ANDREA CAROLINA AGUILAR AGUILAR

ANÁLISIS COMPARATIVO PARA LA APLICACIÓN DE UN MODELO PREDICTIVO EN LA ESTIMACIÓN DE LA CALIDAD DE AGUA

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Maestría en Electrónica y Telecomunicaciones

Popayán
2021

ANDREA CAROLINA AGUILAR AGUILAR

ANÁLISIS COMPARATIVO PARA LA APLICACIÓN DE UN MODELO PREDICTIVO EN LA ESTIMACIÓN DE LA CALIDAD DE AGUA

Trabajo de Grado presentado a la Facultad de
Ingeniería Electrónica y Telecomunicaciones de
La Universidad del Cauca para la obtención del
Título de

Magister en:
ELECTRÓNICA Y TELECOMUNICACIONES

Director:
Mg. Francisco Franco Obando

Co director:
Dr. Pablo Emilio Jojoa Gómez

Popayán
2021

Resumen

El agua es un recurso natural y una fuente vital para el ser humano y otras especies que habitan el planeta. Diversos esfuerzos se suman para garantizar su protección y consumo eficiente haciendo frente a la creciente demanda y contaminación de los últimos años. El agua puede pasar por diferentes procesos antes de llegar a su uso final y en muchas ocasiones se requiere almacenar este líquido en depósitos para su tratamiento, transporte o recolección, por lo cual es necesario conocer y controlar la calidad y el tiempo en el cual se mantienen las características óptimas para su consumo. La calidad del agua se analiza mediante la medición y monitoreo de parámetros fisicoquímicos o biológicos.

Los modelos predictivos permiten estimar comportamientos futuros a partir de información previa extraída de un fenómeno, su aplicación en esta área se centra en estimar la calidad del agua utilizando la información proveniente de los parámetros. El proyecto “análisis comparativo para la aplicación de un modelo predictivo en la estimación de la calidad de agua” tiene como objetivo estimar la calidad del agua en almacenamiento mediante la comparación de modelos de aprendizaje automático utilizando para ello los datos medidos de parámetros fisicoquímicos. Para el experimento se almacenaron diferentes muestras de agua lluvia, de acueducto y envasada sobre las cuales se midieron los parámetros de pH, solidos disueltos totales y temperatura. Se utilizaron tres técnicas para la construcción de los modelos de regresión: arboles de regresión, regresión lineal, y máquinas de vectores de soporte. Luego se dividieron las muestras tanto para la construcción de los modelos como para etapa de validación. Los resultados obtenidos muestran que es posible estimar la calidad del agua en almacenamiento representada en un indicador ICA (escala de o a 100) utilizando como entradas del modelo los parámetros pH, solidos disueltos totales y temperatura, con buenos resultados en la etapa de validación para las métricas MSE, RMSE y R^2.

Palabras Clave: Calidad de agua, Regresión lineal, Arboles de regresión, Máquinas de vectores de soporte, Índice de calidad.

Abstract

Water is a natural resource and a vital source for humans and other species that inhabit the planet. Various efforts are being added to guarantee its protection and efficient consumption, facing the increasing demand and pollution in recent years. The water can go through different processes before reaching its final use and in many cases it is required to store this liquid in tanks for its treatment, transport or collection, for which it is necessary to know and control the quality and the time in which the optimal characteristics for consumption are maintained. Water quality is analyzed by measuring and monitoring physicochemical or biological parameters.

Predictive models allow estimating future behavior from previous information extracted from a phenomenon, its application in this area focuses on estimating water quality using the information from the parameters. The project "Comparative analysis for the application of a predictive model in estimating water quality" aims to estimate the quality of water in storage by comparing machine learning models using the measured data of physicochemical parameters. For the experiment, different samples of rainwater, aqueduct and bottled water were stored. The parameters of pH, total dissolved solids and temperature were measured in the samples. Three techniques were used for the construction of the regression models: regression trees, linear regression, and support vector machines. Then the samples were divided both for the construction of the models and for the validation stage. The results obtained show that it is possible to estimate the quality of the water in storage represented in an index (scale of or to 100) using the parameters pH, total dissolved solids and temperature as model inputs, with good results in the validation stage for the MSE, RMSE and R^2 metrics.

Keywords: Water quality, Linear regression, Regression trees, Support vector machines, Quality index.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Abreviaturas

OMS: Organización mundial de la salud
ICA: Índice de calidad de agua
ICO: Índice de contaminación del agua
MAPE: Media de la desviación porcentual absoluta
MAE: Error absoluto medio
MSE: Error cuadrático medio
RMSE: Raíz del error cuadrático medio
R^2: Coeficiente de determinación
ANFIS: Red neuronal de inferencia difusa
RN: Redes neuronales
MVS: Máquinas de vectores de soporte
RL: Regresión lineal
AR: Árboles de regresión
T: Temperatura
pH: Potencial de hidrogeno
TDS: Solidos disueltos totales
ST: Solidos totales
TSS: Solidos suspendidos totales
DO: Oxígeno disuelto
DQO: Demanda química de oxigeno
DBO: Demanda biológica de oxigeno
TN: Nitrógeno total
TP: Fosforo total
CO_2: Dióxido de carbono
CE: Conductividad eléctrica
AL: Agua lluvia
AA: Agua de acueducto
AE: Agua envasada

1. Introducción

De acuerdo a las directrices de la OMS, se deben apoyar el desarrollo y la ejecución de estrategias de gestión de riesgos que garanticen la inocuidad del agua de consumo humano(WHO, 2019), sin embargo las crecientes problemáticas medioambientales, factores socioeconómicos y demográficos contaminan y afectan la calidad del agua potable, trayendo como consecuencia problemas en la salud de las personas y reduciendo la preservación de los ecosistemas naturales (Nazemi y Madani, 2018). El término calidad de agua se asocia a un conjunto de parámetros físicos, químicos y biológicos cuyas mediciones proporcionan la información sobre el estado en que se encuentra un cuerpo de agua(Gómez-Gutiérrez *et al.*, 2016), estos valores se comparan con los rangos de referencia establecidos para cada parámetro en normas ambientales, definidas por entidades gubernamentales y organizaciones internacionales.

Las limitaciones relacionadas con la disponibilidad y suministro de agua potable, incentivan la creación de mecanismos para aprovechar eficientemente el agua como el almacenamiento y/o reutilización. En cuanto al suministro, los depósitos de agua tratada pueden aliviar los problemas de abastecimiento en lugares de poco acceso al recurso hídrico(Soltani, Kerachian y Shirangi, 2010). Estos depósitos deben contar con la capacidad de proveer y almacenar un volumen adecuado para solventar situaciones de emergencia. El correcto almacenamiento de agua requiere cumplir una serie de condiciones, tanto en la forma y tipo de depósito que se utilice de acuerdo al uso al que se destine el agua.

Uno de los problemas para la gestión del agua en almacenamiento, se relaciona con la conservación de las características apropiadas para su consumo y el tiempo para el cual, es posible conservar los valores deseados de calidad. Las mediciones que se realizan en el monitoreo de diferentes parámetros pueden permitir estimar el tiempo de conservación del agua en almacenamiento(Brentan *et al.*, 2017).

El concepto de predicción se enfoca en la extracción de información de datos reales previos de un proceso a fin de predecir patrones de comportamiento o tendencias de posibles eventos futuros(Espino, 2017), su aplicación se da en diferentes campos de

la ciencia y en fenómenos naturales, no obstante, las tareas de predicción pueden llegar a ser complejas debido al número de variables, fuerte interacciones entre ellas o dinámica desconocida del fenómeno que se estudia.

El proceso que se realiza para la estimación de la calidad del agua, consiste en la aplicación de técnicas predictivas que infieren una variable de salida en función de una o más variables de entrada. Las variables de entrada para el estudio de calidad de agua serán los datos de los parámetros fisicoquímicos o biológicos medidos previamente y la variable de salida puede ser un parámetro o indicador; el cual, brindara información sobre el estado del agua (Cruz, Alonso y Franco, 2017). La inserción de la tecnología y la inteligencia artificial ha permitido desarrollar algoritmos o técnicas de predicción que permiten estimar las condiciones de calidad de un cuerpo de agua, en menor tiempo y con mayor precisión.

En este proyecto se busca realiza la estimación temporal de la calidad en tres tipos de agua almacenada a fin de poder cuantificar el tiempo en el que es posible conservar las propiedades deseadas para consumo humano, a partir de la información de parámetros fisicoquímicos y algoritmos de aprendizaje automático.

En el segundo capítulo se realiza una conceptualización de la problemática de la calidad de agua, aprendizaje automático y las técnicas que se utilizan en la estimación de parámetros e índices de calidad. En el tercer capítulo se expone el experimento y la toma de datos para la construcción de los modelos predictivos. La validación de los resultados obtenidos y las estimaciones temporales en cada caso se muestran en el cuarto capítulo y finalmente se presentan las conclusiones y referencias.

Capítulo 2

Calidad de Agua y Aprendizaje Automático

En este capítulo se presenta un primer acercamiento al campo del aprendizaje automático y su aplicación en la calidad de agua, se resalta la importancia del control de la calidad del agua en almacenamiento, por sus diferentes usos y su contribución a la gestión eficiente del agua. Los trabajos investigativos relacionados con la estimación de calidad de agua, se utilizan como referentes en la selección de técnicas a implementar para el desarrollo de este proyecto.

2.1 Depósitos de Agua Potable

Los depósitos de agua se han convertido en un elemento importante dentro de la gestión y aprovechamiento del recurso hídrico, su abastecimiento se puede dar no solo para el uso doméstico, sino que también es aprovechado por sectores como la agricultura y la industria. La disponibilidad de estos reservorios garantiza el abastecimiento en lugares sin suministro de agua potable y en épocas críticas de verano. En las redes de acueducto, durante las etapas de tratamiento y desinfección previa a su distribución, también se utilizan depósitos de agua, no obstante, los tiempos de almacenamiento pueden variar por lo que suelen estar expuestos a diversos tipos de riesgos por contaminación y reacciones internas (Wu *et al.*, 2014). Otro tipo de almacenamiento en periodos más largos se da en el agua envasada, la cual además de cumplir con los parámetros para consumo humano, deber ser tratada para conservar sus propiedades por un rango de tiempo más amplio.

Otra alternativa viable y sustentable que ha ganado gran importancia como estrategia de beneficio ambiental es la reutilización del agua lluvia, la cual necesita de procesos de captación y purificación que involucra inevitablemente el almacenamiento y tratamiento del líquido. Su limitación está en que no es una fuente abundante en todas las regiones y constante en todas las épocas del año (Duran *et al.*, 2010).

Las actividades de conservación y desinfección de los sistemas de almacenamiento de agua potable deben ser controladas a fin de garantizar los valores óptimos de calidad para consumo humano, sin embargo, no se establece cuanto tiempo pueden permanecer los valores deseados de calidad en un depósito de agua y más si se tiene en cuenta el posible contacto con sustancias químicas u otros elementos utilizados en el mantenimiento y limpieza de depósitos (Janet Gil *et al.*, 2012).

2.2 Evaluación de la Calidad de Agua

La calidad de agua puede clasificarse de acuerdo al uso final al que se destine, de manera más práctica, los análisis de calidad de agua se basan en las mediciones de parámetros sobre fuentes hídricas (ríos, lagos, aguas subterráneas, etc.) por organismos medioambientales de control y se nutren con la información proporcionada por diferentes sectores en una recolección de datos sistémica, que pueden utilizar los indicadores para su representación(Gorde y Jadhav, 2013). El agua que se destina para el consumo humano debe cumplir con los criterios admisibles reglamentados para cada parámetro, en el caso de los índices, por ejemplo, muestran el estado del agua en un rango de 0 a 100; los índices de calidad (ICA) y contaminación del agua (ICO) son los más comunes. La Figura 2.1 muestra los escenarios de evaluación de la calidad de agua(Castro *et al.*, 2015).

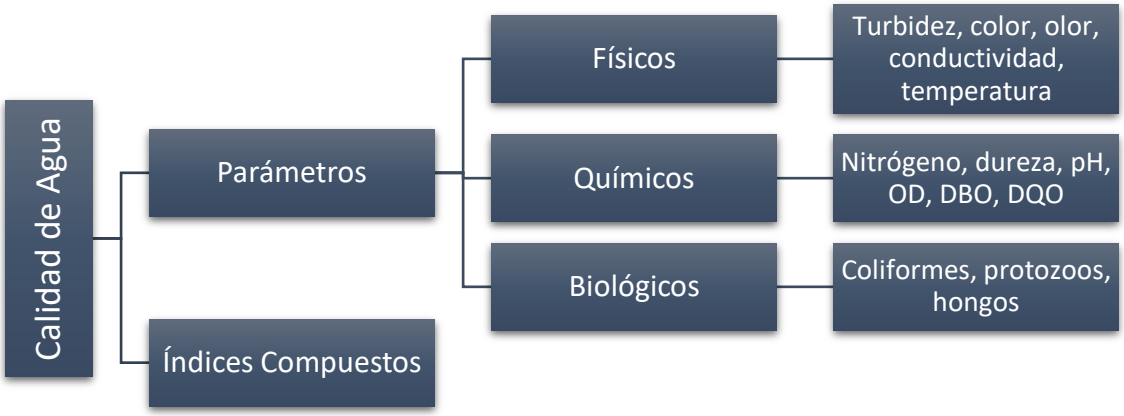


Figura 2.1. Evaluación para calidad de agua.

En cuanto a la normatividad establecida para el control de la calidad del agua para consumo humano, en Colombia la resolución (MPS, 2007) establece los límites máximos que no deben superar los parámetros fisicoquímicos y poder así, ser aptas para consumo humano, en la Tabla 2.1 se relacionan los parámetros de mayor relevancia en el monitoreo de calidad de agua según la norma (MPS, 2007).

Tabla 2.1 Valores de máximos aceptables de parámetros físicos, químicos y biológicos.

Características Físicas	**Unidad**	**Valor Máximo Aceptable**
Color aparente	UPC	15
Olor y sabor	A/NA	A / Aceptable
Turbiedad	UNT	2
Conductividad	microsiemens/cm	1000
Solidos disueltos totales	ppm	640
Temperatura	°C	< 25
Características Químicas	**Unidad**	**Valor Máximo Aceptable**
Cloro residual	mg/l	0.3 a 2.0
pH	pH	6.5 a 9.0
Carbono Orgánico Total	COT	5.0
Nitritos	NO_2^-	0.1
Nitratos	NO_3^-	10
Alcalinidad Total	$CaCo_3$	200
Cloruros	Cl^-	250
Aluminios	Al^{3+}	0.2
Dureza Total	$CaCo_3$	300
Hierro	Fe	0.3
Manganeso	Mn	0.1
Sulfatos	SO_4^{2-}	250
Características Microbiológicas	**Unidad**	**Valor Máximo Aceptable**
Coliformes Totales	UFC/cm^3	0
Escherichia coli	UFC/cm^3	0
Mesófilos	UFC/ml	≤100

2.3 Aprendizaje Automático

Las acciones de monitoreo y control son necesarias para minimizar los riesgos e impactos negativos que se puedan presentar en los sistemas naturales y el bienestar de las personas, no obstante, cuando se tienen sistemas complejos de alta variabilidad o contaminantes de alto riesgo, es necesario poder contar con otro tipo de acciones que permitan anticipar un riesgo potencial y tomar decisiones correctivas con suficiente rango de tiempo.

El procesamiento y análisis de datos que se efectúa en un aprendizaje automático, se lleva a cabo con una alta velocidad y con una mínima intervención humana en la toma de decisiones. Dependiendo de los requerimientos del problema es posible escoger entre distintos métodos y técnicas disponibles, capaces de seguir operando con alto rendimiento incluso cuando se adicionan más valores durante su ejecución (Günnemann, 2017). La Figura 2.2 muestra la clasificación de algunas técnicas de aprendizaje automático bajo los escenarios supervisado y no supervisado en las tareas de regresión, clasificación y agrupamiento.

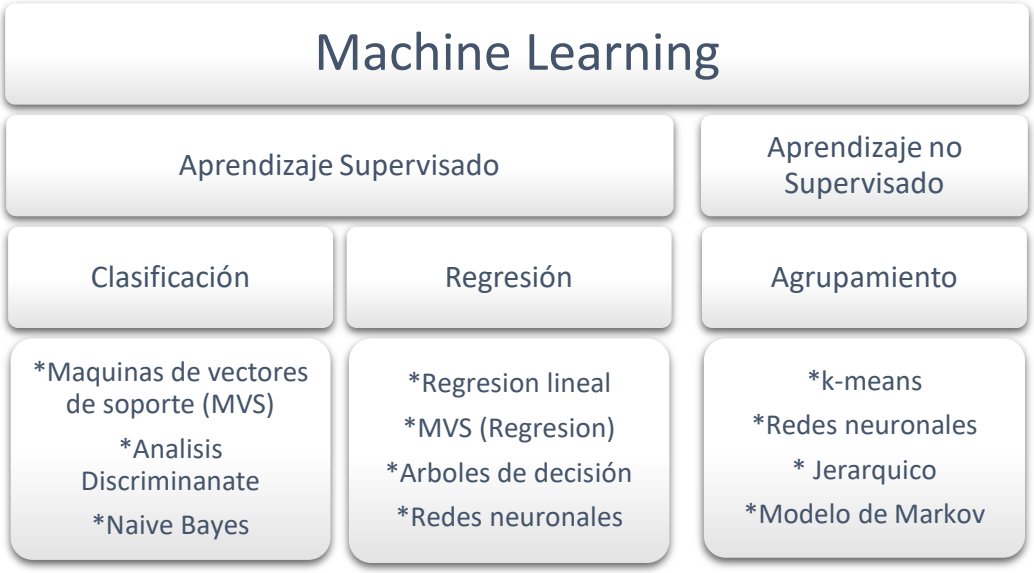


Figura 2.2. Machine Learning – Clasificación.

El aprendizaje supervisado es uno de los más comunes en este campo, se utiliza generalmente cuando se conocen los parámetros de la salida deseada, entre las tareas más frecuentes esta la regresión y clasificación. Los algoritmos no supervisados ajustan su modelo utilizando solo la información de entrada y no están predispuestos operativamente por los valores de salida esperados, permitiendo identificar o agrupar estructuras de un conjunto de datos. Dentro del proceso de aprendizaje para el caso de los algoritmos supervisados, se pueden identificar dos fases en las cuales es necesario dividir el total de datos en dos conjuntos; pruebas y entrenamiento o mejor conocidos como *testing and training*(Harrington, 2012).

Durante la fase de entrenamiento se construye el modelo utilizando uno de los dos conjuntos de datos a fin de supervisar la variable a estimar, de esta manera el modelo aprende sobre las posibles causas que influyen en su comportamiento. En la fase de pruebas se verifica la validez del modelo sobre el otro conjunto, se calcula el error entre las predicciones del modelo y los valores reales. La fase de pruebas también permite evitar el sobre ajuste, que representa un comportamiento muy bueno a los datos para los que se conoce el resultado esperado, pero bajo rendimiento en nuevas estimaciones. Otra estrategia utilizada para evitar el sobre ajuste es la validación

cruzada, en la cual se divide el conjunto de entrenamiento en k subconjuntos, una vez seleccionado un subconjunto k como conjunto de prueba, los datos restantes se utiliza como datos de entrenamiento; repitiendo el proceso para k iteraciones (Sotiropoulos y Tsihrintzis, 2016). Para cuantificar la precisión entre los valores reales y las estimaciones, se utilizan algunas medidas de exactitud como: Media de la desviación porcentual absoluta **(**MAPE), Error absoluto medio (MAE), Error cuadrático medio (MSE), Raíz del error cuadrático medio (RMSE) y Coeficiente de determinación (R^2).

2.3.1 Evaluación del desempeño

Las medidas de exactitud se utilizan para evaluar la precisión en las estimaciones de un modelo predictivo, en donde $\boldsymbol{y}$ es el valor real, $\boldsymbol{y'}$ el estimado y $\boldsymbol{n}$ el número de observaciones, algunas de ellas son:

- Media de la desviación porcentual absoluta (MAPE): mide en términos porcentuales el error absoluto, muy efectivo al momento de identificar diferencias entre modelos, un 0% representa un ajuste perfecto. Se calcula a partir de la ecuación (2.1).

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-y'}{y}\right| \tag{2.1}$$

- Error absoluto medio (MAE): mide el promedio de las medias absolutas entre los valores reales y los estimados. Es un valor lineal y no es muy sensible frente a valores atípicos, está dado por ecuación (2.2).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y'-y|}{n} \tag{2.2}$$

- Error cuadrático medio (MSE): mide el error cuadrado promedio entre el valor estimado y el valor real para cada punto, y su resultado no es negativo, ecuación (2.3), de más utilidad cuando se trata de grades errores puesto que un valor de MSE alto también puede representar un buen ajuste.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y'-y)^2 \tag{2.3}$$

- Raíz del error cuadrático medio (RMSE): para dos conjuntos de datos, el RMSE mide el tamaño del error, es la raíz cuadrada de la suma de errores entre un valor estimado y uno observado o real. Es eficiente al revelar diferencias muy notables y se da en términos de la variable analizada. Esta dada por ecuación (2.4).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'-y)^2} \tag{2.4}$$

- Coeficiente de determinación (R^2): evalúa la calidad del modelo al proporcionar información sobre qué tan bien el modelo se aproxima a los valores observados. Se obtiene de la ecuación (2.5), el numerador representa la suma de cuadrados de los residuos y el denominador corresponde a la suma total de cuadrados. R^2, se da entre 0 y 1 (o en porcentaje de 0 a 100%) donde 1 (100%) denota que las estimaciones de regresión se ajustan perfectamente a los datos.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y' - y)^2}{\sum_{i=1}^{n}\left(y - \frac{\sum_{i=1}^{n} y}{n}\right)^2} \tag{2.5}$$

2.3.2. Aprendizaje automático para calidad de agua

Hoy en día es posible extraer una gran cantidad de información valiosa sobre los fenómenos que ocurren, para el caso de los ecosistemas hídricos, las investigaciones relacionadas con la estimación de variables utilizando técnicas de aprendizaje automático aplicadas al análisis predictivo, se han incrementado en los últimos años, lo que ha permitido obtener avances importantes. Estas estrategias también benefician la captura de datos que en su mayoría se realizan de forma digital y por métodos manuales, facilitando el estudio de cuerpos de agua en lugares remotos. La Tabla 2.2 resume las características y técnicas de los trabajos seleccionados de bases de datos bibliográficas en los últimos cuatro años enfocados en la estimación de la calidad de agua para consumo humano, a partir de técnicas de aprendizaje automático sobre parámetros de calidad de agua y/o índices de calidad.

Tabla 2.2 Investigaciones recientes sobre estimación de calidad de agua.

