no debe ser tan grande como para que sufra de sobre-ajuste. Generalmente el tamaño de D_{nt} es de entre el 60 y 80% del total del conjunto D_n.

2.5.2. Validación cruzada

La validación cruzada (*Cross Validation*, CV) es un método estadístico de remuestreo para evaluar y comparar modelos. En su forma más básica, consiste en dividir los datos en dos partes: la primera, empleada para entrenar el algoritmo, y la segunda utilizada para validar el modelo. En una CV típica, los conjuntos de entrenamiento y validación deben cruzarse en rondas sucesivas de tal manera que se validaría cada muestra de datos.

A diferencia del método de validación simple, los datos son separados en tres subconjuntos, uno de entrenamiento, otro de validación y el tercero es para prueba. Los primeros dos se utilizan para seleccionar el mejor modelo y el tercero para evaluar el comportamiento del modelo contra datos desconocidos. Usualmente los conjuntos de entrenamiento y validación se derivan del subconjunto D_{nt}. En la Figura 2.7 se muestra la comparación entre los métodos de validación simple y CV.

Existen varias formas de CV; a continuación se mencionan las más utilizadas:

- *Leave One Out* (LOO CV):

Es la forma clásica de CV, donde $n_t = n - 1$; en este tipo de CV cada muestra de D_n será separada del conjunto y utilizada para validación (Geiser, 1975).

- *Leave p-Out* (LPO CV):

Es una forma modificada de la LOO CV, donde $n = n_t - p$, y $p \in \{1,2,\ldots, n-1\}$; en esta forma de CV cada posible subconjunto de los datos p es separado del conjunto total y utilizado para validarlo (Shao, 1993). Se debe mencionar que un LPO CV con $p = 1$ coincide con LOO CV.

- *K-Fold CV*:

Esta forma de CV, introducida por Geiser en 1975 (Geiser, 1975), consiste en una partición preliminar del conjunto de datos en K subconjuntos de aproximadamente el mismo tamaño n/K, donde $K \in \{1,2,\ldots, n\}$. En el K-Fold CV cada subconjunto K se utiliza para validación y el resto para entrenamiento. Existe otra versión de este modelo que consiste en repetir K veces las K particiones del CV. En resultado, el consumo computacional aumenta drásticamente, por lo que el valor de K suele ser pequeño (5 o 10). En la Figura 2.7 se muestra el funcionamiento de un K-Fold CV.

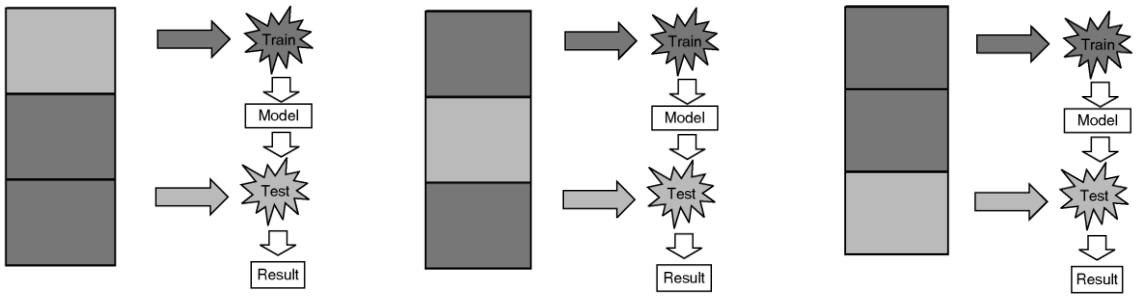


Figura 2.7 Funcionamiento de un K-Fold CV con K=3

2.6. Conclusiones

Se consideran algunos modelos de aprendizaje automático, siendo elegidos las redes neuronales y las máquinas de soporte vectorial, ya que la literatura ha mostrado sus ventajas en procedimientos que requieren manejo de datos de índole biológica.

Aunque se debe tener en cuenta que por la amplia gama de tipo de redes neuronales o máquinas de soporte vectorial, así como de configuraciones de estas, se debe limitar las configuraciones cuando estas rebasen un tiempo determinado de entrenamiento, así como de complejidad de la misma red, siempre y cuando los datos así lo requieran. Lo anterior dependerá de la cantidad de factores que muestre el conjunto de datos, así como del comportamiento que estos tengan.

En caso de las redes neuronales las cantidades de neuronas o capas de neuronas son el factor que aumenta de gran manera el tiempo computacional requerido para resolver algún tipo de problema, en caso de las máquinas de soporte el tamaño de los datos para entrenamiento o prueba incrementaran será el factor a analizar.

Otra variable a considerar será que dependiendo de la complejidad de los datos y la cantidad de muestras que se posean, se deberá aplicar alguna de las diferentes técnicas de procesamiento, como alguna técnica de normalización, las cuales tendrán un nivel de impacto proporcional a la complejidad del conjunto de datos, lo anterior al momento de entrenar los diferentes algoritmos de aprendizaje. De igual manera se deberá de utilizar alguna técnica para la evaluación y selección del modelo, siendo considerada para su utilización la técnica K-Fold CV, ya que presenta el nivel más bajo en el consumo computacional, de las 3 técnicas que presenta este trabajo, dado que K-Fold CV consume menos recursos computacionales (tiempo y memoria) que LOO CV o LPO CV.

3. ANÁLISIS DEL EXPERIMENTO ELECTROQUÍMICO

Para obtener los datos para el experimento computacional se realizó un experimento electroquímico, variando los niveles de 3 parámetros utilizados (ACh, pH y Temperatura). En el capítulo se analiza el comportamiento de los datos resultantes de estas variaciones.

3.1. Experimento Electroquímico

Los siguientes reactivos fueron utilizados en las reacciones:

Acetilcolinesterasa (AChE), 2.5 mg/cm ^{3};
Soluciones de acetiltiocolina (ACh) ($0.2 \leq Cs \leq 1.0$),
Buffer Britton-Robinson ($5 \leq pH \leq 9$).

Los datos experiméntales se obtuvieron utilizando un biosensor amperométrico. La velocidad de rotación del electrodo, así como la concentración enzimática se mantuvieron constantes, 1000 rpm y 2.5 mg/cm ^{3} respectivamente. La concentración de ACh, el pH y la temperatura fueron los parámetros que variaron. Los parámetros, sus notaciones, sus unidades de medición y sus niveles se muestran en la Tabla 3.1.

Tabla 3.1 Parámetros Experimentales

Parámetros	**Notaciones**	**Unidades de Medición**	**Niveles de parámetro**
Concentración del substrato (ACh)	C_s	mM/ L	0.2, 0.4, 0.6, 0.8, 1.0
Acidez	pH	-	5, 6, 7, 8, 9
Temperatura	t	˚C	25, 30, 40, 50, 60, 70
Corriente de salida	I	μA	-

3.2. Influencia de los parámetros

3.2.1. Concentración del Substrato

El comportamiento de la corriente al cambio de la concentración de ACh es completamente ascendente. A mayores concentraciones, se obtienen mejores resultados en función de la corriente. La función corriente I vs concentración del sustrato Cs se describe a través de una hipérbola: $I=I_{max} * C_s/(C_s + Km)$. Km es una constante llamada Constante de Michaelis-Menten. De esta manera, el nivel óptimo de concentración de ACh es el nivel más alto utilizado: 1.0 mM/L. En la Figura 3.2 se observa el comportamiento ascendente que tiene la corriente (z) en base a aumentos de concentración de acetilcolina (y), con cambios en la temperatura (x) y un pH fijo en nivel 7. Los resultados óptimos en función de la concentración de ACh predefinida a 1.0 mM/L, pH 5-9 y temperatura 25 - 70˚C se presentan en la Tabla 3.2.

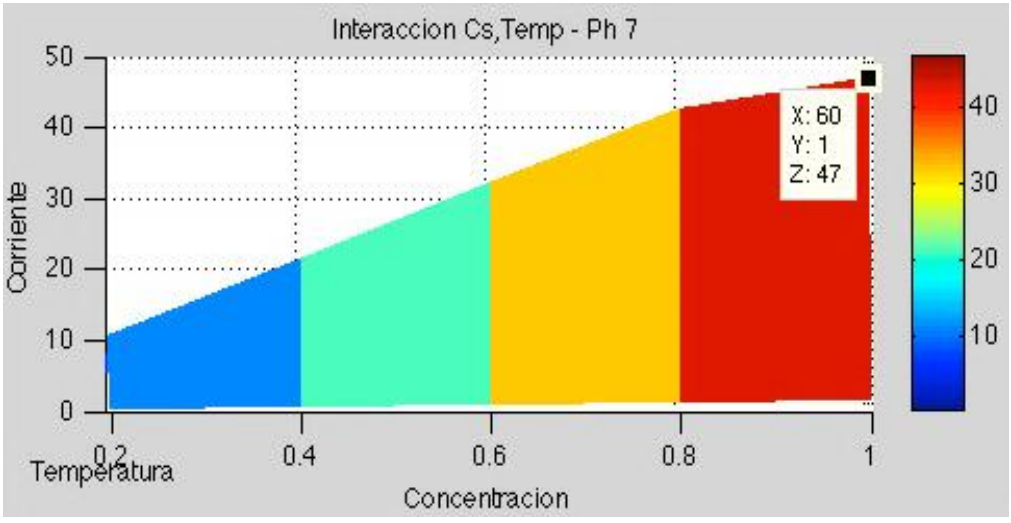


Figura 3.1 El resultado máximo de la corriente resultante (z) en función del pH predefinido a 7, temperatura T (x) 25-70 ˚C, y concentración de la sustancia Cs (y) 0.2-1.0 mM/L

