PREDICCIÓN ADAPTATIVA DE LA CALIDAD DEL AGUA MEDIANTE TÉCNICAS DE INTELIGENCIA COMPUTACIONAL

IVÁN DARÍO LÓPEZ GÓMEZ

Tesis de Maestría en Ingeniería Telemática

Director:
Juan Carlos Corrales
Doctor en Ciencias de la Computación

Universidad del Cauca
Facultad de Ingeniería en Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Investigación e-@mbiente

Popayán, Diciembre de 2015

IVÁN DARÍO LÓPEZ GÓMEZ

PREDICCIÓN ADAPTATIVA DE LA CALIDAD DEL AGUA MEDIANTE TÉCNICAS DE INTELIGENCIA COMPUTACIONAL

Tesis presentada a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para la obtención del Título de

Magíster en:
Ingeniería Telemática

Director:
PhD. Juan Carlos Corrales

Popayán
2015

Quiero agradecer a Dios por permitirme afrontar este gran reto con la mayor dedicación y responsabilidad para finalizar así una etapa más de formación académica y personal.

A mis padres, familiares, compañeros y amigos por su apoyo incondicional.

A mi tutor, el Doctor Juan Carlos Corrales por su permanente disposición y sus contribuciones en el desarrollo de esta tesis, igualmente al Doctor Apolinar Figueroa Casas por sus valiosos aportes dentro del proyecto y a cada una de las personas que de alguna u otra forma participaron en el desarrollo del mismo.

Finalmente a nuestra Alma Mater, la Universidad del Cauca, en especial a la Facultad de Ingeniería Electrónica y Telecomunicaciones por brindar el espacio propicio para este proceso de formación.

Muchas gracias a todos.

Resumen Estructurado

Antecedentes: Colombia es el sexto país con mayor oferta hídrica en el mundo, sin embargo el Ministerio del Medio Ambiente calcula que aproximadamente un 50% de los recursos hídricos presentan problemas de calidad. En muchas ocasiones para controlar unas condiciones apropiadas de calidad del agua, no es suficiente con establecer actividades de monitoreo, además de esto es necesario contar con modelos o mecanismos que permitan anticiparse a la materialización del riesgo con el suficiente rango de tiempo para prevenir los efectos negativos que perturben la calidad del recurso hídrico. En este sentido, la predicción de la calidad del agua desempeña un papel muy importante para muchos sectores socio-económicos que dependen del uso de este líquido.

Objetivos: Proponer un mecanismo de predicción de la calidad del agua mediante un enfoque adaptativo, que soporte los procesos de toma de decisiones sobre diferentes usos del recurso hídrico.

Métodos: Se propone un mecanismo adaptativo de predicción de la calidad del agua empleando técnicas de Inteligencia Computacional; el enfoque principal de este mecanismo es la posibilidad de ser aplicado sobre conjuntos de datos pertenecientes a diferentes usos del agua sin que la precisión de las predicciones se vea afectada de una manera drástica. Este mecanismo consta de un componente de calibración de parámetros, un componente encargado de realizar las operaciones de predicción utilizando técnicas de IC y un componente de adaptación, el cual a su vez implementa el algoritmo de IC que permite ajustar los valores de predicción a los valores reales del uso del agua seleccionado.

Resultados: Tres conjuntos de datos de calidad del agua (piscícola, consumo humano y recreacional), un mecanismo adaptativo de predicción de la calidad del agua basado en técnicas de Inteligencia Computacional.

Conclusiones: La Regresión por Vectores de Soporte configurada con el kernel PUK presentó un mejor desempeño en la precisión de las predicciones respecto a otras técnicas de Inteligencia Computacional. Para que los valores predichos por el mecanismo se aproximaran a los valores reales en diferentes usos del agua, fue necesario utilizar la técnica de Optimización por Nubes de Partículas (PSO).

Palabras Clave: Calidad del Agua, Inteligencia Computacional, Predicción, Sistemas Adaptativos Complejos, Regresión por Vectores de Soporte, Optimización por Nubes de Partículas.

Structured Abstract

Background: Colombia is the sixth country with the greatest water supply in the world, but the Ministry of Environment estimates that approximately 50% of water resources have quality problems. Often to control proper water quality conditions, it is not sufficient to establish monitoring activities in addition to this there is a need models and mechanisms to anticipate the risk materialized with enough time range to prevent negative effects disturbing the quality of water resources. In this sense, predicting water quality plays a very important role for many socio-economic sectors that depend on the use of this liquid.

Goals: Propose a mechanism for water quality prediction through an adaptive approach that supports decision-making processes on different uses of water resources.

Methods: An adaptive prediction mechanism of water quality using Computational Intelligence techniques is proposed; the main focus of this mechanism is the ability to be applied to datasets from different water uses without the prediction accuracy is affected in a drastic way. This mechanism consists of a parameter calibration component, a predictive component, responsible for performing prediction operations using CI techniques and finally an adaptive component, which implements the CI algorithm to adjust the predicted values to actual values in the water use selected.

Results: Three water quality datasets (aquaculture, human consumption and recreational use), an adaptive prediction mechanism of water quality based on Computational Intelligence techniques.

Conclusions: Support Vector Regression configured with PUK kernel presented a better performance in the accuracy of predictions compared to other techniques of Computational Intelligence. In order that the predicted values by the mechanism from approaching the actual values in different uses of water, it was necessary to use the Particle Swarm Optimization (PSO) technique.

Keywords: Water Quality, Computational Intelligence, Forecasting, Complex Adaptive Systems, Support Vector Regression, Particle Swarm Optimization.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Ecuaciones

Capítulo 1

Introducción

1.1. Planteamiento del Problema

La historia de la humanidad ha estado ligada estrechamente al agua. Desde sus albores existen reseñas a grandes inundaciones, sequías y a los esfuerzos del hombre por dominar este recurso vital de la naturaleza para cubrir sus propias necesidades [1]. Tradicionalmente el agua se ha tratado y gestionado como si fuese un recurso ilimitado debido a su naturaleza renovable, sin embargo, el incremento indiscriminado de su uso ha acarreado consigo un acelerado deterioro en su calidad y en ocasiones se ven cambios en su distribución temporal y espacial, con consecuencias que no son previstas por completo, pero de una importante gravedad ambiental, económica, social, entre otras.

La calidad del agua se puede definir como el conjunto de características físicas, químicas, biológicas y radiológicas de los cuerpos de agua superficiales y subterráneos[1] [2]. Así mismo, estas características afectan la capacidad del agua para sustentar tanto a las comunidades humanas como la vida vegetal, animal y microbiana. Por otra parte, los expertos en el manejo del recurso hídrico, indican que no existe una única definición de la calidad del agua, pues esta depende estrictamente del uso al que esté destinado el preciado líquido [3], es así como por ejemplo, el agua que no se puede utilizar para el consumo humano, puede servir para otras actividades como el riego o la piscicultura, entre otros, ya que posee características específicas que la hacen apropiada para dicho uso [4].

A pesar de que Colombia es el sexto país con mayor oferta hídrica en el mundo[2], el Ministerio del Medio Ambiente calcula que aproximadamente un 50% de los recursos hídricos presentan problemas de calidad [5]. Lo anterior obedece en gran medida al crecimiento de la población y de las actividades económicas, siendo necesario un

[1] Para este trabajo se manejará esta definición de Calidad del Agua.
[2] Este ranking se ha ido especializando aún más, actualmente para el análisis se tienen en cuenta aspectos como la calidad y la disponibilidad que la población tiene del recurso. Bajo estos nuevos criterios, Colombia se ubica en el puesto 24 a nivel mundial.

monitoreo y control constante que permita tomar las acciones necesarias para abordar esta problemática con el fin de disminuir su impacto en los procesos naturales y sociales, especialmente en la salud humana [6]. Sin embargo, en muchas ocasiones para controlar unas condiciones apropiadas de calidad del agua, no es suficiente con establecer actividades de monitoreo[3] que brinden acciones correctivas ante determinado tipo de contaminación, sino que además de esto, se hace necesario contar con modelos o mecanismos que permitan anticiparse a la materialización del riesgo de contaminación con el suficiente rango de tiempo para prevenir los efectos negativos que afecten la calidad del recurso hídrico.

En este sentido, la predicción de la calidad del agua desempeña un papel muy importante para muchos sectores socio-económicos que dependen del uso del preciado líquido. Es así como en los últimos años, el campo de la Inteligencia Artificial (IA) [7], ha introducido algoritmos y técnicas de predicción que cuentan con la capacidad de estimar las condiciones futuras de un cuerpo de agua con base en el análisis de los datos que han sido recolectados en el pasado.

Dada la importancia que representa la gestión de la calidad del recurso hídrico para diferentes sectores socio-económicos, varias propuestas investigativas se han enfocado en la predicción como una herramienta útil para anticipar posibles eventos adversos tanto para la producción como para la salud humana. Los trabajos más destacados hacen uso de las Redes Neuronales Artificiales (RNA) como en [8]–[10]. Adicional a esto, otros trabajos plantean modelos de predicción híbridos en los cuales combinan las RNA y los Modelos Auto-regresivos Integrados de Media Móvil, también conocidos como modelos ARIMA[4] [11], [12] que trabajan con series de tiempo, con el objetivo de mejorar la precisión de las predicciones. Por otro lado, trabajos como [13]–[16] permiten optimizar la selección de los parámetros más representativos para tener en cuenta en la predicción (en muchas ocasiones adicionar demasiados parámetros no es conveniente debido a que genera ruido en el modelo y en vez de aumentar la precisión, esta disminuye). Actualmente nuevas estrategias se han planteado para mejorar la precisión del proceso de predicción como en [17] que utiliza Máquinas de Vectores de Soporte (SVM)[5] y [18], donde se

[3] Estas actividades de monitoreo pueden ser mediante la toma de muestras de agua que se realizan de forma manual o automatizada mediante sensores electrónicos.
[4] Acrónimo del inglés, “Auto-Regressive Integrated Moving Average”.
[5] Acrónimo del inglés, “Support Vector Machine”.

hace uso de esta misma técnica adicionando la Optimización por Nubes de Partículas (PSO)[6] obteniendo mejores resultados en la precisión comparada con los trabajos mencionados anteriormente.

Con base en lo anterior, es importante destacar los avances y aportes de estos trabajos dentro del campo de la predicción de la calidad del agua, sin embargo ninguno de ellos considera la aplicación de su modelo de predicción sobre diferentes usos del recurso hídrico, estos se limitan a realizar el proceso de predicción tomando como referencia los datos pertenecientes a un determinado uso del agua, muchas veces sesgando los resultados hacia ese escenario en particular sin tener en cuenta la aplicabilidad que el mismo modelo de predicción pueda tener para otros usos y que puede ser de utilidad para diversos sectores que utilizan el agua en sus procesos productivos. Por consiguiente, en el presente trabajo de investigación se pretende que un mecanismo de predicción de calidad del agua cuente con la característica adaptativa, en el sentido de que pueda ser aplicado en diferentes usos del agua manteniendo la precisión de las predicciones; lo anterior mediante técnicas de Inteligencia Computacional (IC).

Teniendo en cuenta las consideraciones anteriores, en la presente propuesta de investigación se formula la pregunta ¿Cómo realizar una predicción de la calidad del agua sobre diferentes usos, que permita dar soporte a los procesos de toma de decisiones sobre la gestión del recurso hídrico?

1.2. Escenario de Motivación

El agua es uno de los elementos naturales que se encuentra en mayor cantidad en nuestro planeta; además es uno de los elementos que más influyen en la posibilidad de desarrollar distintas formas de vida. Anualmente caen casi 110.000 km^3 de precipitación sin incluir los océanos; de esta cantidad casi dos tercios se evaporan de la tierra, los restantes 40.000 km^3 se convierten en escorrentía superficial (ríos y lagos) y en aguas subterráneas (acuíferos). Parte de esta agua es removida mediante infraestructura instalada por humanos y la mayor parte del agua extraída es posteriormente devuelta al medio ambiente después de que se ha utilizado. La calidad del agua de retorno puede haber cambiado durante el uso [19].

[6] Por su sigla en inglés, "Particle Swarm Optimization"

En este sentido, la toma de decisiones sobre el manejo del recurso hídrico es un proceso de gran importancia para muchos sectores productivos, permitiendo de esta manera identificar posibles acciones que incrementen la rentabilidad y reduzcan al máximo las pérdidas económicas. Sin embargo al hablar de calidad del agua no solo hay que referirse al factor económico sino al impacto ambiental y de salubridad que se puede generar. Dado lo anterior surge la necesidad de contar con un mecanismo de predicción que pueda ser utilizado por distintos entes encargados de tomar decisiones sobre la calidad del agua, es decir, el mecanismo debe brindar predicciones fiables independientemente del uso del agua en el que se aplique.

1.3. Objetivos

2.1.1 Objetivo general

Proponer un mecanismo de predicción de la calidad del agua mediante un enfoque adaptativo, que soporte los procesos de toma de decisiones sobre diferentes usos del recurso hídrico.

2.2.1 Objetivos específicos

1. Seleccionar una o más técnicas de Inteligencia Computacional (RNA, SVM, entre otros) que puedan ser utilizadas en el proceso de predicción de la calidad del agua.
2. Definir el/los algoritmo(s) que permitan adaptar el proceso de predicción a diferentes usos del agua.
3. Evaluar experimentalmente el mecanismo propuesto a través del desarrollo de un prototipo, aplicado en dos contextos del uso del agua.

1.4. Contribuciones

Las principales contribuciones de éste trabajo de maestría son las siguientes.

- Tres conjuntos de datos de calidad del agua pertenecientes a los siguientes sitios: a) estuario de Alviso, b) lago Don Pedro, ambos en el estado de California, USA; y c) estación Puente Juanchito del Río Cauca, Colombia. Estos conjuntos de datos corresponden a los usos: piscícola, recreacional y consumo humano respectivamente; y contienen información de variables

fisicoquímicas como temperatura, conductividad, pH, oxígeno disuelto y turbidez, entre otras.

- Un mecanismo adaptativo de predicción de la calidad del agua basado en técnicas de Inteligencia Computacional (Regresión por Vectores de Soporte configurada con el Kernel Universal de Pearson y la técnica de Optimización por Nubes de Partículas).

- Un artículo expuesto en el VII Congreso Iberoamericano de Telemática (CITA 2015) realizado los días 10, 11 y 12 de Junio de 2015 en Popayán - Colombia, el cual tuvo como principal objetivo desarrollar un mapeo sistemático de la literatura relacionada con la predicción de la calidad del agua haciendo uso de técnicas de Inteligencia Computacional. Este artículo será publicado en el Número 28 de enero-junio de 2016 de la Revista Ingenierías Universidad de Medellín, Clasificada A2 en Publindex-Colciencias, ISSN 1692-3324. (Ver ANEXO A).

1.5. Contenido de la Monografía

El presente trabajo de grado está compuesto por cinco capítulos los cuales se describen a continuación.

- **Capítulo 2. Estado Actual Del Conocimiento**

  Presenta una visión general sobre los trabajos relacionados y los conceptos que giran en torno al problema de investigación identificado.

- **Capítulo 3. Mecanismo de Predicción Adaptativo**

  Explica de forma detallada el mecanismo de predicción adaptativo que se propone en el presente trabajo; algoritmos y técnicas de Inteligencia Computacional que se integran para conformar los componentes del mecanismo (calibración de parámetros, predictivo y adaptativo).

- **Capítulo 4. Experimentación y Evaluación**

  Describe los conjuntos de datos de calidad del agua utilizados y el proceso de implementación del prototipo mediante la utilización de una metodología para el desarrollo de software. Además de lo anterior, este capítulo presenta el

proceso de evaluación y las pruebas ejecutadas al mecanismo de predicción desarrollado, con el objetivo de analizar la calidad de los resultados y su rendimiento.

- **Capítulo 5. Cumplimiento de Objetivos**

  En esta sección se realiza un análisis detallado acerca del cumplimiento de los objetivos del proyecto por medio de un modelo de indicadores, además de exponer los lineamientos de conformación e interpretación de estos indicadores.

- **Capítulo 6. Conclusiones y Trabajos Futuros**

  Por último, en este capítulo se analizan los resultados del trabajo realizado, adicionalmente se detallan las principales contribuciones obtenidas en la ejecución del proyecto y se presenta un conjunto de recomendaciones importantes para el desarrollo de trabajos futuros dentro de la misma línea de investigación.

Capítulo 2

Estado Actual del Conocimiento

El presente proyecto dispone de una amplia referencia documental, la cual permite formar una base para los núcleos de trabajo sobre los cuales se orienta: Predicción de la calidad del agua y Calibración de técnicas de predicción de la calidad del agua; ambos núcleos se encuentran enmarcados dentro de la Inteligencia Computacional y se han combinado dentro de los Conceptos Generales (sección 2.1), de tal manera que se presentan una a una, las técnicas utilizadas en las temáticas mencionadas. Para la sección 2.2 se realiza una división de los trabajos correspondientes a cada uno de los núcleos identificados.

2.1. Conceptos Generales

2.1.1 Sistemas Adaptativos Complejos (SAC)

Un SAC es un tipo especial de sistema complejo, considerándose complejo debido a que está conformado de múltiples y diversos elementos interconectados entre sí; y adaptativo, porque tiene la capacidad de cambiar y aprender a partir de la experiencia. Uno de los principales autores que ha trabajado en el área de los SAC es John Holland [20], quien lo define como una red dinámica de muchos agentes (los cuales pueden representar células, especies, individuos, empresas, naciones) actuando en paralelo, constantemente y reaccionando a lo que otros agentes están haciendo. El resultado total del sistema proviene de un enorme número de decisiones tomadas en algún momento por muchos agentes individuales. Las principales características de un SAC son las que se mencionan a continuación.

- Están compuestos por una red de agentes altamente interconectados y que actúan en paralelo, emergiendo la conducta global coherente del sistema de las conductas cooperativas y competitivas de los agentes que lo componen.

- Tienen muchos niveles de organización, en donde los agentes de un nivel son los bloques con los que se construye el nivel inmediatamente superior.

- Constantemente realizan predicciones basadas en sus modelos internos acerca del mundo.

- Tienen múltiples nichos en los que operar, en los que poder adaptarse, lo que genera cambiar de entornos a fin de optimizar su ajuste con el mismo.

2.1.2 Predicción Científica

La predicción científica puede considerarse como un pronóstico razonable y verificable acerca de un hecho o acontecimiento nuevo o desconocido. Una de sus principales características se enfoca en anticipar lo que va a ocurrir; en cambio la inferencia, trata de explicar o interpretar lo que ya ha ocurrido [21].

De acuerdo con lo anterior, el concepto de predicción científica se centra en una declaración precisa de lo que ocurrirá en determinadas condiciones especificadas; de esta manera, su validez se mide por el éxito o acierto que tengan sus predicciones. Las teorías que generan muchas predicciones que resultan de gran valor (tanto por su interés científico como por sus aplicaciones) se confirman o se falsean fácilmente y, en muchos campos científicos, las más deseables son aquéllas que, con número bajo de principios básicos, predicen un gran número de sucesos. Sin embargo existe una multitud de campos en la ciencia donde la predicción se convierte en una tarea compleja, bien sea por el gran número de variables involucradas o por la misma dinámica desconocida de los fenómenos involucrados en una situación problemática [22].

Fenómenos como las horas de salida y la puesta del sol, los eclipses, el tiempo atmosférico, etc. pueden ser predichos. Las predicciones tienen su base en observaciones, mediciones e inferencias; una predicción que no se sustenta en antecedentes serios no es científica: es una adivinanza o conjetura. Por ejemplo, un meteorólogo fundamenta sus predicciones en una serie de observaciones (datos) que provienen básicamente de dos fuentes: estaciones meteorológicas terrenas y oceánicas, además de satélites meteorológicos. Por otra parte el meteorólogo conoce perfectamente cómo se relacionan los factores del clima y fundamenta sus predicciones en la historia de sus registros y en las regularidades.

2.1.3 Inteligencia Computacional (IC)

La Inteligencia Computacional es una rama de la IA, conformada por una colección de paradigmas computacionales con inspiración biológica y lingüística al hacer uso de algoritmos bio-inspirados que emulan la forma de pensar, el modo de procesar información y la resolución de problemas de los sistemas biológicos. En estos paradigmas se incluye la teoría, el diseño, la aplicación y el desarrollo de redes neuronales, sistemas conexionistas, algoritmos evolutivos, sistemas difusos y sistemas inteligentes híbridos [23], [24]. Recientemente, se han incluido nuevas áreas tales como: desarrollo mental autónomo, bioinformática, bioingeniería, finanzas y economía computacional.

La IC comprende el estudio de los mecanismos de adaptación para permitir o facilitar un comportamiento inteligente en entornos complejos y cambiantes. Estos mecanismos incluyen los paradigmas de la IA que exhiben la capacidad de aprender o de adaptarse a nuevas situaciones, generalizar, abstraer, descubrir y asociar.

Por otra parte, cabe mencionar que las técnicas individuales de estos paradigmas de IC se han aplicado con éxito para resolver problemas del mundo real, sin embargo la tendencia actual es el desarrollo de paradigmas híbridos, ya que ningún paradigma es superior a los demás en todas las situaciones. Al hacerlo, se aprovecha los puntos fuertes de los componentes del sistema de IC híbrido, y se eliminan en gran medida las debilidades de los componentes individuales. A continuación se exponen las principales técnicas de IC que enmarcan el presente trabajo de maestría.

Redes Neuronales Artificiales (RNA): Una Red Neuronal Artificial es un sistema capaz de aprender a partir de un conjunto de datos. Imita el funcionamiento del cerebro humano el cual está conformado por neuronas interconectadas entre sí. Del mismo modo, una Neurona Artificial (NA) es un modelo de una Neurona Biológica (BN). Cada NA recibe señales desde el medio ambiente o de otras NA, reúne estas señales y cuando se dispara, transmite una señal a todas las NAs conectadas. Las señales de entrada son inhibidas o excitadas por medio de pesos numéricos asociados a cada conexión a la NA. El disparo de una NA y la intensidad de la señal que sale se controlan a través de una función, conocida como la función de activación. La NA recoge todas las señales entrantes, y calcula una señal de entrada neta como una función de los respectivos pesos. La señal de entrada neta sirve

como entrada a la función de activación que calcula la señal de salida de la NA [23]. La Figura 1 muestra la estructura de una neurona artificial.

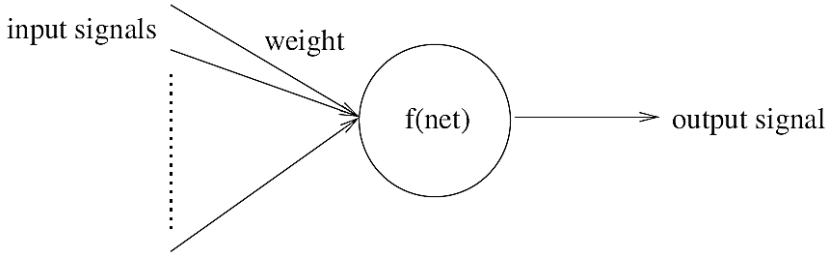


Figura 1. Componentes básicos de una Neurona Artificial. Tomado de [23].

Ahora bien, una RNA es una red de capas de NA que consiste en una capa de entrada, una o más capas ocultas y una capa de salida. Las NA en una capa están conectadas, total o parcialmente, a la NA de la siguiente capa.