ID	Proyecto	Técnica	Parámetro Estimado	Parámetros de Entrada	Validación
1.	Predicción del Índice de calidad de agua en la cuenca del rio Peak – Malasia. (Al-Musawi y Al-Rubaie, 2017)	Red Neuronal	ICA	OD, ST, pH, NH_3-NL, T, CE, Turbidez, D S, TS, NO_3, Cl, PO4, As, Zn, Ca, Fe, K, Mg, Na, OG, E-Coli, Coliformes, Cd, Cr, Pb	R^2 MSE
2.	Estimación del Índice de calidad de agua potable –agua subterránea – Bardaskan. (RadFard *et al.*, 2019)	Red neuronal de inferencia difusa (ANFIS)	ICA	DT, CaH, Turbidez, pH, T, TDS, CE, ALK, Mg, Ca, K, Na, Sulfato, Bicarbonato, Fluoruro, NO_3-, NO_2-, Cl-.	MAE R^2
3.	Predicción de parámetros de calidad para una represa – Cheongpyeong. (Seo, Yun y Choi, 2016)	Red neuronal	Temperatura, OD, pH, CE, TN, TP turbidez y clorofila	T, OD, pH, CE, TN, TP	RMSE R^2

4.	Estimación de parámetros sobre un rio en Irán. (Azad *et al.*, 2019)	ANFIS - Hibrido	CE, TDS, sodio, dureza carbonatos, dureza total	CE, TDS, SAR, CH, TH, pH, Nap, Cl, Carbonato, Sulfato, Mg y Ca	RMSE MAPE R^2
5.	Predicción de fosforo y nitrógeno en los drenajes de un lago en Manzala. (Kang, Gao y Xie, 2017)	ANFIS	Fosforo Nitrógeno	Caudal, pH SST, CE, TDS, T, OD, Turbidez	RMSE R^2
6.	Estimación de la calidad de agua en un embalse. (Chou, Ho y Hoang, 2018)	Red Neuronal MVS Árbol de regresión, Regresión lineal.	Índice de estado trófico de Carlson	T, DBO, SS, DQO, NH3, y variables categóricas, temporada y lugar	RMSE MAE MAPE R^2
7.	Predicción del índice de calidad del agua en el Rio Tigris – Bagdad. (Hussein Ewaid, Ali Abed y Kadhum, 2018)	Regresión lineal	ICA	Turbidez, CE DQO, dureza y pH	R^2
8.	Estimación de parámetros de calidad de agua – Rio Tireh. (Haghiabi, Nasrolahi y Parsaie, 2018)	Red neuronal Máquinas de vectores de soporte Red neuronal – Modificado	Ca, Cl, CE, HCO3, Mg, Na, So4, TDS, pH	Ca, Cl, EC, HCO_3, Mg, Na, So_4, TDS, pH	RMSE R^2
9.	Estimación del Oxígeno disuelto –ríos urbanos –China. (Zhu y Heddam, 2019)	Máquinas de aprendizaje extremo (ELM) Red neuronal perceptrón multicapa	Oxígeno Disuelto	Combinaciones de T, pH, DO, índice de permanganato, NH_3 - N, CE, DQO, TN y TP	RMSE MAE R^2
10.	Estimación de Índice de calidad de agua sobre un lago – China. (Wang, Zhang y Ding, 2017)	Máquinas de vectores de soporte- Hibrido	ICA	pH, HCO_3, TP, TN, DBO, NH_3, -N, Fe, Cu, Zn, fenol, DO, TDS, Cl, SO_4, Na, Ca, Mg, COD, PO_4, Cr	RMSE R^2
11.	Predicción de la demanda bioquímica de oxígeno- Argelia. (Khaled *et al.*, 2018)	ANFIS	DBO	TIN, COD, O_2, TDS, PO_4 Combinación	RMSE MAE R^2
12.	Estimación de CO2 para un reservorio. (Chen, Ye y Huang, 2018)	Red neuronal Modificado	CO2	Clorofila, T carbono orgánico disuelto, TP, CO_2.	RMSE MAE R^2
13.	Estimación de parámetros para evaluar aguas residuales- EE.UU. (Granata *et al.*, 2017)	MVS Arboles de regresión	TSS, TDS, DQO, DBO	TSS, TDS, DQO, DBO	RMSE R^2
14.	Predicción de Bio – indicadores sobre un rio –China. (Fan *et al.*, 2017)	MVS	Bio indicadores	EC, DO, BOD_5, COD NH_3-N, TP, Hidromorfologia	MSE R^2

15.	Estimación microbiana del agua lago Noruega. (Mohammed, Longva y Seidu, 2018)	Redes neuronales MVS	Coliformes	pH, T, CE, Turbidez, color, alcalinidad, coliformes	MSE

La Tabla 2.2 presenta los trabajos de estimación de parámetros e índices de calidad de agua sobre diferentes fuentes, así como las técnicas utilizadas y los parámetros de entrada sobre los cuales se han realizado la toma de datos generalmente en sitio. En (Al-Musawi y Al-Rubaie, 2017), exponen una estrategia para estimar en tiempo real el índice de calidad de agua sobre el rio Peak en Malasia. Como se ha mencionado, una de las ventajas de la estimación es facilitar el acceso y procesamiento de la información, para lo cual no se tienen en cuenta los parámetros de DBO (demanda biológica de oxígeno) y DQO (demanda química de oxígeno), ya que para ellas no es posible obtener un valor por medición directa. En cuanto a las técnicas, se comparan dos arquitecturas: una red neuronal simple y otra formada por múltiples redes neuronales con la cual se consigue mejorar los resultados de desempeño.

Otro ejemplo se da en (Seo, Yun y Choi, 2016), en donde se estima la temperatura, oxígeno disuelto, pH, conductividad, TN, TP, turbidez y clorofila en una represa utilizando una red neuronal. Se obtienen buenos resultados de RMSE y R^2 para siete de los ocho parámetros estimados. En (Azad *et al.*, 2019), se realiza un análisis de correlación para determinar el mejor conjunto de parámetros de entrada para el modelado. Se compararon los resultados entre las técnicas ANFIS y ANFIS hibrida; es decir una combinación con la optimización de enjambre de partículas y colonia de hormigas, evidenciando un mejor desempeño en esta última, ejemplos similares se muestran en (Zhu y Heddam, 2019) y (Khaled *et al.*, 2018).

En cuanto a los índices, formados por dos o más parámetros, en (RadFard *et al.*, 2019) y (Hussein Ewaid, Ali Abed y Kadhum, 2018) se estima el índice de calidad de agua sobre una fuente subterránea y un rio, a partir de diferentes parámetros de entrada aplicando las técnicas ANFIS y regresión lineal respectivamente. En el control de riesgos por contaminación es posible evaluar los drenajes y vertimientos a una fuente de agua. La estimación puede aplicarse tanto en índices de calidad, contaminación o en parámetros específicos que pueden ser importantes para un estudio o como referencia de control; la estimación del fosforo y el nitrógeno en un lago es un ejemplo de este tipo de análisis (Kang, Gao y Xie, 2017).

La calidad del agua también se ve afectada por factores externos que en algunos estudios se toman en consideración a fin de poder mejorar los resultados, estas variables se identifican como categóricas y puede estar relacionadas con la distribución geográfica, estaciones del año, hasta información socioeconómica del sector. En (Chou, Ho y Hoang, 2018), se evalúan diferentes técnicas de aprendizaje computacional como redes neuronales, máquinas de vectores de soporte, árboles de decisión y regresión lineal, para estimar un índice de calidad muy característico en embalses, en el estudio además se comparan los desempeños de diferentes software

de modelado. De forma similar en (Wang, Zhang y Ding, 2017), los datos hiperespectrales de teledetección contribuyen a controlar la calidad de los efluentes en la estimación del índice de calidad, en (Chen, Ye y Huang, 2018) la estimación del CO_2 se da a partir de parámetros y datos categóricos utilizando redes neuronales modificadas.

Estudios comparativos de técnicas de aprendizaje automático, también se consideran en el campo medioambiental, en algunas de ellas se pueden encontrar cambios de la estructura original como es el caso de las redes neuronales que se combinan con otras estrategias para potenciar sus resultados. En (Haghiabi, Nasrolahi y Parsaie, 2018), se realiza un estudio comparativo de redes neuronales, máquinas de vectores de soporte y redes neuronales hibridas, otra comparación se da en (Granata *et al.*, 2017),en donde se estiman parámetros para evaluar la calidad de agua residual de vertimientos en cuencas, al comparar el desempeño de las máquinas de vectores y los árboles de regresión.

La calidad de agua no es exclusiva para el consumo del ser humano, los ecosistemas acuáticos también requieren que el agua cumpla ciertas condiciones que garanticen su conservación, además de los parámetros fisicoquímicos y los bioindicadores pueden proporcionar información valiosa para controlar la calidad de agua dulce. En (Fan *et al.*, 2017), los indicadores se estiman a partir de parámetros fisicoquímicos e información biológica del cuerpo de agua aplicando la técnica de máquinas de vectores. De igual manera, se estima la calidad microbiana de un lago comparando dos técnicas de aprendizaje automático (Mohammed, Longva y Seidu, 2018).

Para analizar el desempeño de las diferentes técnicas, se compararon los resultados de validación en cada caso, la Tabla 2.3 resume los valores obtenidos.

Tabla 2.3 Comparación de características técnicas de predicción.

ID	Algoritmo de predicción	Parámetros de salida	Exactitud				
			RMSE	**MAPE (%)**	**MSE**	**MAE**	**R^2**
1.	Redes Neuronales	ICA – indica de calidad de agua	--	**--**	0.9090	--	0.9340
	Múltiples redes neuronales		--	**--**	0.1740	--	0.1156
2.	ANFIS	ICA	2.89	--	--	0.923	0.2808
3.	Redes neuronales	Temperatura	0.360	--			0.998
4.	ANFIS	CE	4.30	7.73	--	--	0.91
	ANFIS – Hibrido	CE	**3.50**	**4.69**	--	---	**0.97**
5.	ANFIS	Fosforo	0.023	--	--	--	0.94
		Nitrógeno	1.109	--	--	--	0.92
6.	Redes neuronales	Índice de estado trófico de Carlson	4.644	7.721	--	3.622	0.865
	Máquinas de vectores de soporte		5.035	8.090	--	3.814	0.840
	Arboles de regresión		5.080	8.534	--	3.991	0.835
	Regresión Lineal		5.115	8.351	--	3.936	0.835
7.	Regresión lineal	ICA	--	--	--	--	0.974
8.	Redes Neuronales	Calcio	0.295	--	---	---	0.84
	Máquinas de Vectores de Soporte		0.193	---	---	---	0.94
	Redes Neuronales - Modificado		**0.313**	---	---	---	**0.85**

9.	Extreme Machine Learning	OD	0.518	---	---	0.359	0.870
	MLPNN multilayer perceptron neural network		**0.365**	---	---	**0.262**	**0.937**
10.	MVS – Hibrido	ICA	165.91	---	----	---	0.92
11.	ANFIS	DBO	3.2991	---	----	2.3715	0.8906
12.	Red neuronal modificado	CO_2	418.48	--	--	295.34	0.61
13.	MVS	TSS	1049	---	---	---	0.97
	Arboles de regresión		3486	---	---	---	0.906
14.	MVS	Bio indicador	---	---	87.72	--	0.98
15.	Redes neuronales	Coliformes	---	---	84.57	---	---
	MVS		---	---	140.09	---	---

De acuerdo a la Tabla 2.3 en la evaluación del desempeño, las medidas de R^2 y RMSE son las más utilizadas, seguidas por MAE, MAPE y MSE. Los resultados de estimación de las técnicas de aprendizaje automático se encuentran con valores por encima de 0.61 y alcanzando valores de 0.998 para R^2, lo que muestra que es posible estimar parámetros o índices de calidad de agua con muy buena fiabilidad. Se observa también que los resultados de exactitud mejoran en los estudios comparativos en donde se contrasta la estructura original con una hibrida (Al-Musawi y Al-Rubaie, 2017), (Chou, Ho y Hoang, 2018) y (Haghiabi, Nasrolahi y Parsaie, 2018).

Las técnicas más utilizadas son: las redes neuronales (RN), máquinas de vectores de soporte (MVS), ANFIS, regresión lineal (RL) y árboles de regresión (AR). La regresión lineal permite crear un modelo que describe la relación entre una variable de respuesta basada en una o más variables predictoras. En los árboles de regresión, la salida del modelo se estima en base al aprendizaje de las reglas de decisión inferidas de las características de los datos. Las máquinas de vectores de soporte construyen un hiperplano que mejor represente el comportamiento de los datos. La técnica ANFIS, integra las redes neuronales y la lógica difusa, su sistema de inferencia responde a reglas difusas y es ideal para sistemas no lineales (Harrington, 2012).

Las herramientas software diseñadas para facilitar el procesamiento y análisis de datos, implementan algunas de las técnicas anteriormente mencionadas para realizar los procesos de estimación, para este caso se utilizará la herramienta de MATLAB® R2018a la cual dispone de un módulo para el entrenamiento, validación y ajuste de modelos predictivos. De acuerdo a los resultados del análisis bibliográfico, se seleccionan las técnicas de máquinas de vectores de soporte, regresión lineal y árboles de regresión para estimar la evolución temporal de parámetros de calidad de agua en almacenamiento para consumo humano.

Capítulo 3

Análisis Predictivo y Modelado

En este capítulo, se describe el proceso de construcción de los modelos predictivos, así como las técnicas de regresión lineal, máquinas de vectores de soporte y árboles de regresión discriminadas en el capítulo 2. Se expone el procedimiento experimental para la extracción de los datos y la obtención de los modelos a través de una herramienta software, en la que se procesa la información de los experimentos. Se finaliza entonces con los resultados predictivos obtenidos para cada una de las técnicas.

3.1. Análisis Predictivo

El análisis predictivo agrupa diferentes técnicas estadísticas y de aprendizaje automático, cuya finalidad se centra en extraer las características de un fenómeno o evento para crear un modelo capaz de predecir tendencias y patrones probabilísticos de comportamiento futuro. La Figura 3.1 muestra de una manera general las etapas de proceso predictivo, extracción de características, construcción del modelo y las ponderaciones de probabilidad sobre el fenómeno que se analiza.

Figura 3.1. Fases modelado predictivo.

Existe una gran variedad de técnicas o algoritmos aplicables al modelado predictivo y su evolución ha ido acompañada de los beneficios de las herramientas software para hacer el proceso mucho más rápido y eficiente. Para el caso estudio del presente trabajo las técnicas aplicables a la estimación de calidad de agua seleccionadas de acuerdo a la información presentada en el capítulo dos, son: regresión lineal, árboles

de regresión y máquinas de vectores de soporte, las cuales se describen a continuación.

3.1.1. Regresión lineal

La regresión lineal (RL) construye una ecuación que permite describir la relación entre una variable dependiente y (variable respuesta o salida) respecto a un conjunto de variables independientes x (variables predictoras o de entrada).

La representación para el modelo de regresión lineal simple es:

$$y = \beta_0 + \beta_1 \cdot x_1 \tag{3.1}$$

Para el caso de dos o más variables independientes es decir el modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p + \varepsilon \tag{3.2}$$

- y es la variable de respuesta.
- $x_1, \ldots x_p$son las variables independientes o predictoras.
- β_0 es el término independiente. Es el valor esperado de y cuando $x_1, \ldots x_p$ son cero.
- $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes parciales de la regresión, determinan la cantidad en la que cambia y cuando x se incrementa en una unidad:
- ε es el error de observación debido a variables no controladas(Rial, 2014).

Las distancias entre los puntos y la ecuación de regresión representan la porción de la respuesta que no es explicada por la ecuación, es decir la diferencia entre el valor observado y el valor estimado, al cual se le llama residuo. La Figura 3.2 muestra la representación del modelo de regresión simple y múltiple para dos variables predictoras.

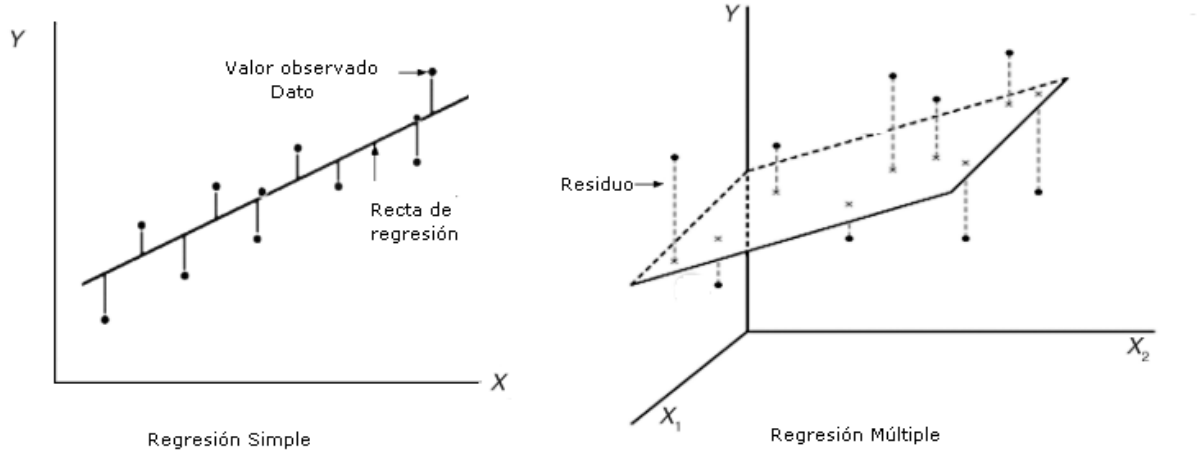


Figura 3.2. Modelos de regresión simple y múltiple.

La regresión lineal utiliza el ajuste por mínimos cuadrados, la cual consiste en calcular el menor valor de la suma de cuadrados de las diferencias entre los valores observados y esperados. (Carrasquilla-Batista *et al.*, 2016).

El método de mínimos cuadrados del modelo de regresión lineal simple (Matlab, 2020b), se define *S* como un sistema de *n* ecuaciones lineales simultaneas de dos incógnitas:

$$S = \sum_{i=1}^{n} \left((y_i - (\beta_1 \cdot x_i + \beta_0) \right)^2 \tag{3.3}$$

Dado que el proceso de ajuste de mínimos cuadrados minimiza la suma de cuadrados de los valores residuales, los coeficientes se determinan diferenciando *S* con respecto a cada parámetro e igualando a cero:

$$\frac{dS}{d\beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - (\beta_1 \cdot x_i + \beta_0) \right) = 0 \tag{3.4}$$

$$\frac{dS}{d\beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_1 \cdot x_i + \beta_0) \,) = 0 \tag{3.5}$$

solucionando para β_1:

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum {x_i}^2 - (\sum x_i)^2} \tag{3.6}$$

para β_0 con el valor de β_1:

$$\beta_0 = \frac{1}{n} (\sum y_i - \beta_1 \sum x_i) \tag{3.7}$$

El procedimiento anterior es ampliable a más de dos variables del modelo.

3.1.2. Árboles de regresión

Uno de los enfoques de aprendizaje supervisado más efectivos en la minería de datos son los árboles de regresión (AR), su estructura es una variante de los árboles de decisión desarrollados para aproximar funciones con valores reales. Su objetivo es desarrollar un modelo que sea capaz de predecir el valor de una variable de respuesta, dadas varias variables predictoras.

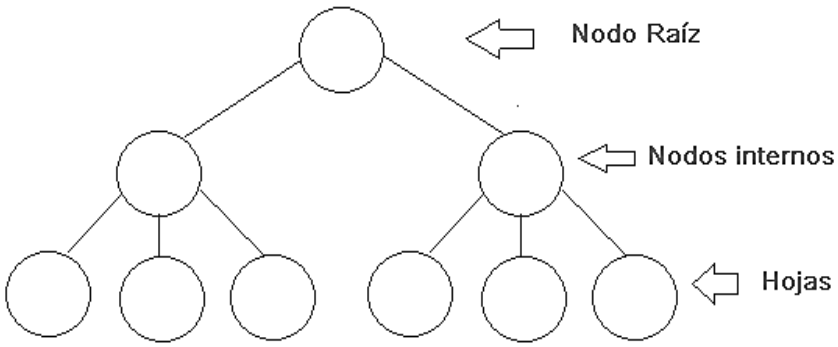


Figura 3.3. Arquitectura árboles de regresión.

El árbol está compuesto por un nodo raíz que contiene todos los datos, un conjunto de nodos internos o divisiones y un conjunto de nodos terminales llamados hojas. El algoritmo utiliza el estándar CART (*Classification And Regression Trees*) (Matlab, 2020a), realizando los siguientes pasos:

1. Inicia con todos los datos de entrada analizando todas las posibles divisiones de cada predictor
2. Selecciona una división con el mejor criterio de optimización, se elige una división que produce el mejor criterio de optimización.
3. Establece la división
4. Repite de forma recursiva para los nodos secundarios

La división se detiene cuando se detecta algunas condiciones por ejemplo si la suma de cuadrados residuales (MSE) tiende a cero en un nodo, este se considera un nodo terminal (hoja), el proceso se detiene si se alcanza el nivel de error más bajo o si se cumplen algunas condiciones de parada como un número límite de capas que pueda tener. El criterio de optimización elije una división para minimizar el MSE de las predicciones en comparación con los datos de entrenamiento.

3.1.3. Máquinas de vectores de soporte

Las máquinas de vectores de soporte (MVS) son modelos de algoritmos de aprendizaje supervisado que analizan datos para clasificación o análisis de regresión. En tareas de clasificación, la idea central es encontrar un hiperplano con un margen máximo de separación que permita dividir el conjunto de datos en dos clases, a mayor margen, mayor seguridad de que se está ante un hiperplano de separación bueno. Para la regresión, por el contrario, se busca seleccionar un hiperplano que mejor se ajuste al conjunto de datos, considerando una distancia o margen entorno al hiperplano esperando que todos los datos se encuentren dentro de ella. La Figura 3.4 muestra las representaciones de las MVS para las tareas de clasificación y regresión.

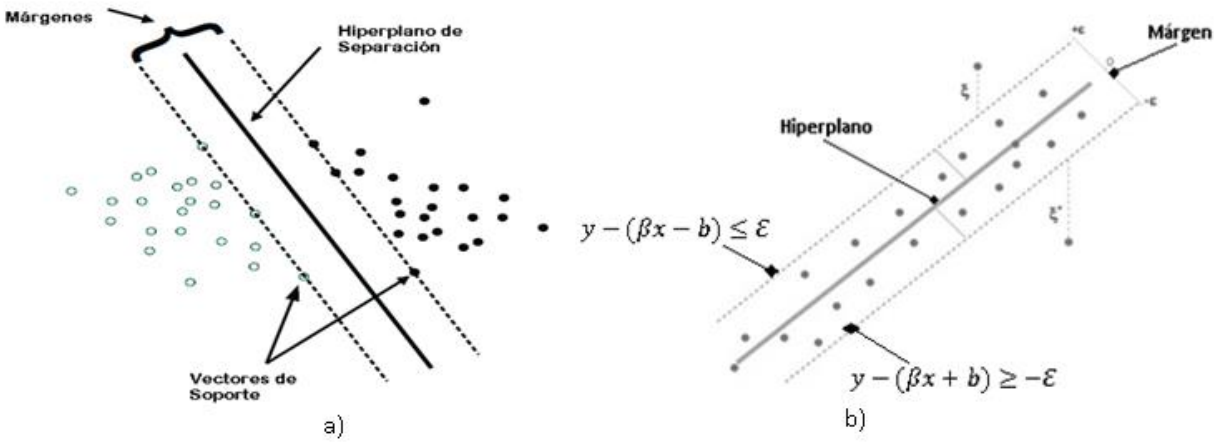


Figura 3.4. Márgenes e hiperplano MVS - a) Clasificación, b) Regresión.

Para el algoritmo de regresión MVS, se parte de un conjunto de entrenamiento cuyos datos se suponen lineales, para lo cual se obtiene un hiperplano que mejor represente el comportamiento de los datos para este caso será una línea cuya fórmula está dada por:

$$y = \beta x + b \tag{3.8}$$

Se considera una distancia de margen ε (distancia $+\varepsilon$, $-\varepsilon$) de modo que se construya una banda entorno al hiperplano, en donde se espera se encuentren todos los datos, las ecuaciones para estos márgenes son:

$$y - (\beta x - b) \leq \varepsilon \tag{3.9}$$

$$y - (\beta x + b) \geq -\varepsilon \tag{3.10}$$

Si las bandas no cubren todos los datos y todavía se tienen puntos fuera de ella, estos datos serían los errores y se deben considerar para el algoritmo (ver Figura 3.4 (b)), siendo ξ las distancias de las bandas y el punto se tiene:

$$y - (\beta x - b) \leq \varepsilon + \xi \tag{3.11}$$

$$y - (\beta x + b) \geq -\varepsilon - \xi^* \tag{3.12}$$

Lo que conlleva a la función objetiva:

$$\text{Minimizar: } \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}(\xi + \xi^*) \tag{3.13}$$

donde β es la magnitud del vector o hiperplano, C una constante mayor a 0, $\xi + \xi^*$ son las variables que controlan el error cometido por la función de regresión al aproximar las bandas y b un término de sesgo en el modelo de regresión MVS.

El procedimiento es extensible a datos no lineales (Matlab, 2020c), para los cuales se implementa un kernel para convertir los datos en una característica de espacio dimensional más alto para hacer posible la separación lineal.

3.2. Casos de Estudio y Recolección de Datos.

El agua para consumo humano está expuesta a diferentes factores desde su fuente hasta su uso final. El agua de acueducto y el agua envasada representan un porcentaje de consumo importante para la población, sin embargo, los sectores que no cuentan con redes de suministro de agua potable pueden encontrar en el agua lluvia una alternativa viable. Para este trabajo se analizaron tres clases de agua, las cuales se describen a continuación.