Tabla 3.2 Resultados óptimos en función de la concentración de ACh predefinida a 1.0 mM/L (pH 5-9 y temperatura 25-70 ˚C)

pH	Corriente, µA					
	25˚C	**30˚C**	**40˚C**	**50˚C**	**60˚C**	**70˚C**
5	8.80	14.06	15.67	17.27	18.88	0.72
6	20.30	32.44	36.14	39.14	**43.56**	1.66
7	21.90	35.00	39.00	**43.00**	**47.00**	1.80
8	17.90	25.09	31.87	35.14	38.41	1.47
9	15.70	25.09	27.95	30.82	33.69	1.29

3.2.2. pH

La respuesta de la corriente al aumento en el nivel de acidez fue de algún modo ascendente. El pH óptimo varía dependiendo de gran manera en los cambios de concentración de ACh y de la temperatura. Si la concentración de la sustancia es menor o igual a 0.6 µmoles y la temperatura es menor o igual a 30 ˚C, el pH óptimo se encuentra en el nivel 7; si la temperatura es mayor a 30 ˚C se encuentra en el nivel 8. Si la concentración de acetilcolina es mayor a 0.6 mM/L, el pH óptimo se ubica en el nivel 7. Los datos obtenidos se presentan en la Tabla 3.3. En la Figura 3.3 se observa el comportamiento de la corriente resultante (z) en función del pH (x), de la concentración de ACh y de un valor fijo de temperatura.

Tabla 3.3 Resultados óptimos en función de la temperatura predefinida a 60 ˚C, pH 5-9, y concentración de ACh 0.2-1.0 mM/L.

Concentración, mM/L	Corriente, µA				
	pH 5	**pH 6**	**pH 7**	**pH 8**	**pH 9**
0.2	8.56	11.22	10.65	15.79	13.13
0.4	13.32	22.27	21.32	28.36	24.55
0.6	15.68	30.98	31.98	32.35	28.62
0.8	18.11	39.85	42.63	36.87	32.39
1.0	18.88	43.56	47.00	38.41	33.69

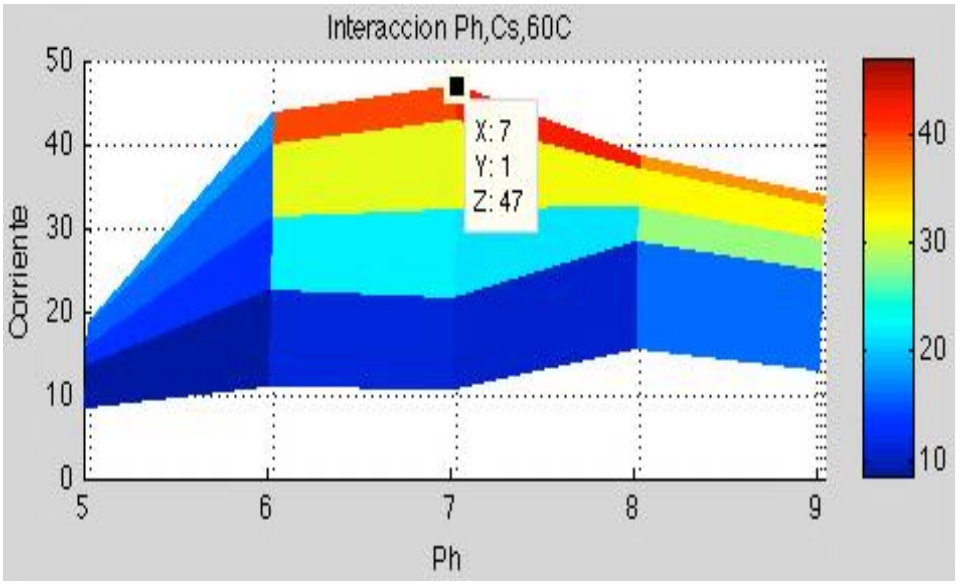


Figura 3.2 Grafica que presenta el resultado máximo de la corriente resultante (z) en función de la Temperatura T predefinida a 60 ˚C, pH (x) 5-9, y concentración de ACh (y) 0.2-1.0 mM/L.

3.2.3. Temperatura

La respuesta de la corriente resultante al aumento de temperatura es de forma ascendente de 4.5 μA hasta 47 μA. Este último es el valor máximo de la corriente, encontrado a una temperatura de 60 ˚C. Cuando se aumenta la temperatura de 60 ˚C a 70 ˚C, la corriente desciende de manera drástica, de tal modo que se obtuvo el valor de 0.382075 μA (70 ˚C, pH 7, 0.2 μmoles), menor que el encontrado a 25 ˚C. En la Figura 3 se observa la influencia de la temperatura (*x*) en la variable respuesta (*z*) con cambios en la concentración de ACh (*y*). La Tabla 4 incluye los datos correspondientes.

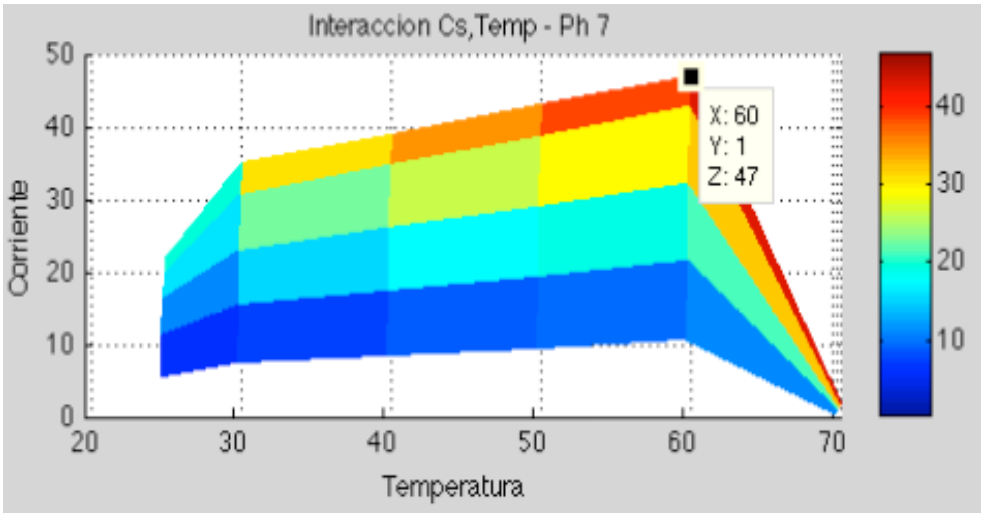


Figura 3.3 Grafica que presenta el resultado máximo de la corriente resultante (z) en función del pH predefinida a 7, temperatura T (x) 25-70 ˚C y concentración de ACh (y) 0.2-1.0 mM/L.

Tabla 3.4 Resultados óptimos en función del pH en nivel 7, temperatura 25-70 ˚C y concentración de ACh Cs 0.2-1.0 mM/L

Concentración, mM/L	**Corriente, µA**					
	25 ˚C	**30 ˚C**	**40 ˚C**	**50 ˚C**	**60 ˚C**	**70 ˚C**
0.2	5.60	7.63	8.68	9.60	10.65	0.38
0.4	11.20	15.26	17.37	19.20	21.31	0.76
0.6	16.30	22.89	26.06	28.80	31.97	1.14
0.8	20.00	30.52	34.75	38.40	**42.63**	1.52
1.0	21.90	35.00	39.00	**43.00**	**47.00**	1.80

3.3. Conclusiones

Después del análisis de los datos, se muestra una clara interdependencia de los tres factores en la variable respuesta.

El aumento de nivel en los diferentes factores incrementan el resultado de la variable repuesta, sin embargo, existen determinadas combinaciones que aunque representan un aumento en los factores, reducen el valor de la respuesta.

La cantidad de variables hace que no se presente claramente una buena visualización de los datos por lo que se muestran desde ciertas perspectivas, en este caso se hicieron utilizando los niveles de factor donde se presentó la variable respuesta en su nivel más alto, solo después de la creación del modelo y de su simulación se conocerá si este nivel máximo de la variable respuesta permanecerá como su máximo global o encontraremos el verdadero máximo.

4. MODELO PARA LA PREDICCIÓN DE PARÁMETROS

El proceso por el cual se desarrolla el modelo de aprendizaje automático consta de varios pasos los cuales deben ser cuidadosamente seguidos y evaluados, en caso contrario no se puede asegurar el óptimo desempeño del modelo, ni que los resultados sean confiables. El capítulo 4 muestra el proceso utilizado para el desarrollo de los modelos de regresión de una Red Neuronal y una Maquina de Soporte Vectorial, así como la selección del mejor de estos modelos.

4.1. Proceso General

Dado que el proceso de desarrollo de los modelos de aproximación y optimización es extenso, se propone separarlo en una serie de pasos, para mejorar la comprensión y facilitar el análisis. A continuación se desglosan los diferentes pasos a utilizar:

Análisis de datos: El análisis de datos conlleva todo el estudio de la información experimental, para determinar si es o no necesario procesar los datos previo a someterlo al entrenamiento de los modelos, además conocer el porcentaje en la que se dividirán en conjuntos de entrenamiento, validación y pruebas, de igual manera considerar si será necesaria alguna técnica de remuestreo y cual técnica sería mejor utilizar.