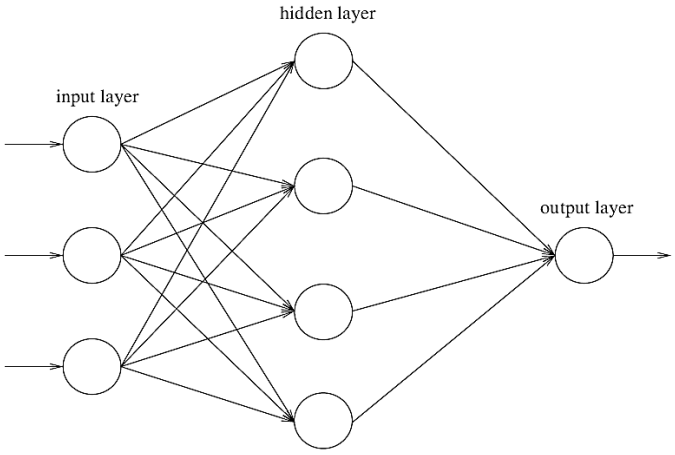


Figura 2. Estructura de una Red Neuronal Artificial. Tomado de [23].

Existen diferentes tipos de RNA, los más conocidos se mencionan a continuación.

- Supervisadas, como las RNA con conexiones hacia adelante (Feedforward) y con conexiones hacia atrás (retro propagación o backpropagation).

- No supervisadas, como los mapas de auto-organización de Kohonen.

Estos tipos de RNA se han utilizado en una amplia gama de aplicaciones, incluyendo el diagnóstico de enfermedades, reconocimiento de voz, la minería de datos, composición de música, procesamiento de imágenes, pronóstico, control de robots,

aprobación de créditos, clasificación, reconocimiento de patrones, estrategias de juego de planificación, de compresión, entre otros.

Máquinas de Vector de Soporte (SVM): Las Máquinas de Vectores de Soporte son un conjunto de métodos de aprendizaje supervisado[7] que se usan para clasificación estadística y análisis de regresión. Las SVM proporcionan soluciones con un balance óptimo entre una representación precisa de los datos existentes y el comportamiento con nuevos datos no utilizados en el proceso de entrenamiento [25]. En contraste con las RNA, los resultados del aprendizaje estadístico tienen complejidad óptima para el conjunto dado de datos utilizados en el aprendizaje y puede tener capacidad de generalización si se seleccionan los parámetros adecuados. Las SVM representan el conocimiento aprendido mediante los puntos o patrones más informativos, llamados vectores soporte. Se aplican con éxito para resolver problemas de clasificación, predicción y detección de novedades.

En 1996 Vladimir Vapnik propuso una nueva versión de SVM para regresión a la cual denominó Regresión por Vectores de Soporte (SVR por su sigla en inglés, de Support Vector Regression); esta técnica es descrita a continuación.

- **Regresión por Vectores de Soporte (SVR).** La idea básica de la SVR consiste en realizar un mapeo de los datos de entrenamiento $x \in X$, a un espacio de mayor dimensión F a través de un mapeo no lineal $\varphi: X \rightarrow F$, donde se puede realizar un proceso de regresión lineal. La Figura 3 muestra la arquitectura general de una SVR.

[7] El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (problemas de regresión) o una etiqueta de clase (problemas de clasificación).

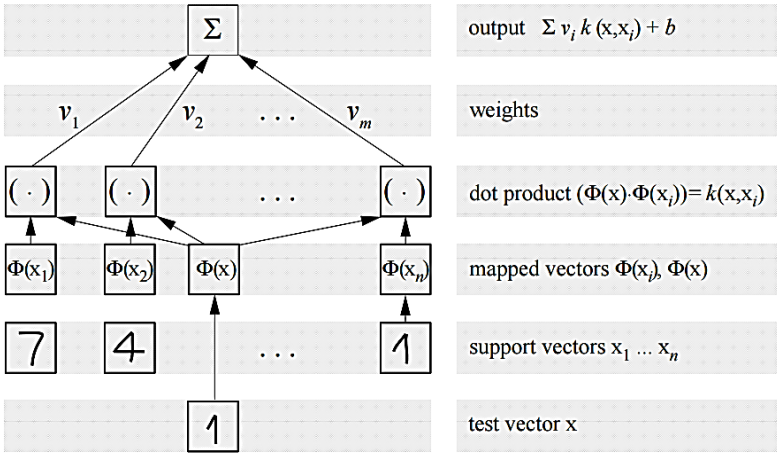


Figura 3. Arquitectura general de una SVR. Tomado de [26].

La figura anterior permite observar que el patrón de entrada (para el cual se desea realizar la predicción) es mapeado dentro de un "*espacio de características*" mediante un vector de mapeo Φ, posteriormente se calculan los productos escalares con las imágenes de los patrones de entrenamiento que entrega el vector de mapeo. Esto corresponde a realizar una evaluación de las funciones kernel en los puntos $k(x_i,x)$. Finalmente los productos escalares son sumados usando los pesos α_i - α_i*, adicionalmente es sumada una constante b al valor final resultado de la predicción [26].

De la misma manera que con el enfoque de clasificación, se busca optimizar los límites de generalización para la regresión; es decir, existe una función que descarta los errores que están situados dentro de la distancia determinada del valor real. Este tipo de función se llama a menudo función de pérdida intensiva épsilon. La Figura 4 muestra un ejemplo de esta función, en el cual sólo los puntos fuera de la región sombreada contribuyen al costo en la medida en que las desviaciones son penalizadas de forma lineal.

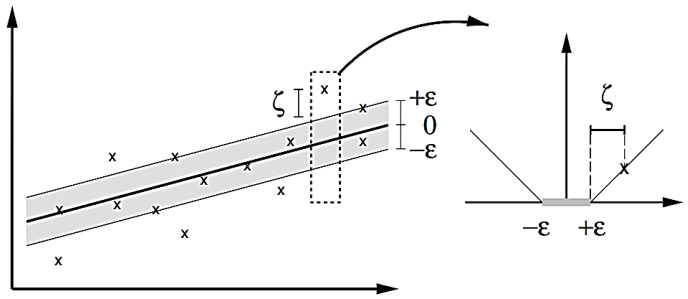


Figura 4. Función de pérdida intensiva épsilon. Tomado de [26].

Comparativa entre RNA y SVM. Teniendo en cuenta lo expuesto acerca de las RNA y las SVM, en la Tabla 1 se presentan, a modo de comparación, las principales características correspondientes a cada una de estas dos técnicas.

RNA	SVM
Capas ocultas transforman a espacios de cualquier dimensión.	Kernels transforman a espacios de dimensión muy superior.
Espacio de búsqueda con múltiples mínimos locales.	El espacio de búsqueda posee sólo un mínimo global.
Entrenamiento costoso.	Entrenamiento muy eficiente.
Clasificación muy eficiente.	Clasificación muy eficiente.
Se puede diseñar el número de capas ocultas y nodos.	Se diseña la función kernel y el parámetro de coste *C*.
Muy buen funcionamiento con problemas típicos.	Muy buen funcionamiento con problemas típicos.
	Extremadamente robusto para generalización, menos necesidad de heurísticas para entrenamiento.

Tabla 1. Comparación entre las RNA y las SVM. Tomado de [27].

Computación Evolutiva (CE): Este paradigma tiene como objetivo imitar los procesos de la evolución natural, en el que el concepto principal es la supervivencia de los más aptos: los débiles deben morir [23]. En la evolución natural, la supervivencia se logra a través de la reproducción. Hijos, reproducidos a partir de dos padres (a veces más de dos). Los algoritmos evolutivos utilizan una población de individuos, donde un individuo es un cromosoma. Un cromosoma define las características de los individuos de la población. Cada característica se conoce como un gen. Para cada generación, los individuos compiten para reproducir descendencia. Aquellos individuos con las mejores capacidades de supervivencia tienen la mejor oportunidad de reproducirse.

Se han desarrollado diferentes clases de algoritmos evolutivos, algunos de los más importantes son los siguientes.

- Algoritmos genéticos con modelo de evolución genética.
- Programación genética basada en algoritmos genéticos, pero los individuos son programas (representados como árboles).
- Programación evolutiva derivada de la simulación de la conducta adaptativa en la evolución (evolución fenotípica).

El presente trabajo de maestría aborda la técnica de algoritmo genético dentro de los trabajos relacionados, de esta manera este tipo de algoritmo se detalla a continuación.

- **Algoritmo Genético (AG).** Este algoritmo es probablemente el primero que simula los sistemas genéticos, propuesto inicialmente por Alex Fraser, pero extendido ampliamente a partir del trabajo de Jhon Holland [28]. La Figura 5 presenta las principales fases de este algoritmo.

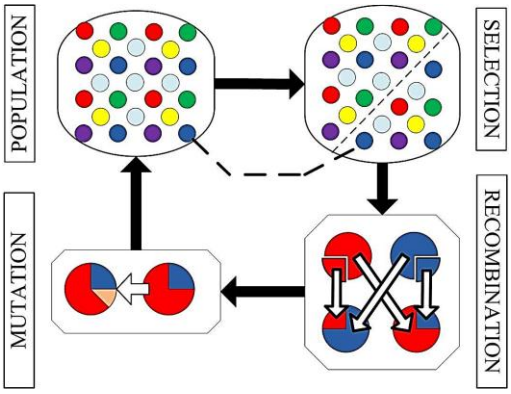


Figura 5. Fases básicas de un algoritmo genético. Tomado de [29].

Este algoritmo funciona dentro del conjunto de soluciones de un problema, este conjunto es denominado fenotipo; y el conjunto de individuos de una población natural, el cual codifica la información de cada solución en una cadena, por lo general binaria, llamada cromosoma; esta cadena está conformada por símbolos llamados genes. Cuando la representación de los cromosomas se hace con cadenas de dígitos binarios se le conoce como genotipo. Los cromosomas evolucionan a través de iteraciones, conocidas

como generaciones. En cada generación, los cromosomas son evaluados usando alguna medida de aptitud. Las siguientes generaciones (nuevos cromosomas), son generadas aplicando los operadores genéticos repetidamente, siendo estos los operadores de selección, cruzamiento, mutación y reemplazo.

2.1.4 Inteligencia de Enjambres (IE)

La inteligencia de enjambre se originó a partir del estudio de las colonias o enjambres de organismos sociales. Los estudios sobre el comportamiento social de los organismos (individuos) en enjambres impulsaron el diseño de algoritmos de optimización y de agrupamiento muy eficientes. Por ejemplo, los estudios de simulación de la coreografía elegante, pero impredecible, de bandadas de aves llevaron al diseño del algoritmo de Optimización por Nubes de Partículas (conocido por su sigla en inglés: PSO, de “Particle Swarm Optimization”), así como estudios sobre el comportamiento de forrajeo de las hormigas resultaron en algoritmos de optimización de colonias de hormigas [21].

Los estudios de las colonias de hormigas han contribuido ampliamente para el conjunto de algoritmos inteligentes. El modelado de deposición de feromona de las hormigas en su búsqueda de los caminos más cortos a las fuentes de alimentos contribuyó con el desarrollo de algoritmos de optimización de la ruta más corta. Otras aplicaciones de la optimización de colonia de hormigas incluyen la optimización del enrutamiento en redes de telecomunicaciones, la programación y la solución del problema de asignación cuadrática, entre otros. Los estudios de la construcción del nido de las hormigas y las abejas dieron como resultado el desarrollo de algoritmos de agrupamiento y de optimización estructural.

Para el presente trabajo de maestría es de gran importancia la técnica de PSO, dado que es el componente principal que permite adaptar las predicciones a diferentes usos del agua. Teniendo en cuenta lo anterior, a continuación se describe esta técnica.

- **Optimización por Nubes de Partículas (PSO).** Este algoritmo es una técnica de búsqueda basada en población. Permite emular el comportamiento social de las aves dentro de un grupo. La intención inicial del concepto de PSO fue simular gráficamente la coreografía curiosa e impredecible de una bandada de

aves, sin embargo este concepto evolucionó en un algoritmo simple y eficiente optimización [30].

PSO permite optimizar un problema a partir de una población de soluciones candidatas las cuales son denotadas como "partículas". Dichas partículas se mueven por todo el espacio de búsqueda de acuerdo a reglas matemáticas que tienen en cuenta la posición y la velocidad de las mismas. En el movimiento de cada partícula influye su mejor posición local hallada hasta el momento, de igual forma también influyen las mejores posiciones globales que han sido encontradas por otras partículas a medida que recorren el espacio de búsqueda (iteraciones). El fundamento teórico de esto es hacer que la nube de partículas converja rápidamente hacia las mejores soluciones. La Figura 6 muestra el espacio de búsqueda (en este caso de tres dimensiones) en el cual el conjunto de partículas realizan la búsqueda de la mejor solución del problema, bien sea un máximo o un mínimo global.

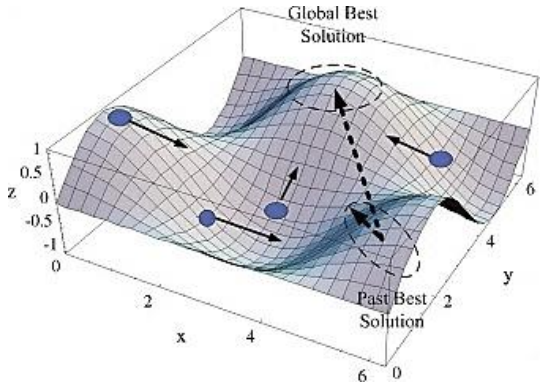


Figura 6. Espacio de búsqueda de PSO para encontrar un valor óptimo global. Tomado de [31].

Finalmente cabe resaltar que PSO es una meta-heurística, por el hecho de asumir pocas o ninguna hipótesis sobre el problema a optimizar. Por otra parte puede aplicarse en grandes espacios de soluciones candidatas. Sin embargo, como toda meta-heurística, esta técnica no garantiza la obtención de una solución óptima en todos los casos.

2.2. Trabajos Relacionados

Para establecer un punto de partida alrededor del conocimiento generado en el campo de la predicción de la calidad del agua, es de gran importancia adoptar ciertos lineamientos en este campo de investigación que permitan definir un método para construir un esquema de clasificación general sobre dicho campo y por consiguiente, identificar el tipo y la cantidad de investigación disponible. Teniendo en cuenta lo anterior, este trabajo de maestría está guiado por la construcción de un Mapeo Sistemático [32] para establecer los núcleos temáticos generales de acuerdo al análisis de tendencias; y una Revisión Sistemática [33] para identificar las brechas específicas de investigación. Vale la pena mencionar que ambas metodologías están basadas en un procedimiento conformado por cinco etapas: A) Definir preguntas de investigación, B) Realizar la búsqueda literaria, C) Seleccionar estudios, D) Clasificar artículos y E) Extraer y realizar la agregación de datos. A continuación se detallan cada uno de los pasos.

2.2.1 Metodología de la Investigación

A. **Preguntas de Investigación.** El objetivo principal de este estudio se centra en conocer las técnicas de IC más utilizadas dentro del campo de aplicación de la predicción de la calidad del agua y cuál es la tendencia en los nuevos trabajos de investigación. Para obtener un conocimiento más detallado y una visión integral del tema, se plantean las siguientes preguntas de investigación:

   RQ.1. ¿Qué temas interesan a la comunidad científica respecto a la predicción de la calidad del agua aplicando técnicas de IC dentro de un marco temporal reciente?

   RQ.2. ¿Qué técnicas de IC se han utilizado para brindar capacidad adaptativa a los mecanismos de predicción?

   RQ.3. ¿Cuáles son las técnicas de IC que aún no han sido exploradas en su totalidad y que pueden ser una importante alternativa dentro de la predicción adaptativa de la calidad del agua?

B. **Fuente de Datos y Estrategia de Búsqueda.** Para implementar la estrategia de búsqueda fue tenida en cuenta la terminología referente a la predicción de la

calidad del agua principalmente en el idioma inglés, en el cual se encuentran escritos la mayor parte de los trabajos. Las cadenas de búsqueda utilizadas son presentadas en la Tabla 1. Por otro lado las bases de datos utilizadas para realizar la búsqueda fueron establecidas a partir de la importancia de cada una dentro de la comunidad científica a nivel internacional; estas bases de datos son: *Google Schoolar, Elsevier, IEEE, ACM y Springer.*

C. **Selección de Estudios.** Dentro del proceso de selección de estudios se establecen dos criterios para determinar la relevancia de los mismos:

Inclusión: *aquellos trabajos que se centran en predicción de la calidad del agua haciendo uso de técnicas de IC y también los que utilizan técnicas de Inteligencia Artificial en general.*

Exclusión: *los trabajos que no contengan el término "predicción", "calidad del agua", "inteligencia computacional", "inteligencia artificial" o alguna técnica de estos campos especificada directamente (ej. Redes neuronales, Algoritmos genéticos, entre otras).*

El criterio de exclusión permite descartar algunas de las cadenas de búsqueda debido a que su utilización presenta problemas como el *High Recall (HR)*, en el cual se obtiene la información relevante pero además una cantidad enorme de información que no es de utilidad para el estudio; y por otro lado, *Low Precision (LP)*, al obtenerse muy pocos trabajos relevantes en comparación con el total de trabajos encontrados. La Tabla 2 muestra las cadenas de búsqueda utilizadas para seleccionar los trabajos relevantes dentro de la temática del presente estudio.

Cadena de Búsqueda	No. Trabajos	Motivo Exclusión	No. Trabajos Seleccionados	Precisión
water quality prediction	~1000	HR, LP		~0%
water quality forecast	~1000	HR, LP		~0%
water quality prediction AND computational intelligence	**20**		**10**	**50%**
water quality prediction AND artificial intelligence	**25**		**18**	**72%**
water quality parameters calibration	347	LP		~0%
water quality parameters optimization	170	LP		~0%
water quality parameters calibration AND computational intelligence	**17**		**6**	**35.3%**
water quality parameters calibration AND artificial intelligence	**13**		**5**	**38.5%**

Total			**39**	

Tabla 2. Cadenas de búsqueda utilizadas (En negrita las tenidas en cuenta en el proceso).

La cadena con la cual se obtuvieron los mejores resultados en cuanto a precisión fue *"water quality prediction AND artificial intelligence"* con un 72% en la precisión. Finalmente, en esta etapa fueron seleccionados 39 trabajos relevantes para el presente estudio.

D. **Clasificación de Estudios.** El proceso de clasificación de trabajos está basado en la agrupación planteada por Engelbrecht [7] en la cual se establecen varios paradigmas de IC ().

Figura 7. Agrupación de paradigmas de IC de acuerdo al esquema propuesto por Engelbrecht. Fuente propia.

Teniendo en cuenta la anterior clasificación, en este paso se resumen los aspectos más relevantes de los trabajos que fueron seleccionados en el paso C y se dividen en 2 subcategorías: predicción de calidad del agua mediante IC y calibración de técnicas de predicción de calidad del agua.

Como se mencionó en el Capítulo 1 del presente trabajo de maestría, la predicción de la calidad del agua ha cobrado una gran importancia en los últimos años y en muchos campos de estudio se reporta una significativa cantidad de trabajos que han abordado este problema a partir de diferentes enfoques. Uno de los campos de la ciencia más prometedores es la Inteligencia Artificial (IA), dentro de esta encontramos dos ramas; la Inteligencia Artificial Convencional (IAC) [7], la cual está basada en el análisis formal y estadístico del comportamiento humano ante diferentes problemas y por otra parte, la

Inteligencia Computacional (IC), la cual implica desarrollo o aprendizaje interactivo y dicho aprendizaje se realiza con base en datos empíricos.

Teniendo en cuenta lo anterior, la IC presenta características importantes para el proceso de predicción, las cuales no son tenidas en cuenta en la IAC. Dichas características están enfocadas en mejorar la inteligencia humana aprendiendo y descubriendo nuevos patrones, relaciones y estructuras complejas en ambientes dinámicos para resolver problemas prácticos [25]. De este modo, la mayor parte de los trabajos que se presentan a continuación, se centran en la predicción de la calidad del agua utilizando diferentes técnicas de IC. Por otra parte, también se consideran los trabajos cuyo objetivo es la calibración de parámetros para incrementar la precisión de los mecanismos de predicción, lo anterior enmarcado dentro la IC. A continuación se presentan los trabajos de investigación.

Predicción de la Calidad del Agua mediante IC

En la actualidad empiezan a cobrar importancia los métodos inspirados en los procesos biológicos de la naturaleza, incluido el ser humano. Estas nuevas técnicas surgen como una alternativa prometedora en muchos campos de aplicación gracias a su capacidad de adaptarse muy bien a los cambios en su entorno, imitando la forma en que lo hacen los diversos sistemas naturales. Estos mecanismos incluyen los paradigmas de la IA que muestran capacidad de aprender o adaptarse a nuevas situaciones, generalizar, abstraer, descubrir y asociar. Las técnicas individuales de estos paradigmas de IC se han aplicado con éxito para resolver problemas del mundo real, sin embargo la tendencia actual es el desarrollo de paradigmas híbridos, ya que ningún paradigma es superior a los demás en todas las situaciones. Al hacerlo, se aprovecha los puntos fuertes de los componentes del sistema de IC híbrido, y se eliminan las debilidades de los componentes individuales [23].

En el área de la IC se destacan las investigaciones realizadas utilizando las técnicas como las Redes Neuronales, Máquinas de Vectores de Soporte, Computación Evolutiva, entre otros. Dentro de las técnicas de Aprendizaje de Máquina se ha trabajado en la predicción de parámetros biológicos a partir de parámetros físico-químicos [34] y viceversa [35]. Se presentan modelos predictivos por cada parámetro, estos modelos están basados en árboles de regresión [36] y se comparan con técnicas como "vecino más cercano" y con la

regresión lineal, obteniéndose resultados competitivos con respecto a la precisión de los enfoques comparados. Además de lo anterior, la interpretación de los resultados se hace más intuitiva y comprensible. En el trabajo propuesto por Džeroski [35], se utiliza un árbol de regresión por cada variable, mientras que en el trabajo realizado por Blockeel [37], se analizan estos árboles y se detectan muchas similitudes entre ellos, lo cual conduce a realizar la pregunta de si es posible tener un solo árbol para predecir más de una variable a la vez sin tener pérdida en la precisión. En este estudio los parámetros biológicos se miden una vez en invierno y otra en verano, mientras que los parámetros físico-químicos se miden con mayor frecuencia. También se utiliza una base de datos proveniente del Instituto Meteorológico de Eslovenia con registros en un lapso de 6 años. Sin embargo solo se trabaja con un conjunto de 1060 muestras debido a que existe una gran pérdida datos en el resto de registros del repositorio y además de esto la eficiencia no es la deseable (elevados tiempos de procesamiento).

Continuando el recorrido por las técnicas de Aprendizaje de Máquina aplicadas a la predicción de la calidad del agua, se encuentran los métodos de ensamble [38], específicamente el enfoque de los Algoritmos Voraces de Selección de Ensamble expuesto en [39], mediante los cuales se agrupan varios conjuntos de datos (tomados a partir de la medición en un punto de monitoreo) y por cada uno de ellos se realiza una predicción del valor de una variable fisicoquímica (temperatura, pH, conductividad y salinidad), para esto se elige un conjunto de regresores que representen los valores óptimos en cada paso local del algoritmo con el fin de llegar a una solución general óptima. Tan [17] propone un Método de Mínimos Cuadrados apoyado de SVM para la predicción de parámetros de calidad del agua. En este trabajo se predice el total de fósforo en el agua a través de un método que presenta algunas ventajas como la fuerte capacidad de predecir el valor verdadero, la optimización global, una buena generalización y una alta velocidad de operación.

Por otra parte, el problema de predicción de parámetros de calidad del agua se ha abordado ampliamente dentro de las técnicas de Redes Neuronales Artificiales (RNA) [40], Computación Evolutiva (CE) [41] y Sistemas de Lógica Difusa (SLD) [42]. El campo de las RNA es uno de los que más se ha explorado con miras de abordar el problema mencionado anteriormente. Dentro de los trabajos realizados en este campo, se puede observar el propuesto por Romero y

Shan [8], en el cual se desarrolla una herramienta de software basada en redes neuronales para la predicción de la temperatura del agua de descarga del canal en una planta de energía de carbón. Las variables consideradas en este sistema incluyen los parámetros de funcionamiento de la planta y las condiciones climáticas locales, incluyendo información de mareas.