- Agua lluvia: La lluvia es un tipo de fenómeno atmosférico que se produce por la condensación del vapor de agua que cae a la tierra, depende de factores como la presión, la temperatura y la humedad atmosférica. Al caer se distribuye de forma irregular por lo que puede ser aprovechada en sectores como la agricultura, el uso doméstico incluso industrial. El agua lluvia es una alternativa viable frente a los problemas de abastecimiento, sin embargo, esta puede contener altas concentraciones de ácidos y elementos contaminantes asociados con las zonas y que requieren de tratamientos y controles de calidad para su consumo.

- Agua acueducto: El agua que se transporta por medio de redes de alcantarillado para el abastecimiento de la población, recibe controles de calidad y tratamientos que se realizan desde el origen o captación en embalses, ríos, pozos y que continúa posteriormente en estaciones de monitoreo y a través de la red hasta llegar finalmente al consumidor. La calidad de agua de acueducto se encuentra reglamentada, así como las frecuencias de muestreo y las técnicas adecuadas para la evaluación de parámetros.

- Agua envasada: El termino agua envasada se designa a aquella proveniente de fuentes subterráneas o de abastecimiento público y que se comercializa envasada en botellas u otros tipos de contenedores. El agua antes de ser embotellada, pasa por tratamientos para cumplir con los valores normativos que permitan su consumo y almacenamiento.

Para calcular el tamaño de la muestra, existen algunas ecuaciones que permiten obtener la cantidad de datos u observación que mejor se ajuste a la población para que los resultados sean representativos (Chow, Shao y Wang, 2008). Para realizar inferencias a valores poblacionales de medias a partir de una muestra en donde no se conoce el tamaño de la población se utiliza la ecuación:

$$n = \frac{\sigma^2 . \ Z^2}{e^2} \tag{3.14}$$

en donde,
n = tamaño de la muestra
σ= desviación estándar estimada (se supone 0.5 si no se puede estimar o no se cuenta con información previa)
Z= nivel de confianza según la distribución normal estándar.
e= precisión con que se desea estimar (de 0.01 a 0.1).

Para este trabajo se desea estimar el tamaño de muestra que permita analizar el comportamiento temporal del agua en almacenamiento, para un nivel de confianza del 95%, ($Z_{95\%}$ corresponde a 1.960 para el estadístico Z), se tiene:

$$n = \frac{(0.5)^2 . \ (1.960)^2}{(0.1)^2} = 96{,}04 \ \cong 96 \text{ observaciones por muestra}$$

De acuerdo al resultado de la ecuación (3.14) se decide medir 100 observaciones por parámetro en cada muestra.

3.2.1. Toma de muestras y recolección de datos

Dentro de la metodología para la toma de muestras y recolección de datos se identifican las siguientes fases:

I. **Selección de parámetros:** Con base en lo estipulado en Organización Mundial de la Salud OMS, que establecen los parámetros de nivel básico para determinar la calidad de agua: pH, solidos disueltos totales (TDS), Temperatura, Turbidez y Conductividad, para el presente trabajo se evaluaron los parámetros de pH, TDS y temperatura sobre cada muestra. Estos parámetros aportan información básica sobre el estado del agua que a su vez mantienen algún grado de incidencia sobre otros parámetros fisicoquímicos, además, se cuenta con la disponibilidad de los equipos para su medición directa sin evaluaciones posteriores por personal y equipos especializados. A continuación, se da una breve descripción de estos parámetros:

- **pH:** Es la medida de la acidez del agua, y es expresada en una escala que va entre 1 y 14 (unidades de pH), las aguas naturales (no contaminadas) exhiben un pH en el rango de 6 a 9.

- **TDS:** Compuestos inorgánicos que se encuentran en el agua, como sales, metales pesados y algunos rastros de compuestos orgánicos que se disuelven en el agua. Su unidad de medición es en partes por millón (ppm), un valor estándar internacional para el agua potable se encuentra alrededor de los 300 ppm.

- **Temperatura:** Uno de los parámetros físicos más importantes, por lo general influye en el retardo o aceleración de la actividad biológica, la absorción de oxígeno, la precipitación de compuestos entre otros. Los valores deseados de la temperatura para aguas superficiales, debe ser inferior a 15°C ya que valores mayores favorecen el desarrollo de microorganismos no deseables para el agua de consumo.

II. **Instrumentos de medición:** Los instrumentos de medición para medir los parámetros de pH, TDS y temperatura son:

- **Medidor de pH digital:** mide valores de pH en un rango de 0 a 14 pH, con una precisión de +-0,1pH y resolución de 0,1 pH, temperatura de operación de 0 a 60 °C, calibración manual mediante reactivos. La Figura 3.5 muestra el dispositivo de medición de pH.

Figura 3.5. Medidor de pH

- **Medidor de TDS digital:** mide TDS en un rango de 0 a 9990ppm, con una precisión de +/- 2% capacidad de medir temperatura en un rango de 0.1 a 80°C, calibración automática. La Figura 3.6 muestra el equipo de medición de TDS.

Figura 3.6. Medidor de TDS

- **Medidor de temperatura:** El dispositivo de medición de TDS cuenta con la capacidad de medir la temperatura en un rango de 0.1 a 80°C además se utilizó un termómetro de mercurio para validar los valores de temperatura. La Figura 3.7 muestra el equipo de medición de temperatura.

Figura 3.7. Medidor de temperatura

III. **Toma de muestras:** Se recolectaron cuatro muestras por cada clase, es decir, cuatro muestras de agua lluvia, cuatro de agua de acueducto y cuatro de agua envasada. Las muestras de agua lluvia se recolectaron en el municipio de Popayán Cauca, en diferentes fechas, para el agua de acueducto se tomaron muestras en distintos puntos de la ciudad de Popayán y se escogieron cuatro marcas de agua envasada. Se asignaron las etiquetas de identificador “AL” para Agua Lluvia, “AA” para Agua de Acueducto y “AE” para Agua Envasada, los números del 1 al 4 corresponden a cada muestra. La Tabla 3.1 presenta los indicadores y la información de los puntos y/o datos de recolección para cada muestra.

Tabla 3.1. Indicadores y descripción toma de muestras.

Clase	Identificador	Descripción
Agua Lluvia (AL)	**AL-1**	Fecha toma de muestra: 28/02/2020
	AL-2	Fecha toma de muestra: 3/03/2020
	AL-3	Fecha toma de muestra: 5/03/2020
	AL-4	Fecha toma de muestra: 6/03/2020
Agua de Acueducto (AA)	**AA-1**	Ubicación geográfica:2° 26’ 51.44” N, 76° 35’ 56.60” O
	AA-2	Ubicación geográfica:2° 26’ 42.65” N, 76° 36’ 43.62” O
	AA-3	Ubicación geográfica:2° 26’ 57.27” N, 76° 37’ 35.40” O
	AA-4	Ubicación geográfica:2° 28’ 55.88” N, 76° 34’ 5.09” O
Agua Envasada (AE)	**AE-1**	Marca 1 - Lote L874CP31
	AE-2	Marca 2 - Lote L7R0003330052
	AE-3	Marca 3 - Lote- L608200052C
	AE-4	Marca 4 – Lote L9F0375016003

IV. Recolección de datos: Durante el experimento, las muestras se conservaron en recipientes plásticos sin tapa de 1000ml y sin exposición directa al aire libre, es decir, dentro de una habitación cerrada. Con condiciones de temperaturas promedio de 19°C, de clima cálido templado y humedad promedio de 84% propias de la zona y las cuales pueden variar de acuerdo a las temporadas del año. Se midieron y se registraron los valores de los parámetros de pH, TDS y temperatura diariamente sobre todas las muestras durante un periodo de tres meses. En la medición los equipos se introducían directamente sobre cada recipiente realizando la limpieza del equipo después de cada medida, con el fin de no alterar los contenidos.

Figura 3.8. Recolección de datos

En total se analizaron doce muestras cuatro para cada clase, producto de la medición y registro diario de la información, se recolectaron 100 datos u observaciones por cada parámetro, por lo que cada muestra cuenta con 300 datos en total para la construcción de los modelos, sin embargo, es importante mencionar que debido a las condiciones de almacenamiento, en las mediciones se obtienen los mismos valores de temperatura para las diferentes muestras, por lo cual se utiliza un solo registro de la temperatura en cada una de ellas. El tamaño total de la información recolectada en este experimento es de 2500 datos. El estudio de más de una muestra permite aplicar el experimento varias veces y reducir la probabilidad de llegar a conclusiones erróneas asociadas a variaciones aleatorias.

3.3. Procedimiento.

Para la presentación de los resultados del experimento en este documento, se seleccionaron tres muestras al azar (una de cada clase), para la estimación de parámetros: AL-3 (agua lluvia), AA-1 (agua de acueducto), AE-2 (agua envasada). De igual manera se escogieron las muestras AA-3, AL-1 y AE-4, para realizar la validación (Capitulo 4). Los procedimientos de estimación se replicaron en las muestras restantes y se pueden encontrar en el Anexo C. Previo a la construcción de los modelos, se realizó un análisis de características estadísticas de los datos, correlación de parámetros y significancia estadística (*valor -p*). Posteriormente se construyeron los modelos respectivos para las técnicas: regresión lineal (RL), arboles de regresión (AR) y máquinas de vectores de soporte (MVS) en la herramienta software para cada muestra de clase.

En el análisis predictivo es posible estimar un parámetro fisicoquímico o un índice de calidad, en este caso, se estimaron inicialmente los parámetros de pH y TDS los cuales tienen mayor importancia dentro de un análisis de calidad de agua, finalmente se estimó un índice de calidad construido a partir de los tres parámetros medidos pH, TDS y temperatura (Capitulo 4), como mecanismo que permitió generalizar los resultados y facilitar su comprensión.

El análisis de características estadísticas permite interpretar la información recolectada, con el objetivo de identificar aspectos representativos del conjunto de datos del fenómeno que se estudia, algunas como la media, desviación estándar valor máximo y mínimo del conjunto de datos medidos se presentan en la Tabla 3.2 para cada una de las muestras.

Tabla 3.2.Características estadísticas de los parámetros.

	ID	pH (Unidades)				TDS (ppm)			
		Min.	Max	Media	DesvEst.	Min.	Max.	Media	DesvEst.
Agua Lluvia	AL-1	7.02	10.12	8.4688	0.9634	12	206	91.930	55.075
	AL-2	7.40	10.20	8.6998	0.7736	14	293	126.46	90.733
	AL-3	7.50	9.65	8.5166	0.4911	11	238	100.43	75.211
	AL-4	7.34	9.07	8.4347	0.3939	5	106	40.55	30.641
Agua Acueducto	AA-1	7.29	9.46	8.6382	0.4567	36	183	103.48	49.674
	AA-2	7.66	9.85	8.7204	0.6193	33	150	80.290	35.959
	AA-3	7.61	9.72	8.7362	0.5504	33	158	82.250	38.146
	AA-4	7.22	9.43	8.4425	0.5963	30	178	92.470	47.297
Agua Envasada	AE-1	7.76	10.30	9.2697	0.6562	84	283	184.27	57.953
	AE-2	7.85	10.30	9.0659	0.6259	68	227	137.75	47.963
	AE-3	7.56	9.52	8.6869	0.5382	122	298	201.41	51.022
	AE-4	7.43	9.45	8.5315	0.5902	75	286	173.33	67.846

	Min.	Max.	Media	DesvEst.
Temperatura (°C)	16	22	18.03	1.13

La media o promedio se interpreta como un punto de equilibrio del conjunto de datos. En la Tabla 3.2 se puede observar que los valores de la media para el pH se encuentran entre 8 y 9 unidades de pH para las diferentes muestras, para el TDS, por el contrario, los valores de la media difieren más entre cada muestra y clase de agua, estando en su mayoría por encima de los 80(ppm). Se tiene un valor de temperatura promedio de 18.03 °C acorde con la temperatura ambiental de la zona.

La medida de dispersión más común es la desviación estándar e indica que tan dispersos están los datos con respecto a la media. De acuerdo la desviación estándar obtenida para el pH, se observa que la distribución de los datos se encuentra cercanas a los valores de la media correspondientes para cada muestra, dado que estos son muy cercanos a cero lo que puede asociarse físicamente con la lenta evolución del pH en el tiempo, con valores máximos y mínimos que varían desde 7 a 10 unidades de pH aproximadamente.

Una mayor distribución de los datos puede apreciarse en los valores de TDS, que presentan valores superiores de desviación estándar, con valores mínimos de TDS que van desde 5 (ppm) hasta 200 (ppm) en algunas muestras, lo que representa una distribución de valores más grande que el pH en el tiempo. La temperatura varía entre los 16°C y alcanza hasta los 22°C, su desviación estándar es baja con un valor de 1.13 con respecto al valor de la media.

Las diferencias entre valores de la media para muestras de una misma clase, pueden asociarse a factores externos como lluvias o contaminantes provenientes de la fuente, estos generalmente se observan en los valores de TDS.

3.4. Correlación de Parámetros

El coeficiente de correlación de Pearson se utiliza para evaluar la asociación lineal estadística entre variables, en tareas predictivas es útil para determinar la selección de predictores o parámetros de entrada para la construcción de un modelo, el coeficiente puede tomar valores en un rango de -1 a +1. Un valor de 0 indica que no existe asociación entre las variables, valores cercanos a +1 representan una asociación positiva y para valores cercanos a -1 una asociación negativa. La correlación de parámetros se realizó sobre cada una de las muestras, la Tabla 3.3 presenta los resultados obtenidos para la muestra AL-3 de agua lluvia.

Tabla 3.3.Coeficientes de correlación – muestras AL-3.

AL-3			
	pH	**TDS**	**T**
pH	1		
TDS	0.8732	1	
T	-0.0044	-0.0208	1

En los resultados de la Tabla 3.3 se observa un valor de correlación positiva de 0.8732 entre el pH y TDS, lo que puede reflejarse en que el agua superficial es baja en contenido mineral haciendo que se presente un menor valor de conductividad eléctrica, lo cual se encuentra directamente relacionado con el TDS. La Tabla 3.4 contiene los valores de correlación para los parámetros de la muestra AA-1 de agua de acueducto.

Tabla 3.4.Coeficientes de correlación – muestras AA-1.

AA-1			
	pH	**TDS**	**T**
pH	1		
TDS	0.9293	1	
T	-0.1175	-0.0310	1

En la muestra AA-1 de la Tabla 3.4, se observa también una correlación positiva entre el pH y TDS y una correlación negativa de estos con la temperatura, lo que puede asociarse con una relación inversa entre los parámetros, es decir las bajas temperaturas favorecen los niveles de pH y TDS en el agua. Los niveles de oxígeno en el agua se reducen si su temperatura aumenta provocando la proliferación de algas que afecta los ecosistemas acuáticos. Los coeficientes de correlación de la muestra AE-2 de agua envasada se presentan en la Tabla 3.5.

Tabla 3.5.Coeficientes de correlación – muestras AE-2.

AE-2			
	pH	**TDS**	**T**
pH	1		
TDS	0.9744	1	
T	-0.1063	-0.0441	1

De manera similar a los casos anteriores, para la muestra AE-2 se evidencia una correlación positiva entre los parámetros TDS y pH, pero negativa con la temperatura. El análisis de correlación para las muestras restantes pude encontrarse en el Anexo A.

3.4.1. Análisis inferencial

El propósito fundamental de los análisis estadísticos de tipo inferencial es el conocimiento de poblaciones a partir del estudio de muestras o subconjuntos representativos y suficientes de dichas poblaciones. Para estimar el tamaño muestral

necesario para el experimento del presente trabajo, se realizó una prueba para calcular el tamaño necesario n_A (García, Reding y López, 2013) para comparar dos muestras a partir de la ecuación:

$$n_A = (\sigma_A^2 + \sigma_B^2)\left(\frac{Z_{1-\alpha}+Z_{1-\beta}}{\mu_A-\mu_B}\right)^2 \quad (3.15)$$

en donde:

n_A= tamaño muestra A
σ_A= desviación estándar muestra A
μ_A= media muestra A
σ_B= desviación estándar muestra B
μ_B= media muestra B
$1-\beta$= poder estadístico
$1-\alpha$ = factor de riesgo error tipo I

Para comprobar el tamaño necesario para una muestra de pH se utilizaron los datos de las muestras AA-3 AL-3, con un poder estadístico $(1-\beta)$ de 0.80 correspondiente a un valor Z estadístico de 0,842 y un riesgo de α de 0.05 correspondiente a un valor Z estadístico de 1.645 para la ecuación (3.15) se tiene:

$\mu_{A=}8{,}51$
$\mu_{B=}8{,}73$
$\sigma_{A=}\ 0{,}49$
$\sigma_{B=}0{,}55$
$k = \frac{n_A}{n_B} = 1$

$$n_A = ((0{,}49)^2 + (0{,}55)^2)\left(\frac{(1{,}645)+(0.842)}{(8{,}51)-(8{,}73)}\right)^2 = 69{,}34 \cong 70 \text{ observaciones}$$

Para comprobar el tamaño necesario para una muestra de TDS se utilizaron los datos de las muestras AL-1 AL-2, con un poder estadístico $(1-\beta)$ de 0.80 correspondiente a un valor Z estadístico de 0,842 y un riesgo de α de 0.05 correspondiente a un valor Z estadístico de 1.645 para la ecuación (3.15) se tiene:

$\mu_{A=}91{,}93$
$\mu_{B=}126{,}46$
$\sigma_{A=}\ 55{,}07$
$\sigma_{B=}90{,}73$
$k = \frac{n_A}{n_B} = 1$

$$n_A = ((55{,}07)^2 + (90{,}73)^2)\left(\frac{(1{,}645)+(0.842)}{(91{,}93)-(126{,}46)}\right)^2 = 59{,}43 \cong 60 \text{ observaciones}$$

Con lo anterior se verifica que el tamaño escogido de 100 observaciones por parámetro en cada muestra está dentro del rango definido.

En el análisis inferencial, se destaca el importante papel que juegan los contrastes de hipótesis. Se refieren a los procedimientos estadísticos mediante los cuales aceptamos o rechazamos una hipótesis nula (***H₀***) lo que automáticamente nos habilita para rechazar o aceptar otra hipótesis denominada hipótesis alternativa (***H₁***). En el desarrollo del proceso de contraste de hipótesis es igualmente relevante el nivel de significación, error tipo I o α (rechazar una Ho cuanto ésta es verdadera). Los valores habituales asumidos para los errores son el 10%, 5% y 1%, siendo, por ende, los niveles de confianza del 90%, 95% y 99%.

Existen dos grandes grupos de pruebas de significación estadística: paramétricas y no paramétricas. Las pruebas de tipo paramétrico están sometidas a determinadas condiciones de aplicación o verificación de supuestos como son: normalidad, homoscedasticidad e independencia (Tejedor y Etxeberria, 2006). El *valor p* (conocido también como ***p-valor*****,** o en inglés *p-value*) en el contraste de hipótesis, ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos, su valor de probabilidad oscila entre 0 y 1. En este trabajo se verificaron los supuestos para las muestras AL-3 y AA-3 en las series de pH y TDS.

I. Prueba de normalidad: La prueba de Kolmogorov-Smirnov para una muestra, es un procedimiento de "bondad de ajuste", que permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica específica. La prueba de Kolmogorov-Smirnov para una muestra se puede utilizar para comprobar si una variable se distribuye normalmente. Las hipótesis a contrastar:

H₀*: Los datos de la muestra siguen una distribución normal*
H₁*: Los datos de la muestra no siguen una distribución normal.*

Para un nivel de significancia α de 0.05 y utilizando el p-valor asociado al estadístico *D* de la prueba Kolmogorov-Smirnov, la regla de decisión para este contraste es:

Si p-valor $\geq \alpha \Rightarrow$ *Aceptar* H_0
Si p-valor $< \alpha \Rightarrow$ *Rechazar* H_0

Los resultados para las series de tiempo de pH de la muestra AA-3 y AL-3 son:

- **pH muestra AA-3**
  Estadístico de Kolmogorov-Smirnov (D): 0,123
  p- valor de la prueba: 0,099
  p≥ 0,05 = Aceptar H_0

- **pH muestra AL-3**
  Estadístico de Kolmogorov-Smirnov (D): 0,091
  p- valor de la prueba: 0,375
  p≥ 0,05 = Aceptar H_0

Los resultados para las series de tiempo de TDS de la muestra AL-1 y AL-3 son:

- **TDS muestra AL-1**
  Estadístico de Kolmogorov-Smirnov (D): 0,098
  p- valor de la prueba: 0,292
  p≥ 0,05 = Aceptar H_0

- **TDS muestra AL-2**
  Estadístico de Kolmogorov-Smirnov (D): 0,143
  p- valor de la prueba: 0,033
  p<0,05 = Rechazar H_0

II. Prueba de Homoscedasticidad: El supuesto de homogeneidad de varianzas, también conocido como supuesto de homocedasticidad, considera que la varianza es constante (no varía) en los diferentes niveles de un factor, es decir, entre diferentes grupos. Para la prueba de Levene las hipótesis a contrastar son:

$$H_0: \text{Las muestras son homogeneas}$$
$$H_1: \text{Las muestras no son homogéneas}$$

Para un nivel de significancia α de 0.05 y utilizando el p-valor asociado al estadístico *F* de la prueba de Levene, la regla de decisión para este contraste es:

$$\text{Si p-valor} > \alpha \Rightarrow \text{Aceptar } H_0$$
$$\text{Si p-valor} < \alpha \Rightarrow \text{Rechazar } H_0$$

Los resultados al evaluar las muestras AA-3 y AL-3 en pH y AL-1 y AL-2 para TDS son:

- **pH muestra AA-3 y AL-3**
  Estadístico de Levene(F): 3.4547
  p- valor de la prueba: 0.064
  p>0,05 = Acepta H_0

- **TDS muestras AL-1 y AL-2**
  Estadístico de Levene(F): 41.2807
  p- valor de la prueba: 0.00001
  p<0,05 = rechaza H_0

III. Prueba de independencia: El contraste de rachas permite verificar la hipótesis nula de que la muestra es aleatoria, es decir, si las sucesivas observaciones son independientes. Este contraste se basa en el número de rachas que presenta una muestra. Una racha se define como una secuencia de valores muestrales con una característica común precedida y seguida por valores que no presentan esa característica. Para la prueba de rachas las hipótesis a contrastar son:

$$H_0: \textit{Las muestras son aleatorias}$$
$$H_1: \textit{Las muestras no son aleatorias}$$

Para un nivel de significancia α de 0.05 y utilizando el p-valor asociado a la prueba de aleatoriedad Z, la regla de decisión para este contraste es:

$$\textit{Si p-valor} > \alpha \Rightarrow \textit{Aceptar } H_0$$
$$\textit{Si p-valor} < \alpha \Rightarrow \textit{Rechazar } H_0$$

Los resultados para las series de tiempo de pH de la muestra AA-3 y AL-3 son:

- **pH muestra AA-3**
  Prueba de aleatoriedad (Z): -9.3473
  p- valor de la prueba: 9.7174e-26
  p<0,05 = Rechaza H_0

- **pH muestra AL-3**
  Prueba de aleatoriedad (Z): -8.1412
  p- valor de la prueba: 1.9489e-18
  p<0,05 = Rechaza H_0

Los resultados para las series de tiempo de TDS de la muestra AL-1 y AL-3 son:

- **TDS muestra AL-1**
  Prueba de aleatoriedad (Z): -9.7494
  p- valor de la prueba: 3.9647e-29
  p<0,05 = Rechaza H_0

- **TDS muestra AL-2**
  Prueba de aleatoriedad (Z): -9.7494
  p- valor de la prueba: 3.9647e-29
  p<0,05 = Rechaza H_0

Debido a que no se cumplen todos los supuestos para una prueba paramétrica, se aplica la prueba no paramétrica Mann Whitney- U (Rivas, Moreno y Talaveraa, 2013) la cual, se utiliza para comparar dos grupos de rangos (medianas) y determinar que la diferencia no se deba al azar (que la diferencia sea estadísticamente significativa). Para la prueba, las hipótesis a contrastar son:

$$H_0: \text{No hay diferencias entre las medianas de las muestras}$$
$$H_1: \text{Hay diferencias entre las medianas de las variables}$$

Para un nivel de significancia α de 0.05 y utilizando el p-valor asociado a la prueba U, la regla de decisión para este contraste es:

$$\text{Si p-valor} > \alpha \Rightarrow \text{Aceptar } H_0$$
$$\text{Si p-valor} < \alpha \Rightarrow \text{Rechazar } H_0$$

- **pH muestra AA-3 y AL-3**
  Estadístico de Mann Whitney (U): 3983.5
  p- valor de la prueba: 0.01304
  $p<0,05$ = rechaza H_0

- **TDS muestras AL-1 y AL-2**
  Estadístico de Mann Whitney (U): 4145.5
  p- valor de la prueba: 0,03691
  $p<0,05$ = rechaza H_0

Existe una diferencia estadísticamente significativa entre las medianas de los datos de pH de las muestras AA-3 y Al-3 y en los datos de TDS de las muestras AL-1 y AL-2.