Aproximación a la función: Involucra la creación del modelo y los procesos necesarios para encontrar su mejor configuración, la que permite aproximar de mejor manera a los datos de la función.

Evaluación de la función de aproximación: Es el análisis cuantitativo y cualitativo de los resultados que tienen los modelos en las pruebas, así como la comparación de los mejores modelos y sus resultados contra los resultados experimentales.

Optimización: Es encontrar el punto máximo que puede tomar el valor de la corriente resultante *I*, a través de la función de aproximación.

Definir el modelo: Involucra la definición de la función objetivo, las restricciones y el dominio del modelo.

4.2. Elección del modelo

En la Figura 4.1 se muestra el proceso utilizado para encontrar el mejor modelo de aproximación.

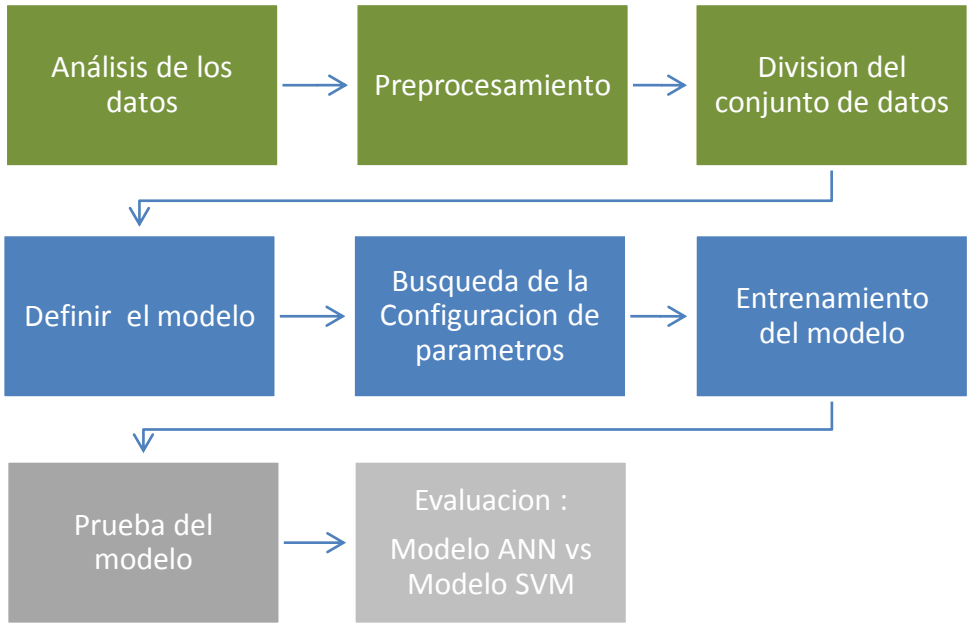


Figura 4.1 Proceso para determinar el modelo de aproximación

Análisis de datos: Consiste en observar el comportamiento de los datos, así como decidir si será necesario aplicar algún método de procesamiento o la manera en la que se dividirá el conjunto de datos (como se está descrito en la sección 2.5).

Pre-procesamiento: Si se considera necesario, se aplica alguna técnica de pre-procesamiento de datos, como un tipo de normalización, con el objetivo de mejorar su desempeño en el entrenamiento de los modelos.

División de datos: Dependiendo del análisis se evalúa el tamaño del conjunto de datos, para conocer de qué manera se dividirán, y si será o no necesario la aplicación de alguna técnica de remuestreo, la cual también influirá en la división del conjunto.

Definir el modelo: Involucra decidir qué algoritmo en su forma de regresión se utilizará para desarrollar el modelo y la forma en la que será entrenado.

Búsqueda de la configuración del modelo: Se aplica la técnica *Grid Search* para buscar la mejor configuración posible del algoritmo para el desarrollo del modelo.

Entrenamiento: Se realiza el aprendizaje del algoritmo utilizando la mejor configuración encontrada, para el desarrollo del modelo de aproximación a la función. En el caso de que se emplea alguna técnica de remuestreo, también conlleva el proceso de validación.

Evaluación: Es el análisis comparativo cualitativa y cuantitativamente del desempeño de los diferentes modelos desarrollados contra el conjunto de datos de pruebas y los datos experimentales correspondientes.

4.3. Pre-procesamiento y división de datos

Se analizaron las 150 diferentes muestras obtenidas durante el experimento electroquímico (Capitulo 3), llegando a la conclusión de que era necesario utilizar un método de normalización ya que las diferencias entre los rangos de los parámetros (*Cs*, *pH*, *T*) podrían influir al entrenamiento de los algoritmos. Por este motivo les fue aplicado un ajuste a un rango de [0, 1] utilizando el método Min-Max (Tabla 2.2). En la Tabla 4.1 se muestran los valores de los datos antes y después de normalizar.

Tabla 4.1 Datos antes y después de normalizar.

Parámetros	Datos originales	Datos normalizados
Cs	[0.2, 0.4, 0.6, 0.8, 1.0]	[0.0, 0.25, 0.50, 0.75, 1.0]
pH	[5, 6, 7, 8, 9]	[0.0, 0.25, 0.50, 0.75, 1.0]
T	[25, 30, 40, 50, 60, 70]	[0.0, 0.1111, 0.3333, 0.5556, 0.778 ,1.0]

Después de normalizar los datos, se procede a dividir todo el conjunto en subconjuntos para entrenamiento, validación y prueba. Dado que el conjunto total de datos es pequeño (150 muestras), se considera que es necesario utilizar alguna técnica de

remuestreo, que además permitirá separar el conjunto principal de datos en 2 subconjuntos en lugar de 3. El primero es un subconjunto asignado al entrenamiento-validación y el segundo es un subconjunto solo para pruebas. La técnica elegida de remuestreo es una versión de la *K*-Fold CV, la cual se repite *k*-veces, de esta manera nos permite entrenar con un mayor número de muestras, las cuales a su vez, serán validadas. Además, esta técnica permite evaluar el modelo de una mejor manera ya que se consideran más elementos para validación, y se refuerza el comportamiento del modelo contra "nuevas" muestras.

4.4. Desarrollo del modelo de regresión

4.4.1. RNA: Entrenamiento y validación

Con el objetivo de encontrar la mejor configuración de parámetros para la Red Neuronal MLP se utiliza un proceso de *Grid Search* para encontrar el mejor número de neuronas, capas, tipos de entrenamiento y aprendizaje. Considerando que el proceso de búsqueda sería largo, se redujo el rango de esta, basándonos en trabajos previos sobre aproximaciones a funciones de características biológicas u obtenidas mediante el uso de biosensores (Guide, et al., 2003).

A continuación se muestra la configuración propuesta de la red neuronal MLP:

Capas	5;
Neuronas	30 por cada capa;
Entrenamiento	Algoritmo Levenberg-Marquadt;
Aprendizaje	Gradiente Descendente;
Técnica de Remuestreo	*k* x *k*-Fold CV, $k \in (5, 10)$;
Medida de Desempeño	Error Cuadrático Medio (*MSE*)

La configuración presentada es la que dio los mejores resultados en base al consumo de tiempo, recursos y el buen desempeño durante el entrenamiento. Las otras configuraciones destacadas aumentan el desempeño de la red de forma poco significativa,

sin embargo tienden a consumir cantidades considerables de tiempo y recursos, por lo cual fueron descartadas.

La configuración seleccionada se programó con 30 neuronas en cada una de las 5 capas. Se utilizó el algoritmo Levenberg-Marquadt para optimizar el aprendizaje del algoritmo. Se usó la *K*-Fold CV como técnica de remuestreo en su forma repetitiva $k \times k$ (Refaeilzadeh, et al., 2008), donde se usaron 2 valores diferentes de *K*, con el objetivo de mejorar la evaluación del modelo $K \in [5,10]$.

4.4.2. MSV: Entrenamiento y validación

Después de la etapa de pre procesamiento y división de los datos se procedió a la aplicación de *Grid Search* con incremento exponencial en los parámetros, como se sugiere en (Guide, et al., 2003), con el objetivo de encontrar la mejor configuración del algoritmo SVR[1] en un menor tiempo. Ya que el desempeño de este algoritmo depende del tipo de la función kernel empleada, se realizaron varias búsquedas, una por cada tipo de kernel utilizado: lineal, polinomial, base radial y sigmoidea. El mejor resultado mostró el SVR con kernel de base radial. En la Tabla 4.2 se muestran los mejores resultados de cada uno de los 4 tipos de kernel.

El proceso de búsqueda utilizando el kernel de base radial solo se realizó en 3 parámetros: el parámetro de penalización del error(C), la función de perdida insensible para la tolerancia al ruido (ε), y el parámetro único del kernel radial (α). Para mejorar el proceso de búsqueda se aplicó una *K*-Fold CV en su forma repetitiva $k \times k$, donde $K \in [5,10]$ para asegurar que la configuración encontrada es la mejor posible.

[1] La aplicación *libsvm* [Chang & Lin, 2011] como implementación de MATLAB para el uso del algoritmo SVR fue utilizada durante el experimento.