Por otro lado, Hatzikos [43], utiliza las redes neuronales con neuronas de activación como herramienta de modelado para la predicción de la calidad del agua del mar. El enfoque propuesto se refiere a predecir si el valor de cada variable aumentará o decaerá en el día siguiente. Los experimentos se centran en cuatro indicadores de calidad, a saber, la temperatura del agua, el pH, la cantidad de oxígeno disuelto y turbidez. Por su parte Palani [9], utiliza las RNA para predecir y pronosticar las características cuantitativas de los cuerpos de agua en una zona costera de Singapur, de tal forma que facilite la rápida evaluación y pronóstico de determinadas variables de calidad del agua como la salinidad, temperatura, oxígeno disuelto, y la clorofila-a.

Por otra parte, el objetivo del trabajo de Aguilera [10] es evaluar la Red Neuronal de Kohonen [40] como una posible herramienta en el proceso de toma de decisiones para predecir el estado trófico de las aguas costeras; esta herramienta se basa en una serie de mapas de activación compuestos por un conjunto de neuronas, las cuales se activan formando diferentes patrones que determinan la calidad del agua. He [44], propone una red neuronal con una conexión feed-forward y un entrenamiento con propagación hacia atrás (en aprendizaje supervisado), para predecir parámetros biológicos como el Indicador de Bacteria Fecal, Coliformes Totales, Coliformes Fecales y Enterococo, a partir de parámetros fisico-químicos como el pH, la temperatura, la conductividad, la turbidez y el oxígeno disuelto. El dominio de aplicación de este enfoque son las aguas para uso recreacional en las zonas costeras de San Diego, USA.

En [45] se estudia la precisión de los pronósticos sobre varios períodos acumulados a partir de un conjunto diverso de posibles modelos de predicción para la gestión de la calidad del agua. Los modelos se caracterizan por sus estructuras de memoria a corto plazo (memoria por retraso o memoria por retroalimentación) y largo plazo (lineal o no lineal). Los experimentos se llevan a cabo como una serie de ciclos de pronóstico, con un origen de un tamaño

constante ajustado. Los modelos son recalibrados en cada ciclo, y los pronósticos son generados por un horizonte de pronóstico de cinco períodos. Los resultados confirman que los modelos de red neuronal JENN y GMNN generalmente son más precisos que los competidores para un acumulado de múltiples períodos de predicción de la calidad del agua. Por ejemplo, los modelos JENN y GMNN reducen los errores de pronóstico acumulados en cinco período hasta en un 50%, en relación a los modelos ARIMA y de suavizado exponencial. Estos hallazgos son importantes en vista de las crecientes consecuencias sociales y económicas de la gestión de la calidad del agua de la cuenca.

En [46] se describe el entrenamiento, la validación y pruebas sobre análisis de incertidumbre en los modelos de regresión general de una red neuronal (GRNN) para la predicción de oxígeno disuelto (OD) en el río Danubio. Los principales objetivos de este trabajo fueron determinar las técnicas óptimas de normalización de datos y selección de entrada, la determinación de la importancia relativa de la incertidumbre en las diferentes variables de entrada, así como el análisis de la incertidumbre de los resultados de los modelos que utilizan la técnica de simulación de Monte Carlo (MCS). Min-max, mediana, z-score, sigmoides y tanh fueron validados como técnicas de normalización, mientras que el factor de inflación de la varianza, análisis de correlación y el algoritmo genético se probaron como las técnicas de selección de entrada. Como entradas, los modelos GRNN utilizan 19 variables de calidad del agua, medida en el agua del río cada mes en 17 sitios diferentes en un período de 9 años. Los mejores resultados se obtuvieron a partir de datos normalizados con min-max y la selección de entrada con base a la correlación entre OD y variables dependientes, que proporcionó el modelo GRNN más preciso, y en combinación el menor número de entradas: Temperatura, pH, HCO3-, SO42 -, NO3-N, Dureza, Na, Cl, conductividad y alcalinidad. Los resultados mostraron que el coeficiente de correlación entre los valores de OD medidos y predichos es 0,85. Las entradas con el mayor efecto en el modelo GRNN (dispuestas en orden descendente) fueron T, pH, HCO3-, SO42- y NO3-N. De todas las entradas, la variabilidad de la temperatura tuvo la mayor influencia en la variabilidad de contenido de OD en el cuerpo del río, con el OD disminuyendo a una tasa similar a la tasa teórica relacionada con la temperatura decreciente. El análisis de la incertidumbre de los resultados del modelo demostró que GRNN puede pronosticar efectivamente el contenido de OD, ya que la distribución de los

resultados del modelo es muy similar a la distribución correspondiente de los datos reales.

En [47] se describe el diseño y aplicación de un modelo de red neuronal feed-forward, conectada totalmente, con un perceptrón de tres capas para calcular el índice de calidad de agua (ICA) en el río Kinta (Malasia). Los esfuerzos de modelado demostraron que la arquitectura de red óptima era 23-34-1 y que las mejores predicciones del ICA se asociaron con un algoritmo de entrenamiento de propagación rápida (QP); una tasa de aprendizaje de 0,06; y un coeficiente de 1,75 QP. Las predicciones del ICA de este modelo fueron significativas, muy alta correlación positiva ($r = 0,977$, $p < 0,01$) con los valores medidos del ICA, lo que implica que las predicciones de los modelos explican alrededor de 95,4% de la variación en los valores medidos del ICA. El enfoque que se presenta en este trabajo ofrece una alternativa útil y poderosa para el cálculo y predicción del ICA, especialmente en el caso de los métodos de cálculo del ICA que implican largos cálculos y uso de diversas fórmulas con sub-índices para cada valor o rango de valores de variables de calidad del agua.

En RNA, finalmente se destacan los trabajos de Gutiérrez [48], García [49] y Saint-Gerons [50], los cuales se desarrollan sobre calidad del agua en ríos Colombianos, Mexicanos y Españoles respectivamente, en el primero se implementa una metodología de RNA como herramienta para la estimación de la calidad del agua en la cuenca alta y media del río Bogotá. En el segundo trabajo se desarrolla un Modelo de RNA para predecir el transporte de contaminantes en aguas subterráneas, para esto hace uso de la retro-propagación en una estructura de 3 capas con datos de entrenamiento, los cuales son evaluados a partir de la ecuación propuesta por Ogata y Banks [51]. Por último, en el tercer trabajo, se monitorea el Oxígeno Disuelto mediante el desarrollo de un sensor software que hace uso de RNA, este utiliza como entradas tres variables medibles: Temperatura, Turbidez y Potencial Redox.

Calibración de Técnicas de Predicción de Calidad del Agua

Los trabajos relacionados que se han encontrado hasta el momento se enfocan primordialmente en la optimización o calibración de modelos de monitoreo de parámetros de calidad del agua, dichos trabajos involucran a los algoritmos bio-

inspirados convirtiéndolos en un amplio espacio de investigación en este eje temático.

QUAL2Kw [52] es un framework o marco de trabajo para la simulación de la calidad del agua en arroyos y ríos; incluye un algoritmo genético para facilitar la calibración del modelo en su aplicación a cuerpos de agua particulares. El algoritmo genético se utiliza para encontrar la combinación entre parámetros de velocidad cinética y constantes, lo que se traduce en un mejor ajuste para un modelo de aplicación en comparación con los datos observados. Además de lo anterior, el usuario tiene la flexibilidad de seleccionar cualquier combinación de parámetros para realizar la optimización y especificar la función más adecuada para el ajuste. En el mismo sentido, Liu [53] propone un algoritmo genético para calibrar un modelo de contaminación difuso. Para este trabajo se implementó una herramienta de indicadores de fósforo (Phosphorus Indicators Tool, PIT), en la cual se utilizaron 78 parámetros para predecir la pérdida anual de fósforo. También se realizó un análisis de sensibilidad para investigar el impacto de los operadores del algoritmo genético sobre su eficacia en la búsqueda de un óptimo global. Una de las principales dificultades de este trabajo es la cantidad de parámetros utilizados en la predicción, a pesar de esto se obtienen tiempos de cálculo razonables.

Continuando con la aplicación de los algoritmos genéticos, el estudio de Huang [54] tiene como objetivo el acoplamiento de un Algoritmo Genético Híbrido (AGH) y una RNA para la calibración multi-objetivo de modelos de calidad de agua superficial. El AGH está propuesto como un algoritmo robusto de optimización por medio de la combinación con un método de búsqueda local. Este enfoque tiene la ventaja de realizar de una forma más eficiente, la evaluación de la función objetivo; sin embargo no siempre se garantiza hallar los valores óptimos de los parámetros.

Por otro lado, Chau [55] presenta la aplicación de un modelo de Optimización por Nubes de Partículas (PSO) por fracción de paso, para entrenar perceptrones con el fin de predecir la dinámica en tiempo real de proliferación de algas en Tolo Harbour, Hong Kong; los resultados indican que en comparación con el algoritmo de propagación hacia atrás de RNA, el algoritmo propuesto alcanza una mayor precisión (en 3 escenarios propuestos, los coeficientes de correlación son

superiores respecto a los valores reales) en un tiempo mucho más corto (el tiempo de entrenamiento fue inferior en aproximadamente un 50%).

En el trabajo de Baltar [56], se aplica un algoritmo Multi-Objetivo de Optimización por Nubes de Partículas (MOPSO, por su sigla en inglés) para encontrar soluciones que permitan minimizar las desviaciones de variables objetivo tales como temperatura, oxígeno disuelto, sólidos totales disueltos y pH. Este algoritmo está implementado como un complemento para Excel y permite encontrar soluciones para cualquier combinación de las cuatro variables mencionadas. Además de lo anterior, fue desarrollado un método gráfico interactivo que permite al usuario tomador de decisiones, identificar la mejor solución para ser aplicada en su modelo o bien soluciones similares que pueden ser de gran importancia en el proceso de toma de decisiones.

Por otra parte, Afshar [57] implementa un modelo de calibración de la calidad del agua determinando parámetros pertenecientes a hidrodinámica y a modelos de calidad del agua. Utiliza la técnica de nubes de partículas como una herramienta de optimización. Finalmente Zhangzan [58] propone un método que combina la técnica de la Percepción Visual (PV) con la Selección Negativa (SN) en la teoría de los Sistemas Inmunes Artificiales. La técnica de PV se utiliza para adquirir el Índice de Calidad del Agua (ICA) y la técnica de la SN es usada para evaluarlo.

2.2.2 Resultados y Análisis

Dentro de esta sección son presentados los resultados del proceso investigativo, los cuales fueron obtenidos a partir de los pasos A, B, C y D correspondientes al mapeo sistemático de la literatura que fueron presentados anteriormente. Cabe aclarar que el paso E de dicho mapeo se presenta en esta sección con el objetivo de resaltar en profundidad los resultados y el análisis de los mismos, lo que conlleva finalmente a establecer las brechas de investigación que guían el desarrollo del presente trabajo de maestría.

E. **Extracción de Datos y Síntesis.** Por cada uno de los paradigmas identificados en la fase anterior (D), se han distribuido los trabajos seleccionados en la tercera fase (C) (ver Tabla 3 y Figura 8).

Paradigma de IC	No. Trabajos	Porcentaje
Redes Neuronales Artificiales (RNA)	15	38.5%
Máquinas de Vectores de Soporte (MVS)	10	25.7%
Computación Evolutiva (CE)	5	12.8%
Inteligencia de Enjambres (IE)	7	17.9%
Sistemas Inmunes Artificiales (SIA)	2	5.1%

Tabla 3. Clasificación de trabajos de acuerdo al paradigma de IC

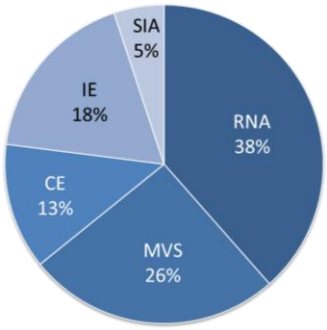


Figura 8. Clasificación de trabajos de acuerdo al paradigma de IC (gráfica tipo pastel). Fuente propia.

Como se puede apreciar en la tabla y gráfica anterior, el paradigma que agrupa mayor número de trabajos es el de RNA. En un menor porcentaje, pero muy significativo, se encuentran los trabajos concernientes a la utilización de las MVS, en los cuales se realiza una comparación con las RNA obteniendo resultados más precisos en cuanto a la predicción. Finalmente una menor cantidad de trabajos se encuentra agrupada en los paradigmas de CE, IE y SIA.

Brechas de Investigación

Finalmente para obtener información detallada y concisa de cada trabajo que ha sido seleccionado y clasificado previamente, se elaboró una planilla de extracción de datos por cada estudio, en esta planilla fue consignada la información principal como el título, el resumen, las palabras clave y lo más importante, los aportes y ventajas así como las falencias que derivan al final en las brechas relacionadas con el problema de investigación.

En esta sección se abordan los principales vacíos y falencias de los trabajos que se presentaron en la sección anterior, de tal forma que se pueda establecer un punto de referencia de este trabajo investigativo para la generación de nuevo conocimiento dentro del campo de la predicción de la calidad del agua.

Los trabajos se han agrupado dentro de un paradigma de IC en particular tal como se muestra en la Tabla 4 y se han establecido las falencias que caracterizan a cada grupo de trabajos.

Núcleo Temático	Enfoques		Brecha
Predicción de Calidad del Agua	RNA	[8], [9], [10], [43], [44], [48], [49], [50]	La precisión de las predicciones depende del conjunto de datos de entrenamiento, con una menor cantidad de datos, la precisión tiende a disminuir. Por otro lado, la tasa de error tiende a aumentar si las predicciones se realizan sobre un intervalo de tiempo más amplio.
	MVS	[17], [34], [35], [37], [39]	
Calibración de técnicas de predicción de calidad del agua	CE	[52], [53], [54]	La capacidad predictiva está ligada a las interrelaciones que presenten algunas variables de calidad del agua en un uso determinado.
	IE	[18], [55], [56], [57]	Estos trabajos se utilizan para incrementar la precisión de las predicciones, sin embargo no se aprovecha totalmente su potencial teniendo en cuenta la capacidad adaptativa de las técnicas propuestas, permitiendo beneficios como la posibilidad de aplicar el mecanismo de predicción a diferentes usos del agua.
	SIA	[58]	

Tabla 4. Brechas Existentes

Teniendo en cuenta lo anterior, el presente trabajo de maestría plantea atacar el problema de la poca adaptabilidad de modelos predictivos de la calidad del agua en diferentes usos del recurso hídrico, es decir, si un modelo de predicción obtiene buenos resultados de precisión al ser aplicado a determinado uso del agua, estos mismos resultados deberían ser buenos si se aplica dicho modelo en otro uso. En la gran mayoría de los trabajos revisados no se tiene en cuenta esta característica, lo cual no brinda el carácter adaptativo al modelo, es por esto que el presente trabajo de maestría intenta dar solución a este importante aspecto haciendo uso de la IC y la calibración automática de parámetros de calidad del agua.

2.3. Resumen

En este capítulo fueron presentados los conceptos teóricos correspondientes a tres conceptos principales como los Sistemas Adaptativos Complejos, la Predicción Científica y la Inteligencia Computacional, este último concepto engloba diferentes técnicas de predicción como Redes Neuronales Artificiales y Máquinas de Vector de Soporte; y otras técnicas de optimización como la Computación Evolutiva y la Inteligencia de Enjambres.

Posteriormente fueron expuestos los trabajos relacionados respecto al problema de investigación definido en el Capítulo 1, haciendo uso de una metodología para la construcción del mapeo y revisión sistemática. Los enfoques de estos trabajos fueron orientados en dos nichos de investigación, el primero referente a la predicción de la calidad del agua haciendo uso de técnicas de inteligencia computacional; y el segundo enmarcado dentro de la calibración de técnicas de predicción de la calidad del agua a partir de algoritmos de optimización.

Capítulo 3

Mecanismo de Predicción Adaptativo

Tomando como base la revisión documental realizada en la sección anterior, este estudio plantea abordar el problema referente a la escasa capacidad adaptativa de los modelos de predicción de la calidad del agua, es decir, si un modelo de predicción obtiene buenos resultados en cuanto a precisión al ser aplicado a determinado uso del agua, por ejemplo el tratamiento para consumo humano, estos mismos resultados deberían ser buenos si dicho modelo es aplicado en otro uso como el agua que es utilizada en una estación piscícola o un centro recreacional, entre otros. Teniendo en cuenta que las variables de calidad del agua en estos usos no son las mismas que maneja el tratamiento de agua para consumo humano, el mecanismo debe contar con la capacidad de generalización para adaptarse a diferentes usos, entregando predicciones confiables en cada uno de estos.

El desarrollo de este estudio depende ampliamente del conjunto de datos de calidad del agua, es por esto que gestionar adecuadamente dichos datos se convierte en un aspecto de gran importancia. Dado lo anterior durante el proceso de construcción del mecanismo de predicción adaptativo fue utilizada la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [59], la cual es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Minería de Datos.

3.1. Descripción General del Mecanismo

Dentro del análisis documental, los trabajos revisados orientan la predicción sobre un uso del agua específico, lo cual deja de lado el carácter adaptativo del modelo, es por esto que la presente propuesta intenta dar solución a este importante aspecto haciendo uso de la IC y la calibración automática de parámetros de calidad del agua.

Con el objetivo de guiar la construcción y modelado del mecanismo, se optó por la aplicación de una metodología que estuviera orientada específicamente hacia el proceso de predicción, el cual es el principal enfoque del mecanismo. Esta metodología es denominada SVM-UP (Support Vector Machine – Unified Process) [60] y está desarrollada para trabajar con modelos regresivos para predicción de

series temporales, los autores han instanciado esta metodología para emplear la Regresión por Vectores de Soporte como método de predicción. En la Figura 9 es presentada la estructura planteada en SVM-UP.

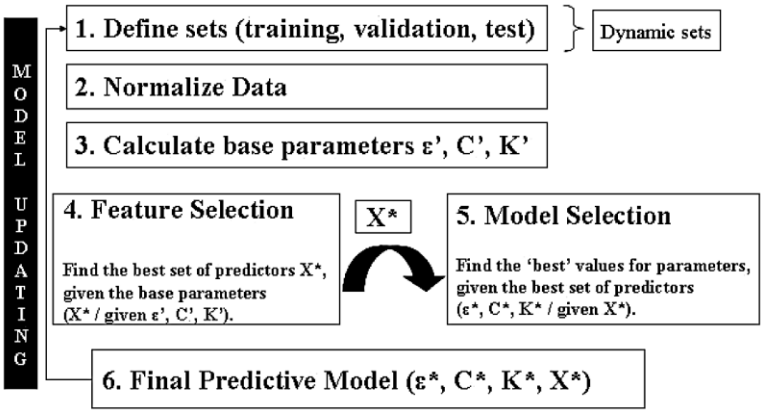


Figura 9. SVM-UP - Metodología para el desarrollo de modelos predictivos. Tomado de [60].

Como se observa en la figura anterior, inicialmente se deben definir y dividir los datos en tres subconjuntos (entrenamiento, validación y prueba). Los datos de entrenamiento son utilizados para construir el modelo, los datos de validación para la selección del modelo y la selección de características, y los datos de prueba son un subconjunto totalmente independiente, útil para proporcionar una estimación del nivel de error del modelo.

En esta metodología además se calculan los valores iniciales de los parámetros base (ε', C' y K), los cuales serán descritos más adelante en este capítulo. Posteriormente se realiza la selección de características para obtener el mejor conjunto de variables a predecir (X^*). Y por último se vuelve a actualizar el modelo con el objetivo de mejorar su rendimiento en el proceso de predicción. Por lo tanto, al final, el modelo predictivo está determinado por los parámetros ε', C', K y X^*.

El mecanismo de predicción adaptativo que se propone en esta tesis de maestría, está conformado por los componentes que se observan en la Figura 10 y fue desarrollado teniendo en cuenta cada una de las etapas de la metodología planteada anteriormente.

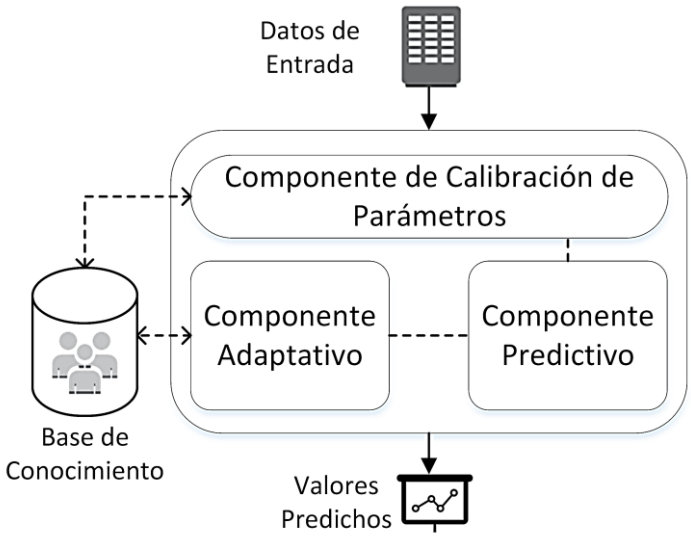


Figura 10. Diagrama general de los componentes del mecanismo de predicción adaptativo. Fuente propia.

3.2. Componente de Calibración de Parámetros

Teniendo en cuenta la complejidad de ciertas aplicaciones de minería de datos, como por ejemplo la regresión, se hace necesario seleccionar las características más importantes para la construcción del modelo respectivo. El componente de calibración de parámetros tiene como objetivo obtener un subconjunto de características (en este caso denominadas variables de calidad del agua) a partir del conjunto de datos original, el cual puede pertenecer a un uso del agua diferente. Dado lo anterior, la selección de características ofrece beneficios como los que se mencionan a continuación.

- Mejoras en la precisión del modelo.
- Reducción de los tiempos de cálculo en la construcción del modelo.
- Facilidad de visualización de datos y comprensión del modelo.
- Reducción de riesgo de sobreajuste[8].

[8] También conocido como "overfitting" en el idioma inglés. Es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. El algoritmo de aprendizaje debe alcanzar un estado en el que será capaz de predecir el resultado en otros casos a partir de lo aprendido con los datos de entrenamiento, generalizando para poder resolver situaciones distintas a las acaecidas durante el entrenamiento.

Para realizar este refinamiento, el componente se apoya en dos estrategias de selección; la primera de ellas está orientada hacia métodos que ya han sido definidos por otros autores para realizar este proceso, Guyon y Elisseeff [61] los clasifican en tres categorías: métodos de filtro, métodos de envoltura o "*wrapper"* y métodos embebidos; y la segunda es utilizar una base de conocimiento que está conformada por información asociada a los diferentes usos del agua. Respecto a la primera estrategia, en la presente tesis de maestría se optó por utilizar el método wrapper de selección hacia adelante (forward selection) dado que es uno de los más utilizados y de mejor rendimiento de acuerdo al sourvey realizado en [62].

En la Figura 11 es presentado el esquema general de la selección de características hacia adelante; la idea básica de este proceso es obtener un subconjunto de variables a partir del conjunto inicial de datos $(x_1, \ldots, x_n)$, también conocido como predictores. Este subconjunto debe contener aquellas variables que mejor representen al conjunto inicial y por consiguiente permitan obtener un modelo de regresión más preciso al evitar incluir información redundante o de poca relevancia para el modelo (también conocida como *"ruido"*). En el mismo sentido, debe definirse un número máximo de predictores a seleccionar (m), donde $m \leq n$. Por cada iteración se combinan diferentes tuplas de predictores, esto es, en una primera iteración se evalúa el modelo con una sola variable, en la segunda iteración se evalúan todas las posibles combinaciones entre 2 variables y de esta forma hasta la iteración m para finalmente encontrar la mejor o las mejores tuplas de variables que hagan más preciso el modelo de predicción.