3.5. Generación de Modelos

Para la construcción de los modelos de regresión, se ingresaron los datos medidos en el software Matlab utilizando las funciones integradas de la caja de herramientas estadísticas y de aprendizaje automático *Regression Learner APP*, ver Figura 3.9. Principalmente se configuran en la herramienta las variables que servirán como entradas y cuál será la salida del sistema, es decir la variable a predecir.

Para asegurar la capacidad de generalización y reproducibilidad de los resultados se configuro la opción de validación cruzada (*Cross- validation*, Figura 3.5) de 5 veces como método de re-muestreo, donde cinco subconjuntos de datos se mantienen como conjuntos de validación (no se utilizan durante el entrenamiento); los

parámetros del modelo se ajustan en los datos restantes (conjunto de entrenamiento); el modelo entrenado se utiliza posteriormente para hacer predicciones sobre el conjunto de validación.

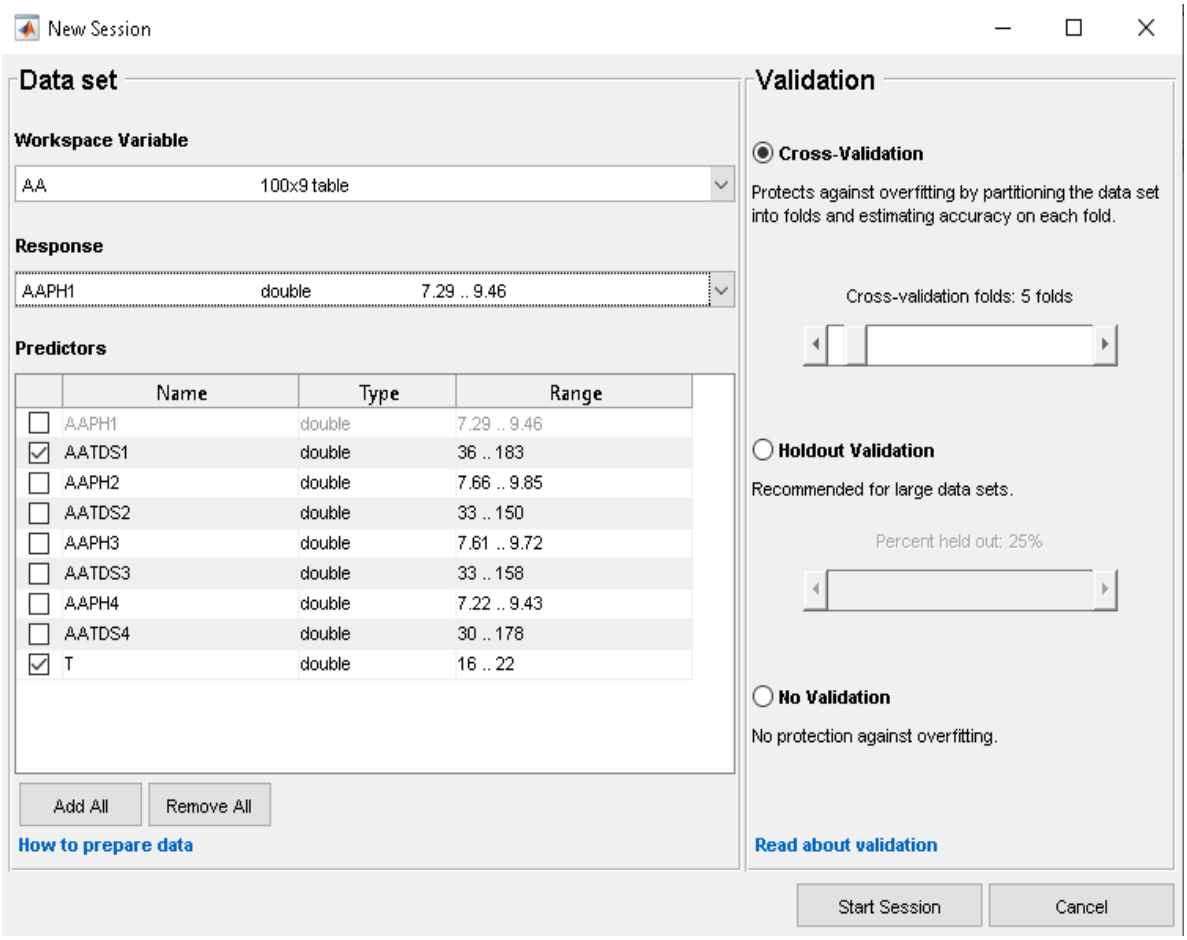


Figura 3.9.Configuración herramienta de aprendizaje automático Matlab.

Una vez cargados los datos, seleccionado las entradas y salidas para el modelo, se estimarán los parámetros pH y TDS para las muestras analizadas: AL-3 agua lluvia, AA-1 agua de acueducto y AE-2 agua envasada, para cada muestra se construirán tres modelos correspondientes a las técnicas de regresión lineal (RL), arboles de regresión (AR), y máquinas de vectores de soporte (MVS). Los resultados de estas estimaciones se presentan a continuación.

3.5.1. Modelos para la muestra AL-3 (agua lluvia)

Para la estimación del pH, se selecciona este parámetro como salida y como predictores o variables de entrada, los parámetros TDS y temperatura. El modelo obtenido para la técnica RL presenta las siguientes características:

	Estimate	SE	pValue
(Intercept)	7.7675	0.3894	4.2475e-36
ALTDS3	0.0057044	0.00032281	4.3723e-32
T	0.0097716	0.021443	0.64963

Figura 3.10.Coeficientes del modelo RL – estimación pH

En la Figura 3.10, la columna ***Estimate*** contiene los coeficientes para cada término del modelo, ***SE***, los errores estándar de los coeficientes y ***pValue,*** el p-valor estadístico para la prueba de hipótesis. Para aun nivel de significancia de 0.05, se observa que el intercepto y el parámetro TDS, son menores al 5% y por lo que son significativos dentro del modelo. Por lo anterior, de la ecuación (3.2) en la sección 3.1.1 del modelo RL para el pH es:

$$pH = 7.7675 + 0.0057044\,TDS + 0.0097716\,T \tag{3.14}$$

La representación gráfica de los valores estimados y reales para el pH se muestra en la Figura 3.11.

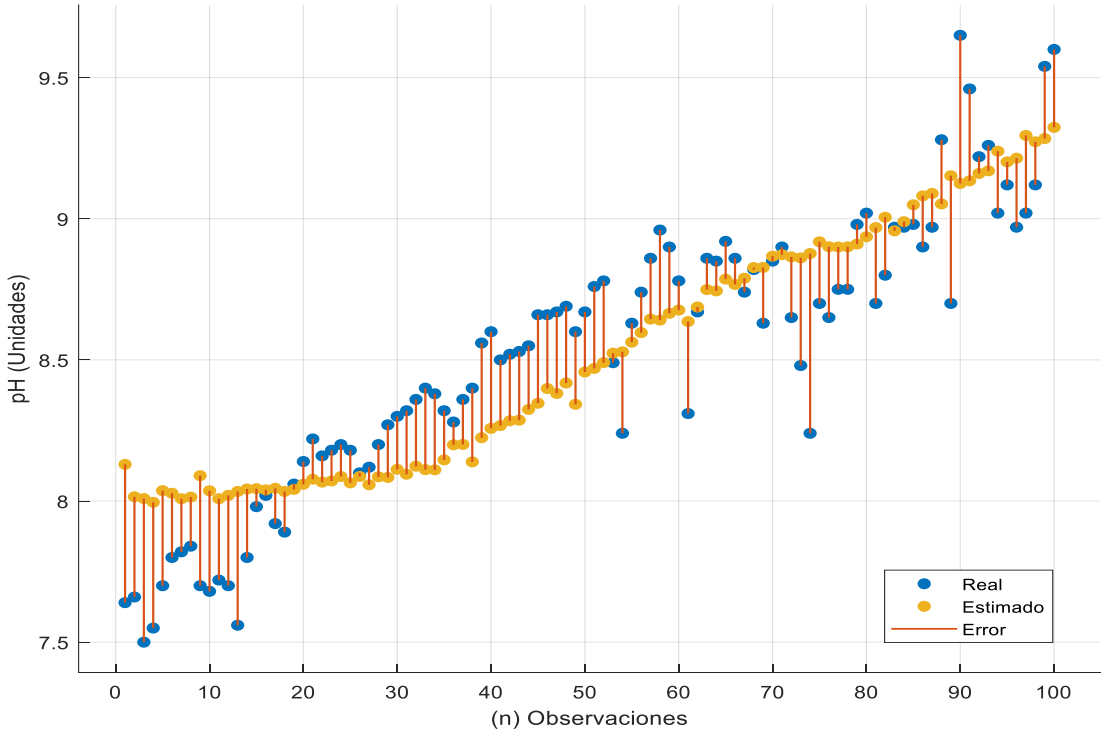


Figura 3.11.Representación gráfica modelo RL – estimación pH muestra AL-3.

El eje horizontal representa el número de observaciones o datos medidos en el experimento, el eje vertical las unidades de la variable de salida para este caso el pH, los puntos azules representan los valores reales y los estimados de color amarillo. La diferencia o error se muestra con una línea roja. El modelo obtenido para la técnica AR en la estimación de pH muestra la arquitectura de la Figura 3.12.

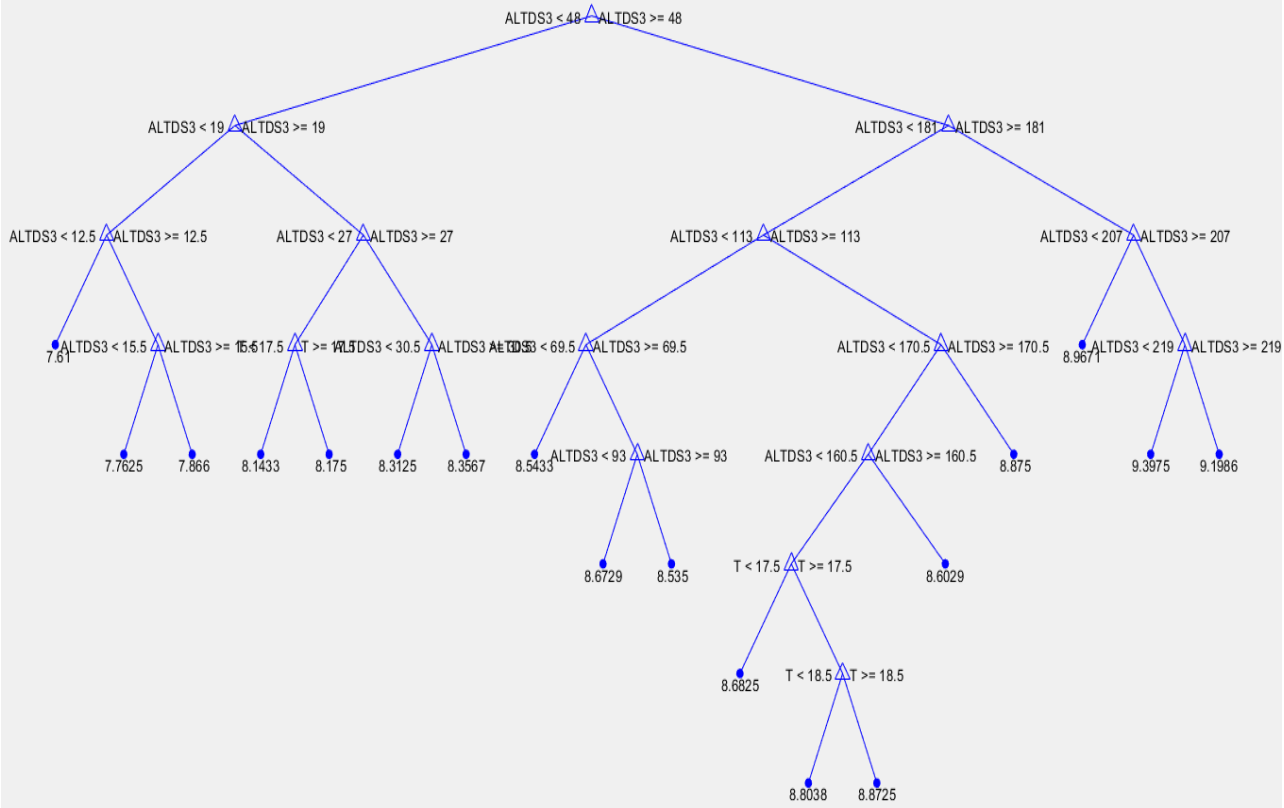


Figura 3.12.Arquitectura del modelo AR – estimación pH muestra AL-3.

La Figura 3.12 muestra la arquitectura del modelo AR y las correspondientes condiciones de decisión para los 35 nodos totales. La representación de los valores estimados y reales para el modelo AR se observan en la Figura 3.13.

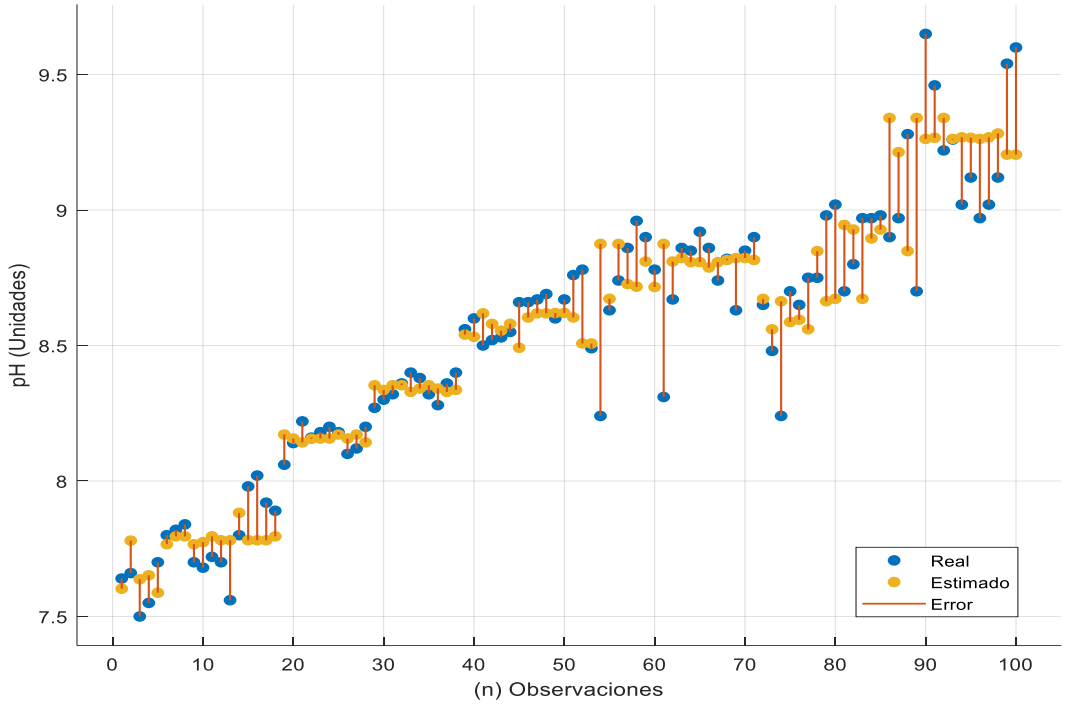


Figura 3.13.Representación gráfica modelo AR – estimación pH muestra AL-3.

El modelo resultante para la técnica MVS en la estimación del pH presenta las propiedades de la Figura 3.14.

```
RegressionSVM
         PredictorNames: {'ALTDS3'  'T'}
           ResponseName: 'Y'
  CategoricalPredictors: []
      ResponseTransform: 'none'
                  Alpha: [84×1 double]
                   Beta: [0.2274  0.0178]
                   Bias: 8.5596
        KernelParameters: [1×1 struct]
                     Mu: [100.4300  18.0300]
                  Sigma: [75.2111  1.1322]
        NumObservations: 100
         BoxConstraints: [100×1 double]
       ConvergenceInfo: [1×1 struct]
       IsSupportVector: [100×1 logical]
                 Solver: 'SMO'
```

Figura 3.14.Propiedades del modelo MVS – estimación pH muestra AL-3.

De acuerdo a la ecuación 3.8 de la sección 3.1.3, se relacionan los valores de α (Alpha en, Figura 3.14) que corresponde al número de vectores de soporte del modelo, β (Beta) los coeficientes asociados a los dos predictores y b (Bias) el termino de sesgo del modelo. La representación de los valores estimados y reales para el modelo MVS se muestra en la Figura 3.15.

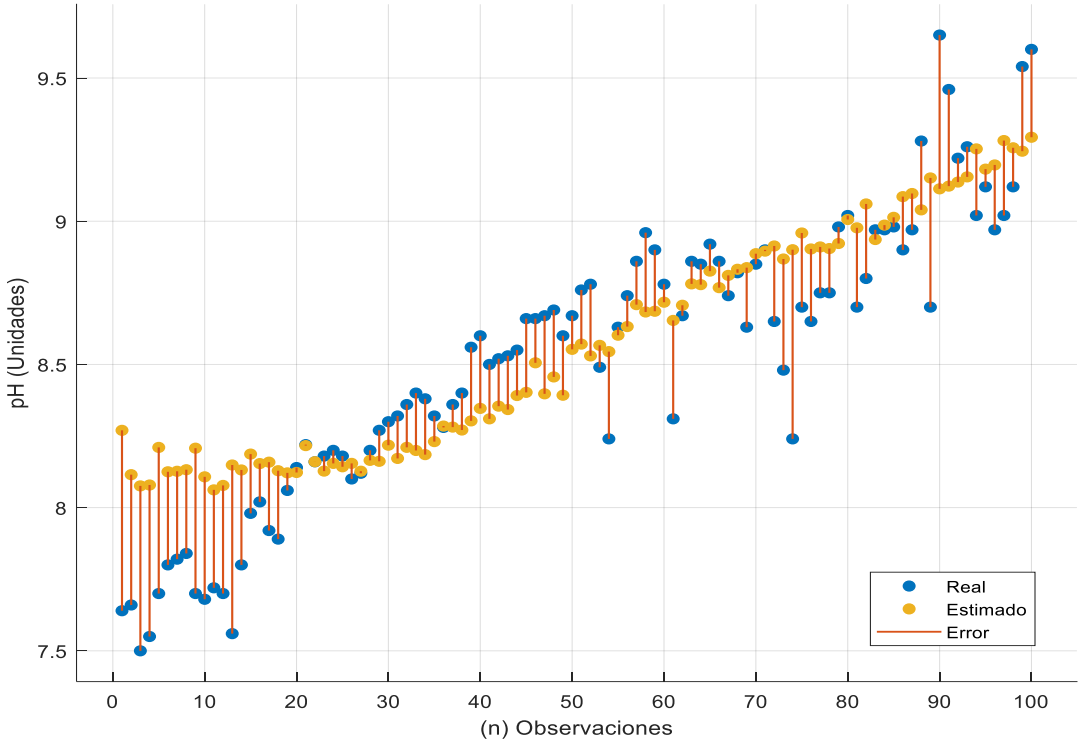


Figura 3.15.Representación gráfica modelo MVS – estimación pH muestra AL-3.

Para la estimación del TDS se establecen como variables de entrada el pH y la temperatura T y se estiman los modelos de RL, AR y MVS. Los coeficientes del modelo, error y p-valor, se presentan en la Figura 3.16.

	Estimate	SE	pValue
(Intercept)	-1009.3	87.404	6.4149e-20
ALPH3	133.76	7.5692	4.3723e-32
T	-1.6344	3.2829	0.61972

Figura 3.16. Representación gráfica modelo MVS – estimación TDS muestra AL-3.

Con un nivel de significancia de 0.05, se observa que para la estimación del TDS el parámetro pH es estadísticamente significativo a diferencia de la T cuyo p-valor está por encima del 5%. A partir de los coeficientes de la Figura 3.16, el modelo resultante para la estimación del TDS es:

$$TDS = -1009.3 + 133.76\, pH - 1.6344\, T \tag{3.15}$$

La Figura 3.17 contiene la representación gráfica de las estimaciones del modelo RL para el TDS.

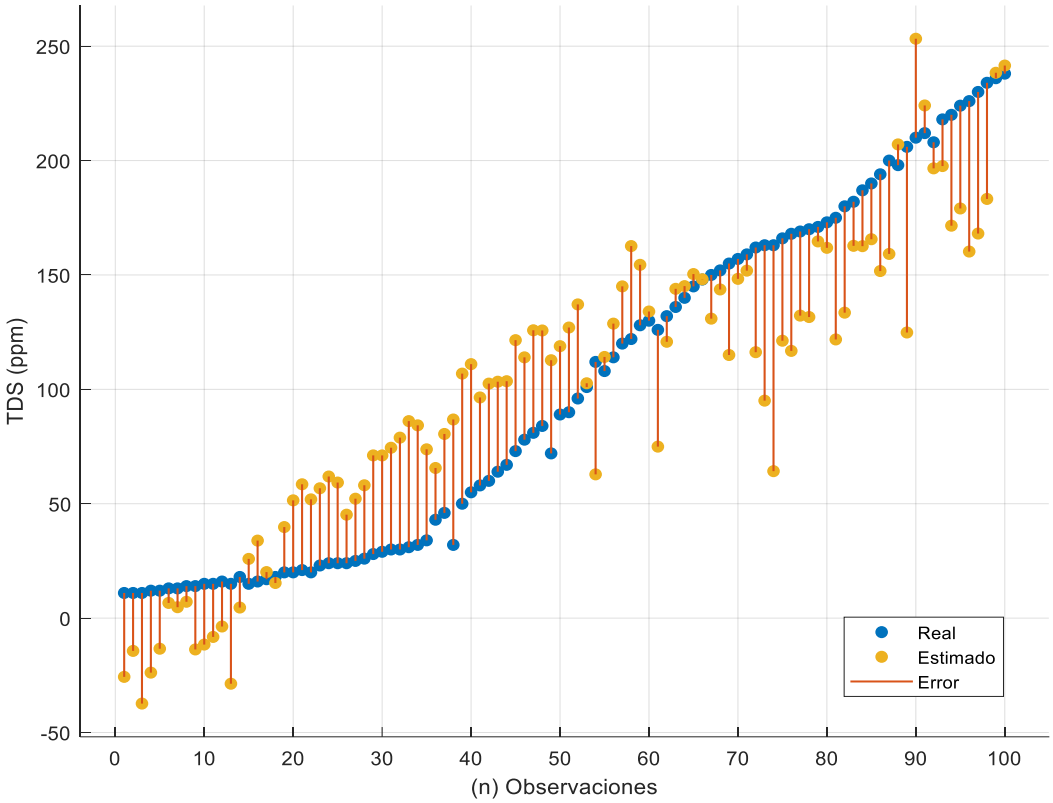


Figura 3.17.Representación gráfica modelo RL– estimación TDS muestra AL-3

La arquitectura el modelo de AR para la estimación del TDS, se presenta en la Figura 3.18

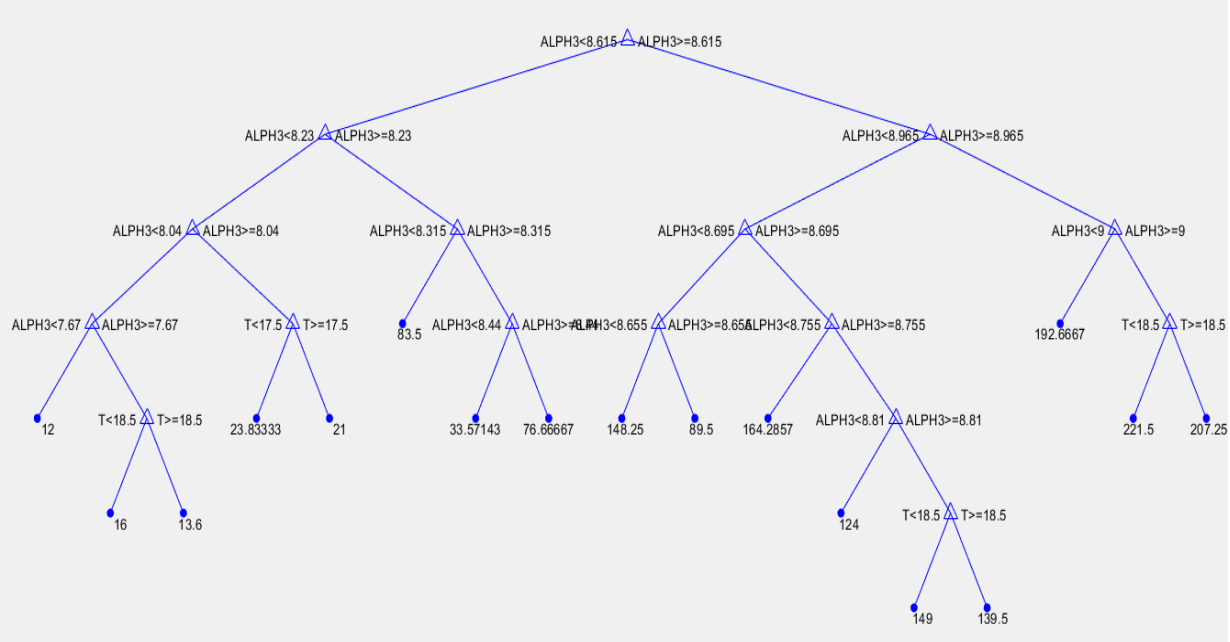


Figura 3.18. Arquitectura del modelo AR – estimación TDS muestra AL-3

La Figura 3.18 presenta la arquitectura del modelo AR y las correspondientes condiciones de decisión para los 33 nodos totales. La representación de los valores estimados y reales para el modelo AR se muestran en la Figura 3.19.