Tabla 4.2 Mejores resultados del algoritmo SVR con diferentes kernel

	Lineal	**Polinomial**	**Función de base Radial**	**Sigmoideo**
Configuración	$c = 2^9$ $\varepsilon = 2^{2.21}$	$c = 2^9$ $\varepsilon = 2^{2.21}$ $\alpha = 0.075$ $d = 2$	$c = 512$ $\varepsilon = 4.6268$ $\alpha = 0.075$	$c = 2^9$ $\varepsilon = 2^{2.21}$
MSE 5-Fold CV	40.53	26.75	5.41	7.83
MSE 10-Fold CV	84.72	53.78	10.84	15.67

La comparación del desempeño entre los algoritmos arriba mencionados asegura que la configuración presentada es la mejor posible utilizando el algoritmo SVR.

La configuración de la SVM que se propone se muestra a continuación:

C $2^9 = 512$;

α $2^{2.21} = 4.6268$;

ε 0.075;

Kernel Función de Base Radial;

Técnica de Remuestreo k x k-Fold CV, $k \in [5,10]$;

Medida de Desempeño MSE

Posteriormente se procede a entrenar el modelo con la mejor configuración para evaluarlo utilizando el conjunto de prueba y a su vez compararlo con los resultados del modelo de red neuronal desarrollado.

4.5. Evaluación del modelo

4.5.1. Pruebas: ANN

Las pruebas realizadas al modelo ANN-MLP descrito en la sección 4.3, consiste en hacer una simulación de los parámetros C_S, pH y t utilizando el conjunto de datos de prueba y el modelo ANN-MLP seleccionado durante el proceso *Grid Search.* En la Tabla 4.3 se muestra la comparación de los datos de prueba y su simulación usando el modelo ANN-MLP.

La Figura 4.2 muestra la comparación de los datos originales contra los datos predichos por el modelo ANN-MLP.

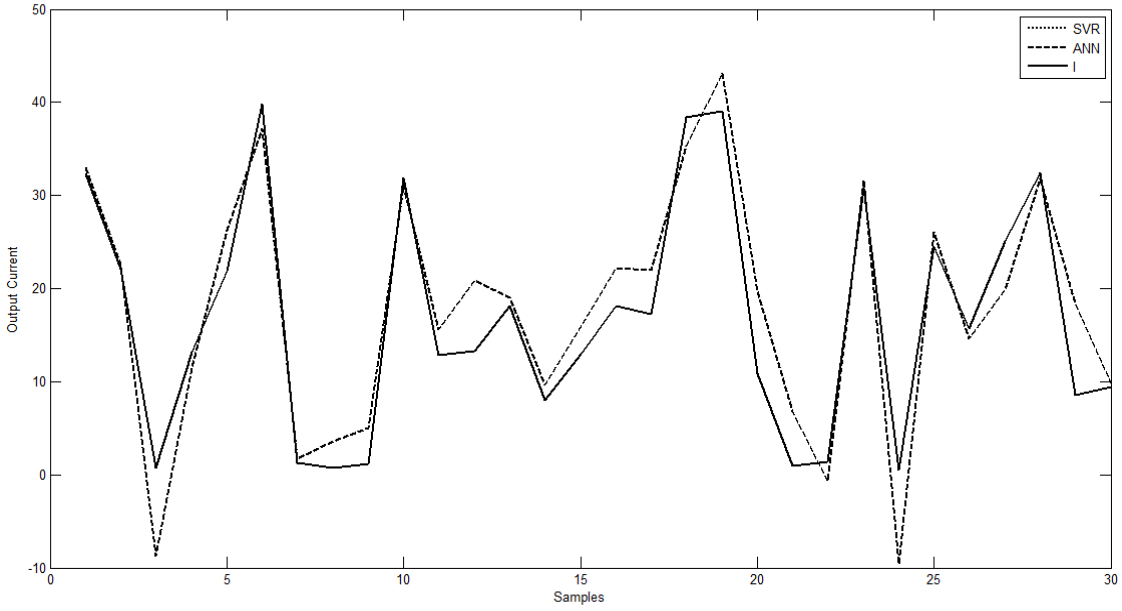
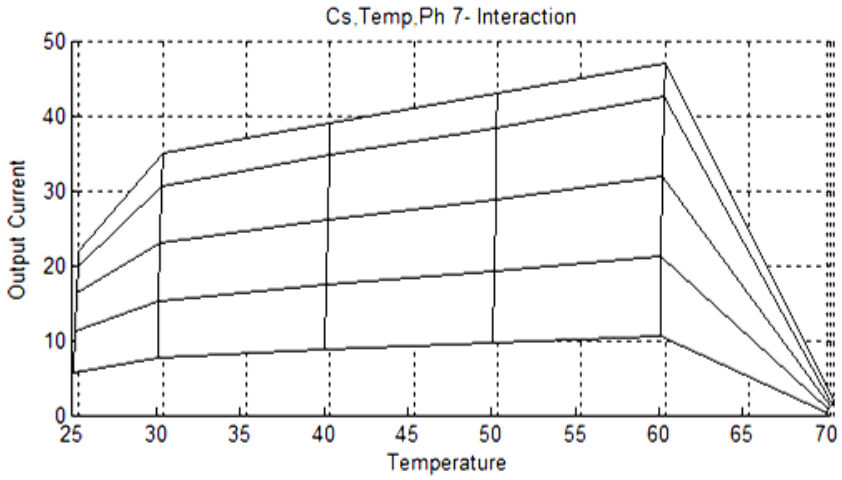
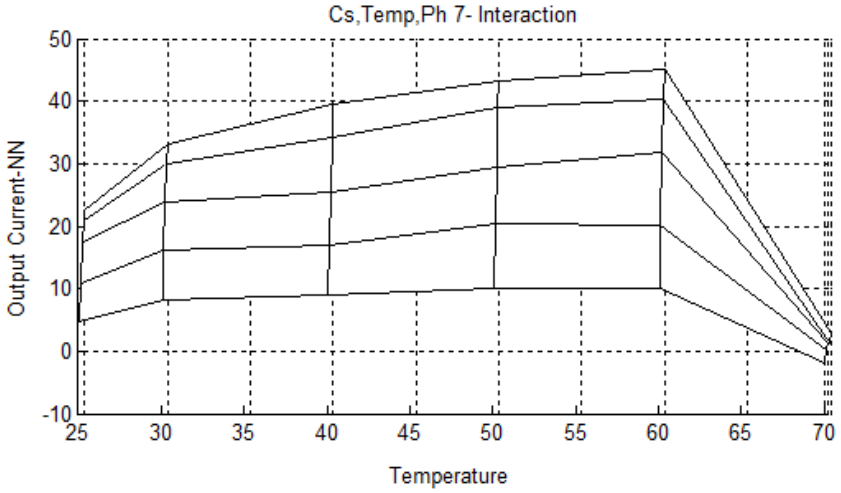


Figura 4.2 Gráficas de la corriente resultante de los datos de prueba contra los datos predichos por el modelo ANN-MLP.



(a)



(b)

Figura 4.3 Gráficas comparativas de la simulación de datos realizada por el modelo ANN-MLP y de los datos experimentales: (a) Gráfica de los datos experimentales; (b) Gráfica de los datos predichos por la ANN-MLP.

4.5.2. Pruebas: SVM

La realización de las pruebas al modelo SVR desarrollado consistió en la simulación del modelo previamente entrenado con la mejor configuración encontrada utilizando el conjunto de pruebas, y comparar los resultados obtenidos contra los resultados reales del experimento. En la Tabla 4.3 se muestra una parte de la comparación de los datos de prueba y la simulación usando el modelo SVR. En el Anexo F se presenta la comparación de todo el conjunto de pruebas.

Tabla 4.3 Mejores resultados del algoritmo SVR

No	Datos de prueba *I*, μA	Simulación *I*, μA	Error
1	32.35	31.7865191	0.56348089
2	22.11	22.8484843	0.73848435
3	0.72	-3.47229928	4.19229928
4	13.13	14.0527819	0.9227819
5	21.9	24.0532271	2.1532271
6	39.85	40.5384729	0.68847285
7	1.32	1.76999286	0.44999286
8	0.74	3.11594287	2.37594287
9	1.16	1.50589061	0.34589061

La comparación del modelo SVR contra el parámetro de salida de los datos originales de prueba se muestra en la Figura 4.4. Una simulación de los datos originales y de los datos predichos por el modelo SVR se expone en las Figuras 4.5 a y 4.5 b respectivamente.