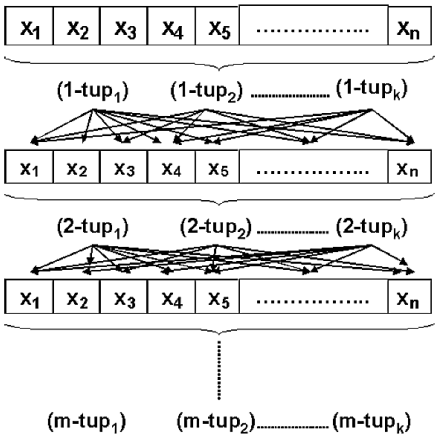


Figura 11. Selección de características de tipo wrapper - selección hacia adelante (forward selection). Tomado de [60].

Por otra parte la estrategia de selección de características de este trabajo de maestría, dispone de una base de conocimiento, la cual está conformada por información asociada a los diferentes usos del agua; lo anterior con el propósito de establecer cuales parámetros o variables de calidad del agua aportan información relevante orientada hacia el uso del agua seleccionado.

Un ejemplo de la aplicación de este proceso es presentado en la Figura 12, en la cual inicialmente por medio del método *wrapper – forward selection* se eligen 3 variables de calidad del agua, *{temperatura, turbidez, amoniaco NH3}* (considerando que este número de variables corresponde a la tupla que presentó un mejor rendimiento sobre una primera etapa de predicción) y a partir de este subconjunto se realiza una nueva selección de variables, esta vez teniendo en cuenta una base de conocimiento de expertos, la cual está estructurada de tal forma que por cada uso del agua sean caracterizadas un conjunto de variables representativas para ese uso. En este ejemplo se asume que el uso del agua establecido corresponde a la piscicultura y que el conjunto de variables relevantes para este uso es *{temperatura, turbidez}*. Como resultado de esta nueva selección de variables se obtiene el subconjunto *{temperatura, turbidez},* dado que es la intersección de los dos subconjuntos de variables anteriormente definidos.

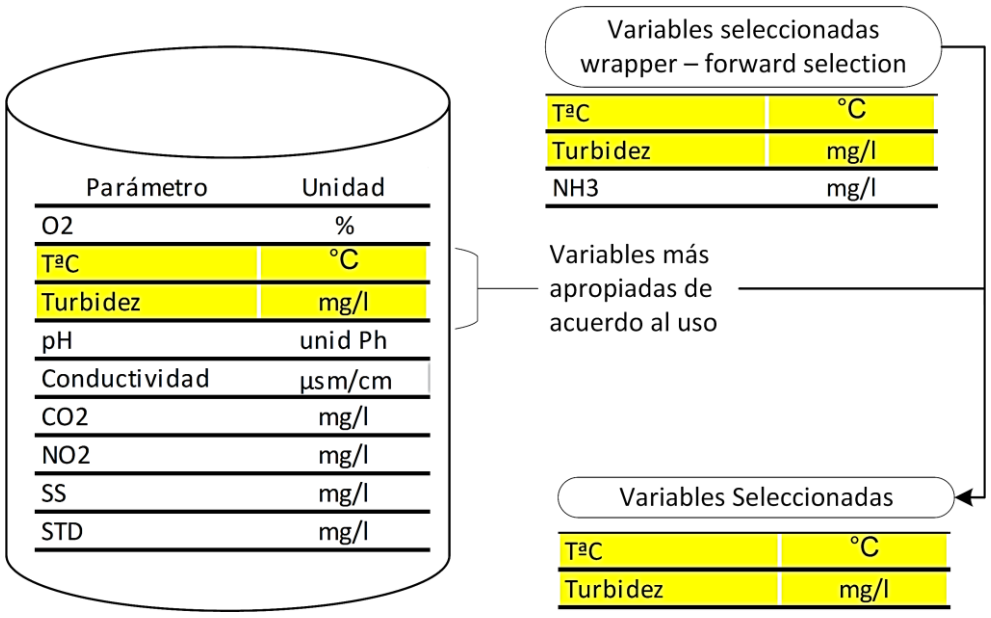


Figura 12. Ejemplo de selección de variables mediante refinamiento con la base de conocimiento del uso del agua. Fuente propia.

3.3. Componente Predictivo

El componente predictivo, como su nombre lo indica, es el encargado de realizar las operaciones de predicción utilizando las Máquinas de Vector de Soporte para Regresión o SVR (esta técnica fue seleccionada durante la primera etapa de esta investigación por medio de pruebas de efectividad las cuales son presentadas en el Capítulo 4). Este componente utiliza el subconjunto de variables de calidad del agua que retorna la calibración de parámetros y su objetivo se centra en obtener un conjunto de posibles valores futuros (valores predichos) para cada una de las variables seleccionadas. La Figura 13 muestra el esquema general del componente predictivo.

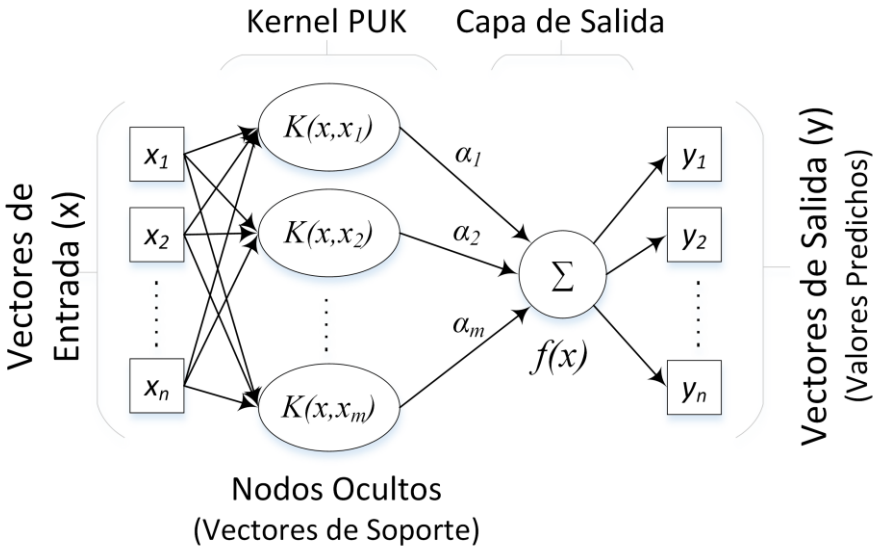


Figura 13. Componente de predicción – Regresión por Vectores de Soporte con el kernel PUK. Fuente propia.

Este componente está conformado por la arquitectura base de una SVR, la cual ha sido modificada para que sus entradas correspondan a uno o más vectores que contienen los datos de cada una de las variables de calidad del agua que pertenecen a un determinado uso y que han sido previamente seleccionadas por el componente de calibración de parámetros. La función de regresión $\mathrm{f(x)}$ está dada por la Ecuación (**3.1**.

$$f(x) = \langle \alpha, x \rangle + b \qquad (\alpha, x \in R^d). \tag{3.1}$$

3.3.1 Pearson VII Universal Kernel (PUK)

Los algoritmos basados en kernel operan mediante la transformación aplicada a los datos de entrada sobre un espacio n-dimensional, es decir, un espacio compuesto por funciones de distancia basadas en vectores de valores reales, que representan entidades físicas; lo anterior se conoce como el "*espacio de características*". En dicho espacio es posible resolver el problema en cuestión de una forma lineal.

Teniendo en cuenta el párrafo anterior, los vectores de entrada son denotados por $x = \{x_1, x_2, \ldots, x_n\}$ y están conectados directamente a una o varias funciones kernel PUK, encargadas de mapear el espacio de entradas X a un nuevo espacio de características de una mayor dimensión. La idea básica de la SVR consiste en realizar un mapeo de los datos de entrenamiento $x \in X$, a un espacio de mayor dimensión $F = \{\alpha(x) | x \in X\}$ a través de un mapeo no lineal donde sea posible realizar una regresión lineal, formalmente denotado como se observa en la Ecuación (3.2.

$$x = \{x_1, x_2, \ldots, x_n\} \rightarrow \alpha(x) = \{\alpha(x)_1, \alpha(x)_2, \ldots, \alpha(x)_n\}. \quad (3.2)$$

La función del kernel PUK utilizada en el módulo de predicción está dada por la Ecuación (3.3.

$$K(x, x_i) = \frac{H}{\left[1 + \left(\frac{2(x - x_i)\sqrt{2^{(1/\omega)} - 1}}{\sigma}\right)^2\right]^{\omega}} \quad (3.3)$$

Donde H es la altura del pico desde su centro x_0, y x la variable independiente. Los parámetros σ y ω controlan el ancho de la curva (también llamado ancho de Pearson) y el factor de asimetría del pico. La razón principal para usar la función de Pearson VII para ajuste de curvas es su flexibilidad para cambiar, variando el parámetro x, a partir de una forma gaussiana (donde x se aproxima a infinito) a una forma de Lorentziana ($x = 1$) [63]. Lo anterior se puede observar con mayor detalle en la Figura 14.

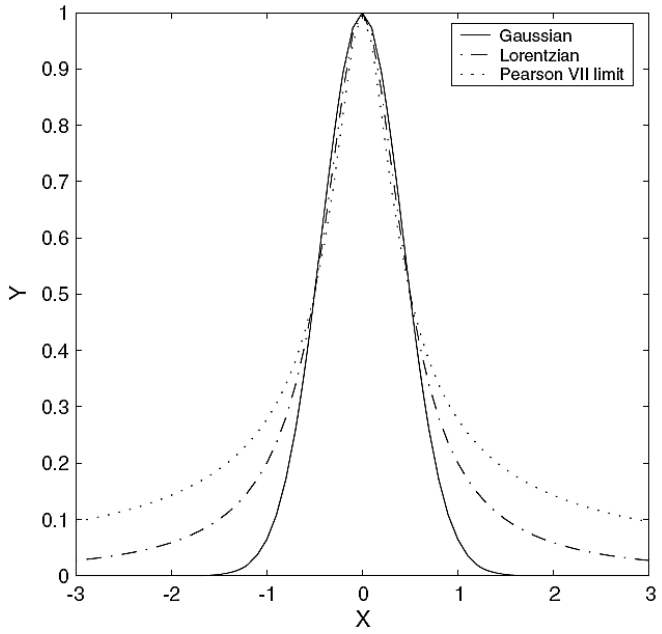


Figura 14. Curva de la función de kernel PUK.

La configuración seleccionada para los parámetros de la función de kernel PUK está dada por: $\sigma=0.8$, $\omega=1$, $H=1$, teniendo en cuenta la evaluación experimental y los resultados que se presentan más adelante en el capítulo 4.

Una de las ventajas de la función de Pearson, es que posibilita la adaptación de valores a picos con una variedad de anchos y formas de línea diferentes. Esta importante propiedad permite que PUK se adopte como una función alternativa de núcleo genérico. En conclusión, teniendo en cuenta su flexibilidad para variar entre una función Gaussiana y una Lorentziana, PUK se convierte en un núcleo universal que puede sustituir un amplio conjunto de funciones de kernel comúnmente usadas como la función lineal, la función polinómica y la función de base radial, entre otras; lo anterior seleccionando el ajuste apropiado para cada parámetro.

La salida del componente predictivo es un vector por cada variable de calidad del agua, el cual contiene los valores predichos mediante la aplicación de la SVR y que posteriormente son utilizados por el componente adaptativo.

Con el fin de ilustrar de una forma más detallada al lector el funcionamiento de este componente, a continuación se presenta un ejemplo que continua con el desarrollo del ejemplo presentado en el numeral 3.2. De esta forma el componente predictivo recibe como entradas los datos correspondientes a las variables seleccionadas

(temperatura y turbidez) en un vector diferente por cada variable. Cada uno de los datos contenidos en un vector, se transfieren a una función de kernel PUK, la cual retorna el valor de x que es pasado como parámetro a la función de regresión (ver Ecuación (**3.1**) junto con el valor α que varía en el intervalo $[0, 1]$ y el valor b que representa el desplazamiento u "offset" de la función (tanto el parámetro α como b son estimados internamente por la SVR mediante la minimización de una función de costo), para finalmente obtener los valores futuros predichos. La muestra de forma gráfica el proceso descrito en este párrafo.

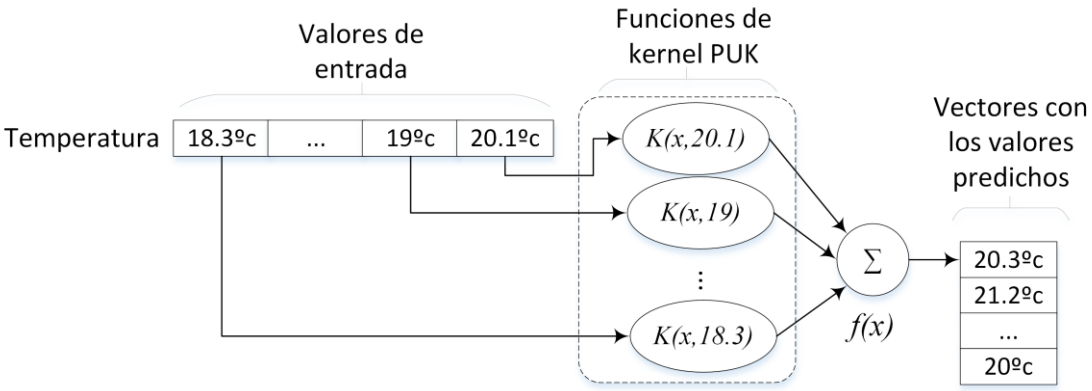


Figura 15. Ejemplo del proceso de predicción para la variable temperatura (es realizado de forma análoga para la turbidez o cualquier otra variable).

3.4. Componente Adaptativo

El componente predictivo que fue descrito en el numeral anterior, mantiene una estrecha relación con el componente adaptativo detallado en la Figura 16, el cual a su vez implementa la técnica de Optimización por Nubes de Partículas (PSO) como principal enfoque para ajustar los valores de predicción a los valores reales del uso del agua seleccionado, es decir, el objetivo de la optimización para el presente estudio está centrado en minimizar la diferencia entre las curvas de valores predichos y valores reales.

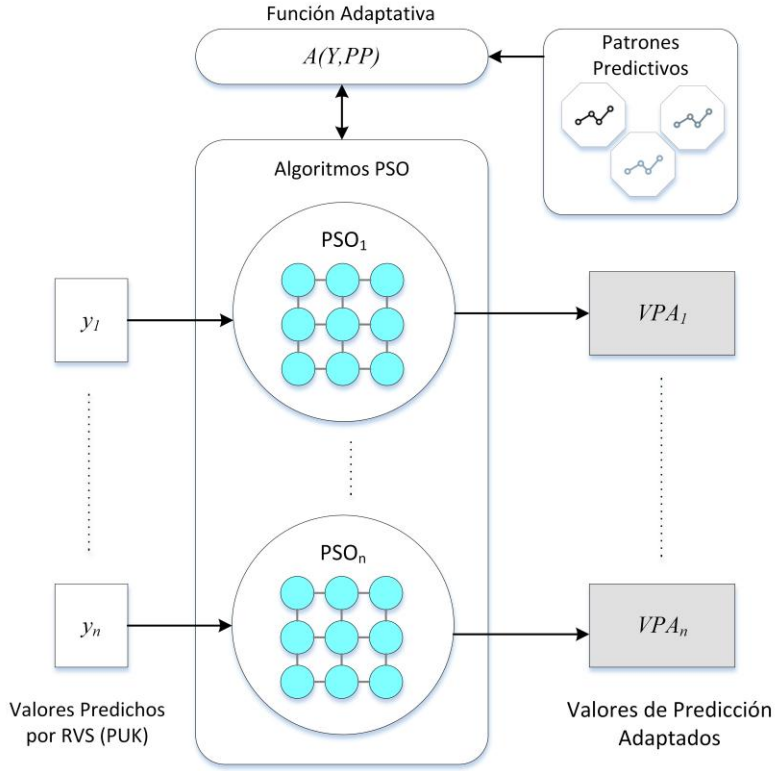


Figura 16. Componente adaptativo – algoritmo PSO para obtener los valores de predicción adaptados

3.4.1 Algoritmos PSO

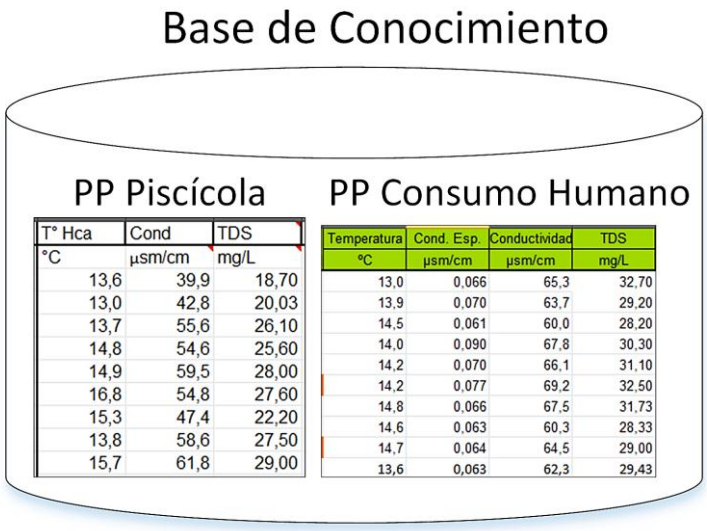
Como se mencionó en el Capítulo 2, PSO es una técnica de IC basada en población la cual busca emular el comportamiento social de ciertos grupos de individuos como por ejemplo las bandadas de aves en vuelo o algunas familias de peces; dado lo anterior, el principal enfoque de esta técnica es proporcionar un algoritmo simple y eficaz para optimización. Los individuos son conocidos como partículas moviéndose a través de un espacio de búsqueda de múltiples dimensiones; y cada una de estas partículas posee una velocidad y una posición; por tanto los cambios en una partícula dentro del grupo se ven influenciados por la experiencia o el conocimiento de sus vecinos.

Las entradas de este componente son los vectores que contienen los valores predichos retornados por el componente predictivo, por cada uno de estos vectores es creado un conjunto de partículas, las cuales mediante su interacción y cooperación se encargan de encontrar nuevos valores que ajustan cada valor predicho de una variable de calidad del agua específica. Cada conjunto de partículas se apoya de una función adaptativa (función objetivo) que es la encargada de

obtener el valor de predicción adaptado de cada uno de los valores predichos. Dado que esta función presenta una relación directa con un Patrón Predictivo (PP), este se describe primero y posteriormente la función adaptativa.

3.4.2 Patrón Predictivo (PP)

Un patrón predictivo corresponde a un conjunto de valores que han sido establecidos por expertos en el monitoreo de la calidad del agua sobre diferentes usos para poder establecer un punto de referencia en la adaptación de las predicciones (ver Figura 17). Estos PP se encuentran dentro de una base de conocimiento la cual contiene los respectivos valores de acuerdo al uso del agua requerido.



T° Hca	Cond	TDS
°C	µsm/cm	mg/L
13,6	39,9	18,70
13,0	42,8	20,03
13,7	55,6	26,10
14,8	54,6	25,60
14,9	59,5	28,00
16,8	54,8	27,60
15,3	47,4	22,20
13,8	58,6	27,50
15,7	61,8	29,00

Temperatura	Cond. Esp.	Conductividad	TDS
°C	µsm/cm	µsm/cm	mg/L
13,0	0,066	65,3	32,70
13,9	0,070	63,7	29,20
14,5	0,061	60,0	28,20
14,0	0,090	67,8	30,30
14,2	0,070	66,1	31,10
14,2	0,077	69,2	32,50
14,8	0,066	67,5	31,73
14,6	0,063	60,3	28,33
14,7	0,064	64,5	29,00
13,6	0,063	62,3	29,43

Figura 17. Almacenamiento de los patrones predictivos (ejemplo) por cada uso del agua. Fuente propia.

3.4.3 Función Adaptativa

La función adaptativa recibe como entradas un vector de valores predichos Y, y un Patrón Predictivo (PP), el cual corresponde a un uso del agua determinado. Esta función está basada en el concepto de Intervalos de Predicción [64], los cuales brindan un intervalo dentro del cual se espera ubicar con una probabilidad especificada, el valor predicho y_i (ver Ecuación (**3.4**, donde ip es el intervalo de predicción).

$$ip(Y) = \hat{y}_t \pm m\hat{\sigma} \tag{3.4}$$

Por otro lado, $\hat{\sigma}$ es una estimación de la desviación estándar de la distribución de los datos pronosticados. Vale la pena mencionar que en los procesos de predicción, es común para calcular intervalos entre 80% y 95%, aunque cualquier porcentaje puede ser utilizado. El valor m es un multiplicador que determina el porcentaje del intervalo de predicción, estos valores son presentados en la Tabla 5.

Porcentaje	Multiplicador
50	0.67
55	0.76
60	0.84
65	0.93
70	1.04
75	1.15
80	1.28
85	1.44
90	1.64
95	1.96
96	2.05
97	2.17
98	2.33
99	2.58

Tabla 5. Multiplicadores usados en los intervalos de predicción. Tomado de [64].

Por ejemplo, suponiendo que los errores de pronóstico no están correlacionados y distribuidos normalmente, con un intervalo de predicción del 95% en una serie de tiempo, se tendría un intervalo dado por: $\hat{y}_t \pm 1.96\hat{\sigma}$.

De esta manera la función adaptativa está dada por la Ecuación (**3.5**, teniendo en cuenta que se busca establecer una correspondencia entre la curva de valores reales (los cuales están representados por los patrones predictivos) y los valores retornados por el componente predictivo. Por tanto, las partículas del algoritmo PSO, están encargadas de buscar un valor dentro del intervalo de predicción que se ajuste a estas dos curvas sin que modifique de una forma drástica el valor predicho.

$$A(Y, PP) = \frac{pso(ip(Y)) + PP}{2}. \qquad (3.5)$$

De esta forma, cada valor obtenido mediante el algoritmo PSO, es promediado con los respectivos valores de los patrones predictivos para finalmente obtener los nuevos vectores que contienen los Valores de Predicción Adaptados (APV), con los

cuales se busca brindar una mayor correspondencia con los valores reales de un uso del agua específico (lo anterior es presentado en el siguiente capítulo de experimentación y evaluación).

3.5. Resumen

Este capítulo presentó de forma detallada cada uno de los componentes del mecanismo de predicción adaptativo de la calidad del agua, el cual es el eje central del presente trabajo de maestría. Este mecanismo está conformado por el componente de calibración de parámetros (selección de variables de calidad del agua relevantes para la predicción), el componente predictivo (obtención de posibles valores futuros de las variables de calidad del agua) y el componente adaptativo (ajuste de los valores predichos de acuerdo al uso del agua).

Con el objetivo de brindar un enfoque formal y estructurado a la construcción de dicho mecanismo, fue aplicada la metodología SVM-PU, que está orientada al desarrollo de modelos predictivos sobre series temporales, específicamente instanciada para predicción mediante la Regresión por Vectores de Soporte.

Capítulo 4

Experimentación y Evaluación

En el presente proyecto fueron realizadas diferentes pruebas para determinar la precisión del mecanismo de predicción adaptativo. El proceso de evaluación fue dividido en dos fases, la primera de ellas consistió en seleccionar el algoritmo que permitiera obtener los resultados más precisos en cuanto a la predicción sobre series temporales con datos de calidad del agua. En la segunda fase fue definido y ajustado un algoritmo que permitiera adaptar los valores predichos sobre diferentes conjuntos de datos pertenecientes a varios usos del agua.

A continuación se detalla el proceso de desarrollo del prototipo que implementa el mecanismo de predicción propuesto; posteriormente se presenta una descripción de los datos, el área de estudio y las pruebas realizadas en cada una de las dos fases mencionadas.