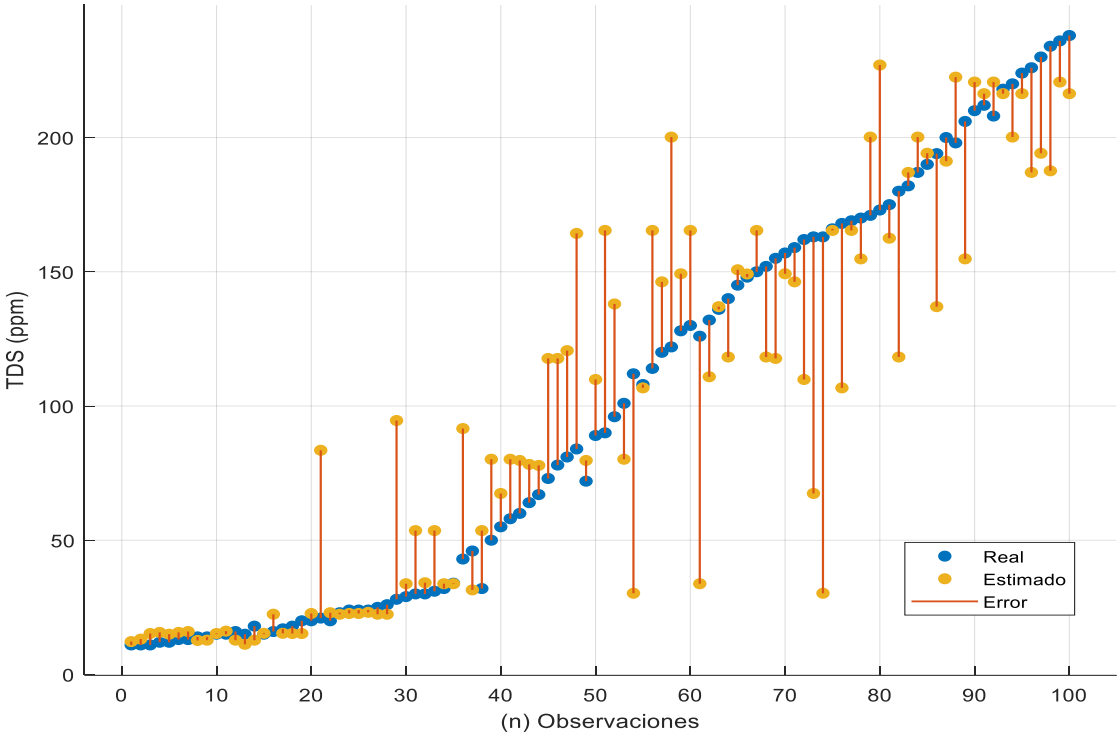


Figura 3.19. Representación gráfica modelo AR– estimación TDS muestra AL-3.

El modelo resultante para la técnica MVS en la estimación del TDS presenta las propiedades de la Figura 3.20.

```
RegressionSVM
          PredictorNames: {'ALPH3'  'T'}
            ResponseName: 'Y'
       CategoricalPredictors: []
        ResponseTransform: 'none'
                   Alpha: [82×1 double]
                    Beta: [51.0867 -4.8742]
                    Bias: 99.1609
         KernelParameters: [1×1 struct]
                      Mu: [8.5166 18.0300]
                   Sigma: [0.4911 1.1322]
          NumObservations: 100
           BoxConstraints: [100×1 double]
         ConvergenceInfo: [1×1 struct]
         IsSupportVector: [100×1 logical]
                  Solver: 'SMO'
```

Figura 3.20. Propiedades del modelo MVS – estimación TDS muestra AL-3

La representación de los valores estimados y reales para el modelo MVS se muestra en la Figura 3.21.

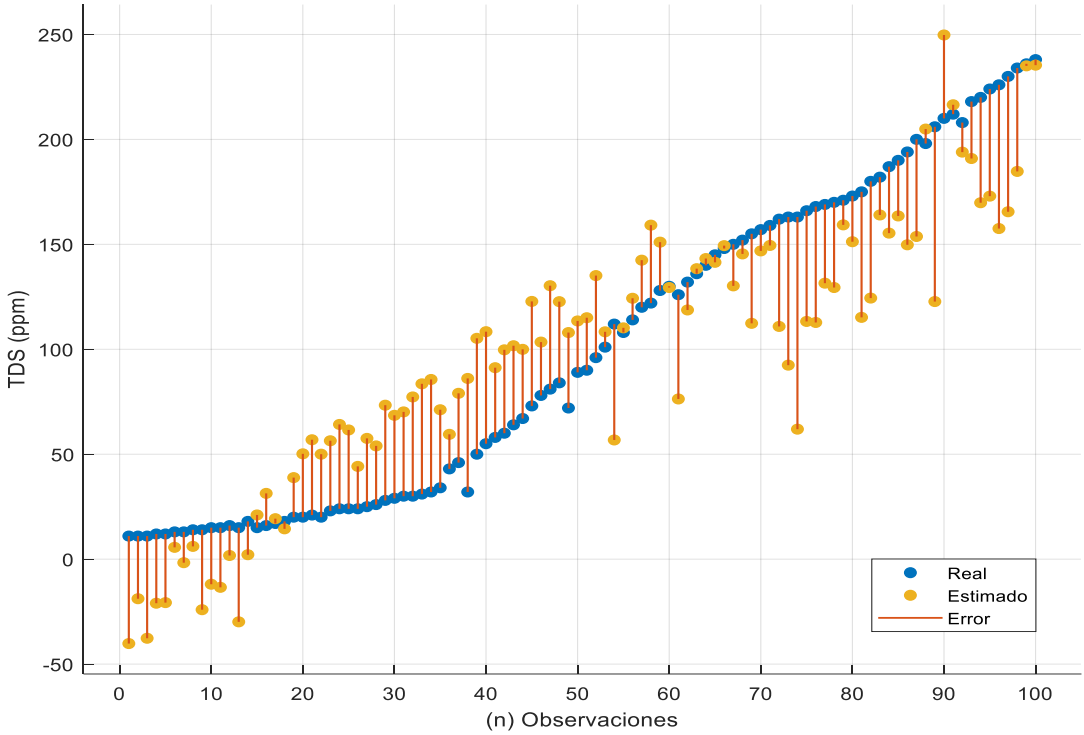


Figura 3.21.Representación gráfica modelo MVS – estimación TDS muestra AL-3.

Como se puede observar, solo es posible obtener una ecuación cerrada del modelo para la regresión lineal, para las otras técnicas la herramienta entrega una serie de características generales de una especie de “caja negra” que contiene los modelos, esta podría ser entonces una desventaja tanto para los AR como para MVS, lo que dificulta su comprensión (Mohammed, Longva y Seidu, 2018). Además, dado que se utiliza un método de re-muestreo, específicamente una validación cruzada de 5 iteraciones, no hay un conjunto único de parámetros para cada modelo, sino un conjunto de parámetros para cada iteración del proceso de validación cruzada. Por lo que los coeficientes por ejemplo para el caso de la RL, pueden cambiar dada la partición aleatoria de los datos. En este contexto, se prestará más atención al desarrollar un modelo que predice con precisión los valores de respuesta futuros que al comprender la importancia de los predictores o los valores de los parámetros del modelo (como en las tareas de modelado explicativo).

Para la evaluación del desempeño se compararon las métricas MAE, RMSE y R^2 (sección 2.3.1), para los modelos RL, AR y MVS en la estimación tanto del pH como de TDS. Los valores muestran que el modelo AR tuvo un mejor desempeño en la predicción de pH para la muestra AL-3. En la estimación del TDS el modelo de AR también presento un desempeño superior, la Tabla 3.6 contiene los valores de desempeño para cada modelo sobre la muestra analizada de agua lluvia.

Tabla 3.6.Resultados estimación pH y TDS - muestra AL-3.

Parámetro	**pH**			**TDS**		
Técnica	**RL**	**AR**	**MVS**	**RL**	**AR**	**MVS**
MAE	0.2059	0.1342	0.2010	31.559	22.625	31.946
RMSE	0.2431	0.1905	0.2542	36.866	34.586	37.642
$\mathbf{R^2}$	0.75	**0.85**	0.73	0.76	**0.79**	0.75

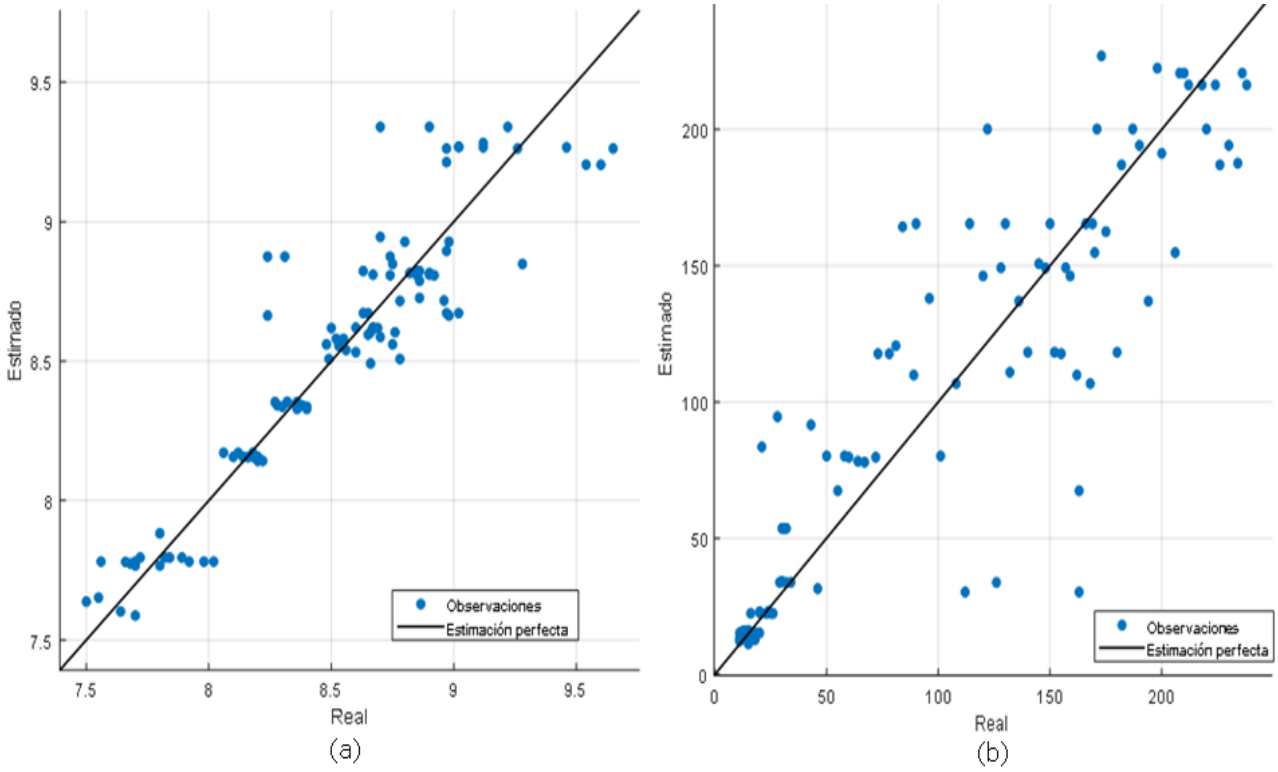


Figura 3.22.Gráficos de dispersión valores estimados vs reales para:(a) Estimación de pH - modelo AR, (b) Estimación de TDS – modelo AR - (muestra AL-3)

La Figura 3.22 muestra la representación gráfica de la dispersión entre los valores estimados y reales para los modelos de mejores resultados en la predicción de pH y TDS para el agua lluvia en la muestra AL-3. La línea diagonal corresponde a la línea de identidad (Estimado = Real, es decir, predicción perfecta).

Los términos *sobrestimación* y *subestimación* se utilizan en análisis de regresión para determinar los predichos que excedan el valor real (*sobreestimado* y *subestimado*) si el resultado es inferior a este. Al observar la Figura 3.22 se puede decir que para el modelo AR de pH, la mayoría de los valores se encuentran sobre la línea diagonal con algunos puntos a mayor distancia sobre ella. Para la Figura 3.22 (b) los puntos se encuentran más dispersos alrededor de la diagonal con mayor sobreestimación de los valores, lo que refleja la capacidad del modelo para explicar el 79% de la varianza.

3.5.2. Modelos para la muestra AA-1 (agua de acueducto)

Siguiendo el procedimiento realizado sobre la muestra AL-3 en la sección anterior, se construyen los modelos RL, AR y MVS, para la estimación del pH en la muestra AA-1 para agua de acueducto estableciendo como entradas los parámetros TDS y temperatura respectivos.

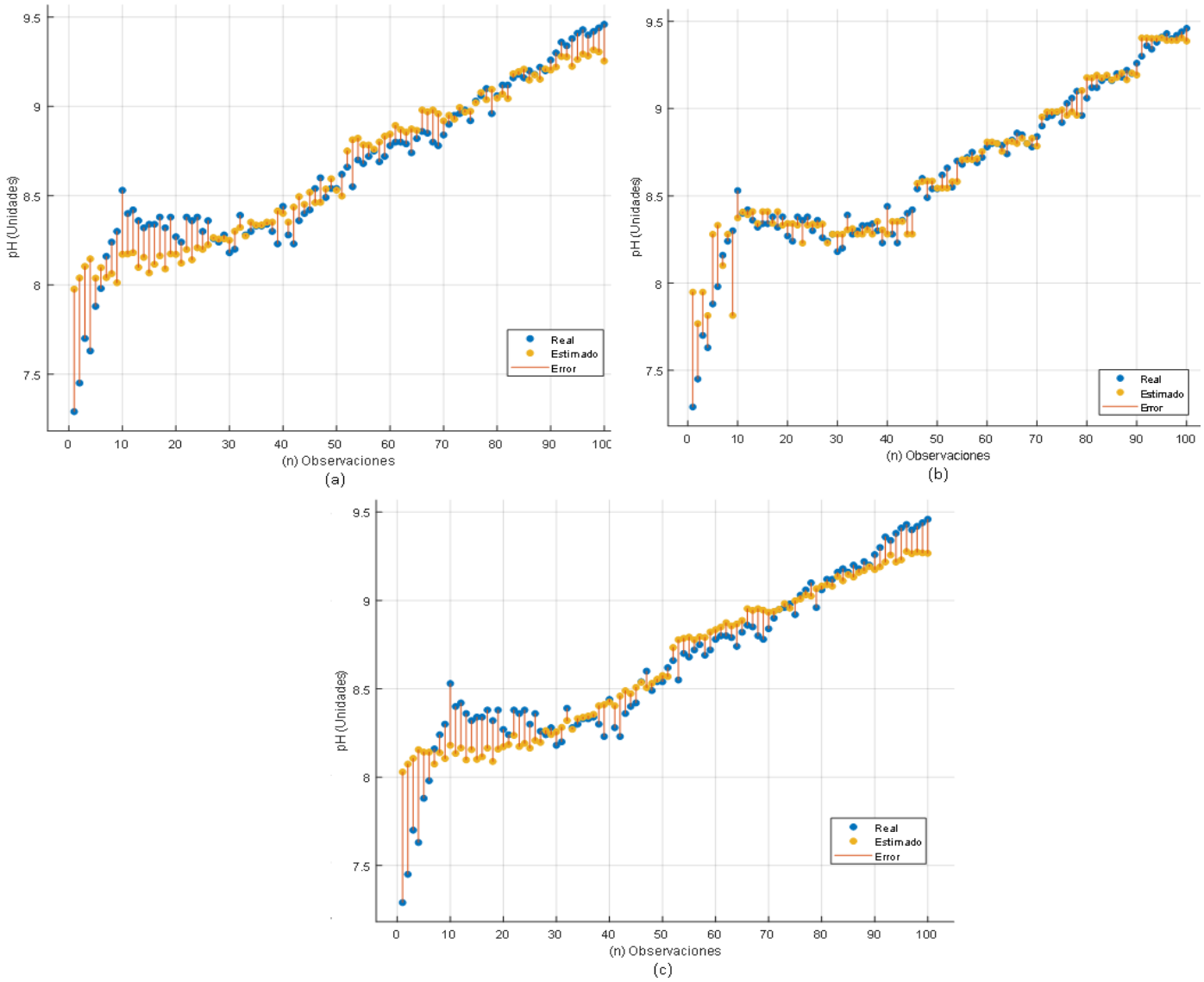


Figura 3.23.Representación gráfica modelos: (a) RL, (b) AR, (c) MVS – estimación pH muestra AA-1.

Los modelos obtenidos para la estimación del TDS, teniendo como parámetros de entrada pH y temperatura sobre la muestra AA-1 se presentan en la Figura 3.24.

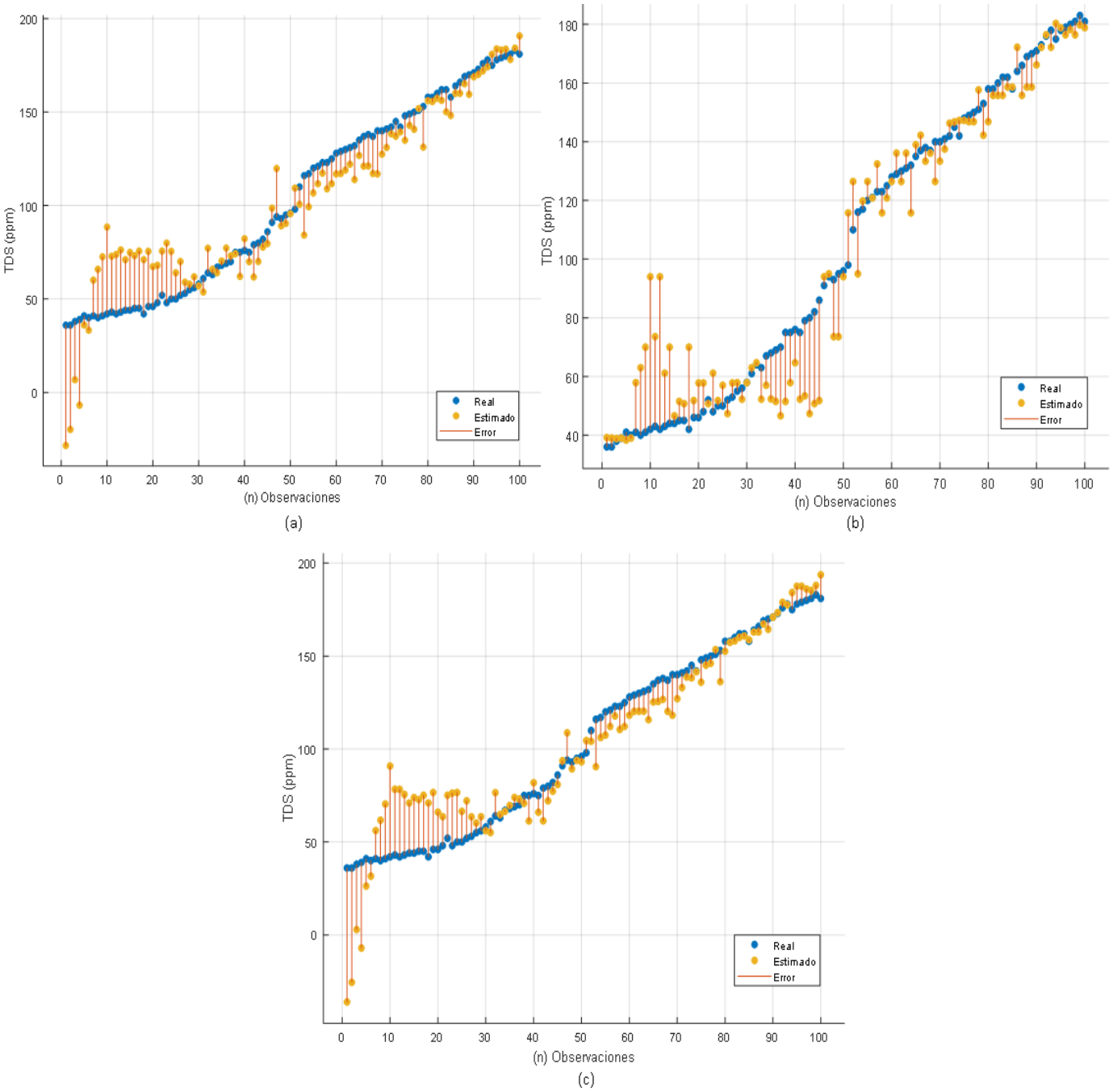


Figura 3.24.Representación gráfica modelos: (a) RL, (b) AR, (c) MVS – estimación TDS muestra AA-1.

Los valores de desempeño de los modelos obtenidos para la estimación de los parámetros se presentan en la Tabla 3.7, en la cual, los modelos de AR fueron superiores en desempeño tanto para pH y como para TDS.

Tabla 3.7.Resultados estimación pH y TDS - muestra AA-1.