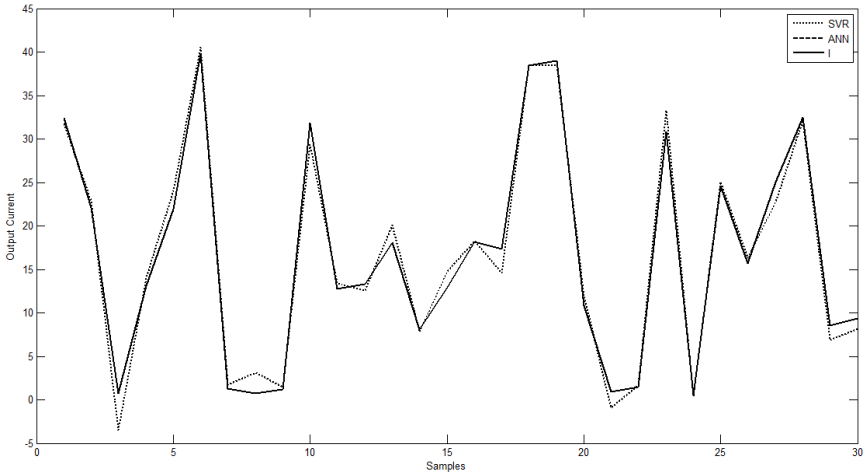
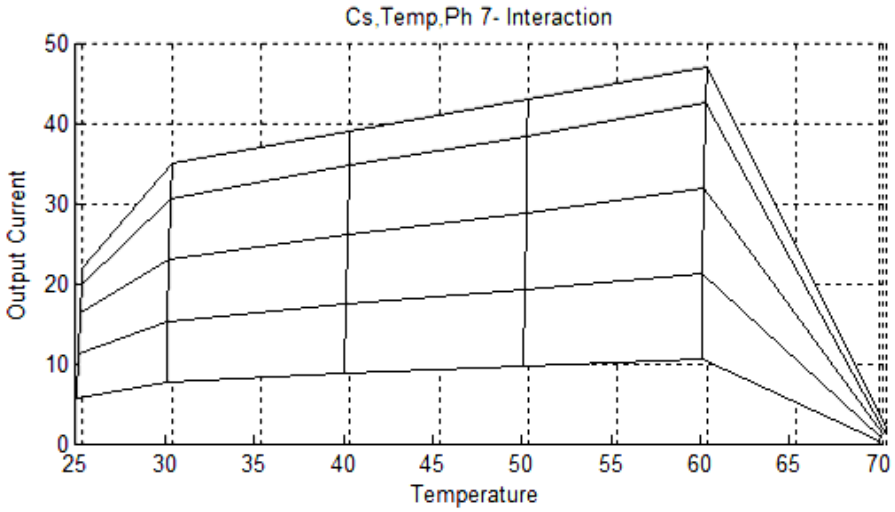
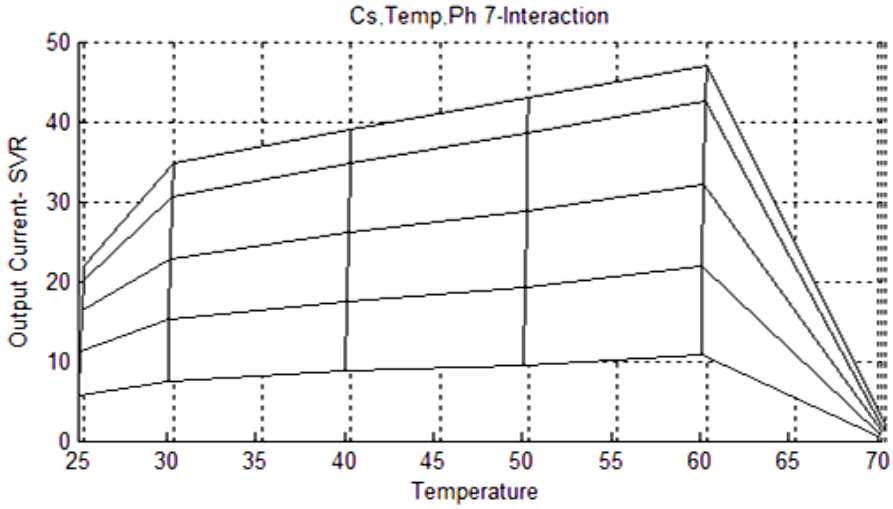


Figura 4.4 Datos de la corriente resultante del conjunto de prueba contra los datos predichos por el modelo SVR.



(a)



(b)

Figura 4.5 Gráficas comparativas de la simulación de datos realizada por el modelo SVR y de los datos experimentales: (a) Gráfica de los datos experimentales; (b) Gráfica de los datos predichos por la SVR.

4.5.3. Análisis Comparativo

El análisis comparativo se realizó observando las gráficas de los datos de prueba modelados por los algoritmos, así como la comparación de los errores que tienen las simulaciones de las muestras de prueba y los MSE que tiene cada modelo en total. La Figura 4.6 muestra las gráficas de los datos experimentales de prueba y los datos simulados por los modelos ANN MLP y SVR.

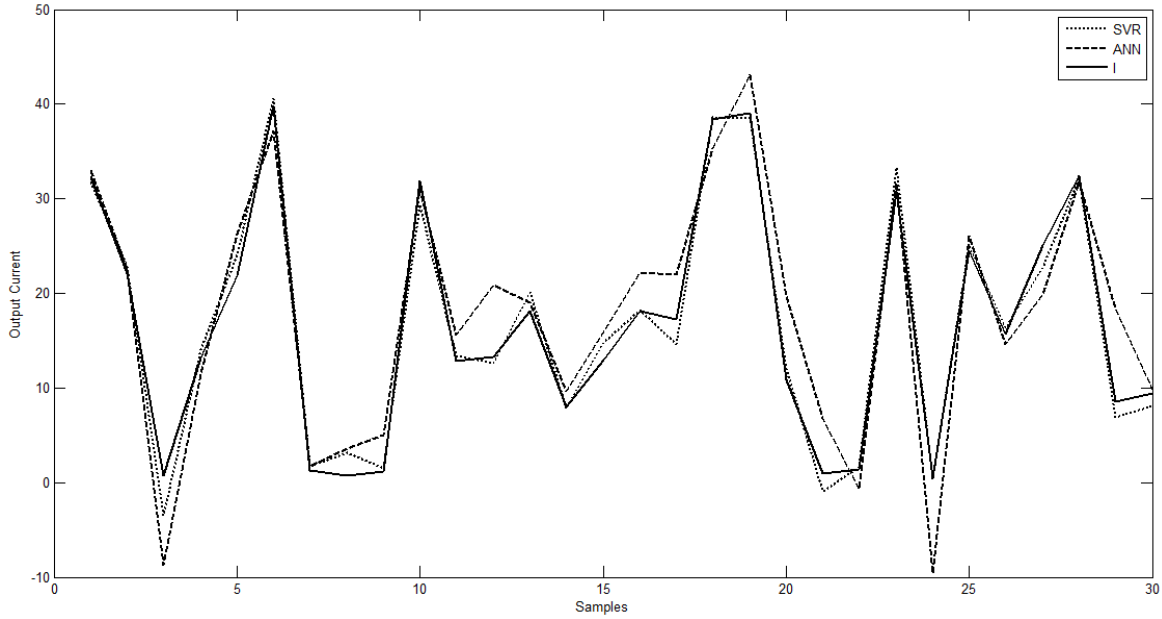


Figura 4.6 Comparación de los datos experimentales de prueba contra los datos predichos por los modelos ANN-MLP y SVR

La Tabla 4.4 muestra la comparación cuantitativa entre los datos de prueba simulados por los modelos y los datos experimentales, basados en el MSE de cada modelo; se tomaron 3 muestras al azar del conjunto de datos de prueba. La Figura 4.6 y la Tabla 4.4 demuestran que ambos modelos son buenos aproximadores a la función, pero el modelo SVR es el mejor que los ajusta sin llegar a caer en el sobreajuste, por lo que se considera el mejor. En el Anexo G se presenta el análisis completo de los ambos modelos y el conjunto de pruebas.

Tabla 4.4 Comparación Cuantitativa de los modelos

No	Datos de prueba				ANN MLP		SVR	
	C_S, μmol	pH	t, ˚C	I, μA	I, μA	*Error,*	I, μA	*Error,*
1	0.2	6	30	8.03	9.68	1.65	7.84	0.19
2	0.8	7	50	18.15	22.21	4.06	18.26	0.09
3	0.4	7	40	38.41	35.27	3.14	38.56	0.15
Desempeño (MSE), %					-	2.95	-	0.143

4.6. Conclusiones

Después del análisis cuantitativo y cualitativo de los resultados obtenidos por las pruebas, se muestra claramente la superioridad de la Máquina de Soporte Vectorial sobre los resultados de la Red Neuronal. Ambos modelos mantienen las tendencias de los datos originales, por lo que se concluye que ambos fueron desarrollados adecuadamente. Además, ninguno de los resultados sugiere que alguno de los modelos sufra de problemas al intentar simular nuevas muestras.

El modelo permite simular cualquier combinación de Cs, pH y Temperatura, siempre y cuando se encuentre dentro de los rangos establecidos en el Capítulo 3 Sección 1, lo anterior es solo porque las pruebas realizadas al modelo eran factores cuyos niveles estaban dentro de los rangos establecidos.

5. SIMULACIÓN Y RESULTADOS

Se muestra el procedimiento necesario para la simulación del modelo de regresión generado en el capítulo anterior. Los resultados de la simulación son analizados y comparados con los datos experimentales.

5.1. Modelo Obtenido

El modelo de regresión se desprende de la función de regresión de cada algoritmo de aprendizaje utilizado, en este caso, la máquina de soporte vectorial es su forma de regresión utilizando una función kernel de base radial.

A continuación se presenta el desarrollo matemático del modelo de regresión.

La función de regresión del modelo computacional MSV es:

$$f(x) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i *)k(\boldsymbol{X}_i, \boldsymbol{X}) + b,$$

aplicando el Kernel de Base Radial (en términos de gamma)

$$k(\boldsymbol{X_i}, \boldsymbol{X_j}) = \exp(-\gamma||\boldsymbol{X}_i - \boldsymbol{X}_j||^2),$$

obtenemos

$$f(x) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i *)\exp(-\gamma||\boldsymbol{X}_i - \boldsymbol{X}_j||^2) + b,$$

dónde

$$\gamma = 4.6589;$$

$$0 < \exp(-\gamma||\boldsymbol{X}_i - \boldsymbol{X}_j||^2) < 1,$$

La variable γ la obtenemos durante el proceso de búsqueda de la mejor configuración del algoritmo.