4.1. Desarrollo del Prototipo

Para la construcción del prototipo que implementa el mecanismo de predicción adaptativo, en primera instancia se utilizó el software WEKA (Waikato Environment for Knowledge Analysis, en español “Entorno para Análisis del Conocimiento de la Universidad de Waikato, Nueva Zelanda), en su versión 3.7.12, el cual es una herramienta de libre distribución desarrollada en java, que permite la implementación de algoritmos para el análisis de datos, aprendizaje automático y tareas de minería de datos así como la visualización y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades [65].

Los algoritmos utilizados para la construcción del mecanismo se detallan en la Tabla 6.

Componente del Mecanismo	Algoritmo	Implementación WEKA
Calibración de Parámetros	Wrapper Forward Selection	weka.attributeSelection.WrapperSubsetEval
Predictivo	Máquinas de Vector de Soporte para Regresión	weka.classifiers.functions.SMOreg

Adaptativo	Optimización por Nubes de Partículas (PSO)	weka.attributeSelection.PSOSearch

Tabla 6. Algoritmos implementados en WEKA utilizados para la construcción del mecanismo de predicción. Fuente propia.

Los algoritmos presentados en la tabla anterior se enmarcan dentro de los respectivos componentes de calibración de parámetros y predictivo que fueron descritos en el capítulo anterior. Para implementar el componente adaptativo fue necesario complementar dichos algoritmos con el desarrollo de un programa java en el entorno integrado de desarrollo de Eclipse, en el cual se construye la interrelación existente entre cada componente; de esta forma con la utilización del algoritmo PSO se obtiene el resultado final, es decir, los valores de predicción adaptados.

Además de lo anterior, con el objetivo de realizar una evaluación más exhaustiva y poder comparar el desempeño de diferentes algoritmos evolutivos, se utilizó el framework Evolving Objects 1.3.1, conocido también como EO, el cual es una biblioteca orientada a la computación evolutiva basada en ANSI-C++, que permite escribir algoritmos de optimización estocásticos en un tiempo de desarrollo relativamente rápido. Los algoritmos evolutivos forman parte de una familia de algoritmos inspirados en la teoría de la evolución y también se encuentran entre los métodos de Inteligencia Computacional. Como se mencionó en el capítulo 2, estos permiten evolucionar un conjunto de soluciones a un problema determinado, con el fin de producir los mejores resultados. Estos algoritmos son estocásticos, dado que iterativamente utilizan procesos aleatorios para su ejecución. La gran mayoría de estos métodos se usan para resolver problemas de optimización, y pueden ser también llamados "meta-heurísticas".

4.1.1. Interacción entre Componentes

Los diagramas de secuencia son utilizados para modelar la interacción entre objetos en una aplicación a través del tiempo. Un diagrama de secuencia contiene detalles de implementación del escenario, incluyendo los objetos y clases que se usan para implementar el escenario y mensajes intercambiados entre los objetos [66]. Estos diagramas se muestran en la Figura 18.

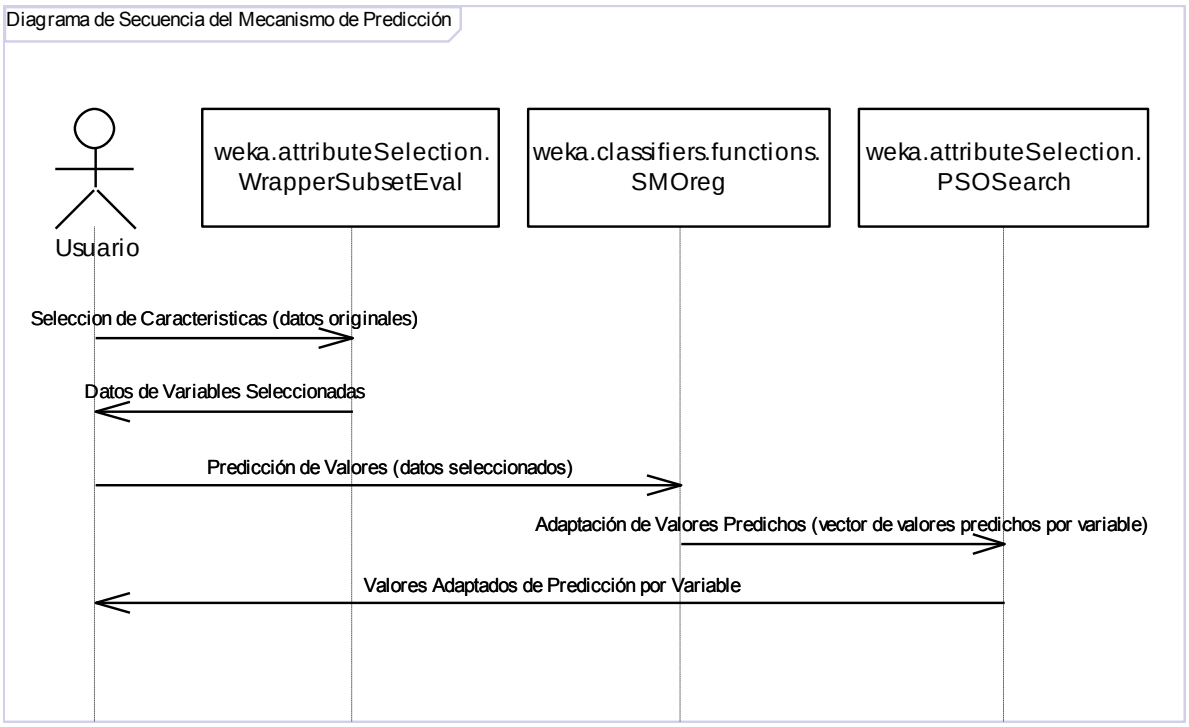


Figura 18. Diagrama de secuencia para la interacción entre los componentes del mecanismo de predicción. Fuente propia.

Como se presenta en la figura anterior, el usuario del sistema debe contar con cierto grado de experiencia en el manejo y tratamiento de datos, dado que la versión del mecanismo de predicción propuesta no cuenta con una interfaz gráfica de usuario que permita una gran facilidad en su uso. Considerando lo anterior, el diagrama de secuencia mostrado, expone la interacción en el tiempo de los principales componentes del mecanismo de predicción propuesto (en términos de implementación), que fue descrito en el capítulo 3.

4.1.2. Base de Conocimiento

En la Figura 19 se muestra la estructura de la base de conocimiento utilizada en el mecanismo de predicción; la cual involucra los elementos descritos en la Tabla 7. Cabe resaltar que esta información fue establecida y refinada en un trabajo conjunto con personal experto del Grupo de Estudios Ambientales de la Universidad del Cauca (GEA) y el Instituto de Investigación y Desarrollo en Abastecimiento de Agua, Saneamiento Ambiental y Conservación del recurso Hídrico (CINARA).

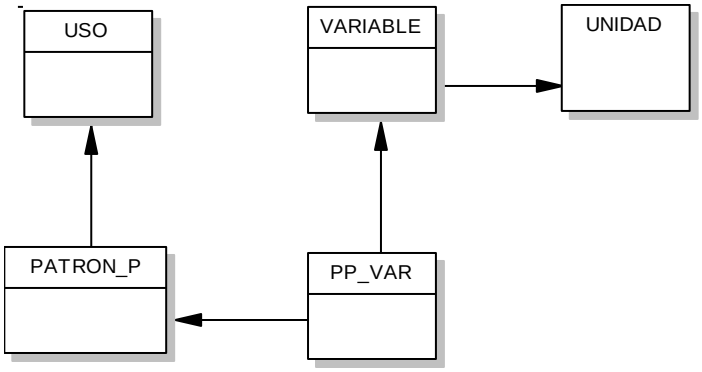


Figura 19. Esquema de la base de conocimiento utilizada en el mecanismo de predicción adaptativo. Fuente propia.

Elemento	Descripción
USO	Almacena los usos del agua disponibles
VARIABLE	Almacena las variables de calidad del agua en general
UNIDAD	Almacena las unidades correspondientes a cada variable de calidad del agua
PATRON_P	Almacena una serie de valores que representan los valores reales de calidad del agua para determinada variable
PP_VAR	Relaciona un patrón predictivo con una determinada variable de calidad del agua

Tabla 7. Descripción de los elementos de la base de conocimiento.

4.2. Datos y Área de Estudio

El acceso a datos fiables de calidad del agua no es una tarea simple, muchos portales de diferentes organizaciones que trabajan con la gestión del recurso hídrico, presentan restricciones y solo permiten consultas de los datos a personal autorizado que pertenezca a la respectiva organización. Sin embargo existen organismos que brindan la posibilidad de obtener datos que cuentan con dos características importantes: calidad (los registros han sido revisados por expertos) y cantidad (conjunto de datos superior a 10.000 registros).

4.2.1. Conjunto de datos del USGS

Uno de estos organismos es el Servicio Geológico de los Estados Unidos [67] o USGS por sus siglas en inglés (United States Geological Survey), es una agencia científica del gobierno federal de los Estados Unidos, la cual se divide en 4 disciplinas científicas mayores: biología, geografía, geología e hidrología; su grupo de científicos estudian el terreno, los recursos naturales y los peligros naturales

que los amenazan. El USGS es una organización investigadora sin responsabilidades reguladoras de modo que el suministro oportuno de datos fiables de calidad del agua para el público es una de sus misiones clave, es por esto que los datos están disponibles al público desde miles de sitios a través de la nación (Estados Unidos) en los 50 estados y territorios.

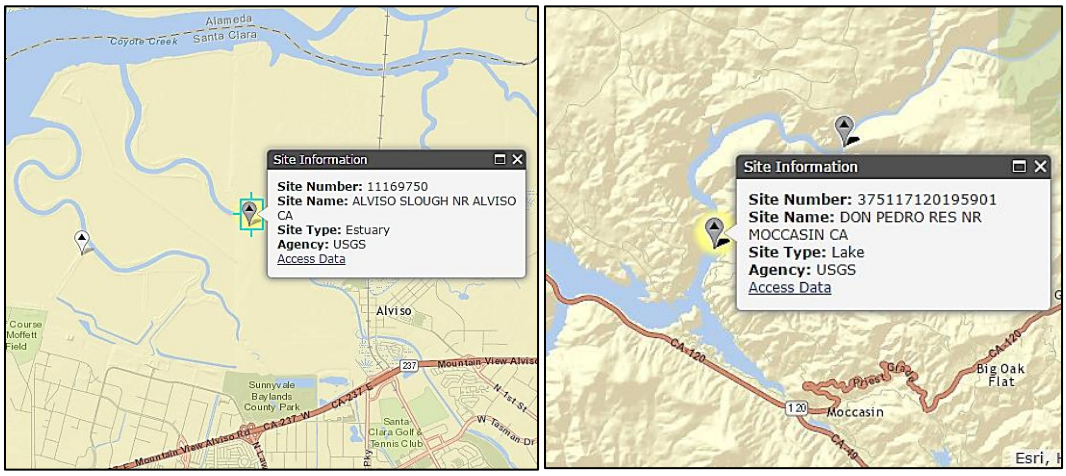
El USGS ofrece diferentes servicios de consulta de datos a través de internet, entre los más destacados se encuentra el Water-Quality Watch, que despliega datos de calidad del agua en tiempo real recolectados remotamente mediante sensores instalados en ríos, lagos y otros cuerpos de agua. Por otra parte el Portal de Calidad del Agua o WQP por sus siglas en inglés (Water Quality Portal), es un servicio cooperativo que integra datos públicos de calidad del agua disponibles desde la bodega de datos de recuperación y almacenamiento de la EPA (STORET, por sus siglas en inglés EPA STOrage and RETrieval) y la base de datos del Sistema de Información Nacional de Calidad del Agua del USGS o NWIS por sus siglas en inglés (USGS National Water Information System); este último permite consultar datos discretos de muestras de agua y datos correspondientes a series de tiempo (estos datos pertenecen a aproximadamente a 1.5 millones de sitios en todos los 50 estados de los Estados Unidos). Los resultados de 5 millones de muestras de agua con 90 millones de resultados de calidad del agua están disponibles a partir de una amplia variedad de métodos de recuperación incluyendo interfaces de mapas estándar y personalizados.

Los datos consultados están etiquetados en dos categorías generales: Datos aprobados para publicación (A), los cuales han sido procesados y revisados completamente por el personal del USGS; y Datos provisionales sujetos a revisión (P) que no cuentan con aprobación del personal de revisión. Teniendo en cuenta lo anterior, en este estudio fue necesario realizar una selección de aquellos datos con la etiqueta (A) que son utilizados como entrada en el mecanismo de predicción adaptativo propuesto.

Los datos incluyen variables como la Temperatura del Agua, pH, Conductividad Específica, Turbidez, Oxígeno Disuelto y/o Nitrato dependiendo del sitio de muestreo de donde se extrae el cuerpo de agua. Para el caso del NWIS, el tipo de sitio de muestreo determina el uso del agua ya que un tipo de sitio es un lugar generalizado en el ciclo hidrológico, o una característica hecha por el hombre que puede afectar a las condiciones hidrológicas medidas en un sitio (algunos

ejemplos del tipo de sitio de muestreo son: estuario, lago, planta hidroeléctrica, planta de tratamiento de suministro de agua, entre otros). Para realizar la consulta de datos fueron tenidos en cuenta parámetros como el tipo de sitio, el estado o territorio (corresponde al nombre de un estado o territorio de los Estados Unidos), el tipo de parámetros de calidad del agua (fisicoquímicos) y finalmente el rango de fechas (el más amplio hasta el desarrollo de esta investigación comprende desde 2007/10/01 a 2015/03/30).

Para la obtención de los datos de prueba, en este estudio fueron seleccionadas dos zonas del estado de California, la primera comprende el territorio de Alviso, la cual es una pequeña comunidad de San José, Condado de Santa Clara. El río Guadalupe y el arroyo Coyote terminan en el humedal de Alviso por medio de un estuario que desemboca en la bahía de San Francisco; una de las principales actividades es la pesca. La segunda zona de estudio es el lago Don Pedro, ubicado en el condado de Mariposa, cubre un área de 32.56 km^2 donde una de los principales usos es el recreacional (natación, paseo en barco y otros deportes acuáticos).



(a) Estuario de Alviso (b) Lago Don Pedro

Figura 20. Localización geográfica del área de estudio. Tomado de [67].

4.2.2. Datos del PMC Fase II

Dentro del trabajo de pasantía desarrollado como parte de esta investigación (ver ANEXO B), se estableció como insumo principal para el mecanismo de predicción, un subconjunto de datos de calidad del agua pertenecientes al Proyecto de Modelación del Río Cauca (PMC Fase II) el cual comprende el valle geográfico del Río Cauca, abarcando el tramo desde la represa de La Salvajina hasta el

municipio de La Virginia. Es importante resaltar que la cuenca del Río Cauca es la segunda arteria fluvial más importante de Colombia y atraviesa alrededor de 183 municipios pertenecientes a los departamentos de Cauca, Valle del Cauca, Quindío, Risaralda, Caldas, Antioquia, Córdoba, Sucre y Bolívar, en donde habitan más de 19 millones de personas que representan aproximadamente el 41 % del total de la población colombiana [68].

Los principales usos del agua que circula por este río han sido la pesca, la recreación, la generación de energía, la extracción de materiales del lecho y el agua además es captada para consumo humano, riego y la industria. La calidad del agua se ha visto afectada debido a que el Río Cauca se ha usado como fuente receptora de residuos sólidos y vertimientos de aguas residuales.

La zona de estudio de donde se obtuvieron los datos de calidad del agua es el tramo Hormiguero - Mediacanoa, específicamente la estación de Puente Juanchito, muy cercana a la ciudad de Santiago de Cali. Este es el tramo en el cual se presenta la mayor contaminación por materia orgánica asociada con las múltiples descargas de aguas residuales que son vertidas a la cuenca en este tramo. La Figura 21 muestra las estaciones de monitoreo pertenecientes la cuenca del Río Cauca en el tramo Salvajina – La Virginia, las cuales están a cargo de la Corporación Regional Autónoma del Valle del Cauca (CVC).

Dentro del PMC II fueron registrados datos de Caudal, Temperatura, OD, DBO5, DQO, pH, Sólidos Suspendidos Totales y Conductividad. El periodo en que se realizaron las muestras fue de 5 días y la frecuencia varió, dependiendo de los recursos disponibles, entre 1 muestra por día y 24 por día.

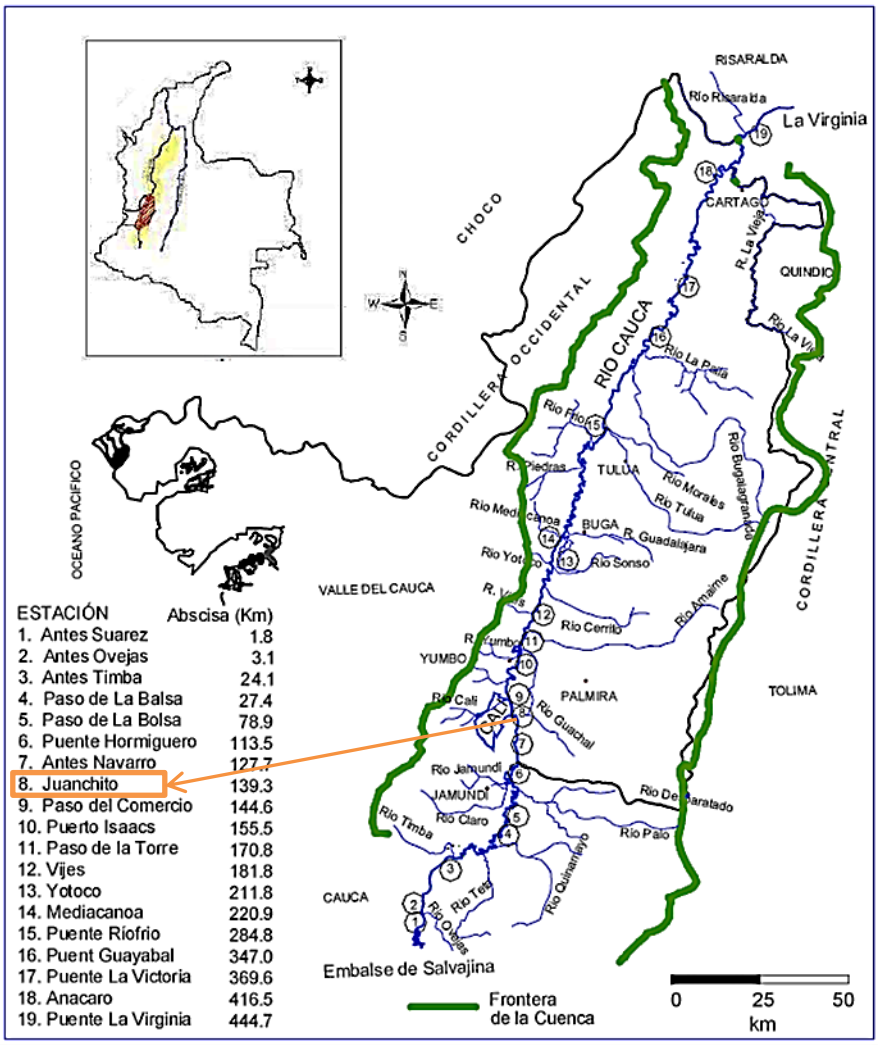


Figura 21. Estaciones de monitoreo de la calidad del agua del Río Cauca a cargo de la CVC – resaltada la estación de Puente Juanchito. Tomado de [68].

4.3. Selección del Algoritmo de Predicción

Para determinar el algoritmo dentro de la IC que ofrezca mejores resultados en la predicción de la calidad del agua, es necesario contar con una base teórica que permita establecer los algoritmos referentes para ser comparados. Los estudios [11], [12], [17], [18] y [69], hacen referencia a la utilización de las RNA (perceptrón multicapa) y la Regresión por Vectores de Soporte (SVR) como estrategias para la obtención de buenos resultados en la predicción de la calidad del agua a partir de valores de series temporales. Adicionalmente se evaluó el método de Regresión Lineal (RL), el cual ha sido utilizado tradicionalmente para este tipo de problemas [70].

Inicialmente se obtuvo un conjunto de datos de temperatura del agua correspondiente a 96 instancias (registros de un día con mediciones cada 15 minutos). Este conjunto de datos fue la entrada para cada uno de los 3 algoritmos de predicción, los cuales fueron configurados para predecir un solo valor (el software utilizado para aplicar los algoritmos fue WEKA 3.7.12).

Con el objetivo de evaluar la efectividad de los algoritmos de predicción, se aplicaron las siguientes métricas de precisión.

Error Absoluto Medio (MAE)

El Error Absoluto Medio (MAE por su sigla en inglés, Mean Absolute Error) permite medir la cercanía que hay entre una predicción y el valor real de un conjunto de datos [71], y se define a través de la Ecuación (**4.1**.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - v_i| \tag{4.1}$$

Donde, p_i es el valor predicho, v_i el valor real y n el número de datos. En caso de que $MAE = 0$, se interpreta que la predicción es perfecta.

Porcentaje de Error Medio Absoluto (MAPE)

En ocasiones, resulta más útil calcular los errores de pronóstico en términos de porcentaje y no en cantidades. El Porcentaje de Error Medio Absoluto (MAPE por su sigla en inglés, Mean Absolute Percentaje Error) se calcula encontrando el error absoluto en cada periodo, dividiendo éste entre el valor real observado para ese periodo y después promediando estos errores absolutos de porcentaje. Se puede utilizar el MAPE para comparar la precisión de la misma u otra técnica sobre dos series de tiempo completamente diferentes. La Ecuación (**4.2** muestra el cálculo del MAPE.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - v_i|}{p_i} \tag{4.2}$$

Error Cuadrático Medio (RMSE)

Esta medida estadística conocida en inglés como Root Mean Square Error, representa la diferencia entre el valor predicho y el valor observado, mediante la media cuadrática [72], y está definida en la Ecuación (**4.3**.

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(p_i - v_i)^2}}{n} \quad (4.3)$$

En la Tabla 8 son presentadas las métricas de precisión que determinaron la efectividad de cada algoritmo. En esta primera etapa de evaluación solo fueron utilizadas MAE y MAPE; en la segunda etapa se aplicó adicionalmente la medida RMSE.

Métrica de Precisión	RL	RNA	SVR
MAE	0.051	0.058	0.044
MAPE	0.2398	0.2741	0.2068

Tabla 8. Métricas de precisión para los tres algoritmos de predicción (1 valor predicho).

La anterior tabla indica que con la técnica de SVR se obtienen resultados más precisos respecto a las RNA y a la Regresión Lineal, dado que el Error Absoluto Medio fue menor que los demás (0.044), así mismo el Porcentaje de Error Absoluto Medio de 0.2068 → 20.68%. Sin embargo este porcentaje de error aún es considerablemente alto para obtener un buen modelo de predicción. Con base en lo anterior, en este trabajo se propuso realizar una nueva evaluación de los 3 algoritmos, pero en este caso aumentando la cantidad de datos de entrada (480 instancias) y el número de valores a predecir (5). Las métricas de precisión para esta nueva prueba se consignan en la Tabla 9.

Métrica de Precisión	RL	RNA	SVR
MAE	0.0537	0.0586	0.0501
MAPE	0.2816	0.3095	0.2635

Tabla 9. Métricas de precisión para los tres algoritmos de predicción (5 valores predichos).

A partir de la tabla anterior se puede inferir que nuevamente con las SVR se obtienen mejores resultados que con las RNA y la Regresión Lineal. Teniendo en cuenta esto, fue necesario realizar una tercera evaluación configurando la SVR de acuerdo a su método kernel, el cual permite construir algoritmos de aprendizaje genéricos que pueden utilizarse sobre cualquier tipo de dato (vectorial o no).

Para realizar esta evaluación se tuvieron en cuenta los siguientes métodos kernel mostrados en la Tabla 10 los cuales están implementados en WEKA y se procedió con el proceso de predicción de la SVR (480 instancias para predecir 5 valores de temperatura del agua).