Parámetro	pH			TDS		
Técnica	**RL**	**AR**	**MVS**	**RL**	**AR**	**MVS**
MAE	0.1220	0.0759	0.1225	13.816	9.8574	13.109
RMSE	0.1690	0.1248	0.1720	18.594	14.439	18.603
$\mathbf{R^2}$	0.86	**0.93**	0.86	0.86	**0.91**	0.86

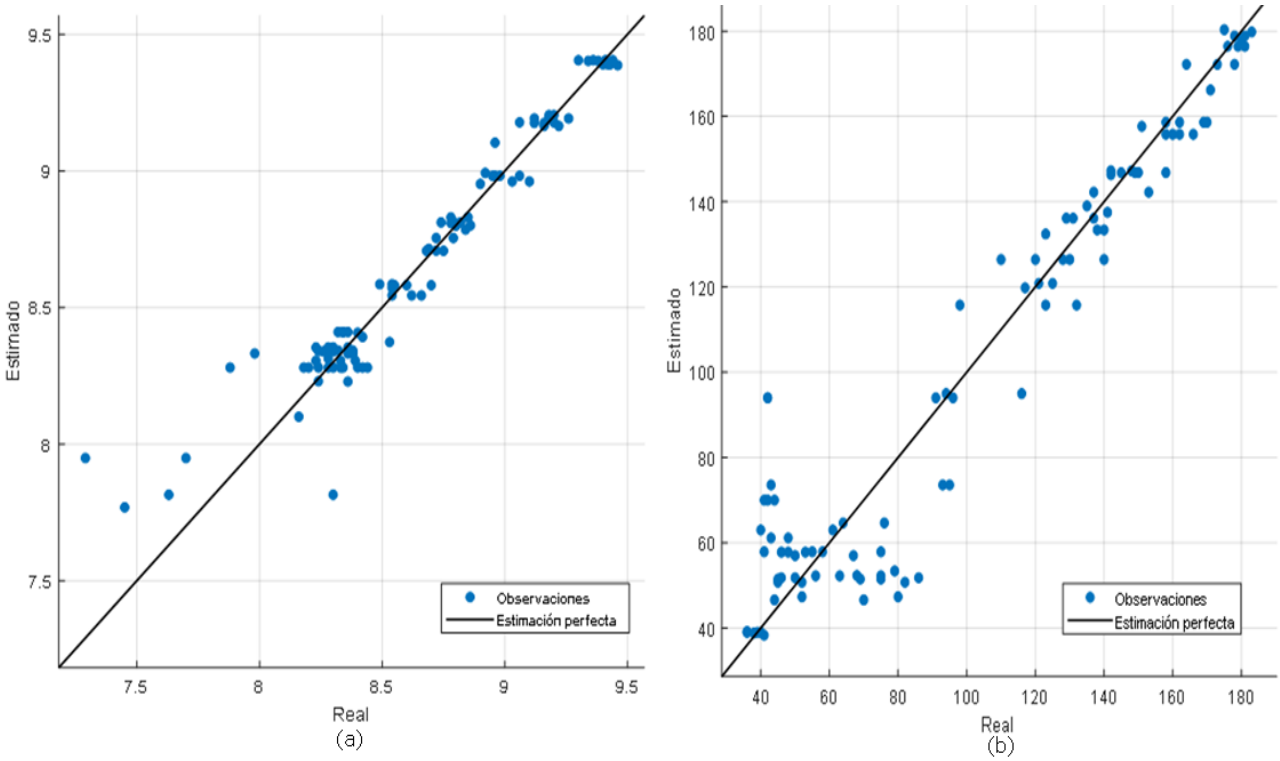


Figura 3.25.Gráficos de dispersión valores estimados vs reales para:(a) Estimación de pH - modelo AR, (b) Estimación de TDS – modelo AR - (muestra AA-1).

En el gráfico de dispersión de la Figura 3.25 (a) se aprecian algunos pocos puntos alejados de la diagonal de estimación perfecta, los valores restantes sin embargo se ubican sobre la línea. Para la Figura 3.25 (b) los valores se encuentran cercanos a la diagonal con mayor subestimación en algunos puntos de la figura.

3.5.3. Modelos para la muestra AE-2 (agua envasada)

Para la muestra AE-2 de agua envasada, se construyeron los modelos para la estimación del pH como para el TDS. Los modelos obtenidos en la estimación del pH sobre esta muestra se presentan en la Figura 3.26.

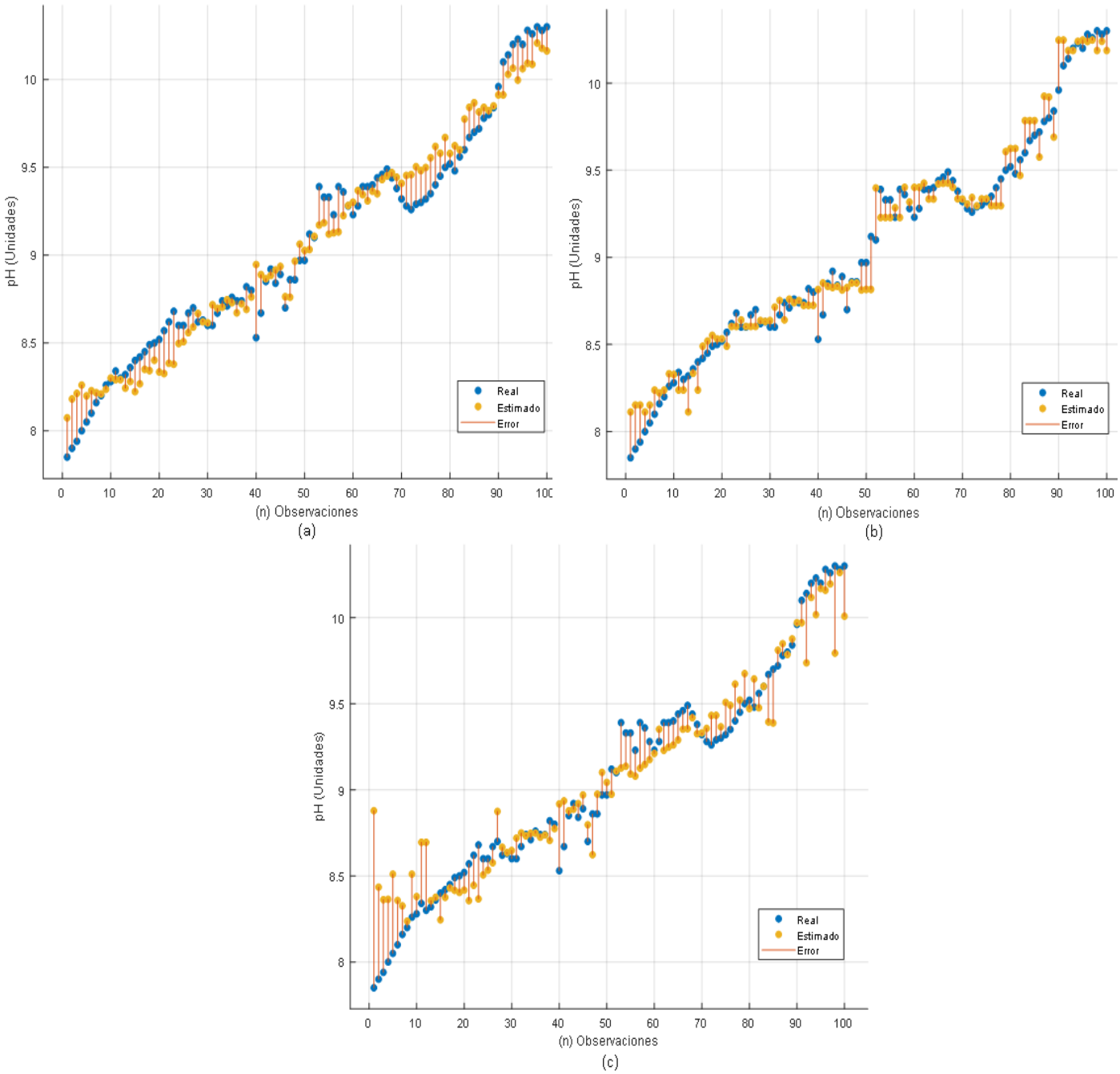


Figura 3.26.Representación gráfica modelos: (a) RL, (b) AR, (c) MVS – estimación pH muestra AE-2.

Los modelos obtenidos para la estimación del TDS, teniendo como parámetros de entrada pH y temperatura sobre la muestra AE-2 se presentan en la Figura 3.27.

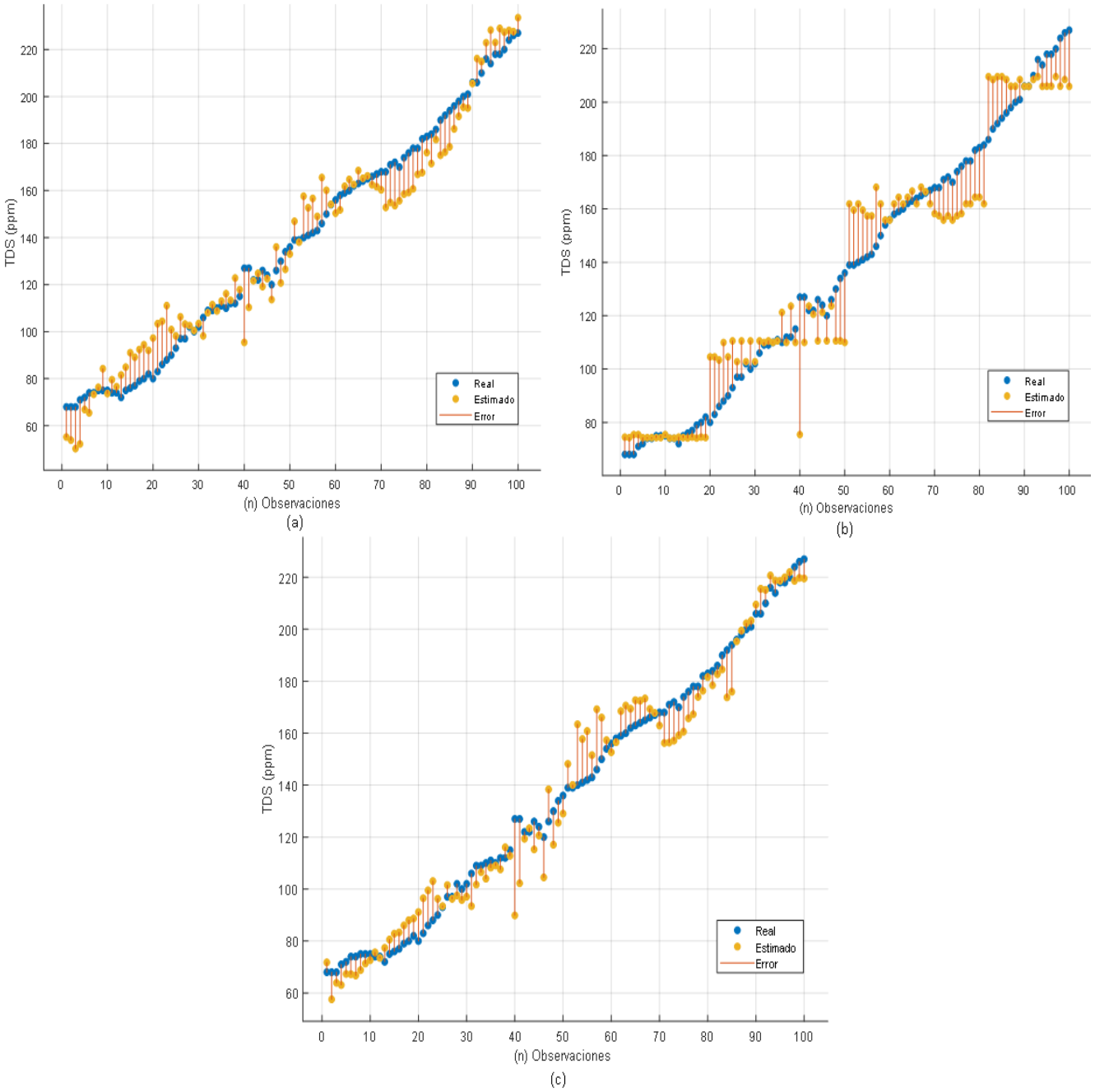


Figura 3.27.Representación gráfica modelos: (a) RL, (b) AR, (c) MVS – estimación TDS muestra AE-2.

Los valores de desempeño de los modelos para la estimación del pH y TDS se muestran en la Tabla 3.8 en donde los modelos de AR para pH y MVS en TDS tuvieron el mejor desempeño.

Tabla 3.8.Resultados estimación pH y TDS - muestra AE-2.

Parámetro	pH			TDS		
Técnica	**RL**	**AR**	**MVS**	**RL**	**AR**	**MVS**
MAE	0.1149	0.0889	0.1490	8.6846	9.9589	7.5823
RMSE	0.1404	0.1139	0.2110	10.707	13.274	9.7623
$\mathbf{R^2}$	0.95	**0.97**	0.89	0.95	0.93	**0.96**

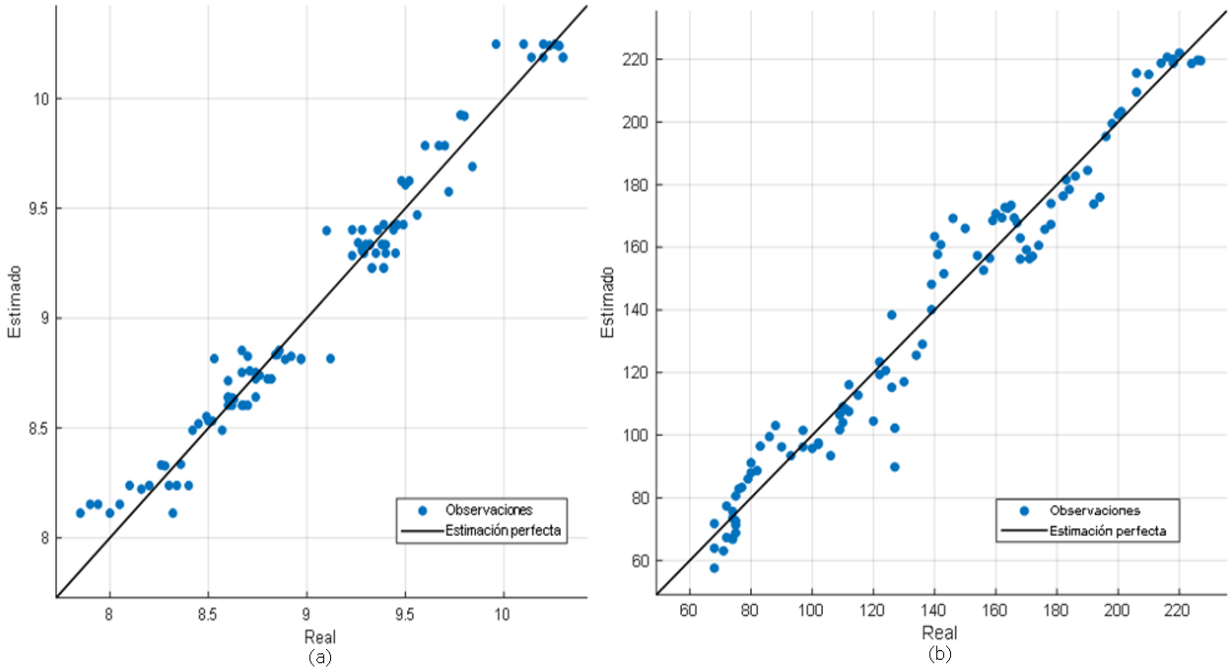


Figura 3.28.Gráficos de dispersión valores estimados vs reales para:(a) Estimación de pH - modelo AR, (b) Estimación de TDS – modelo MVS - (muestra AE-2)

En el gráfico de dispersión de la Figura 3.28 se observan los puntos muy cercanos a línea diagonal en los gráficos (a) y (b) lo que se ajusta a la los valores de R^2 de 97% y 96% respectivamente.

Para la estimación de pH y TDS sobre las muestras seleccionadas, se evaluaron 18 modelos en total, obteniendo valores para $\mathbf{R^2}$ superiores a 0.73 en los casos anteriores, siendo la técnica de árboles de regresión superior en cinco de los seis casos evaluados alcanzando valores de 97% para $\mathbf{R^2}$, seguido por la técnica MVS que supero a la RL y AR en la estimación de TDS en la muestra AE-2.

La superioridad de los modelos de AR en esta fase, puede asociarse al procedimiento de poda aplicado por el árbol de regresión durante el entrenamiento, el cual elimina las divisiones redundantes para optimizar la dimensión del árbol. Esto reduce la complejidad del modelo y el ajuste excesivo, y mejora las capacidades de predicción y generalización del modelo (Granata *et al.*, 2017).

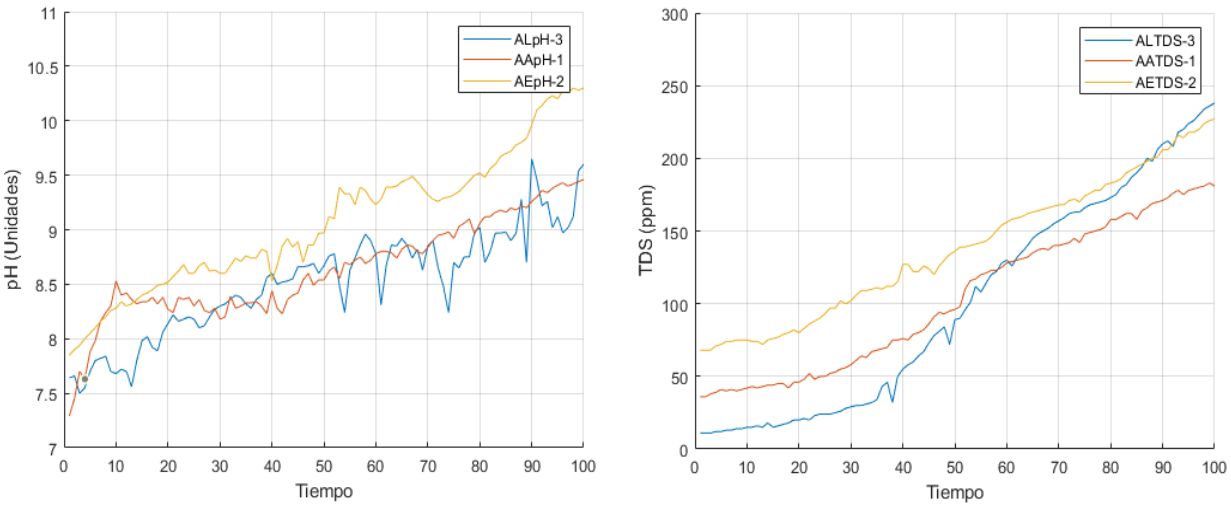
De manera general se puede observar que los modelos y técnicas evaluadas son capaces de predecir los parámetros de pH y TDS sin mayor dificultad para los

diferentes tipos de agua almacenada, los valores de significancia (p-valor) observados en los coeficientes y características del modelo RL aportan confiabilidad a las estimaciones realizadas en este experimento. Resultados similares se obtuvieron en las muestras restantes para agua lluvia, acueducto y envasada (Anexo C), lo cual brinda mayor confiabilidad al procedimiento y resultados.

3.5.4. Análisis de series de tiempo

Las series de tiempo permiten analizar el comportamiento y evolución de los parámetros de calidad en una muestra de agua. La Figura 3.29 presenta una comparativa de las series de tiempo (días) para los parámetros pH y TDS de las muestras AL-3, AA-1 y AE-2, en la Figura 3.29 a), se observan valores iniciales del pH de 7.50, 7.29 y 7.85 alcanzando los valores de 9.65, 9.46 y 10.30 respectivamente, desde del día 1 al 100, de acuerdo al análisis de características estadísticas de las muestras presentadas en la Tabla 3.2, es posible apreciar una lenta evolución del pH ya que sus valores de desviación estándar son bajos menores a 1 lo que indica que sus datos se encuentran poco dispersos y cercanos al valor medio.

Para el TDS por el contrario (Figura 3.29 b)), sus valores se extienden sobre un rango más amplio, lo que coincide con sus altos valores de desviación estándar (Tabla 3.2), el TDS presenta valores iniciales de 11, 36 y 68 (ppm) llegando a valores de 238, 183 y 227 (ppm) para el día 100, un periodo aproximado de tres meses (tiempo de recolección de datos).



a) Series Temporales pH b) Series Temporales TDS

Figura 3.29.Series Temporales - Muestras AL-3, AA-1, AE-2.

Como se ha mencionado, las mediciones de temperatura se realizaron sobre las muestras bajo las mismas condiciones por lo que se obtuvo un solo rango de valores para este parámetro. La Figura 3.30 muestra la variación de la temperatura

diariamente para un periodo aproximado de tres meses, se observa que los valores oscilan entre los 16°C y 22°C, con un valor medio de 18°C y con una baja dispersión de acuerdo a su desviación estándar de 1.13 (Tabla 3.2).

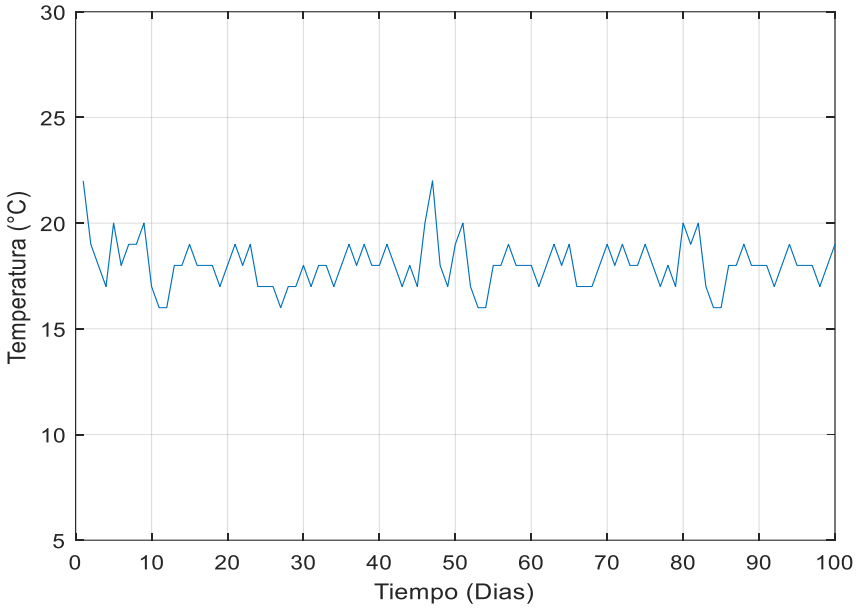


Figura 3.30.Series Temporales - Temperatura.

En las Figuras 3.29 y 3.30 se observa el comportamiento de los parámetros sobre una escala de tiempo diaria, cuya tendencia es creciente para las diferentes muestras. Si bien los cambios en los parámetros tienen una evolución temporal diferente, los modelos construidos son capaces de ajustar las estimaciones y lograr buenos resultados. Para obtener un valor general de la variación de calidad del agua de las muestras almacenadas, en el siguiente capítulo se utiliza un índice conformado por los parámetros pH, TDS y temperatura, el cual permitió evaluar las condiciones del agua en una escala general de 0 a 100.

Capítulo 4

Validación de Resultados

En este capítulo, se tratará la construcción del índice de calidad de agua ICA a partir de los parámetros pH, TDS y temperatura para las muestras del caso de estudio. El indicador ICA permite analizar la evolución temporal de la calidad del agua de una manera más general y práctica. En este sentido, se muestran los pasos para la construcción del ICA y la aplicación de las técnicas de regresión lineal, máquinas de vectores de soporte y árboles de regresión en la predicción de este índice, permitiendo tener así, un estimado temporal de la calidad del agua en almacenamiento. Finalmente se seleccionará el modelo de estimación ICA de mejor desempeño para el proceso de validación, el cual consiste en la predicción del índice con datos de otras muestras y clases de agua.

4.1. Índice de Calidad de Agua (ICA)

Un índice permite cuantificar e interpretar los cambios de un fenómeno a través de los datos que lo describen. Para el caso del recurso hídrico, existen diferentes indicadores que aportan información sobre el estado del agua, entre los más comunes se encuentran los índices de calidad (ICA) e índices de contaminación (ICO) del agua.

En la construcción del índice ICA se pueden identificar tres pasos generales (Lisseth Guzmán, Nava y Díaz, 2015):

- ***Selección de parámetros:*** se deben seleccionar dos o más parámetros (fisicoquímicos – biológicos).
- ***Determinación de los valores escalados para cada parámetro:*** los valores escalados para cada parámetro se calculan a partir de los datos medidos de acuerdo a sus respectivas curvas de calidad, este valor es multiplicado por un peso según el grado de importancia del parámetro para el caso que se analiza.
- ***Determinación del índice mediante formula de agregación:*** en donde finalmente se aplica la fórmula de agregación que suma los resultados de los valores obtenidos para todos los parámetros.

Los valores que el indicador puede llegar a tomar, se encuentran clasificados en categorías. En una escala de 0 a 100 se ha asignado un color como señal de alerta

y las categoría van desde *Muy Mala* para valores cercanos a 0, hasta *Excelente* para valores cercanos a 100(Wills y Irvine, 1996). La escala de clasificación para ICA se presenta en la Tabla 4.1.

Tabla 4.1.Escala de clasificación ICA.

Descriptor	Escala numérica	Color
Excelente	91-100	Azul
Buena	71-90	Verde
Media	51-70	Amarillo
Mala	26-50	Naranja
Muy mala	0-25	Rojo

La fórmula de agregación para calcular el índice (ICA) (Lisseth Guzmán, Nava y Díaz, 2015) se expresa matemáticamente como sigue:

$$ICA = \sum_{i=1}^{n} Q_i \cdot w_i \tag{4.1}$$

donde:

El subíndice i, identifica a cada uno de los parámetros.
w_i, pesos relativos asignados a cada parámetro i y ponderados entre 0 y 1 de tal forma que la sumatoria sea igual a uno.
Q_i, Es el valor calculado del parámetro i (Obtenido al aplicar la curva de calidad correspondiente).