Después del entrenamiento del algoritmo utilizando la configuración antes mencionada obtenemos:

- las variables primales $(-\alpha_i + \alpha_i *)$ que lo expresaremos como $\boldsymbol{W}$,
- el sesgo *b*, donde

  $\boldsymbol{W}$= [-79.1969, 7.6299, 3.2541],

  b=1.4888,

aplicando las variables a la ecuación # obtenemos nuestro modelo de regresión:

$$f(X) = \sum_{i=1}^{l} \boldsymbol{W} \exp(-4.6589 \left|\left| \boldsymbol{X}_i - \boldsymbol{X}_j \right|\right|^2) + 1.4888.$$

5.2. Simulación

Con el objetivo de que la simulación cuente con características más reales, se genera un conjunto de muestras, las cuales, estarán conformadas por la escala mínima medible de ACh, pH y temperatura (Tabla 5.1).

Tabla 5.1 Escala de factores para simulación

Factor	**Escala**	**Incremento**
Concentración del sustrato (C_s)	$0.2 \leq C_s \geq 1.0$	0.01
Acidez (pH)	$5 \leq pH \geq 9$	0.01
Temperatura (t)	$25 \leq t \geq 70$	0.1

Además de que la generación de muestras debe tener en cuenta lo siguiente:

- El rango de datos de las nuevas muestras deben ser los mismos que el de los datos originales.

- Las muestras generadas deben ser normalizadas de la misma forma que los datos experimentales, utilizando la misma configuración de las ecuaciones.

Lo anterior para mantener una estructura definida de los datos y evitar problemas de extrapolación, lo que provocaría una mala generalización de las nuevas muestras y por lo tanto un mayor porcentaje de error en la simulación.

Recordando que la cantidad de muestras utilizadas para el desarrollo del modelo eran 150, con los nuevos niveles la simulación se realiza con 15, 467, 031 muestras. Sin embargo, se demostró en capítulos anteriores que la variable respuesta es completamente dependiente de la C_s, a mayores concentraciones, el amperaje era mayor y viceversa, de esta forma, y con el objetivo de que se pueda analizar cualitativamente la simulación, el proceso se realiza desde la perspectiva de la temperatura y el pH, dejando a la C_s constante a un nivel de 1.0 μmol/ L. Por este motivo las muestras se reducen a 180, 851, disminuyendo el tiempo de generación de muestras, su normalización y su predicción por el modelo de regresión.

5.3. Resultados

5.3.1. Generales

Para evaluar la simulación se realiza un análisis comparativo entre las muestras simuladas y los datos experimentales. En la Figura 5.1 y 5.2 se muestra la simulación del modelo de regresion utilizando la escala minima de factores manejados por el biosensor, con un valor de C_s definido a 1.0 μmol/ L.

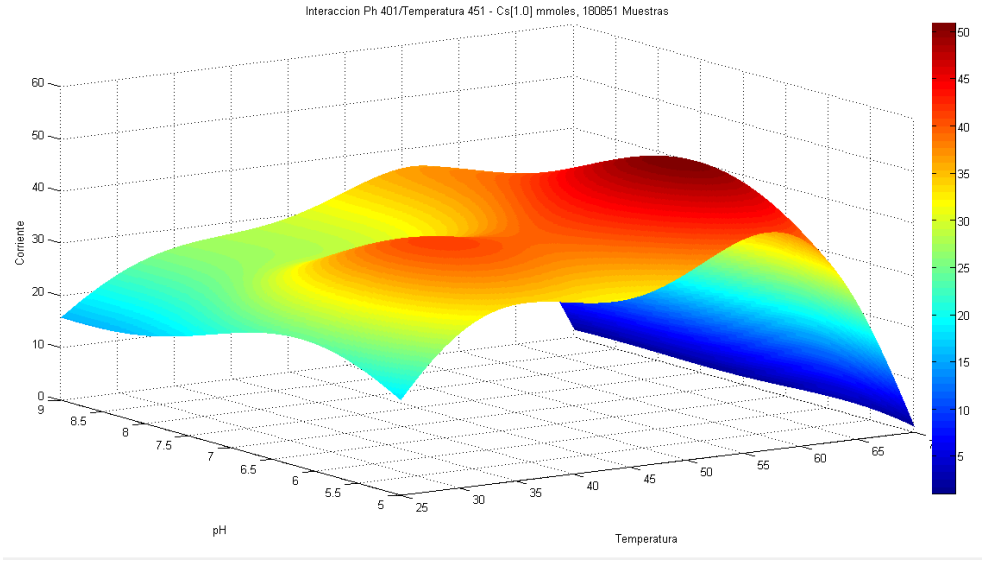


Figura 5.1 Simulación utilizando la escala mínima de factores.

Los resultados muestran que la función mantiene el comportamiento de los datos experimentales con mínimas irregularidades las cuales no comprometen su comportamiento.

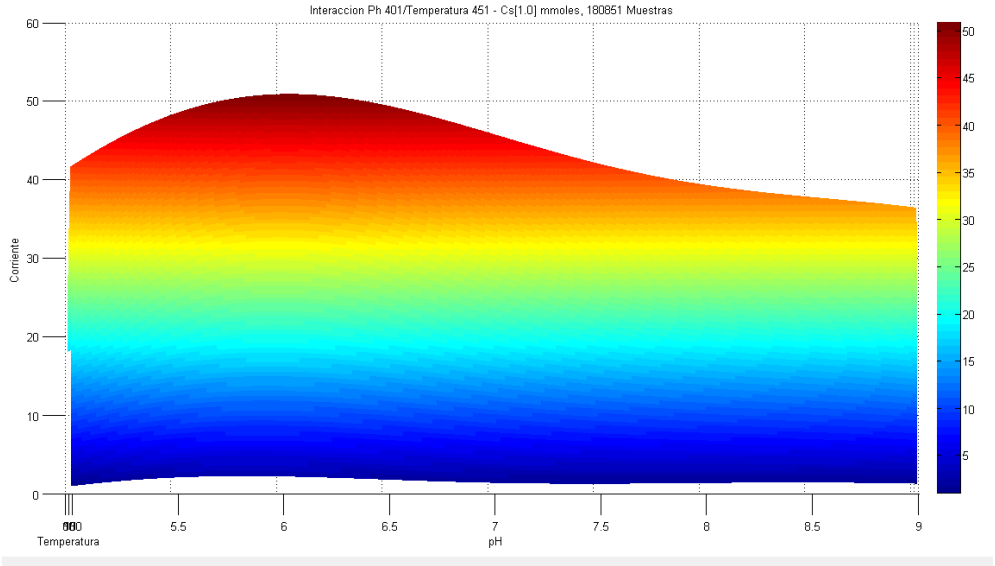


Figura 5.2 Simulación perspectiva pH.

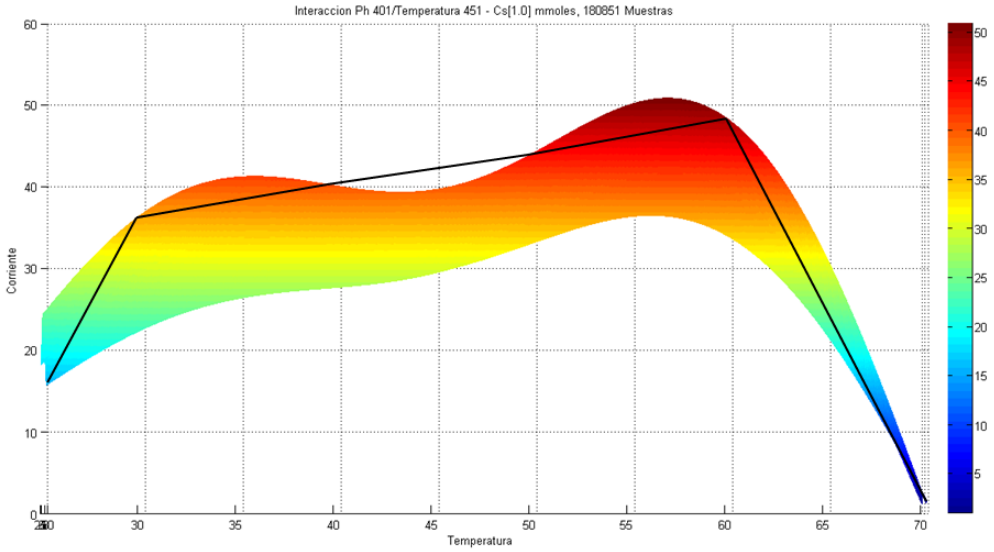


Figura 5.3 Comparación perspectiva Temperatura: Simulación vs datos originales

La Figura 5.3 muestra la comparación de los comportamientos del modelo de regresión y de los datos experimentales, desde la perspectiva de la temperatura, mostrando el comportamiento de ambos y demostrando la buena generalización hecha por el modelo.

5.3.1. Optimo

Con el objetivo de encontrar la muestra que genere el mayor valor de la variable respuesta se realizó una búsqueda a través de todo el conjunto de muestras de simulación. Y se realizó la localización de las posibles zonas donde se encuentra el máximo valor de la corriente (Figura 5.4).