Método Kernel	MAE	MAPE
Normalized Poli Kernel	0.1672	0.8438
Poli Kernel	0.0501	0.2635
PUK	0.0216	0.1138
RBF Kernel	0.0979	0.5145

Tabla 10. Métricas de precisión aplicadas a la SVR variando el método de kernel.

Con base en la tabla anterior, la Figura 22 permite apreciar el Porcentaje de Error Absoluto perteneciente a los 4 métodos de kernel utilizados, de los cuales el menor porcentaje de error fue obtenido mediante el método PUK (Pearson VII Universal Kernel), que tiene la posibilidad de cambiar fácilmente, mediante la adaptación de sus parámetros, a partir de una función Gaussiana en una Lorentziana, es decir, esta propiedad hace posible el uso de Puk como un kernel genérico que puede sustituir a los demás métodos de kernel evaluados [63].

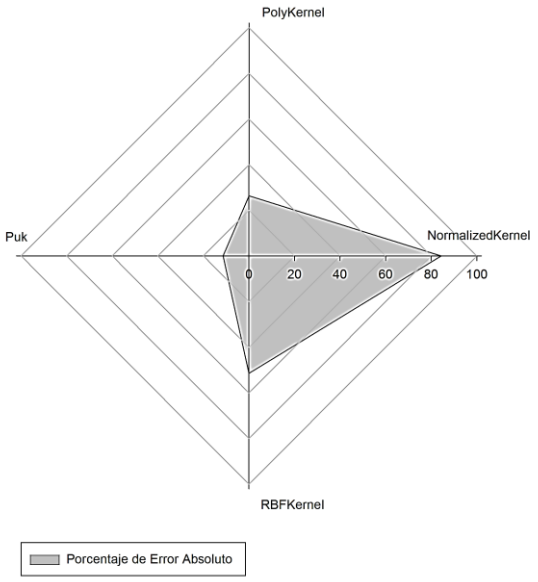


Figura 22. Porcentaje de Error Absoluto Medio para los cuatro métodos de kernel utilizados en la configuración de la SVR.

La Figura 23 presenta los valores de la variable temperatura que fueron predichos mediante el uso de la SVR y el método de kernel PUK. Teniendo en cuenta el conjunto de pruebas realizado durante la primera fase de evaluación, se estableció

la SVR configurada con el kernel PUK como la técnica de predicción utilizada en este estudio, la cual a su vez, es el componente principal del módulo de predicción del mecanismo propuesto.

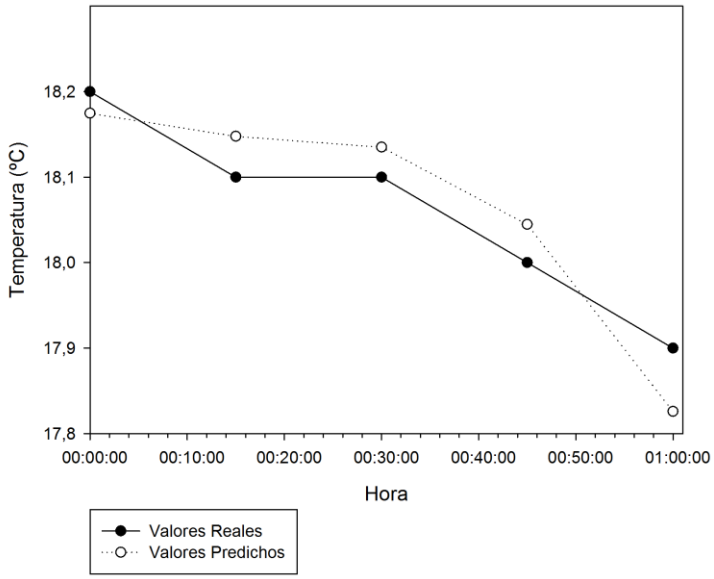


Figura 23. Valores predichos usando SVR con el kernel PUK, comparados con los valores reales de la variable temperatura.

4.4. Ajuste del Algoritmo de Predicción

En la segunda etapa de evaluación, inicialmente se presentan los valores predichos mediante la Regresión por Vectores de Soporte (SVR) configurada con el kernel PUK como se estableció en el numeral anterior. Las pruebas fueron realizadas utilizando el mismo conjunto de datos correspondiente al estuario de Alviso, el cual representa el uso correspondiente a la piscicultura; el Lago Don Pedro, correspondiente al uso recreacional; y la estación Puente Juanchito dentro del uso para consumo humano.

Con base en lo anterior los conjuntos de datos pre-procesados contienen diferentes valores para distinto número de variables fisicoquímicas; esto se debe a que dentro del componente de calibración de parámetros se realiza un proceso de selección de parámetros que en este caso corresponden a las variables físico-químicas. Esta selección permite escoger automáticamente el subconjunto de variables que mejor representan al uso del agua de cada conjunto de datos de entrada. En la Tabla 11 se muestra las variables seleccionadas para los dos conjuntos de datos (USGS y PMC II).

Conjunto de Datos	Variables Seleccionadas
USGS	Temperatura (ºC)
	Conductancia (µsm/cm)
PMC II	Temperatura (ºC)
	Conductancia (µsm/cm)
	Oxígeno Disuelto (%)
	pH

Tabla 11. Variables seleccionadas por cada conjunto de datos.

El número de valores predichos corresponde a un (1) día de medición, es decir 96 datos, teniendo en cuenta que las mediciones se efectúan cada 15 minutos. Las curvas con los valores predichos son mostradas en la Figura 24 y Figura 25 para el estuario de Alviso.

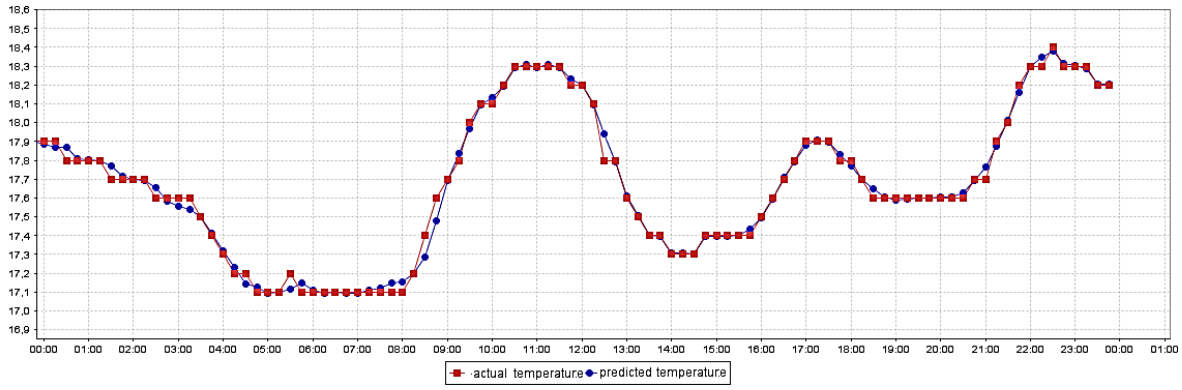


Figura 24. Valores predichos de temperatura del agua usando SVR con el kernel PUK (datos de piscicultura – estuario de Alviso).

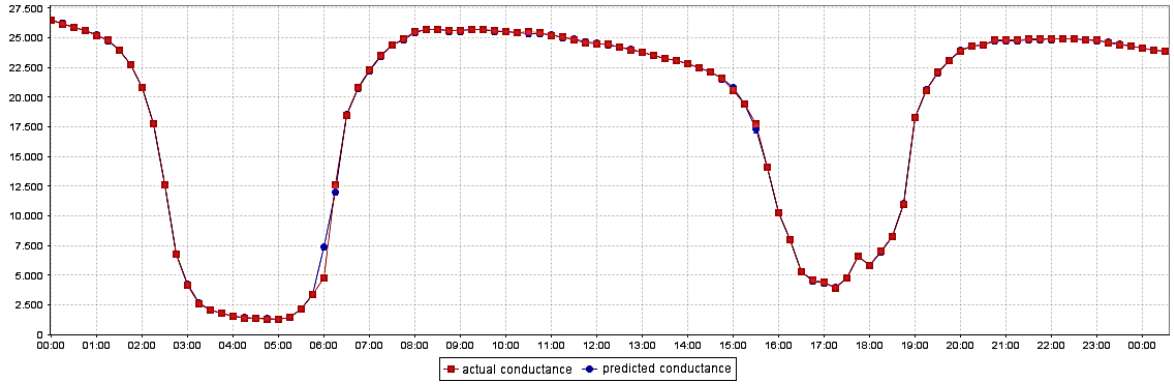


Figura 25. Valores predichos de conductividad del agua usando SVR con el kernel PUK (datos de piscicultura – estuario de Alviso).

En las dos curvas anteriores puede observarse que los valores predichos se aproximan a los valores reales para ambas variables. Sin embargo, teniendo en cuenta que el enfoque de este estudio es mantener la precisión de las predicciones cuando se utilice un conjunto de datos perteneciente a otro uso del

agua, en la Figura 26 y Figura 27 se observa el comportamiento de estas mismas variables correspondientes al uso recreacional del agua (Lago Don Pedro).

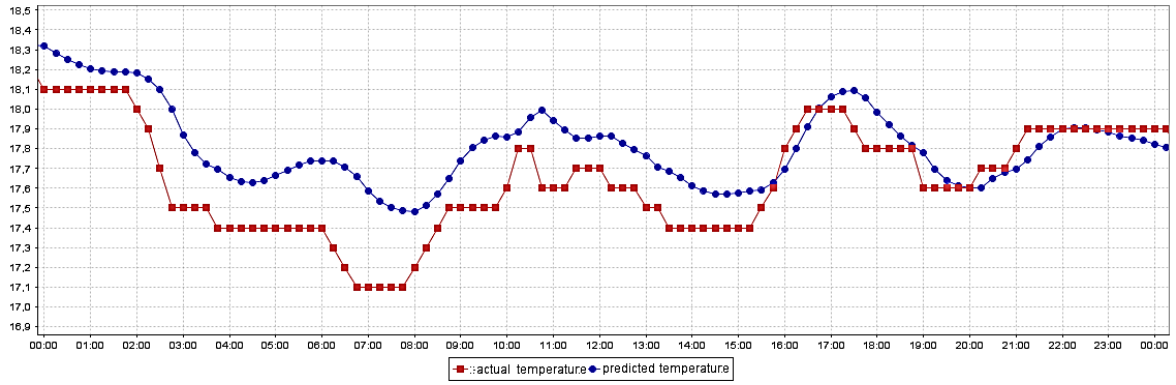


Figura 26. Valores predichos de temperatura del agua usando SVR con el kernel PUK (datos de uso recreacional – lago Don Pedro).

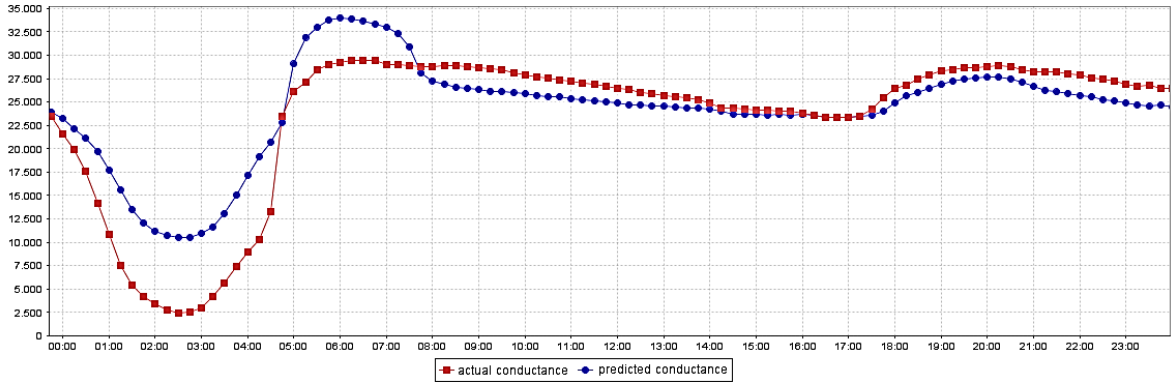


Figura 27. Valores predichos de conductividad del agua usando SVR con el kernel PUK (datos de uso recreacional – lago Don Pedro).

Por otra parte, el comportamiento de la técnica SVR-PUK aplicada sobre el conjunto de datos pertenecientes al PMC II estación Puente Juanchito se muestra desde la Figura 28 a la Figura 31.

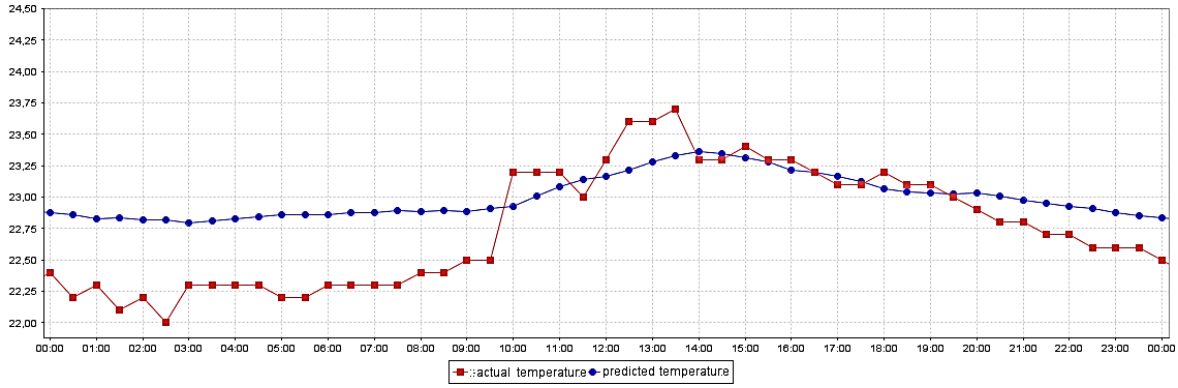


Figura 28. Valores predichos de temperatura del agua usando SVR con el kernel PUK (datos de uso para consumo humano – estación Puente Juanchito).

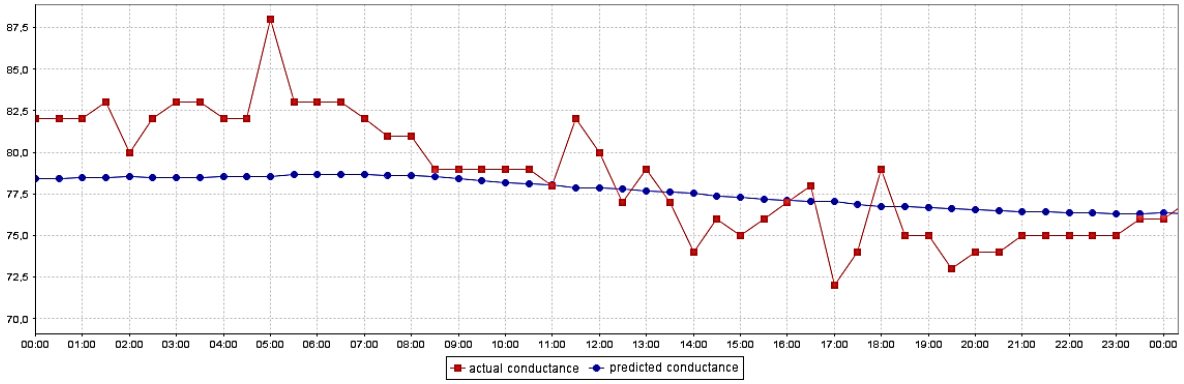


Figura 29. Valores predichos de conductividad del agua usando SVR con el kernel PUK (datos de uso para consumo humano – estación Puente Juanchito).

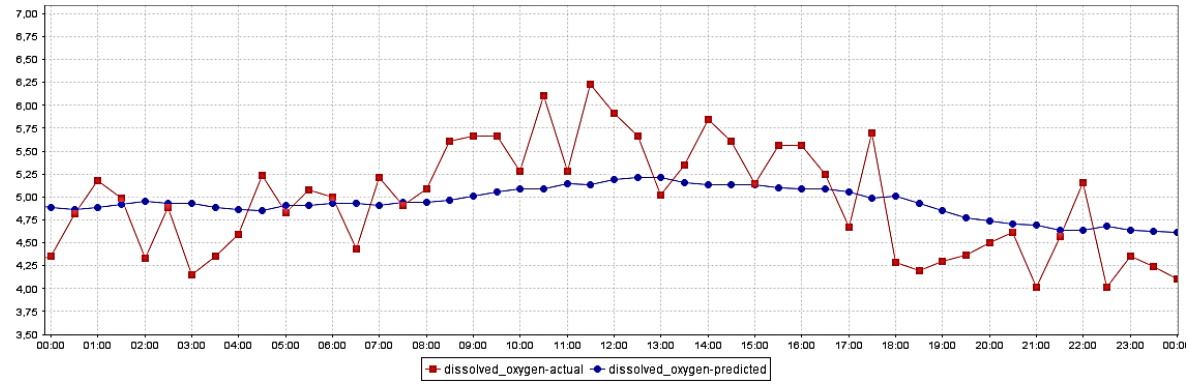


Figura 30. Valores predichos de oxígeno disuelto en el agua usando SVR con el kernel PUK (datos de uso para consumo humano – estación Puente Juanchito).

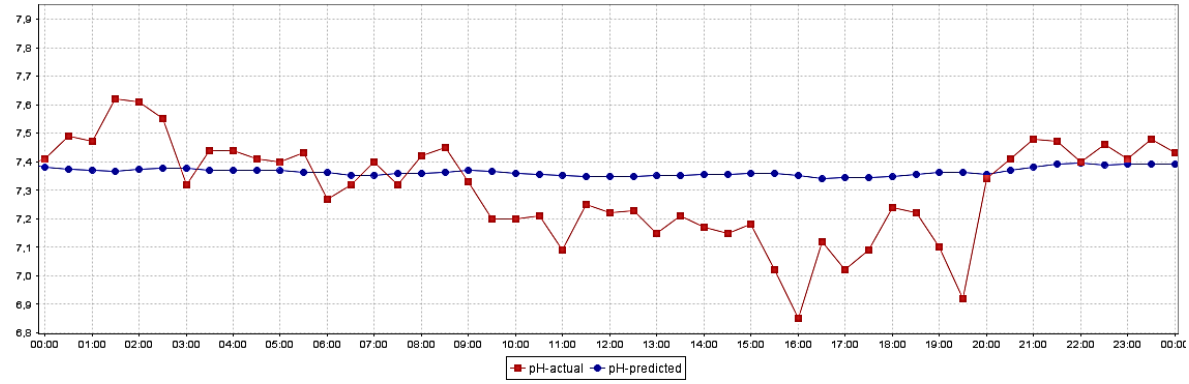


Figura 31. Valores predichos de pH en el agua usando SVR con el kernel PUK (datos de uso para consumo humano – estación Puente Juanchito).

La Tabla 12 presenta la precisión del algoritmo SVR-PUK para cada una de las variables predichas en los tres sitios definidos.

Conjunto de datos	Sitio de Muestreo	Variable de Calidad del Agua	MAE	RMSE	MAPE
USGS	Estuario de Alviso	Temperatura	0.0216	0.0398	0.1138
		Conductividad	2653.28	2942.71	0.1689

	Lago Don Pedro	Temperatura	0.1971	0.2423	0.1938
		Conductividad	2054.327	2737.510	0.3552
PMC II	Puente Juanchito	Temperatura	0.3612	0.4368	0.5868
		Conductividad	2.243	2.7903	0.6523
		Oxígeno Disuelto	0.3759	0.4589	0.6931
		pH	0.0992	0.1362	0.5554

Tabla 12. Métricas de precisión para el proceso de predicción mediante SVR-PUK. Variables de calidad del agua: Temperatura y Conductividad en el estuario de Alviso y el lago Don Pedro; Temperatura, Conductividad, Oxígeno Disuelto y pH en la estación Puente Juanchito.

Como puede observarse en las 8 curvas anteriores y de forma general en la Tabla 12, los valores predichos no presentan una alta coincidencia con los valores reales de sus respectivos conjuntos de datos. Al aplicar el algoritmo de predicción sobre datos de uso recreacional (lago Don Pedro) y datos de calidad de agua para consumo humano (estación Puente Juanchito), el porcentaje de error en la predicción aumenta considerablemente tanto para la temperatura del agua como para la conductancia, el Oxígeno Disuelto y el pH respecto al porcentaje de error que se obtiene sobre estos parámetros en el estuario de Alviso. De esta manera fue necesario realizar un ajuste sobre la técnica de predicción utilizada (SVR-PUK).

Para realizar dicho ajuste, se optó por el uso de algoritmos evolutivos, debido a la naturaleza del problema, en el cual se busca optimizar los valores predichos correspondientes a los parámetros de calidad del agua de determinado uso, de tal forma que al realizar un nuevo proceso de predicción sobre parámetros de otro uso del agua, la precisión no se vea afectada o tienda a disminuir drásticamente. De acuerdo con lo anterior, en los estudios presentado en [73] y [74], se realiza una comparación respecto a la efectividad de este tipo de algoritmos y sus ventajas en la resolución de problemas en entornos altamente cambiantes y dinámicos. De esta manera se seleccionaron dos de las técnicas más usadas y de mejor rendimiento: Los Algoritmos Genéticos y la Optimización por Nubes de Partículas. En la Tabla 13 y en la Figura 32 se muestran los resultados de la evaluación realizada para cada algoritmo variando el número de Valores Predichos (VP) entre 1 y 10.

Algoritmo Evolutivo	**MAPE**									
	VP=1	**VP=2**	**VP=3**	**VP=4**	**VP=5**	**VP=6**	**VP=7**	**VP=8**	**VP=9**	**VP=10**
Algoritmo Genético	0,0568	0,0579	0,0612	0,1804	0,2282	0,2845	0,3491	0,4303	0,4962	0,5608
Optimización por Nubes de Partículas	0.0127	0.0441	0.0967	0.1384	0.1567	0.1601	0.2515	0.2790	0.3125	0.3782

Tabla 13. Porcentajes de error para las técnicas AG y PSO variando el número de valores predichos.

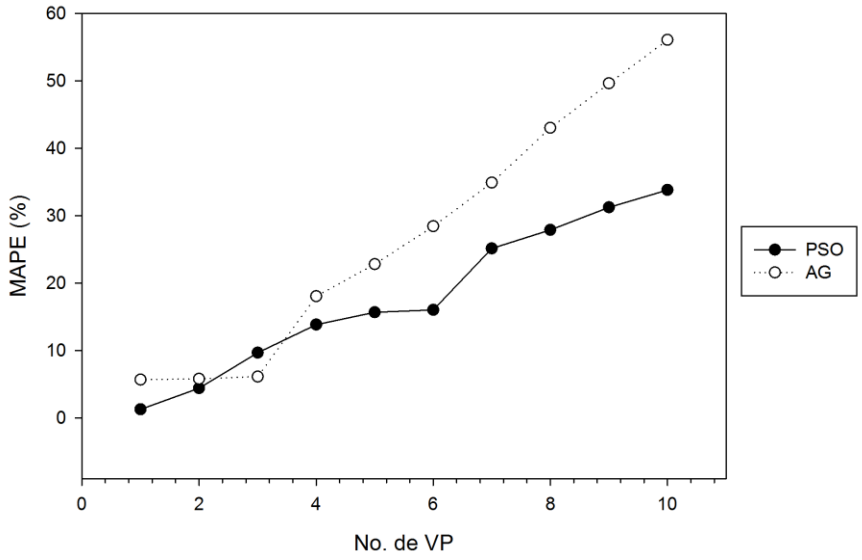


Figura 32. Comportamiento de los porcentajes de error para las técnicas AG y PSO variando el número de valores predichos.