Para determinar el ICA en las muestras de estudio, se utilizan los parámetros de pH, TDS y temperatura, por lo que $i = 1,2,3$ con $n = 3$. Los pesos relativos se han seleccionado teniendo en cuenta los análisis previos de correlación y significancia estadística, la Tabla 4.2 contiene la asignación correspondiente.

Tabla 4.2.Pesos relativos para cada parámetro del ICA.

i	Parámetro	Peso de importancia (w)
1	pH	0.4
2	TDS	0.4
3	Temperatura	0.2

Las curvas de calidad ampliamente utilizadas en la construcción de ICA y desarrolladas en (Lisseth Guzmán, Nava y Díaz, 2015), indican en el eje de la ordenada la calidad del agua en una escala de 0 a 100; en la abscisa se define la escala en unidades del parámetro en particular. Las curvas de pH, TDS y temperatura, necesarias para el cálculo de ICA se presentan en las Figuras 4.1, 4.2 y 4.3

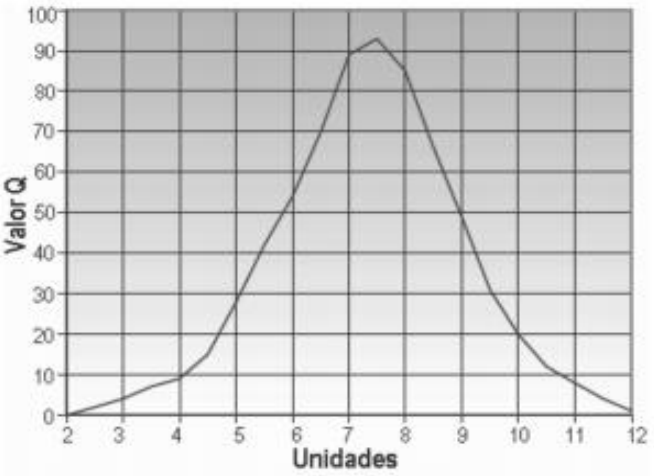


Figura 4.1.Curva de calidad pH,
Tomado de *Lisseth Guzmán, Nava y Díaz, 2015*

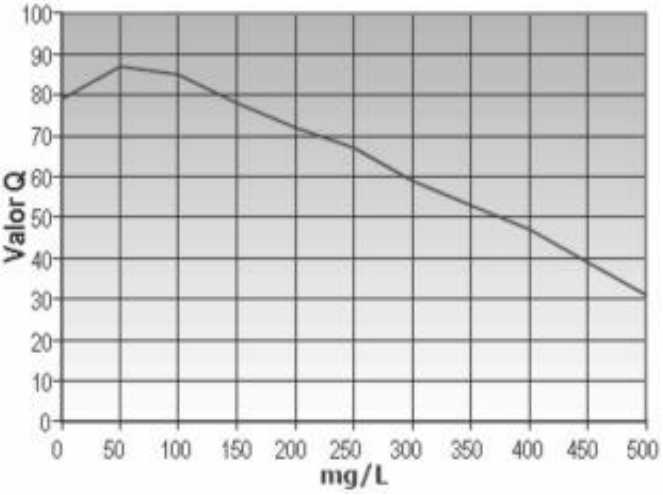


Figura 4.2.Curva de calidad TDS.
Tomado de *Lisseth Guzmán, Nava y Díaz, 2015*

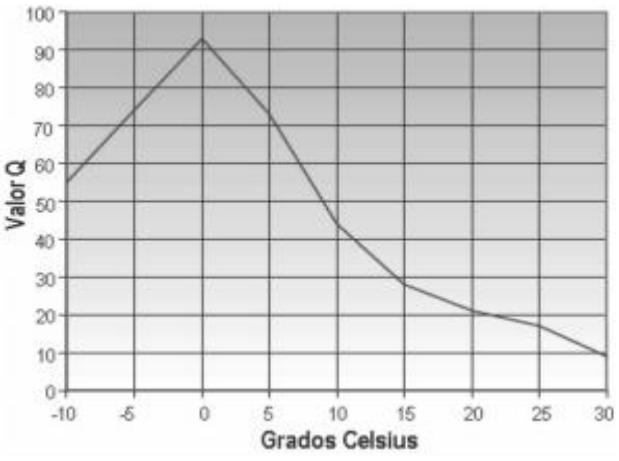


Figura 4.3.Curva de calidad Temperatura.
Tomado de *Lisseth Guzmán, Nava y Díaz, 2015*

Para cada valor medido de los parámetros, se aplica el procedimiento general para calcular los Q_i del ICA que se describe a continuación:

Si el valor de pH es menor o igual a 2 unidades el (Q_1) es igual a 2, sí el valor de pH es mayor o igual a 10 unidades el (Q_1) es igual a 3. Si el valor de pH esta entre 2 y 10 se busca el valor en el eje de (X) en la Figura 4.1 y se procede a interpolar al valor en el eje de las (Y). El valor encontrado es el (Q_1) de pH y se procede a multiplicarlo por el peso w_1.

Teniendo en cuenta que en las unidades de TDS 1ppm equivale a 1mg/L, Si los Sólidos disueltos Totales son mayores de 500 mg/L el (Q_2) es igual a 3, si es menor de 500 mg/L, buscar el valor en el eje de (X) en la Figura 4.2 y se procede a interpolar al valor en el eje de las (Y). El valor encontrado es el (Q_2) y se procede a multiplicarlo por el peso w_2.

Si el valor de temperatura es mayor de 15°C el (Q_3) es igual a 9. Si el valor obtenido es menor de 15°C, buscar el valor en el eje de (X) en la Figura 4.3 y se procede a interpolar al valor en el eje de las (Y). El valor encontrado es el (Q_3) de temperatura y se procede a multiplicarlo por el peso w_3.

Una vez calculados los valores de Q_i y multiplicados por los pesos w_i, se suman los resultados para encontrar el ICA correspondiente, por ejemplo, el ICA de la muestra AL-3 para el día uno a partir de los datos medidos de los tres parámetros se muestra en la Tabla 4.3

Tabla 4.3.Ejemplo - cálculo de ICA para la muestra AL-3 (día 1).

i	**Parámetro**	**Datos medidos día 1**	Q_i	w_i	$Q_i \cdot w_i$
1	pH	7.64 (Unid. de pH)	95.82	0.4	38.328
2	TDS	11 (ppm)	83.447	0.4	33.379
3	T	22(°C)	21.155	0.2	4.231
ICA (día1) $\sum Q_i \cdot w_i$					**75.938**

El ICA obtenido para el día 1 es entonces de 75.94 con lo cual se clasifica en la categoría de *Buena* calidad según el rango de valores de la Tabla 4.3. De esta manera se procede a realizar el cálculo de los ICA para cada día con los datos correspondientes y así poder estudiar el comportamiento de calidad del agua en el tiempo.

4.2. ICA para Muestras de Agua

Con la ayuda del software Matlab para procesar los datos y calcular los valores Q_i de las gráficas, se aplicó el procedimiento anterior para el cálculo del ICA sobre las muestras AL-3 de agua lluvia, AA-1 de agua de acueducto y AE-2 de agua envasada, las Figuras 4.4, 4.5 y 4.6 muestran los valores de calidad de agua para cada día

representando una distribución temporal de la calidad del agua en las muestras a lo largo del experimento.

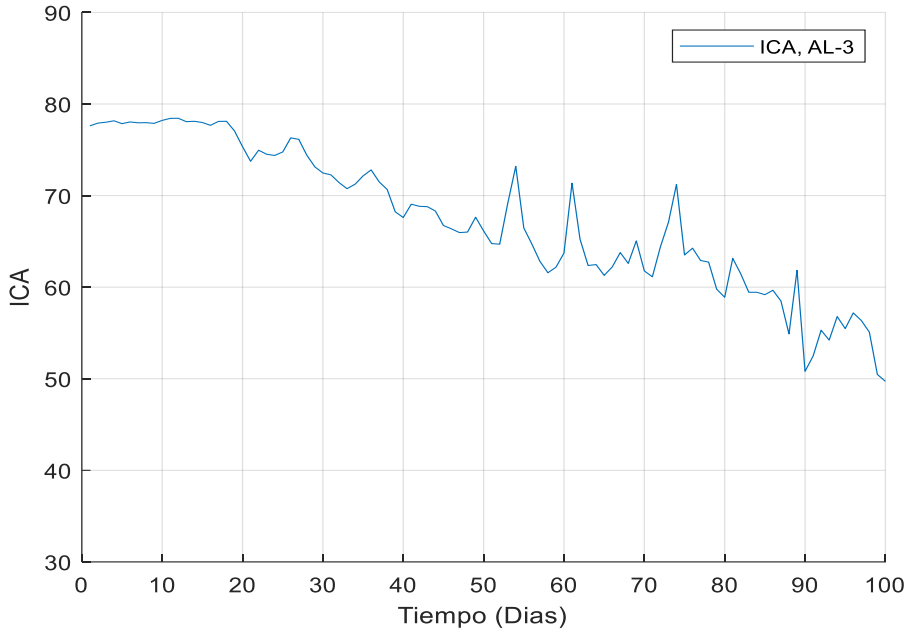


Figura 4.4. Serie de tiempo ICA - muestra AL-3.

Como se observa en la Figura 4.4, la muestra de agua lluvia se clasifica como *Buena* con valores de ICA aproximado de 76 para los primeros días disminuyendo hasta alcanzar un valor de 49 que representa una clasificación de calidad *Mala* para el día N° 100. La Figura 4.5 representa el comportamiento del ICA para la muestra AA-1.

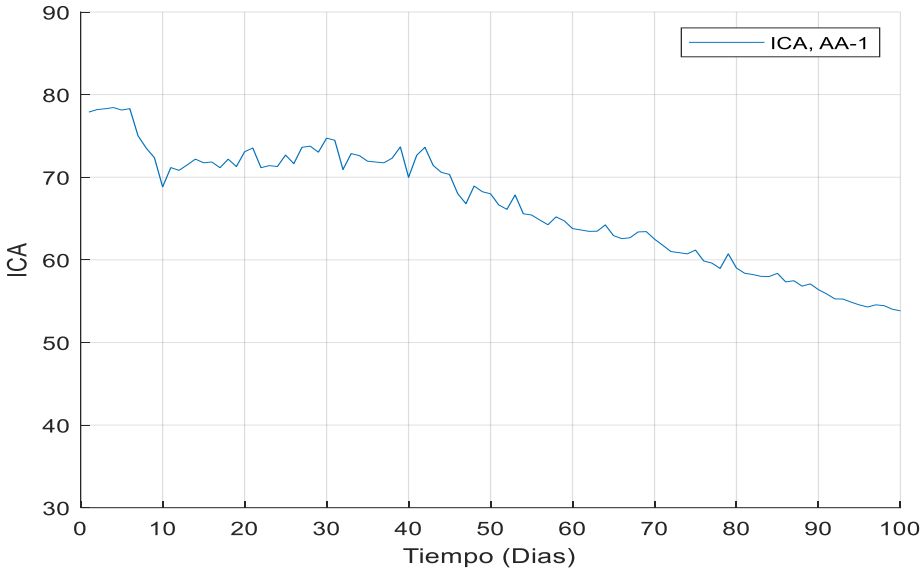


Figura 4.5. Serie de tiempo ICA- muestra AA-1.

En la figura anterior se observa que el ICA pasa de una clasificación *Buena* con valores de índice de 77 aproximadamente, a *Media* con un valor de 53 hacia el final

del experimento. El comportamiento de la calidad del agua para la muestra AE-2 se presenta en la Figura 4.6.

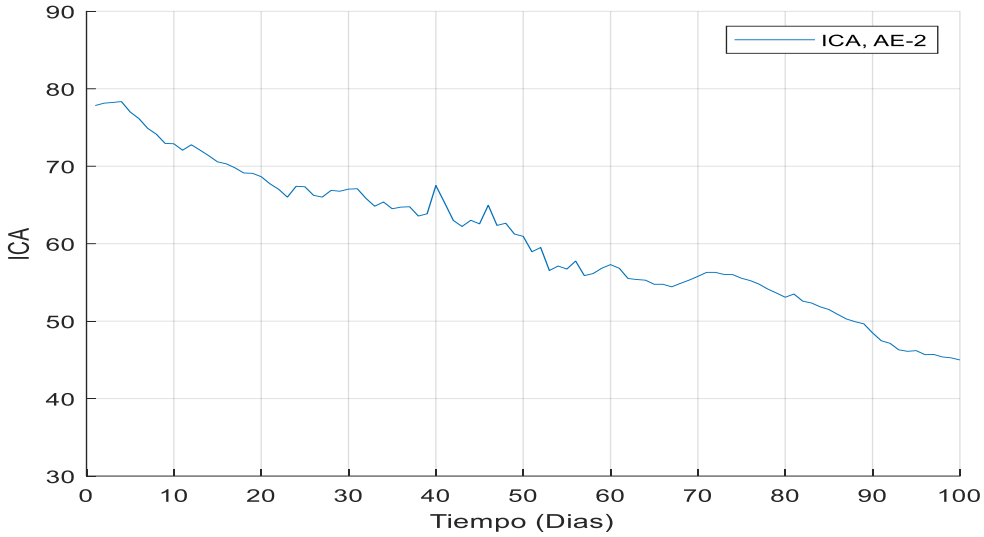


Figura 4.6. Serie de tiempo ICA - muestra AE-2.

El comportamiento del ICA en la muestra de agua envasada pasa también de una clasificación *Buena* a *Mala*, con valores de ICA cercanos a 45 hacia el final del experimento.

4.3. Estimación del Índice de Calidad de Agua (ICA)

Para la estimación de ICA y siguiendo la metodología del Capítulo 3 en la predicción del pH y TDS, se estimará un índice por cada muestra y se utiliza como entradas del modelo, los valores de los parámetros pH, TDS y temperatura en el software Matlab para construir los modelos de regresión lineal, arboles de regresión y máquinas de vectores de soporte. La comparación de sus desempeños se realiza mediante las métricas de evaluación MAE, RMSE y R^2. Utilizando una validación cruzada de 5 iteraciones, la Tabla 4.4 presenta los resultados de estimación del ICA para la muestra AL-3.

Tabla 4.4.Resultados estimación del ICA - muestra AL-3.

Parámetro	**ICA**		
Técnica	**RL**	**AR**	**MVS**
MAE	0.7918	0.8839	2.5265
RMSE	1.3614	1.2533	4.0978
$\mathbf{R^2}$	0.97	0.97	0.73

De acuerdo a los valores obtenidos, el mejor desempeño se obtuvo con los modelos de RL y AR con valores de R^2 de 0.97, RMSE 1.3614 y 1.2533, y MAE 0.7918 y 0.8839, respectivamente. La Figura 4.7 muestra la representación gráfica de los

valores reales y estimados del ICA para el modelo AR que presentó un mejor desempeño con un menor valor de RMSE.

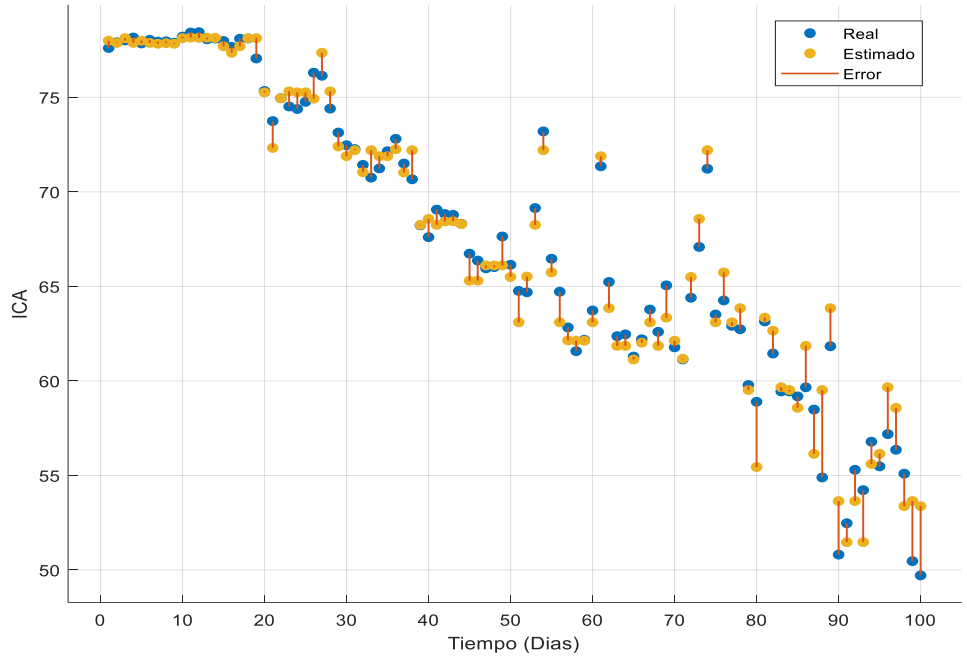


Figura 4.7. Estimación ICA muestra AL-3- Modelo AR.

Dado que el tiempo está representado en días, en la Figura 4.7 se observa que el ICA para la muestra de agua lluvia, alcanza una clasificación *Media* aproximadamente para el día 40 y una clasificación de Mala hacia el día 80.

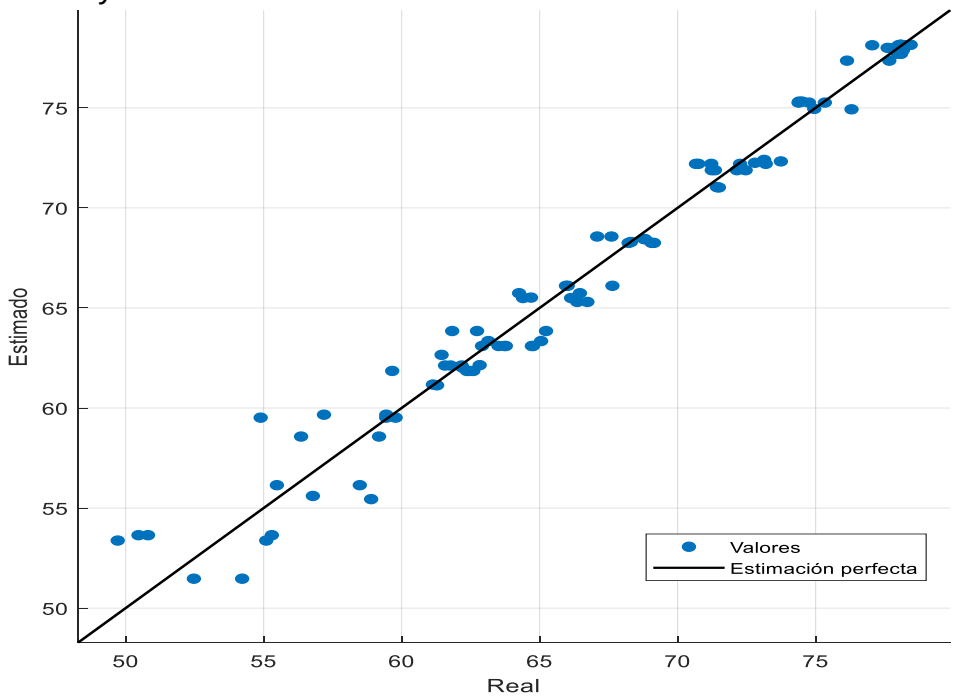


Figura 4.8. Gráficos de dispersión valores IC A estimados vs reales- Modelo AR muestra AL-3.

De acuerdo a la Figura 4.8, el modelo sobreestima el 43% de los datos y subestima el 57%, sin embargo, los valores estimados se encuentran muy cercanos a línea diagonal lo que reduce el error de predicción y refleja el valor de 97% de R^2. La Tabla 4.5 contiene los resultados de los modelos RL, AR y MVS para la muestra AA-1.

Tabla 4.5.Resultados estimación del ICA - muestra AA-1.

Parámetro	ICA		
Técnica	**RL**	**AR**	**MVS**
MAE	0.5071	0.7151	1.9534
RMSE	1.2963	0.9823	3.403
$\mathbf{R^2}$	0.97	0.98	0.77

En la estimacion del ICA para la muestra AA-1 el modelo AR tuvo un mejor desempeño con valores de 0.98 para R^2, 0.9257 para RMSE y 0.6627 para MAE. En la Figura 4.9 se observa la representación gráfica de los valores reales y estimados del ICA para el modelo AR que obtuvo un desempeño superior.

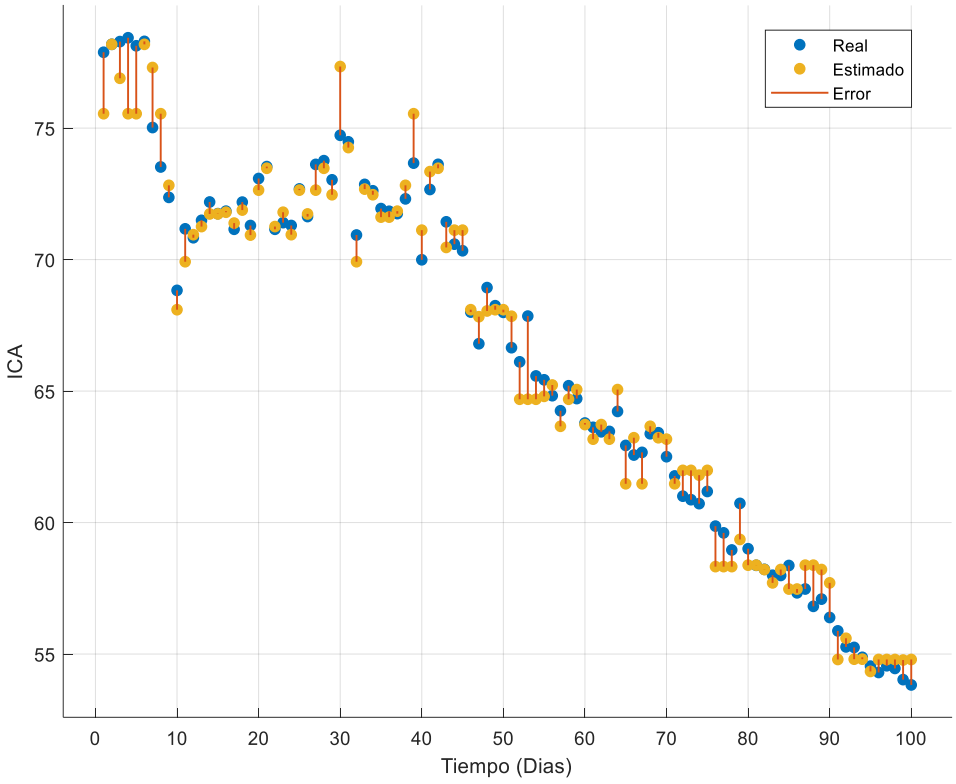


Figura 4.9. Estimación ICA para muestra AA-1.

El ICA para la muestra de agua de acueducto alcanza una clasificación *Media* hacia el día 45 y *Mala* alrededor del día 85.

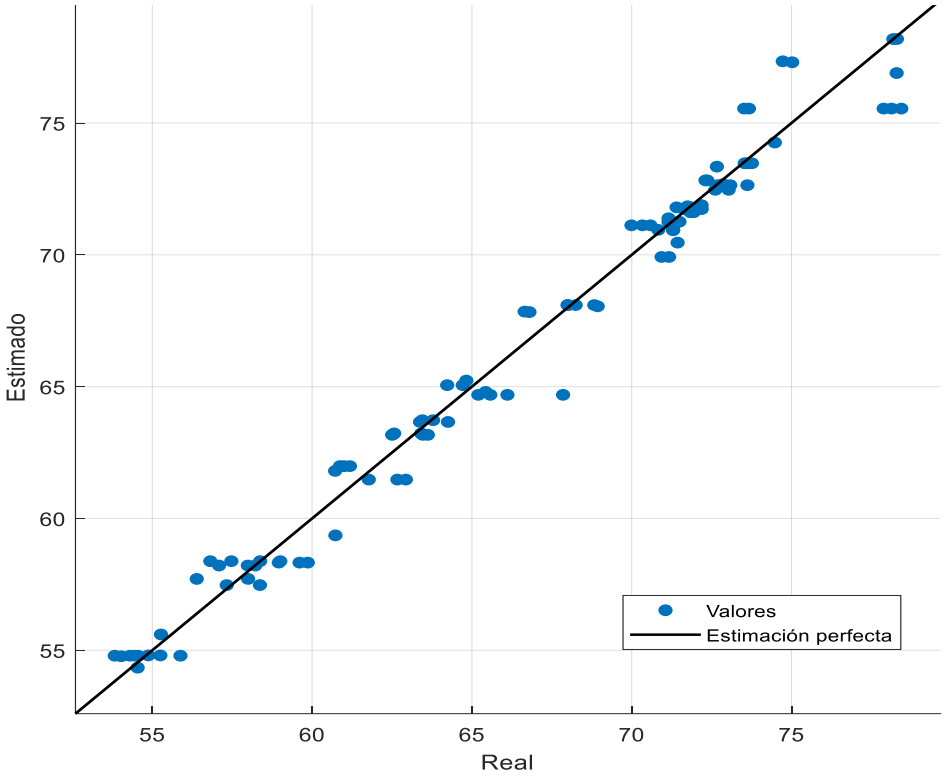


Figura 4.10.Gráficos de dispersión valores ICA estimados vs reales- Modelo AR muestra AA-1.