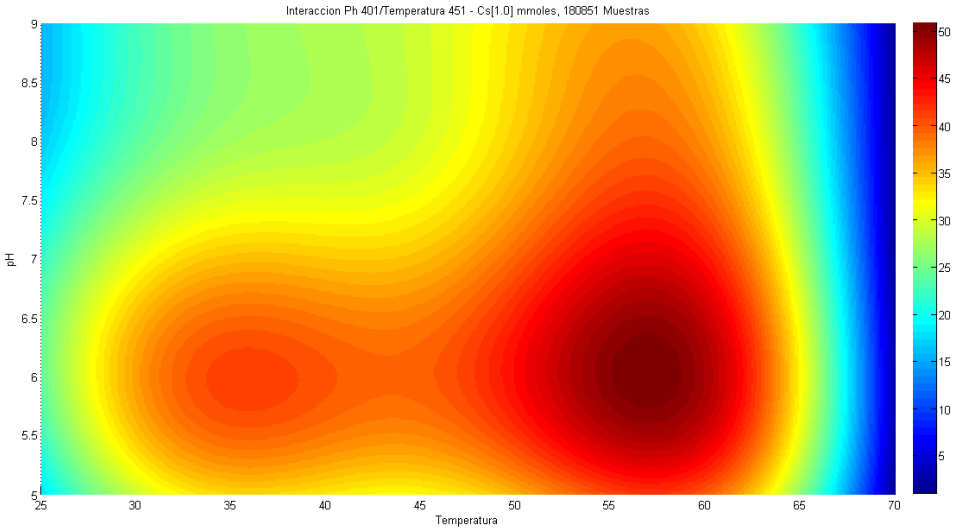


Figura 5.4 Simulación zonas de valores resultantes.

La figura muestra las diferentes zonas producidas por los niveles de la variable respuesta, donde claramente se observa la zona donde se localiza tanto el mínimo valor de la corriente (T = [68 – 70 ˚C], pH = [4-9]), como el máximo (T = [55 – 58 ˚C], pH = [5.5 – 6.5]). De la misma manera demuestra que el valor máximo provisto por el modelo de regresión y el de los datos experimentales son muy cercanos.

Utilizando la figura anterior se generan diferentes intervalos para la localización del valor máximo de la variable respuesta, los cuales se muestran en la Tabla 5.2.

Tabla 5.2 Intervalos de localización del óptimo.

Intervalo (μA)	Temperatura [0.1]	pH [0.01]	# de muestras
±0.001	56.9	**6.03**, 6.04	2
±0.01	56.8≤ t ≤ 57.1	6.00≤ pH ≤ 6.07	24
±0.02	56.7≤ t ≤ 57.2	5.99≤ pH ≤ 6.08	48
±0.05	56.5≤ t ≤ 57.3	5.96≤ pH ≤ 6.11	117
±0.1	56.3≤ t ≤ 57.5	5.93≤ pH ≤ 6.14	232
±0.2	56.0≤ t ≤ 57.8	5.88≤ pH ≤ 6.19	472
±0.5	55.5≤ t ≤ 58.3	5.79≤ pH ≤ 6.29	1,117
±1.0	54.8≤ t ≤ 58.8	5.69≤ pH ≤ 6.41	2353
±2.0	53.8≤ t ≤ 59.6	5.55≤ pH ≤ 6.58	4810
±5.0	51.40≤ t ≤ 61.1	5.28≤ pH ≤ 6.98	12,896

Las dos muestras presentadas en el intervalo ±0.001 presentan el máximo valor de la variable respuesta a un nivel perceptible, ya que el máximo se alcanza con la combinación:

$\mathbf{C_s}$ = 1.0 mMol/ L,

t = 56.9 ºC,

pH = 6.03,

que genera una corriente de 50.9502 μA, y la segunda muestra que tiene la misma configuración a excepción de tener un pH = 6.04, genera aproximadamente la misma corriente, con una diferencia de 0.000004 μA.

En la Figura 5.5 se remarca la zona con el intervalo ±1.0 μA y se localiza el máximo valor de la corriente expresado a través de la simulación de las muestras.

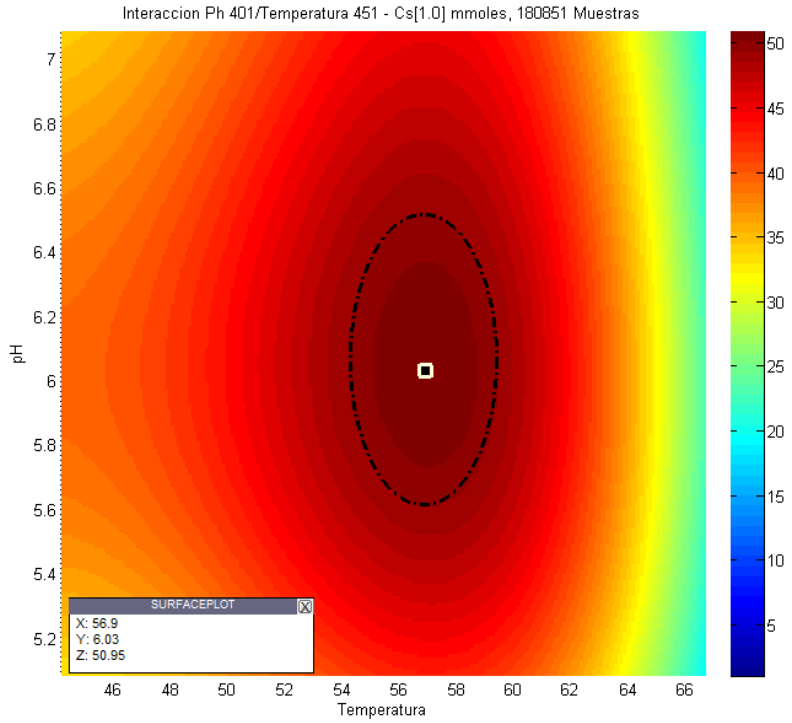


Figura 5.5 Localización del óptimo.

5.4. Conclusiones del capitulo

Se considera que los objetivos de la simulación fueron alcanzados, ya que muestra el mismo comportamiento que tiene los datos experimentales, mientras nos permite conocer las secciones de datos que no eran alcanzables por la tasa de incremento de los factores.

Aunque se considera que un proceso de simulación de datos, no debería contar con las muestras con las que el simulador fue desarrollado, por la cantidad de muestras simuladas (más de 225,000 en caso de las gráficas) no es considerado relevante (150 muestras experimentales).

Se descubre un nuevo máximo global (sección 5.3.1), el cual no era observable en los datos originales, sin embargo, su localización estaba en una zona muy cercana al antiguo máximo.

6. CONCLUSIONES

Esta investigación resuelve las dudas inconclusas encontradas en la literatura acerca del aprendizaje de funciones biológicas provistas por la utilización de sensores electroquímicos. Manteniendo las virtudes de estos, se presenta un proceso detallado en el desarrollo de los diferentes modelos de aprendizaje, y el procedimiento para la evaluación de sus resultados.

Se analizan los datos experimentales a través de las diferentes perspectivas (C_s, pH t), así como sus comportamientos y posibles zonas de encontrar los valores extremos de la corriente resultante. Lo anterior, con el objetivo de encontrar la combinación de parámetros que maximicen la sensibilidad de la determinación de la ACh.

Según el análisis del estado del arte, el algoritmo más utilizado para generar la función de aptitud (*fitness*) con el propósito de predecir y analizar los parámetros de un biosensor es la Red Neuronal. En la tesis se aplica y analiza un algoritmo computacional basado en redes neuronales.

De la misma forma, se implementa un algoritmo basado en Maquinas de Soporte Vectorial, el cual es un algoritmo computacional ampliamente recomendado por la literatura, por su buen desempeño en problemas de clasificación o regresión de índole biológica.

Se ha desarrollado el procedimiento utilizado para entrenar los dos algoritmos. De igual manera se ha creado un modelo que simula el comportamiento de la función de aptitud; la cual es evaluada usando procedimientos estadísticos recomendados por la literatura, tales como el error cuadrático medio y la validación cruzada, así como comparación grafica basada en la visualización del comportamiento de los modelos.

Aunque los resultados obtenidos utilizando Redes Neuronales fueron muy buenos, las máquinas de soporte vectorial muestran las mejores aproximaciones, tanto en las pruebas realizadas a los algoritmos, como en la simulación de los modelos, manteniendo de la mayor forma posible, el comportamiento mostrado en el análisis de los datos experimentales. Sin embargo ninguno de los modelos podrá comprobar cuantitativamente

los resultados de la simulación sin producir nuevas muestras de datos a través de experimentos electroquímicos para su evaluación.

La alta exactitud de la aproximación obtenida por la simulación del modelo basado en máquinas de soporte vectorial permitió la localización del área del óptimo de la corriente, siendo diferente al valor óptimo encontrado en los datos experimentales.

7. PRODUCTOS DE INVESTIGACIÓN

1. García E. R., Burtseva L., Stoytcheva M., González F. F. *Predicting the behavior of the interaction of acetylthiocholine, pH and temperature of an acetylcholinesterase sensor*. I. Batyrshin & G. Sidorov (Eds.): MICAI 2011, Part I, LNAI 7094, pp. 583–591, 2011. Springer-Verlag, Berlin-Heidelberg.