De acuerdo con los resultados expuestos en la anterior tabla se optó por seleccionar el algoritmo PSO como principal elemento del componente adaptativo del mecanismo de predicción propuesto. A continuación se presentan los pasos básicos de este algoritmo.

Algoritmo 1. Optimización por Nubes de Partículas (PSO)

Entradas: $S \in \mathbb{Z}$ $(número\ de\ partículas)$
Salida: $g \in \mathbb{R}$ $(mejor\ solución\ global\ en\ el\ espacio\ de\ búsqueda)$

1: Inicializar parámetros ω y φ
2: **Para cada partícula** $(\ i = 1;\ \ i \leq S;\ \ i = i + 1\ )$
3: Inicializar la posición de la partícula mediante un vector aleatorio uniformemente distribuido: $x_i \sim U(b_{lo}, b_{up})$, donde b_{lo} y b_{up} son respectivamente el límite inferior y el límite superior del espacio de búsqueda.
4: Inicializar la mejor posición conocida de la partícula a su posición inicial: $p_i \leftarrow x_i$
5: **Si** $f(p_i) < f(g)$ **entonces:**
Actualizar la mejor posición global conocida: $g_i \leftarrow p_i$
6: Inicializar la velocidad de la partícula: $v_i \sim U(b_{lo} - b_{up}, b_{lo} - b_{up})$
7: **Fin Para**
8: **Mientras no se cumpla el criterio de parada** (por. ej. límite máximo de iteraciones, encontrada una solución satisfactoria), **repetir**
9: **Para cada partícula** $(\ i = 1;\ \ i \leq S;\ \ i = i + 1\ )$
10 **Para cada dimensión** $(\ d = 1;\ \ d \leq n;\ \ d = d + 1\ )$
11: Elegir números aleatorios: $r_p, r_g \sim U(0,1)$

12: Actualizar la velocidad de la partícula:
$v_i \leftarrow \omega v_{i,d} + \varphi_p, r_p(p_{i,d} - x_{i,d}) + \varphi_g, r_g(g_d - x_{i,d})$
13: **Fin Para**
14: Actualizar la posición de la partícula: $x_i \leftarrow x_i + v_i$
15: **Si** $f(p_i) < f(g)$ **entonces:**
16: Actualizar la mejor posición conocida de la partícula: $p_i \leftarrow x_i$
17: **Si** $f(p_i) < f(g)$ **entonces:** actualizar la mejor posición global: $g \leftarrow p_i$
18: **Fin Si**
19: **Fin Para**
20: **Fin Mientras**
21: Devolver g como la mejor solución encontrada.

En el anterior algoritmo el parámetro ω se denomina “peso de inercia”, y puede ser interpretado como la fluidez del medio en el que se mueven las partículas. De esta manera un alto valor ω representa un entorno de "baja fricción" adecuado para la exploración del espacio, mientras que un valor bajo representa un entorno de "alta fricción", más apropiado para la explotación del espacio (orientado a una búsqueda local). Los parámetros φ_p y φ_g, por otra parte, representan la atracción hacia la mejor posición de una partícula y hacia la mejor posición del grupo, respectivamente.

Partiendo de la elección del algoritmo PSO para realizar el ajuste sobre el algoritmo de predicción, en la Figura 33 a la Figura 40 se presenta el comportamiento de los valores predichos para cada variable (temperatura, conductividad, oxígeno disuelto y pH), haciendo uso del mecanismo híbrido planteado (SVR-PUK-PSO) con los conjuntos de datos de los tres usos del agua establecidos (piscícola, recreacional y consumo humano).

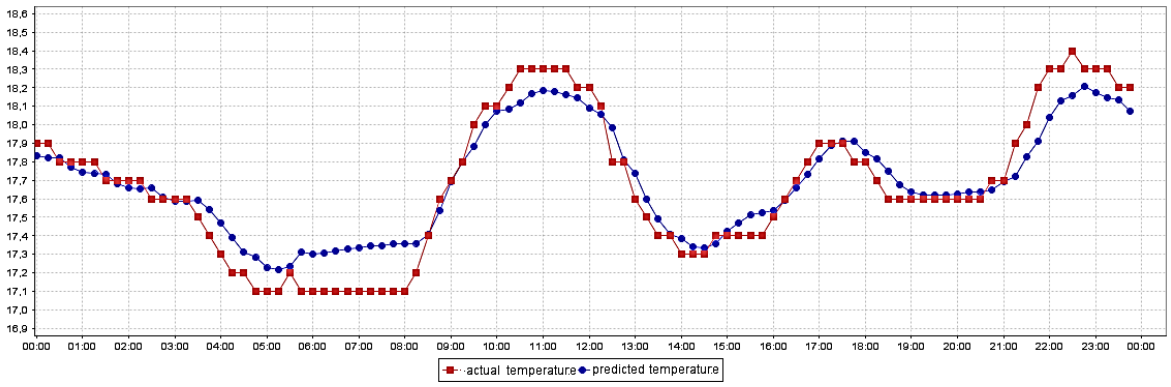


Figura 33. Valores predichos de temperatura del agua usando SVR con el kernel PUK y la técnica PSO (datos de piscicultura – estuario de Alviso).

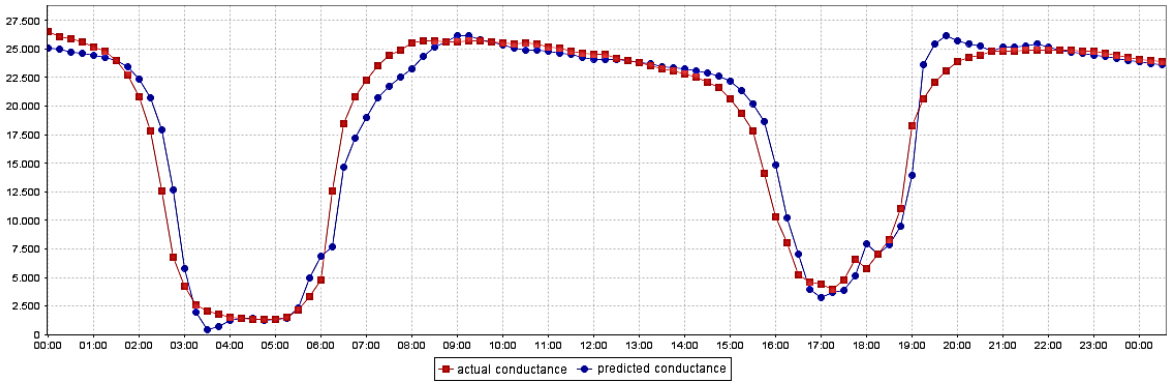


Figura 34. Valores predichos de conductividad del agua usando SVR con el kernel PUK y la técnica PSO (datos de piscicultura – estuario de Alviso).

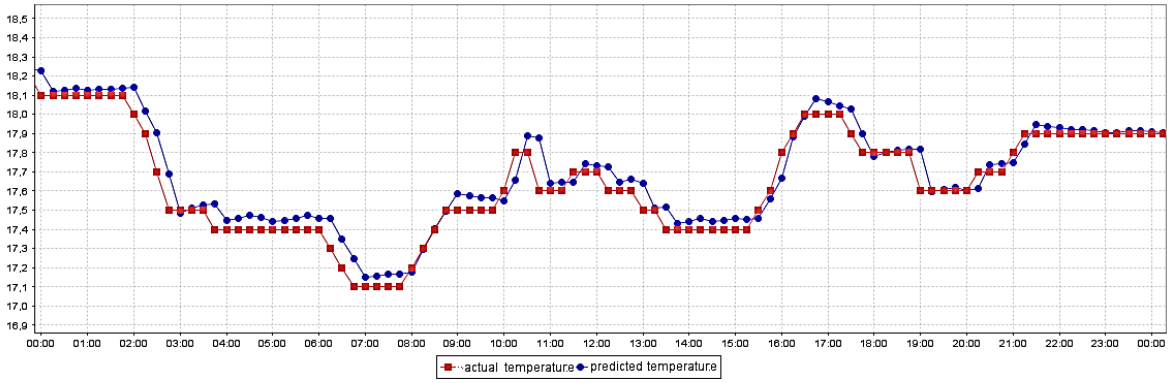


Figura 35. Valores predichos de temperatura del agua usando SVR con el kernel PUK y la técnica PSO (datos de uso recreacional – lago Don Pedro).

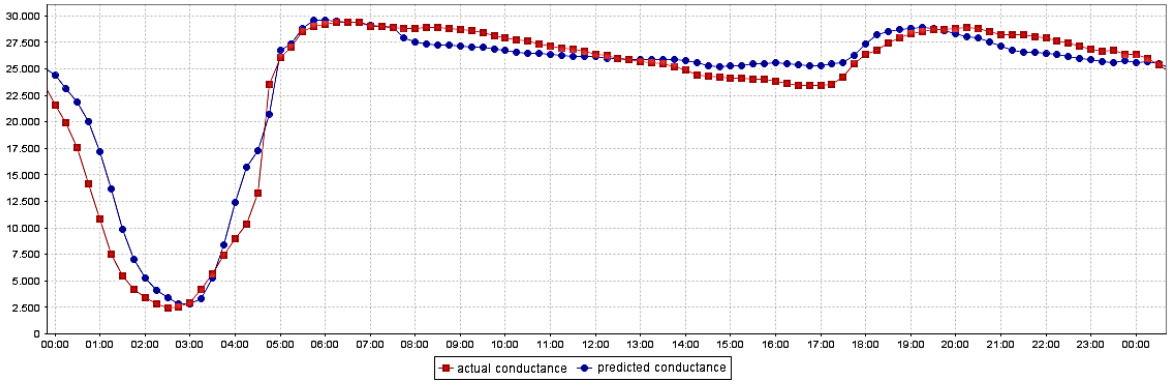


Figura 36. Valores predichos de conductividad del agua usando SVR con el kernel PUK y la técnica PSO (datos de uso recreacional – lago Don Pedro).

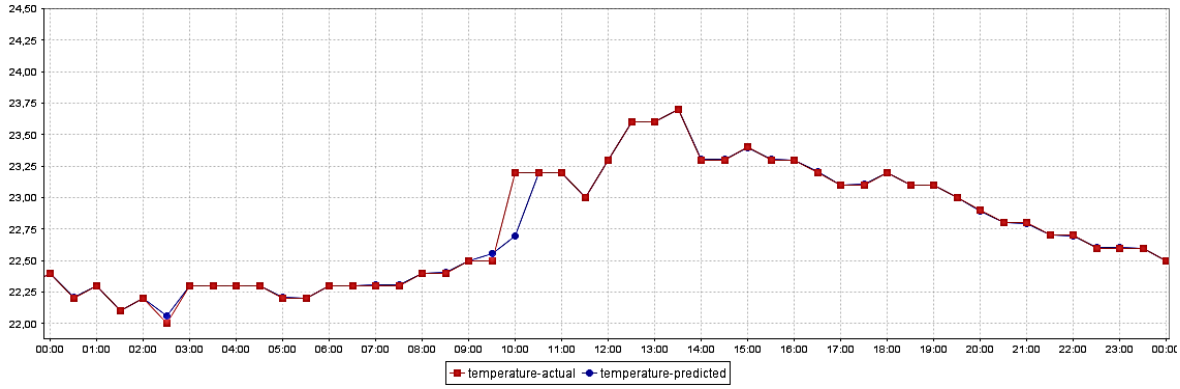


Figura 37. Valores predichos de temperatura del agua usando SVR con el kernel PUK y la técnica PSO (datos de uso para consumo humano – estación Puente Juanchito).

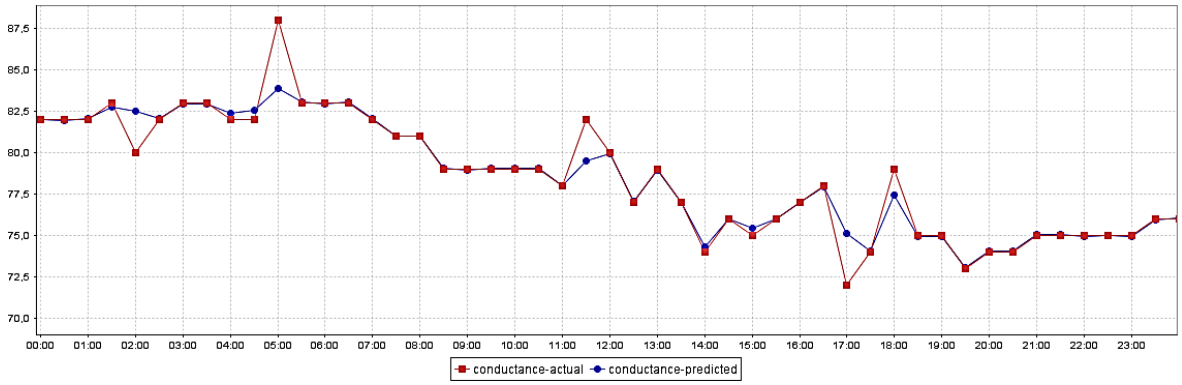


Figura 38. Valores predichos de conductividad del agua usando SVR con el kernel PUK y la técnica PSO (datos de uso para consumo humano – estación Puente Juanchito).

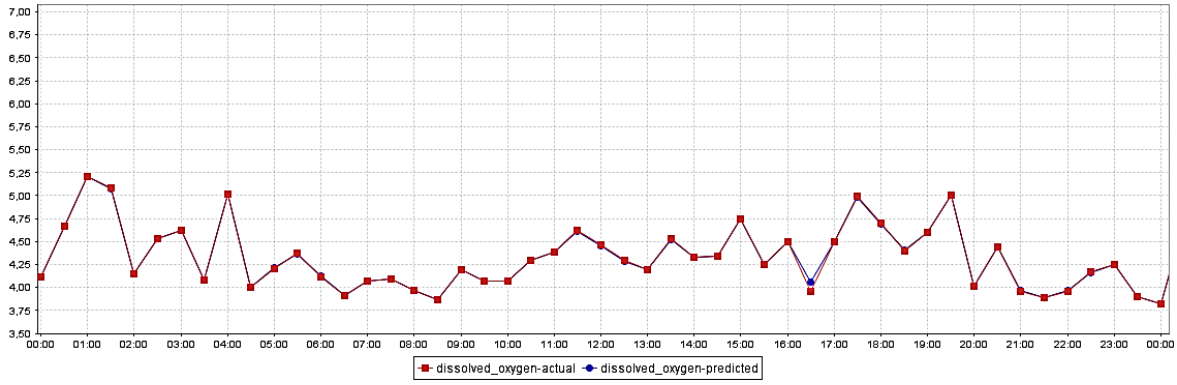


Figura 39. Valores predichos de Oxígeno Disuelto en el agua usando SVR con el kernel PUK y la técnica PSO (datos de uso para consumo humano – estación Puente Juanchito).

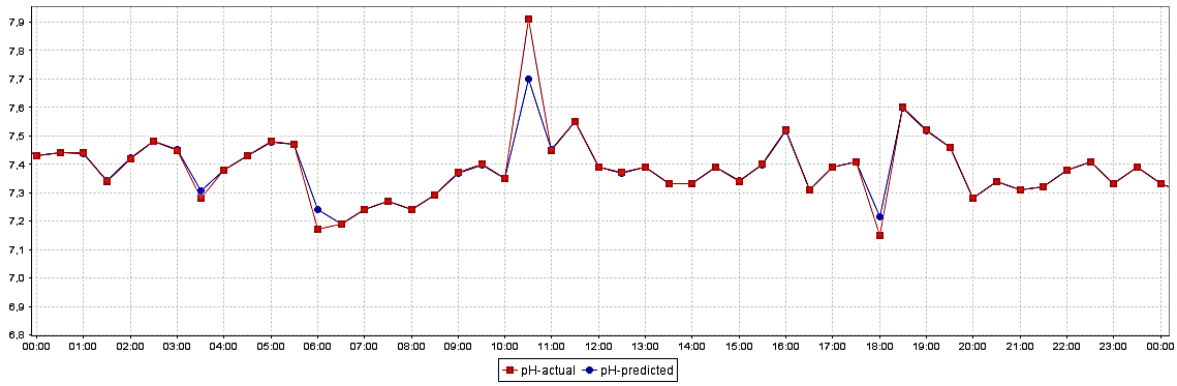


Figura 40. Valores predichos de Oxígeno Disuelto en el agua usando SVR con el kernel PUK y la técnica PSO (datos de uso para consumo humano – estación Puente Juanchito).

La Tabla 14 presenta la precisión del algoritmo SVR-PUK ajustando los valores de cada una de las variables predichas en los dos sitios definidos mediante la técnica de PSO.

Conjunto de datos	Sitio de Muestreo	Variable de Calidad del Agua	MAE	RMSE	MAPE
USGS	Estuario de Alviso	Temperatura	0.1672	0.2556	0.2438
		Conductividad	1302.56	1715.28	0.2239
	Lago Don Pedro	Temperatura	0.0586	0.0789	0.3095
		Conductividad	1632.875	2056.435	0.2862
PMC II	Puente Juanchito	Temperatura	0.0108	0.0499	0.0473
		Conductividad	0.2805	0.8305	0.3589
		Oxígeno Disuelto	0.004	0.0079	0.0901
		pH	0.0038	0.0179	0.0514

Tabla 14. Métricas de precisión para el proceso de predicción mediante SVR-PUK-PSO. Variables de calidad del agua: Temperatura y Conductividad en el estuario de Alviso y el lago Don Pedro; Temperatura, Conductividad, Oxígeno Disuelto y pH en la estación Puente Juanchito

Considerando los resultados que son presentados en la Tabla 14, puede observarse que mediante el ajuste del algoritmo de predicción, se obtienen resultados similares en la precisión de los valores predichos, si bien la precisión disminuye un poco con el conjunto de datos de piscicultura en el estuario de Alviso, puede considerarse que dicha disminución no es significativa si se tiene en cuenta que la precisión de los valores predichos con el conjunto de datos de uso recreacional en el lago Don Pedro, aumentó considerablemente respecto a los valores que se habían obtenido con la aplicación del algoritmo SVR-PUK, en otras palabras, el porcentaje de error disminuyó aproximadamente en un 70% para la variable temperatura del agua; y en un 50% para la variable conductancia del agua.

Además de lo anterior, el mecanismo de predicción mostró un grado de precisión muy alto con el conjunto de datos PMC II – Puente Juanchito para las variables Temperatura, Oxígeno Disuelto y pH. Por otro lado la predicción de valores para la variable Conductividad no fue tan precisa, sin embargo se encuentra dentro de valores aceptables de precisión teniendo en cuenta el carácter adaptativo del mecanismo propuesto; una hipótesis acerca de las posibles causas por las cuales la conductividad no presentó una mayor precisión se debe al hecho de contar con diferentes escalas de esta variable, dependiendo de la salinidad que posea determinado cuerpo de agua [75], en este caso el estuario de Alviso por ser la confluencia entre un río y el mar presenta una salinidad mayor; de igual forma el lago Don Pedro muestra altos niveles de salinidad, lo que contrasta con la baja presencia de niveles de sal en el río Cauca estación Puente Juanchito.

4.5. Resumen

En este capítulo fueron presentadas las pruebas correspondientes a la experimentación y evaluación del mecanismo de predicción adaptativo. Las medidas de precisión que se aplicaron fueron: MAE, MAPE, RMSE. En la primera fase de evaluación fue seleccionado como técnica de predicción la Regresión por Vectores de Soporte, en la segunda fase fue definida la técnica de adaptación, la cual ajusta los valores retornados por el algoritmo de predicción a un determinado uso del agua, aquí se determinó que la Optimización por Nubes de Partículas era la técnica más apropiada para realizar este proceso. Finalmente fueron presentadas las curvas que permiten visualizar el comportamiento del mecanismo propuesto sobre diferentes usos del agua.

Capítulo 5

Cumplimiento de Objetivos

Para determinar el cumplimiento de los objetivos, a continuación se presenta un modelo de indicadores que permite evaluar de manera objetiva el cumplimiento de los mismos en el presente proyecto, de esta manera se exponen los lineamientos de conformación e interpretación de los indicadores propuestos.

5.1. Lineamientos de Conformación e Interpretación de los Indicadores

Con el fin de expresar los resultados finales de cada uno de los objetivos se presenta a continuación una explicación de los tipos de indicadores utilizados en la evaluación de los resultados y la forma correcta de interpretarlos.

Los indicadores de desempeño que se evalúan, básicamente adoptan la forma de un cociente, en el cual, el denominador es un valor numérico que ayuda a efectuar la comparación con el logro obtenido así:

$$indicador = \left(\frac{numerador}{denominador}\right) * factor\ escala. \qquad (5.1)$$

De esta forma se definen los siguientes modelos de indicadores que se deben personalizar y aplicar a los actores, productos, funciones, etc. dependiendo del contexto del objetivo evaluado.

5.1.1. Indicador de Cobertura (IC)

Determina la cantidad de elementos cobijados por un producto o estrategia.

$$cobertura = \left(\frac{número\ de\ nodos\ beneficiados\ con\ el\ servicio}{número\ de\ nodos\ que\ se\ espera\ servir}\right) * 100. \qquad (5.2)$$

5.1.2. Indicador de Eficacia (IE)

Permite analizar el cumplimiento con los requisitos definidos.

$$eficacia = \left(\frac{recursos\ ejercidos}{recursos\ asignados}\right) * 100. \quad (5.3)$$

5.1.3. Indicador de Eficiencia (IF)

Permite identificar la relación que existe entre las metas alcanzadas, el tiempo y los recursos consumidos con respecto a un estándar. Representa el buen uso de los recursos.

$$eficiencia = \left(\frac{metas\ alcanzadas}{recursos\ consumidos}\right) * 100. \quad (5.4)$$

5.1.4. Indicador de Calidad (IQ)

Están orientados a medir la satisfacción de los beneficiarios.

$$calidad = calificación\ entre: [1: Mala(0\%), 2: Regular(50\%), 3: Buena(75\%), 4: Excelente(100\%)]. \quad (5.5)$$

Con el modelo de indicadores aquí presentado, se desarrolló un conjunto de indicadores que permiten evaluar adecuadamente el nivel de cumplimiento de cada uno de los objetivos. A continuación se presenta la evaluación realizada.

5.2. Descripción y alcance del cumplimiento de los objetivos

En la Tabla 15, Tabla 16 y Tabla 17, se especifican los objetivos comprometidos en el proyecto, los productos esperados derivados de cada objetivo, los resultados obtenidos, los indicadores que evalúan el objetivo, los medios de verificación de los resultados y finalmente, las observaciones que permiten aclarar los resultados en cada objetivo.

Se desarrolla una tabla por cada objetivo específico comprometido en la propuesta del presente proyecto.

No. Objetivo	1
Descripción del objetivo	Seleccionar una o más técnicas de Inteligencia Computacional (RNA, SVM, entre otras) que puedan ser utilizadas en el proceso de predicción de la calidad del agua.
Productos esperados	1. Documento donde se consigne las pruebas experimentales que condujeron a la selección de la(s) técnicas de IC.
Resultados obtenidos	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.3: Selección del algoritmo de Predicción.
Indicadores **(Escala * 100)**	**Eficacia** $$IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{1}{1} * 100 = 100\%$$ **Calidad** *IQ1 = ¿La(s) técnicas de IC seleccionadas permiten realizar una predicción de la calidad del agua aceptable sobre un uso del agua específico?* *R* = De acuerdo con el proceso de experimentación realizado aplicando 2 de los más utilizados algoritmos de IC como las RNA y las SVR, además de la Regresión Lineal tradicional, el desempeño de las SVR fue el que obtuvo una mayor precisión para predecir nuevos valores a partir de series temporales de calidad del agua; por lo tanto esta técnica fue seleccionada para ser parte del componente predictivo del mecanismo. $IQ1 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 1 = \frac{100}{1} = 100\%$
Medios de verificación	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.3: Selección del algoritmo de Predicción.
Estrategias, problemas y/u observaciones	Se estableció la SVR configurada con el kernel PUK como la técnica de predicción utilizada en este estudio, la cual es el componente principal del módulo de predicción del mecanismo propuesto. Uno de los inconvenientes encontrados durante la selección de la técnica de predicción fue el ruido encontrado en el conjunto de datos, es decir la presencia de valores atípicos, valores faltantes, variables no relacionadas, entre otros; lo anterior llevó a proponer el componente de calibración de parámetros en el cual se realiza un pre-procesamiento del conjunto de datos para que el componente predictivo pueda obtener mejores resultados.