2. García Curiel E. R., Burtseva L., Stoytcheva M. *Optimización de parámetros de un biosensor de acetilcolinesterasa aplicando modelado matemático*. ELECTRO-2010, Chihuahua, Chih., V. XXXII, P. 398-403.

3. García Curiel E. R., Burtseva L., Stoytcheva M. *Optimización de parámetros de un sensor de Acetilcolinesterasa*. Segundo Congreso Nacional de Estudiantes de Postgrado del Instituto de Ingeniería, Mexicali, B. C., 2010, 7 p.

4. *Predicting the behavior of the interaction of acetylthiocholine, pH and temperature of an acetylcholinesterase sensor*. Presentación impartida en el 10th Mexican International Conference on Artificial Intelligence, MICAI 2011 Puebla, México, Noviembre/Diciembre 2011.

5. *Optimización de funciones Bioelectricas utilizando Modelos Computacionales.* Conferencia impartida en el Simposio de la Facultad de Ingeniería y Negocios Campus Cd. Guadalupe Victoria de la U. A. B. C., 14 de Noviembre, 2011.

6. *Aplicación de modelos de aprendizaje automatizado para el modelado de parámetros de un biosensor electroquímico*. Presentación impartida en el 1er Encuentro de Experiencias de Estudiantes UABC en la Investigación, UABC, Tecate, B. C., 13 de Noviembre, 2012.

8. TRABAJO FUTURO

Dada la variedad de formas de resolver este trabajo de investigación y el uso primordial de las MSV se considera que además de lo realizado, se debe tener en cuenta:

- Desarrollar una función kernel específica para estos datos.
- Probar otros subtipos de MSV.
- El desarrollo de una forma de visualización que permita observar los vectores de soporte, para problemas de más de 3 dimensiones.

El posible desarrollo de un kernel particular puede arrojar mejores resultados, sin la tendencia de caer en problemas de sobreajuste. El probar otros tipos de MSV, además de los diferentes métodos que utilizan para generalizar, corroboraría el destacado desempeño que tienen usando datos de tipo biológico. Durante la investigación jamás se encontró una forma de visualizar los vectores de soporte en problemas de regresión de más de 3 dimensiones, sin embargo, existe una gran demanda de este tipo de visualización, puesto que ayudaría a realizar un mejor análisis cualitativos de los resultados.

REFERENCIAS

1. Alonso, G., Istamboulie, G., Ramirez-Garcia, A., Noguer, T., Marty, J., Muñoz, R., Artificial neural network implementation in single low-cost chip for the detection of insecticides by modeling of screen-printed enzymatic sensor response. *Computers and electronics in Agriculture*. Vol 74, pp. 223-229, 2010.
2. Betancourt, G., Las Máquinas de Soporte Vectorial. *Scientia et Technica*. Año XI, No 27, April, 2005.
3. Bishop, C., Neural Networks for Pattern Recognition. Clarendon Press, Oxford: 482 P. (1995).
4. Britton, H. T. K. & R. A. Robinson., CXCVIII-Universal buffer solutions and the dissociation constant of veronal. *Journal of the Chemical Society*. 1456-1462. (1931)
5. Chang, Ch.-Ch. & Lin Ch.-J., LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf., 2011, consultado 30/03/2012
6. Cortes C., Vapnik, V., Support-vector network. *Machine Learning*, V. 20, pp. 273–297, (1995).
7. L. Devroye & T.J. Wagner. The L1 convergence of kernel density estimates, *Annals of Statistics*, V.. 7, pp. 1136-1139, 1979.
8. FAOSTAT: http://faostat.fao.org/site/424/default.aspx#ancor (accessed on 09.01.2011)
9. Fukuto, R. Mechanism of action of organophosphorus and carbamate insecticides. *Environmental Health Perspectives*, V. 87, pp. 245-254, (9901).
10. Geisser, S. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, V. 70, pp. 320–328. (1975).
11. Guide C.-W. Hsu, C.-C. Chang, C.-J. Lin. : A practical guide to support vector classification . Technical report, Department of Computer Science, National Taiwan University. (2003).
12. Gunn, S. R., Support Vector Machines for classification and regression. University of Southampton. Mayo, (1998).
13. Gupta, R. C. (ed.), Toxicology of organophosphate & carbamate compounds., Elsevier Academic Press, London (2005).

14. Hamel, L., Knowledge Discovery with Support Vector Machines. University of Rhode Island. (2009).
15. Jeannot, R., Dagnac, T. In: Chromatographic analysis of the environmental. Nollet L. (Ed.), 841-889, CRC Press, Boca Raton, London, New York, 2006.
16. Jones, M.T., Artificial Intelligence: A Systems Approach. Jones and Bartlett Publishers , LLC, 2009, 498 p., Sudbury, MA, USA.
17. Luger, G. F., Artificial Intelligence: structures and strategies for complex problem solving. Pearson Addison-Wesley, pp. 754, (2009)
18. Matsumura, F., Toxicology of insecticides, Plenum Press, NY (1985), 598 p.
19. Mitchel, T. M., Machine Learning. McGraw-Hill Science/Engineering/Math, pp. 432, (1997).
20. Pedroza, G., Aplicación de las Maquinas de Soporte Vectorial a Reconocimiento de Hablantes. Universidad Autónoma Metropolitana. México. (2007)
21. Refaeilzadeh, P., Tang, L., Liu, H., Cross-Validation: Encyclopedia of Database Systems Springer (2008).
22. Rangelova, V., Tsankova, D., Dimcheva, N., Soft Computing Techniques in Modeling the pH and Temperature on Dopamine Biosensor, Intelligent and Biosensors, pp. 386, Croatia (2010).
23. Russell, S. J., Norvig, P., Artificial Intelligence: A Modern Approach. Prentice Hall, 1132 P., (2010).
24. Rodriguez-Mozaz, S., Marco, M-P., Lopez de Alda, M. J., Barcela, D.: Biosensors for environmental applications. Future development trends. Pure and Applied Chemistry, 76, 4, 723-752 (2004).
25. Schlecht, P. C.; O'Connor, P. F., (Eds.), NIOSH manual of analytical methods. 4th ed., DHHS (NIOSH) Publication pp. 94-113, 1994.
26. Schomburg, D., Schomburg, I., Chang, A., Springer handbook of enzymes (ed.) V. 9, Springer, Berlin (2003).
27. Shao, J., Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association* Vol. 88, No. 422, pp. 486-494 (Jun., 1993),
28. Sumathi, S., Sivanandam, S. N., Introduction to Data Mining and its Applications. Springer-Verlag Berlin Heidelberg, 828 P., (2006).

29. Smola, A. J., Scholkopf, B., A tutorial on Support Vector Regression. Septiembre, (2003), 24 p., http://alex.smola.org/papers/2003/SmoSch03b.pdf consultado 3.05.2012.
30. Stoytcheva, M., Zlatev, R., Velkova, Z., Valdez, B.: Organophosphorus Pesticides Determination by Electrochemical Biosensors. In: Pesticides-Strategies for Pesticides Analysis (M. Stoytcheva ed.), InTech, Croatia, pp. 359-372 (2011).
31. Thévenot, D. R., Tóth, K., Durst, R. A., Wilson, G. S.: Electrochemical biosensors: recommended definitions and classification. Pure and Applied Chemistry, 71, 12, 2333-2348 (1999).
32. Vapnik,V., Statistical Learning Theory. Wiley, New York. 736 P., (1998).
33. Wolfgang, E., Introduction to Artificial Intelligence. Springer, 316 P., 2011

ANEXOS

Anexo A. Estructura de Archivos Anexados

En el Disco se muestra el archivo "Estructura.docx", el cual contiene la descripción de todos los archivos del disco.

Anexo B. Conjunto Original de Datos

En el Disco se muestra el archivo "AChE.xls", el cual contiene el conjunto de datos experimentales en su estado original (sin normalizar).

Anexo C. Conjunto de Datos de Entrenamiento/Validación.

En el Disco se muestra el archivo "AChETrain.xls", el cual contiene el conjunto de los datos utilizados para el entrenamiento y validación de los algoritmos de aprendizaje, en su estado original y normalizados (min-max 0-1).

Anexo D. Conjunto de Datos de Prueba

En el Disco se muestra el archivo "AChETest.xls", el cual contiene el conjunto de los datos utilizados para las pruebas de los algoritmos de aprendizaje, en su estado original y normalizados (min-max 0-1).

Anexo E. Predicción de Conjunto de Pruebas: Red Neuronal

En el Disco se muestra el archivo "NNPredict.xls", el cual contiene el conjunto de datos utilizados para pruebas, las predicciones hechas por la Red Neuronal y el error que existe entre estos.

Anexo F. Predicción de Conjunto de Pruebas: SVM

En el Disco se muestra el archivo "SVMPredict.xls", el cual contiene el conjunto de datos utilizados para pruebas, las predicciones hechas por la Maquina de Soporte Vectorial y el error que existe entre estos.

Anexo G. Simulador de las funciones de regresión: SVM y Red Neuronal (Matlab)

En el Disco cuenta con un simulador de las funciones de regresión formadas por la SVM y la Red Neuronal. Es una aplicación que cuenta con una interfaz gráfica, así como con una interfaz por consola, desarrollada en Matlab. Si desea utilizar el simulador de la función de la SVM es necesario descargar la librería Libsvm (Chang & Lin, 2011) de la página http://www.csie.ntu.edu.tw/~cjlin/libsvm/libsvm-3.12.zip