Tabla 15. Cumplimiento del primer objetivo específico.

No. Objetivo	2
Descripción del objetivo	Definir el/los algoritmo(s) que permitan adaptar el proceso de predicción a diferentes usos del agua.
Productos esperados	1. Documento donde se consigne las pruebas experimentales que condujeron a la definición del algoritmo o los algoritmos para adaptar el proceso de predicción a diferentes usos del agua.

	2. Documento donde se presente la definición y descripción del mecanismo de predicción adaptativo.
Resultados obtenidos	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.4: Ajuste del Algoritmo de Predicción. 2. Monografía Capítulo 3: Mecanismo de Predicción Adaptativo.
Indicadores **(Escala * 100)**	**Eficacia** $$IE2 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{2}{2} * 100 = 100\%$$ **Calidad** *IQ1 = ¿Se definió el/los algoritmo(s) para adaptar el proceso de predicción a diferentes usos del agua?* *R* = Teniendo en cuenta el proceso de experimentación y evaluación realizado aplicando 2 de los algoritmos evolutivos como los Algoritmos Genéticos y la Optimización por Nubes de Partículas, se pudo determinar un mejor desempeño por parte del segundo algoritmo, el cual permitió adaptar los valores predichos a diferentes usos del agua. $IQ1 = 4 = 100\%$ *IQ2 = ¿Se desarrolló el mecanismo adaptativo de predicción de la calidad del agua?* *R* = A partir de la experimentación y evaluación desarrollada en el presente trabajo de investigación se seleccionaron y ajustaron diferentes algoritmos que mediante su interacción, componen el mecanismo de predicción adaptativo. $IQ2 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 2 = \frac{100}{1} = 100\%$
Medios de verificación	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.4: Ajuste del Algoritmo de Predicción. 2. Monografía Capítulo 3: Mecanismo de Predicción Adaptativo.
Estrategias, problemas y/u observaciones	Se construyó el mecanismo adaptativo de predicción de la calidad del agua mediante la selección y ajuste de varios algoritmos de IC. El principal inconveniente para la evaluación del mecanismo se centró en la poca disponibilidad de los conjuntos de datos pertenecientes a diferentes usos del agua o el restringido acceso que se tiene a estos datos; de esta forma se obtuvieron datos para uso piscícola y recreacional de libre acceso a través de internet (plataforma del USGS).

Tabla 16. Cumplimiento del segundo objetivo específico.

No. Objetivo	3
Descripción del objetivo	Evaluar experimentalmente el mecanismo propuesto a través del desarrollo de un prototipo, aplicado en dos contextos del uso del agua.
Productos esperados	1. Documento donde se establezca la implementación del prototipo que instancia el mecanismo de predicción adaptativo.

Resultados obtenidos	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.1: Desarrollo del Prototipo.
Indicadores **(Escala * 100)**	**Eficacia** $$IE2 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{1}{1} * 100 = 100\%$$ **Calidad** *IQ1 = ¿Se implementó el prototipo que instancia el mecanismo adaptativo de predicción de la calidad del agua?* *R* = Los algoritmos que hacen parte del prototipo fueron implementados mediante el software WEKA 3.7.12, Eclipse Luna y el framework Evolved Objects 1.3.1. $IQ1 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 3 = \frac{100}{1} = 100\%$
Medios de verificación	1. Monografía Capítulo 4: Experimentación y Evaluación, numeral 4.1: Desarrollo del Prototipo (el código fuente del prototipo es anexado digitalmente).
Estrategias, problemas y/u observaciones	Se implementó un prototipo que instancia el mecanismo propuesto utilizando tecnologías de libre distribución. Con el objetivo de evaluar conjuntos de datos pertenecientes a otros usos del agua, en la pasantía de investigación realizada en el instituto CINARA, se obtuvieron datos de calidad de agua para consumo humano. A partir de los dos conjuntos iniciales (piscicultura y uso recreacional) y este nuevo conjunto de datos, fue posible validar el mecanismo propuesto a través del prototipo.

Tabla 17. Cumplimiento del tercer objetivo específico.

Capítulo 6

Conclusiones y Trabajos Futuros

Este capítulo describe inicialmente las principales conclusiones de la presente tesis de maestría a las cuales se llegó durante su desarrollo, posteriormente presenta las recomendaciones, y finalmente propone los trabajos futuros que pueden generarse a partir de este estudio.

6.1. Conclusiones

A continuación se presentan las principales conclusiones obtenidas con la definición de dicho mecanismo.

- Para identificar las brechas de investigación que condujeron al desarrollo de la presente investigación, fue muy importante realizar un mapeo sistemático de la literatura relacionada con la predicción de la calidad del agua haciendo uso de técnicas de Inteligencia Computacional. Cabe resaltar que esta temática es bastante amplia y el análisis documental se abordó desde el planteamiento de 3 preguntas de investigación que permitieron explorar de forma sistemática el tema de investigación.

    - En la pregunta que hizo referencia a los temas de mayor interés para la comunidad científica en torno a la predicción de la calidad del agua, se estableció que el tema de mayor importancia es la construcción de modelos de predicción híbridos, principalmente para cubrir las desventajas de algunas técnicas y aportar mejoras con la incorporación de otras técnicas como refuerzo.

    - Respecto a las técnicas de IC que han sido utilizadas como parte del componente adaptativo de un mecanismo de predicción han sido el algoritmo de proyección y el algoritmo de mínimos cuadrados, sin embargo estas técnicas no son utilizadas en la adaptación de dicho mecanismo a diferentes usos del agua. No obstante existen técnicas que a pesar de no haber sido exploradas a fondo dentro de un

componente adaptativo, poseen características importantes como la capacidad de adaptarse a entornos cambiantes, tal es el caso de PSO y la Computación Evolutiva en especial los algoritmos genéticos.

- Experimentalmente fueron evaluadas las dos técnicas de IC que de acuerdo a la literatura obtenían mejores resultados de precisión en las predicciones (RNA y SVR). Las SVR presentaron un mejor desempeño respecto a las RNA utilizando un conjunto de datos perteneciente al uso piscícola de la zona de Alviso en el estado de California, USA. Para incrementar el desempeño en las predicciones realizadas por la SVR, fueron evaluadas varias configuraciones del método de kernel (Normalized Poli Kernel, Poli Kernel, PUK y RBF Kernel) y se determinó que la configuración con el kernel PUK obtiene resultados más precisos.

- En una segunda fase de evaluación se optimizó la técnica de predicción (SVR-PUK) mediante la Optimización por Nubes de Partículas (PSO) la cual hace parte de los paradigmas de Inteligencia de Enjambres. Esta técnica permitió que los valores predichos se aproximaran a los valores reales en diferentes usos del agua (disminución del porcentaje de error en las predicciones), ya que al utilizar otro conjunto de datos (uso recreacional en el lago Don Pedro, California, USA y uso para consumo humano en la estación Puente Juanchito del río Cauca), la precisión de las pre-dicciones decaía ostensiblemente.

- El mecanismo de predicción adaptativo presentó un comportamiento aceptable para series de tiempo que presentaban cambios drásticos en la distribución de sus valores; teniendo en cuenta que para el desarrollo del componente adaptativo se realizaron pruebas con varios algoritmos adaptativos, PSO permitió obtener mejores resultados sobre este tipo de series de tiempo.

6.2. Trabajos Futuros

Este estudio tuvo como principal objetivo proponer, desarrollar y evaluar un mecanismo de predicción adaptativo de la calidad del agua por medio de técnicas de Inteligencia Computacional. El principal enfoque de este mecanismo es su capacidad para adaptar sus predicciones a diferentes conjuntos de datos pertenecientes a distintos usos del agua sin que la precisión se vea afectada drásticamente.

Teniendo en cuenta lo anterior, y con el objetivo de complementar el mecanismo de predicción, se proponen los siguientes trabajos futuros:

- Detección automática del uso del agua al que corresponda un determinado conjunto de datos, esto es, implementar un módulo que permita identificar a qué uso del agua corresponde un conjunto de datos de acuerdo a sus características con el fin de mejorar la capacidad de adaptación del mecanismo.

- Detección de las causas de contaminación del agua a partir de un análisis del contexto con el objetivo de brindar recomendaciones al personal encargado de la gestión del recurso hídrico.

- Integración de datos de clima y suelo con los datos de calidad de agua; dada la interrelación fuerte que existe entre estos tres elementos, se puede generar no solo un mecanismo de predicción más preciso, si no también definir nuevos modelos computacionales que faciliten la toma de decisiones en el sector agrícola.

Referencias

[1] P. J. and J. S. Claudia Pahl-Wostl y C. Pahl-Wostl Paul Jeffrey, and Jan Sendzimir., *Adaptive and integrated management of water resources*. publisherNameCambridge University Press, 2011.

[2] N. Diersing, «Water Quality: Frequently Asked Questions». Florida Brooks National Marine Sanctuary, Key West, FL, 2009.

[3] Comunidad Autónoma de Extremadura, *Agentes Forestales de Extremadura. Legislacion Basica Ebook*. MAD-Eduforma, 2003.

[4] C. Carbó, *Genética, patología, higiene y residuos animales*, vol. 4. Mundi-Prensa Libros, 1995.

[5] Isis Beleño, «El 50% del agua en Colombia es de mala calidad», *UN Periódico*, Universidad Nacional de Colombia, Bogotá D.C., 12-feb-2011.

[6] IDEAM, «Calidad del Agua Superficial en Colombia», en *Estudio Nacional del Agua*, 2010, pp. 231-277.

[7] E. Kumar, *Artificial Intelligence*. I.K. International Publishing House Pvt. Limited, 2008.

[8] C. E. Romero y J. Shan, «Development of an Artificial Neural Network-based Software for Prediction of Power Plant Canal Water Discharge Temperature», *Expert Syst Appl*, vol. 29, n.º 4, pp. 831–838, nov. 2005.

[9] S. Palani, S.-Y. Liong, y P. Tkalich, «An ANN application for water quality forecasting», *Mar. Pollut. Bull.*, vol. 56, n.º 9, pp. 1586-1597, sep. 2008.

[10] P. A. Aguilera, A. G. Frenich, J. A. Torres, H. Castro, J. L. M. Vidal, y M. Canton, «Application of the kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality», *Water Res.*, vol. 35, n.º 17, pp. 4053-4062, dic. 2001.

[11] D. Ömer Faruk, «A hybrid neural network and ARIMA model for water quality time series prediction», *Eng. Appl. Artif. Intell.*, vol. 23, n.º 4, pp. 586-594, jun. 2010.

[12] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, y J. A. Moncada-Herrera, «A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile», *Atmos. Environ.*, vol. 42, n.º 35, pp. 8331-8340, nov. 2008.

[13] J. I. P. Di Blasi, J. Martínez Torres, P. J. García Nieto, J. R. Alonso Fernández, C. Díaz Muñiz, y J. Taboada, «Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain)», *Ecol. Eng.*, vol. 60, pp. 60-66, nov. 2013.

[14] J. Kibena, I. Nhapi, y W. Gumindoga, «Assessing the relationship between water quality parameters and changes in landuse patterns in the Upper Manyame River, Zimbabwe», *Phys. Chem. Earth Parts ABC*.

[15] Q. Chen, W. Wu, K. Blanckaert, J. Ma, y G. Huang, «Optimization of water quality monitoring network in a large river by combining measurements, a numerical model and matter-element analyses», *J. Environ. Manage.*, vol. 110, pp. 116-124, nov. 2012.

[16] M. A. T. Koçer y H. Sevgili, «Parameters selection for water quality index in the assessment of the environmental impacts of land-based trout farms», *Ecol. Indic.*, vol. 36, pp. 672-681, ene. 2014.

[17] G. Tan, J. Yan, C. Gao, y S. Yang, «Prediction of water quality time series data based on least squares support vector machine», *Procedia Eng.*, vol. 31, pp. 1194-1199, 2012.

[18] S. Liu, L. Xu, D. Li, Q. Li, Y. Jiang, H. Tai, y L. Zeng, «Prediction of dissolved oxygen content in river crab culture based on least squares support vector regression optimized by improved particle swarm optimization», *Comput. Electron. Agric.*, vol. 95, pp. 82-91, jul. 2013.

[19] FAO, «Usos del agua», *Sitio web AQUASTAT*, 2015. [En línea]. Disponible en: http://www.fao.org/nr/water/aquastat/water_use/indexesp.stm. [Accedido: 16-abr-2015].
[20] J. Holland, «Sistemas adaptativos complejos», 1996.
[21] W. J. González, *La predicción científica: concepciones filosófico-metodológicas desde H. Reichenbach a N. Rescher*. Montesinos, 2010.
[22] N. Rescher, *Predicting the Future: An Introduction to the Theory of Forecasting*. SUNY Press, 1998.
[23] A. P. Engelbrecht, *Computational Intelligence: An Introduction*. Wiley, 2007.
[24] IEEE, *IEEE Connections, the Newsletter of the IEEE Computational Intelligence Society*, vol. 1. 2003.
[25] J. M. Pérez, *Inteligencia computacional inspirada en la vida*. Servicio de Publicaciones e Intercambio Científico de la Universidad de Málaga, 2010.
[26] A. J. Smola y B. Schölkopf, «A tutorial on support vector regression», *Stat. Comput.*, vol. 14, n.º 3, pp. 199-222, ago. 2004.
[27] J. L. Alba, «Máquinas de Vectores Soporte (SVM)». Curso de Doctorado: Decisión, Estimación y Clasificación., 2013.
[28] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, 1975.
[29] LUREG, «Lancaster University Renewable Energy Group - Wave Energy», *Optimisation of Collector Form and Response*. [En línea]. Disponible en: http://www.engineering.lancs.ac.uk/lureg/group_research/wave_energy_research/Collector_Shape_Design.php. [Accedido: 10-jul-2015].
[30] R. Poli, J. Kennedy, y T. Blackwell, «Particle swarm optimization», *Swarm Intell.*, vol. 1, n.º 1, pp. 33-57, ago. 2007.
[31] D. Varadi, «Social Learning Algorithms: Particle Swarm Optimization (PSO)», *CSSA*, 2013. .
[32] K. Petersen, R. Feldt, S. Mujtaba, y M. Mattsson, «Systematic Mapping Studies in Software Engineering», en *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, Swinton, UK, UK, 2008, pp. 68–77.
[33] B. Kitchenham y S. Charters, «Guidelines for performing Systematic Literature Reviews in Software Engineering», Keele University and Durham University Joint Report, UK, EBSE 2007-001, 2007.
[34] J. Grbović y S. Džeroski, «Knowledge discovery in a water quality database», *Proc 1st Intl Conf Knowl. Discov. Data Min. KDD95 AAAI Press Menlo Park CA 1995*.
[35] S. Džeroski, D. Demšar, y J. Grbović, «Predicting Chemical Parameters of River Water Quality from Bioindicator Data», *Appl. Intell.*, vol. 13, n.º 1, pp. 7–17, jul. 2000.
[36] L. Breiman, *Classification and regression trees*. Chapman & Hall, 1984.
[37] H. Blockeel, S. Dzeroski, y J. Grbovic, *Simultaneous prediction of multiple chemical parameters of river water quality with TILDE*. 1999.
[38] T. G. Dietterich, «Ensemble Methods in Machine Learning», en *Multiple Classifier Systems*, Springer Berlin Heidelberg, 2000, pp. 1-15.
[39] I. Partalas, G. Tsoumakas, E. V. Hatzikos, y I. Vlahavas, «Greedy regression ensemble selection: Theory and an application to water quality prediction», *Inf. Sci.*, vol. 178, n.º 20, pp. 3867-3879, oct. 2008.
[40] K. Gurney, *An Introduction to Neural Networks*. Taylor & Francis, 2003.
[41] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.
[42] E. Cox, *The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems*. San Diego, CA, USA: Academic Press Professional, Inc., 1994.

[43] E. Hatzikos, J. Hätönen, N. Bassiliades, I. Vlahavas, y E. Fournou, «Applying adaptive prediction to sea-water quality measurements», *Expert Syst. Appl.*, vol. 36, n.º 3, Part 2, pp. 6773-6779, abr. 2009.
[44] L.-M. L. He y Z.-L. He, «Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA», *Water Res.*, vol. 42, n.º 10-11, pp. 2563-2573, may 2008.
[45] D. West y S. Dellana, «An empirical analysis of neural network memory structures for basin water quality forecasting», *Int. J. Forecast.*, vol. 27, n.º 3, pp. 777-803, jul. 2011.
[46] D. Antanasijević, V. Pocajt, A. Perić-Grujić, y M. Ristić, «Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis», *J. Hydrol.*, vol. 519, Part B, pp. 1895-1907, nov. 2014.
[47] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, y M. F. Ramli, «Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors», *Mar. Pollut. Bull.*, vol. 64, n.º 11, pp. 2409-2420, nov. 2012.
[48] J. Gutiérrez, W. Riss, y R. Ospina, «Bioindicación de la calidad del agua con macroinvertebrados acuáticos en la sabana de Bogotá, utilizando redes neuronales artificiales», *Caldasia*, vol. 26, n.º 1, pp. 151-160, 2004.
[49] I. García, J. G. Rodríguez, F. López, y Y. M. Tenorio, «Transporte de contaminantes en aguas subterráneas mediante redes neuronales artificiales», *Inf. Tecnológica*, vol. 21, n.º 5, pp. 79-86, 2010.
[50] A. I. Saint-Gerons y J. M. Adrados, «Desarrollo de una Red Neuronal para estimar el Oxígeno Disuelto en el agua a partir de instrumentación de E.D.A.R.», Navarra, España, 2004.
[51] A. Ogata y R. B. Banks, «A solution of the differential equation of longitudinal dispersion in porous media», United States Government Printing Office, Washington, DC, USA, Geological Survey, 1961.
[52] G. J. Pelletier, S. C. Chapra, y H. Tao, «QUAL2Kw – A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration», *Environ. Model. Softw.*, vol. 21, n.º 3, pp. 419-425, mar. 2006.
[53] S. Liu, D. Butler, R. Brazier, L. Heathwaite, y S.-T. Khu, «Using genetic algorithms to calibrate a water quality model», *Sci. Total Environ.*, vol. 374, n.º 2-3, pp. 260-272, mar. 2007.
[54] Y. Huang y L. Liu, «Multiobjective Water Quality Model Calibration Using a Hybrid Genetic Algorithm and Neural Network–Based Approach», *J. Environ. Eng.*, vol. 136, n.º 10, pp. 1020-1031, 2010.
[55] K. Chau, «A Split-Step PSO Algorithm in Prediction of Water Quality Pollution», en *Advances in Neural Networks – ISNN 2005*, J. Wang, X.-F. Liao, y Z. Yi, Eds. Springer Berlin Heidelberg, 2005, pp. 1034-1039.
[56] A. M. Baltar y D. G. Fontane, «A generalized multiobjective particle swarm optimization solver for spreadsheet models: application to water quality», *Proc. Twenty Sixth Annu. Am. Geophys. Union Hydrol. Days*, pp. 20-22, 2006.
[57] A. Afshar, H. Kazemi, y M. Saadatpour, «Particle Swarm Optimization for Automatic Calibration of Large Scale Water Quality Model (CE-QUAL-W2): Application to Karkheh Reservoir, Iran», *Water Resour. Manag.*, vol. 25, n.º 10, pp. 2613-2632, ago. 2011.
[58] J. Zhangzan, X. Gang, C. Jiujun, y G. Fei, «Anomaly detection of water quality based on visual perception and V-detector», *Inf. Control*, vol. 1, p. 026, 2011.
[59] J. C. Pete Chapman, «CRISP-DM 1.0: Step-by-Step Data Mining Guide», 1999.

[60] J. Guajardo, R. Weber, y J. Miranda, «A Forecasting Methodology Using Support Vector Regression and Dynamic Feature Selection», *J. Inf. Knowl. Manag.*, vol. 05, n.º 04, pp. 329-335, dic. 2006.
[61] I. Guyon y A. Elisseeff, «An Introduction to Variable and Feature Selection», *J Mach Learn Res*, vol. 3, pp. 1157–1182, mar. 2003.
[62] J. Hamon, «Combinatorial optimization for variable selection in high dimensional regression: Application in animal genetic», Université des Sciences et Technologie de Lille - Lille I, 2013.
[63] B. Üstün, W. J. Melssen, y L. M. C. Buydens, «Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel», *Chemom. Intell. Lab. Syst.*, vol. 81, n.º 1, pp. 29-40, mar. 2006.
[64] R. J. Hyndman y G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2014.
[65] Waikato University, «Machine Learning Project at the University of Waikato in New Zealand». [En línea]. Disponible en: http://www.cs.waikato.ac.nz/ml/index.html. [Accedido: 01-dic-2015].
[66] M. Sliger y S. Broderick, *Agile Methodologies, in The Software Project Manager's Bridge to Agility*, K. Gettman, Editor. United States: Addison-Wesley Professional, 2008.
[67] U. S. Geological Survey, «U. S. Geological Survey», *U. S. Geological Survey*, 07-ene-2014. [En línea]. Disponible en: http://www.usgs.gov/. [Accedido: 14-abr-2015].
[68] CVC, «Segunda campaña de muestreo con propositos de calibracion del modelo de calidad del agua del rio cauca», *Corporacion Auton. Reg. Val. Cauca Caracterizacion Model. Mat. Río Cauca - PMC- Fase II Conv. Interadministrativo 0168 Noviembre 27 2002*, vol. XV, 2005.
[69] W. Thoe, S. H. C. Wong, K. W. Choi, y J. H. W. Lee, «Daily prediction of marine beach water quality in Hong Kong», *J. Hydro-Environ. Res.*, vol. 6, n.º 3, pp. 164-180, sep. 2012.
[70] S. J. Langan, A. J. Wade, R. Smart, A. C. Edwards, C. Soulsby, M. F. Billett, H. P. Jarvie, M. S. Cresser, R. Owen, y R. C. Ferrier, «The prediction and management of water quality in a relatively unpolluted major Scottish catchment: current issues and experimental approaches», *Sci. Total Environ.*, vol. 194-195, pp. 419-435, feb. 1997.
[71] R. J. Hyndman y A. B. Koehler, «Another look at measures of forecast accuracy», *Int. J. Forecast.*, vol. 22, n.º 4, pp. 679-688, oct. 2006.
[72] J. Armstrong y F. Collopy, «Error measures for generalizing about forecasting methods: Empirical comparisons», *Int. J. Forecast.*, vol. 8, n.º 1, pp. 69-80, 1992.
[73] N. Hansen, A. Auger, R. Ros, S. Finck, y P. Pošík, «Comparing Results of 31 Algorithms from the Black-box Optimization Benchmarking BBOB-2009», en *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, New York, NY, USA, 2010, pp. 1689–1696.
[74] W. A. Higashino, M. A. M. Capretz, y M. B. F. D. Toledo, «Evaluation of Particle Swarm Optimization Applied to Grid Scheduling», en *Proceedings of the 2014 IEEE 23rd International WETICE Conference*, Washington, DC, USA, 2014, pp. 173–178.
[75] Fondriest Environmental Inc, «Conductivity, Salinity & Total Dissolved Solids», *Environmental Measurement Systems*, 2015. [En línea]. Disponible en: http://www.fondriest.com/environmental-measurements/parameters/water-quality/conductivity-salinity-tds/. [Accedido: 29-ene-2016].