Los valores estimados se distribuyen muy cercanos a la línea diagonal de estimación perfecta demostrando la buena aproximación a los datos reales con un R^2 de 98%, el modelo sobreestima el 44% y subestima el 56% de los datos. Los resultados de la estimación del ICA para la muestra AE-2 se presentan en la Tabla 4.6.

Tabla 4.6.Resultados estimación del ICA - muestra AE-2.

Parámetro	**ICA**		
Técnica	**RL**	**AR**	**MVS**
MAE	0.8633	1.7495	3.1682
RMSE	1.3021	2.1797	5.507
$\mathbf{R^2}$	0.98	0.94	0.62

En la Tabla 4.6 se observa que el modelo que presento un mejor desempeño fue RL con valores de 0.98 para R^2,1.3019 para RMSE y 0.8611 para MAE. La Figura 4.11 muestra la representación gráfica de los valores reales y estimados del ICA para el modelo RL.

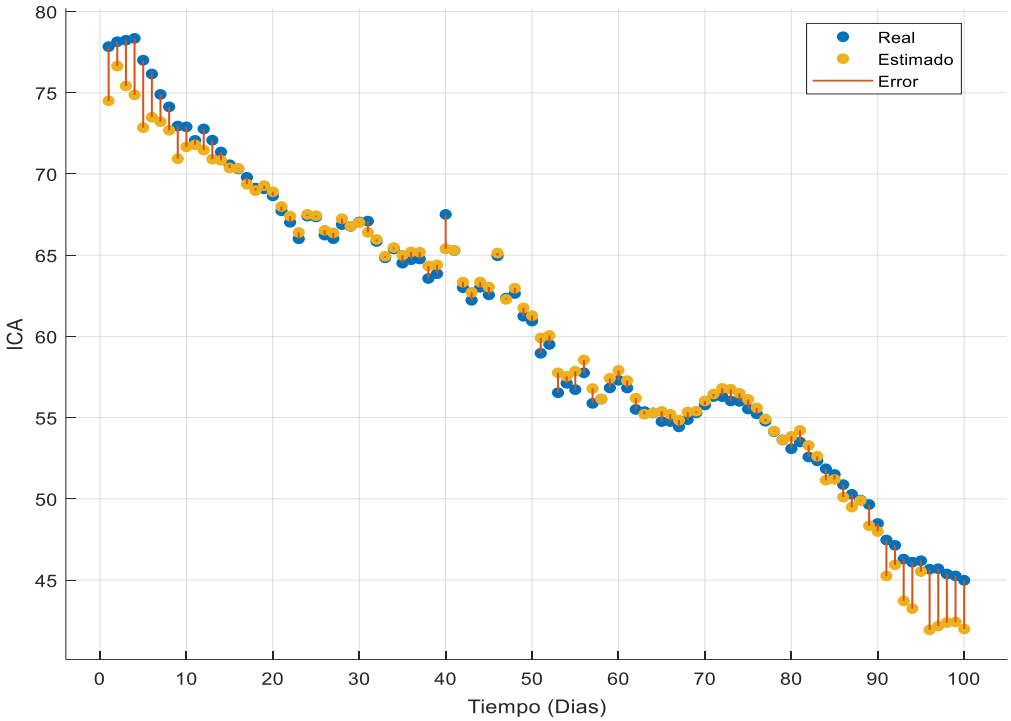


Figura 4.11. Estimación ICA para muestra AE-2.

De la Figura 4.11 se puede analizar, que se obtiene un ICA con clasificación *Media* hacia el día 15 y *Mala* alrededor del día 85.

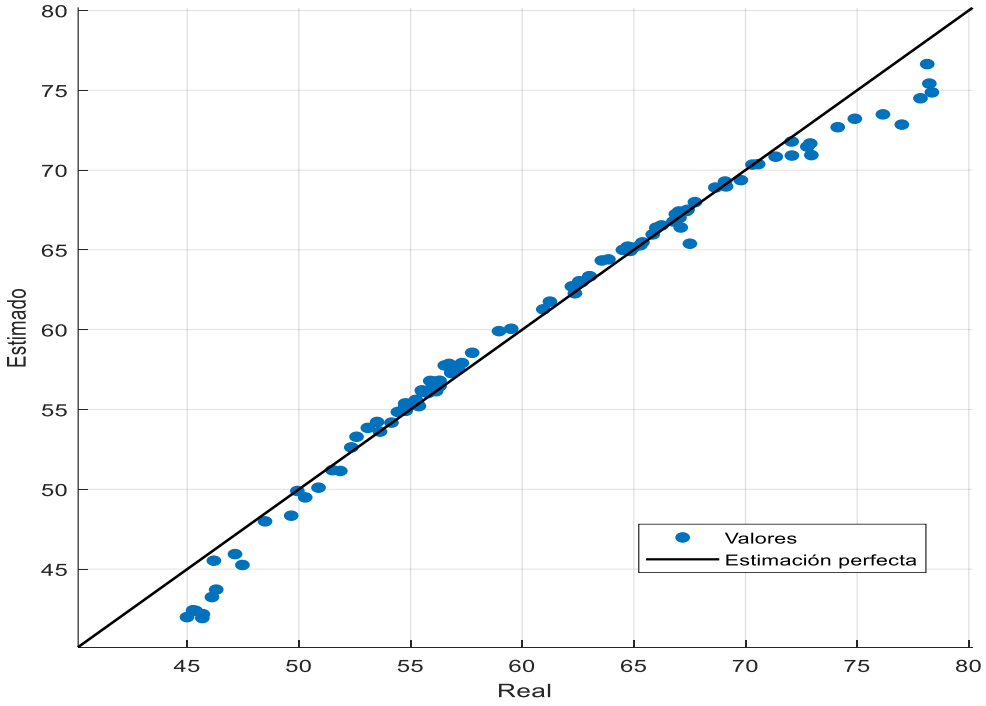


Figura 4.12. Gráficos de dispersión valores IC A estimados vs reales- Modelo RL muestra AE-2.

Los valores estimados presentan un error bajo ubicándose muy cerca a lo largo de la línea diagonal lo que explica los pequeños valores de error y el R^2 de 98%, el modelo RL subestima el 40% y sub estima el 60% de la información.

Las técnicas de aprendizaje automático permiten en este caso estimar un índice a partir de los datos medidos de parámetros fisicoquímicos. Si bien el índice requiere de un proceso matemático, en la construcción de los modelos de regresión no se ingresaron los valores de los parámetros escalados en la herramienta software, por lo que, para el uso del modelo con nuevos datos, sería posible estimar la calidad del agua almacenada en una escala de 0 a 100 ingresando solo los valores de pH, TDS y temperatura al modelo.

4.3.1. Factor de inflación de la varianza (FIV)

En el análisis de datos, es habitual encontrar observaciones con una influencia desproporcionada en los resultados del ajuste de dichos modelos, lo cual puede causar graves problemas en las estimaciones y la inferencia; también se pueden encontrar situaciones en las cuales exista poca variabilidad observada en las variables regresoras o relaciones de dependencia lineal entre ellas, lo cual puede conducir a problemas de multicolinealidad que frecuentemente producen una inflación artificial en la varianza de los coeficientes estimados del modelo (Palacio y Castaño, 2016).

El factor de inflación de la varianza (FIV), se obtiene a partir de la regresión de cada una de las variables explicativas sobre el resto de variables y se analizan los coeficientes de determinación de cada regresión (Kutner, Nachtsheim y Neter, 2004). Si alguno o algunos de estos coeficientes son altos se estaría señalando la posible existencia de un problema de multicolinealidad. El FIV se define por la ecuación 4.2:

$$FIV_j = \frac{1}{\left(1-R_j^2\right)} \tag{4.2}$$

Siendo R_j^2 el coeficiente de determinación de la regresión de la variable X_j sobre el resto de las variables explicativas. Cuanto más se acerque R_j^2 a la unidad, es decir, cuanto mayor sea la colinealidad de la variable X_j con el resto, mayor es el valor de FIV_j y mayor es la varianza del coeficiente estimado. Si $FIV_j > 10$, entonces se concluye que la colinealidad de X_j con las demás variables es alta.

Se calcula el FIV_j para los parámetros de pH, TDS y temperatura de las muestras de AL-3 agua lluvia, AA-1 agua de acueducto y AE-2 agua envasada. Los resultados de simulación obtenidos con los datos de la muestra AL-3 se presentan en la Figura 4.13.

```
lm1 =
Linear regression model:
    ALPH3 ~ 1 + ALTDS3 + T

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    __________    ______    __________

    (Intercept)       7.7675       0.3894    19.947    4.2475e-36
    ALTDS3         0.0057044    0.00032281    17.671    4.3723e-32
    T              0.0097716      0.021443    0.4557       0.64963


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 0.242
R-squared: 0.763,  Adjusted R-Squared: 0.758
F-statistic vs. constant model: 156, p-value = 4.74e-31
lm2 =
Linear regression model:
    ALTDS3 ~ 1 + ALPH3 + T

Estimated Coefficients:
                   Estimate      SE       tStat        pValue
                   ________    ______    ________    __________

    (Intercept)     -1009.3     87.404     -11.547    6.4149e-20
    ALPH3            133.76     7.5692      17.671    4.3723e-32
    T               -1.6344     3.2829    -0.49785       0.61972


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 37
R-squared: 0.763,  Adjusted R-Squared: 0.758
F-statistic vs. constant model: 156, p-value = 4.64e-31
lm3 =
Linear regression model:
    T ~ 1 + ALPH3 + ALTDS3

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    __________    ______    __________

    (Intercept)       16.325       3.8159    4.2781    4.4218e-05
    ALPH3            0.21862      0.47974    0.4557       0.64963
    ALTDS3        -0.0015594    0.0031323    -0.49785       0.61972


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.14
R-squared: 0.00257,  Adjusted R-Squared: -0.018
F-statistic vs. constant model: 0.125, p-value = 0.883
```

Figura 4.13. Resultados de regresión lineal muestra AL-3 – FIV

Los FIV calculados en la ecuación 4.2 con los datos de simulación para cada parámetro se muestran en la Tabla 4.7.

Tabla 4.7. FIV parámetros muestra AL-3

Variables Muestra AL-3	R^2	FIV
pH	0.763	4.219
TDS	0.763	4.219
T	0.00257	1.002

Los resultados de regresión obtenidos con los datos de la muestra AA-1 se muestran en la Figura 4.14.

```
lm1 =
Linear regression model:
    AAPH1 ~ 1 + AATDS1 + T

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    __________    _______    __________

    (Intercept)       8.4026       0.2687     31.271     1.842e-52
    AATDS1         0.0085176    0.00033485     25.437     1.0542e-44
    T                -0.035819     0.014691    -2.4382       0.01658


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 0.165
R-squared: 0.871,  Adjusted R-Squared: 0.869
F-statistic vs. constant model: 329, p-value = 6.21e-44
lm2 =
Linear regression model:
    AATDS1 ~ 1 + AAPH1 + T

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    __________

    (Intercept)     -841.21      47.908    -17.559     7.0142e-32
    AAPH1             102.1      4.0138     25.437     1.0542e-44
    T                  3.48      1.6189     2.1496       0.034071


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 18.1
R-squared: 0.87,  Adjusted R-Squared: 0.867
F-statistic vs. constant model: 324, p-value = 1.16e-43
lm3 =
Linear regression model:
    T ~ 1 + AAPH1 + AATDS1

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    _______    __________

    (Intercept)       30.604       5.1337     5.9614     4.0437e-08
    AAPH1             -1.6122      0.66122    -2.4382       0.01658
    AATDS1         0.013067     0.0060786     2.1496       0.034071


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.11
R-squared: 0.0587,  Adjusted R-Squared: 0.0392
F-statistic vs. constant model: 3.02, p-value = 0.0533
```

Figura 4.14. Resultados de regresión lineal muestra AA-1 – FIV

Los FIV calculados en la ecuación 4.2 con los datos de simulación para cada parámetro se muestran en la Tabla 4.8.

Tabla 4.8. FIV parámetros muestra AA-1

Variables Muestra AA-1	R^2	**FIV**
pH	0.871	7.751
TDS	0.87	7.692
T	0.0587	1.062

La Figura 4.15 presenta los resultados de regresión obtenidos con los datos de la muestra AE-2.

```
lm1 =
Linear regression model:
    AEPH2 ~ 1 + AETDS2 + T

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    _________    _______    _________

    (Intercept)       7.9513      0.22404     35.491    2.2803e-57
    AETDS2           0.01268    0.00028594     44.345    3.2927e-66
    T               -0.035059      0.012113    -2.8943     0.0046934


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 0.136
R-squared: 0.954,  Adjusted R-Squared: 0.953
F-statistic vs. constant model: 995, p-value = 2.29e-65
lm2 =
Linear regression model:
    AETDS2 ~ 1 + AEPH2 + T

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    _________

    (Intercept)     -589.54      24.036    -24.527    2.2569e-43
    AEPH2            75.157      1.6948     44.345    3.2927e-66
    T                2.5471     0.93696     2.7184      0.007771


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 10.5
R-squared: 0.953,  Adjusted R-Squared: 0.952
F-statistic vs. constant model: 985, p-value = 3.62e-65
lm3 =
Linear regression model:
    T ~ 1 + AEPH2 + AETDS2

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    _________    _______    _________

    (Intercept)       34.758      5.7399     6.0555    2.6497e-08
    AEPH2             -2.2675     0.78343    -2.8943     0.0046934
    AETDS2           0.027793    0.010224     2.7184      0.007771


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.1
R-squared: 0.0813,  Adjusted R-Squared: 0.0623
F-statistic vs. constant model: 4.29, p-value = 0.0164
```

Figura 4.15. Resultados de regresión lineal muestra AE-2 – FIV

Los FIV calculados en la ecuación 4.2 con los datos de simulación para cada parámetro se muestran en la Tabla 4.9.

Tabla 4.9. FIV parámetros muestra AE-2

Variables Muestra AE-2	R^2	**FIV**
pH	0.954	21.739
TDS	0.953	21.27
T	0.813	5.347

De acuerdo al análisis anterior, se observa que se obtienen valores de FIV superiores a 10 para los parámetros de la muestra AE-2 y menores a 10 para las muestras AA-1 y AL-3, con lo que se concluye que existe problemas de multicolinealidad entre los parámetros de la muestra de agua envasada.

4.4. Validación del Modelo

Los modelos de regresion construidos pueden ser utilizados posteriormente para estimar la variable para la que fueron creados a partir de nuevos datos. De acuerdo a los resultados obtenidos en la sección anterior, en donde se construyeron diferentes modelos para estimación del índice ICA, comparando los resultados de los tres mejores modelos de cada clase de agua, el modelo de árboles de regresión de la muestra AA-1 (**AR-AA1**) presento un desempeño superior con valores de 0.98 para R^2, 0.9257 para RMSE y 0.6627 para MAE. Este modelo se utilizara entonces para el proceso de validacion y se aplicara sobre otras muestras (AA-3, AL-1 y AE-4), cuyos datos de pH, TDS y temperatura se ingresaran al modelo para generar las nuevas predicciones del indice ICA.

Para contrastar los valores estimados, se calculo el indice con los datos reales para cada muestra, es importante resaltar que se esta evaluando el desempeño del modelo construido de una muestra de agua de acueducto sobre muestras de agua lluvia y agua envasada. Además se ingresan solo los valores de pH, TDS y temperatura y no ecuaciones o valores del indice, por lo que el modelo ademas de estimar un parametro, no requiere de procesos adicionales para entregar las estimaciones en la escala del indice.

Las nuevas predicciones no se relizaron directamente en la aplicacion *Regression Learner APP* de Matlab, ya que esta herramienta solo permite crear los modelos y almacenarlos, el procesamiento de los datos de validación puede verse más detalladamente en el Anexo D.

El modelo AR-AA1, se validó inicialmente sobre otra muestra de la misma clase, en este caso sobre la muestra AA-3. Ingresando los datos de los parametros al modelo y contrastando con los valores calculados reales, la Figura 4.16 presenta los valores estimados y reales del ICA para esta muestra.

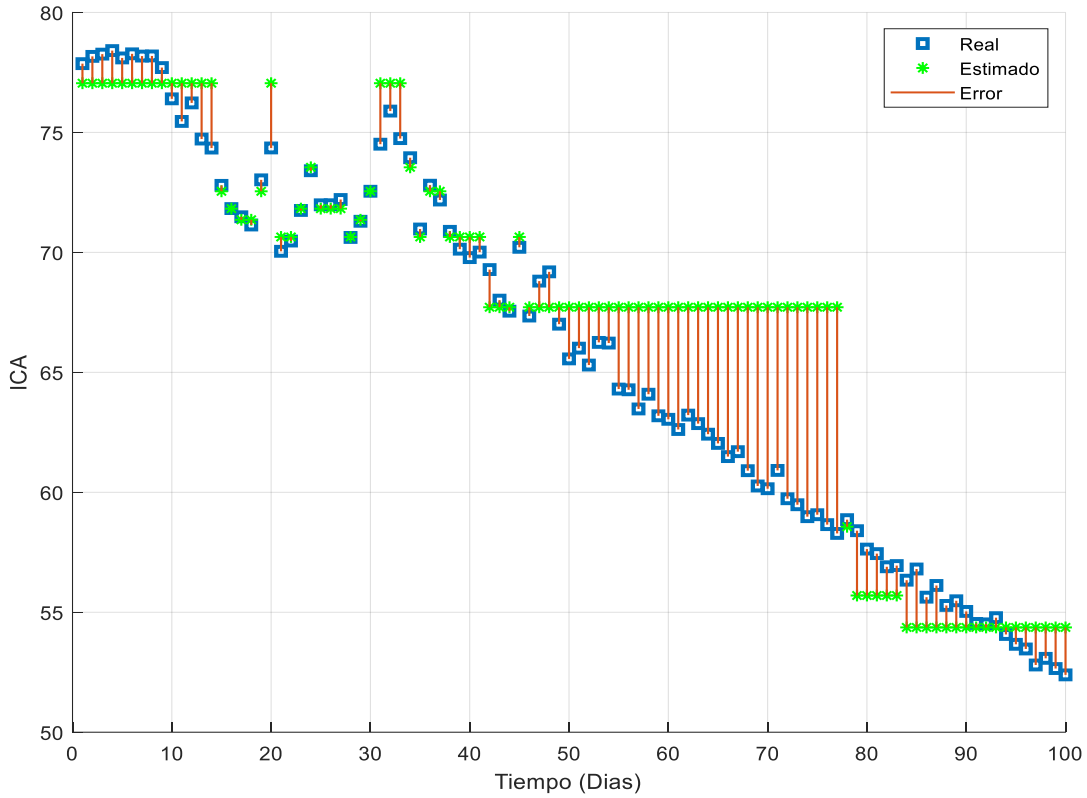


Figura 4.16. Validación ICA para muestra AA-3.

En las Figuras 4.16, 4.17 y 4.18 los cuadros de color azul representan los valores reales calculados de ICA, los valores estimados se identifican con puntos verdes y el error o diferencia entre ellos se simboliza con una linea de color rojo. Los resultados de las metricas de evaluación para la validación anterior fueron:

Tabla 4.10.Resultados Validación del ICA - muestra AA-3.

Modelo	**AR-AA1**
MAE	2.1893
RMSE	3.2899
R^2	0.8235

Como se aprecia en la Figura 4.16 y los resultados de desempeño de la Tabla 4.10, el modelo es capaz de estimar el índice de una muestra con un coeficiente de determinación $\mathbf{R^2}$ de 0.82 utilizando nuevos datos, para una estimación perfecta equivalente al 100% los resultados representan un porcentaje del 82% de ajuste hacia los datos reales.

Se validó el modelo **AR-AA1** sobre la muestra AL-1 de agua lluvia, siguiendo el mismo procedimiento y calculando el ICA con los datos reales de esta muestra para poder contrastarlo con los valores estimados, los resultados se presentan en la Figura 4.17 y la Tabla 4.11.

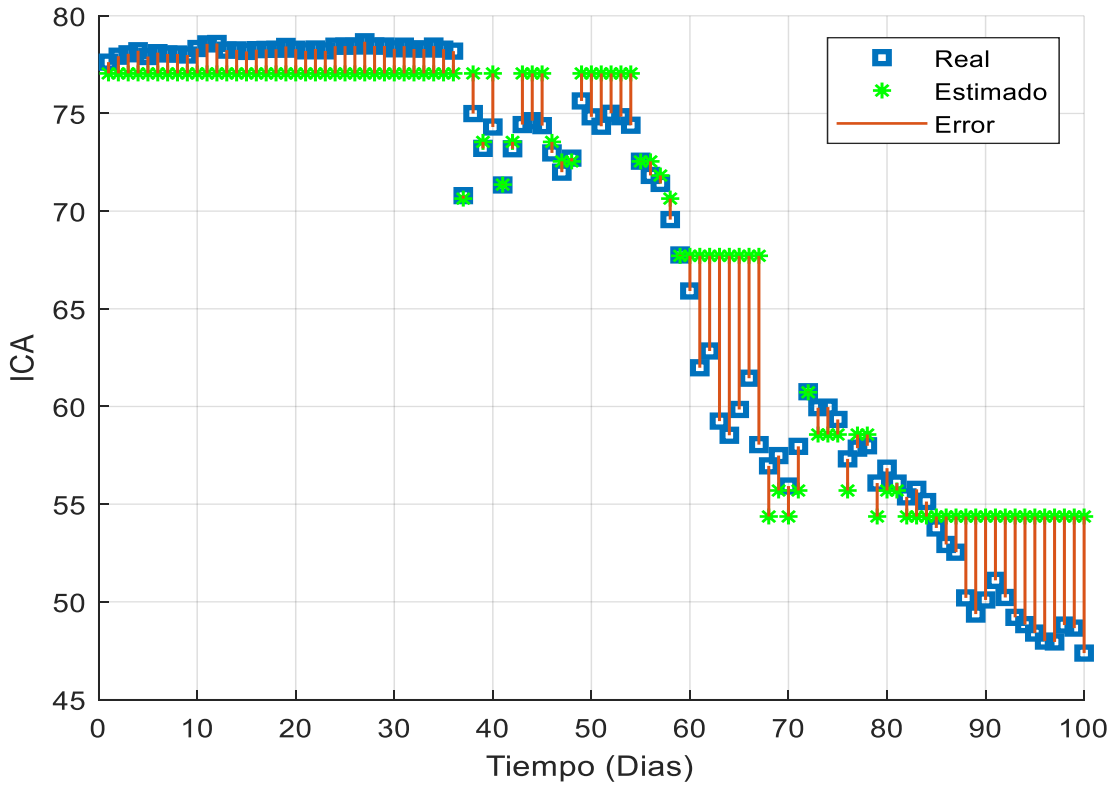


Figura 4.17. Validación ICA para muestra AL-1.

Tabla 4.11.Resultados Validación del ICA - muestra AL-1.

Modelo	**AR-AA1**
MAE	2.2122
RMSE	3.0731
R^2	0.9224

Los resultados obtenidos para la validación anterior fueron de 2.2122 para MAE, 3.0731 para RMSE y un R^2 de 0.9224 en la estimación del índice con nuevos datos de la clase agua lluvia, lo que representa una buena aproximación a los valores del índice reales con un porcentaje de ajuste del 92%. Los pequeños valores de RMSE y MAE también representa la cercanía entre los valores estimados y reales.

Finalmente se validó el modelo **AR-AA1** en una muestra de agua envasada siguiendo el procedimiento aplicado en los casos anteriores, los resultados obtenidos sobre la muestra AE-4 se presentan en la Figura 4.18 y la Tabla 4.12.

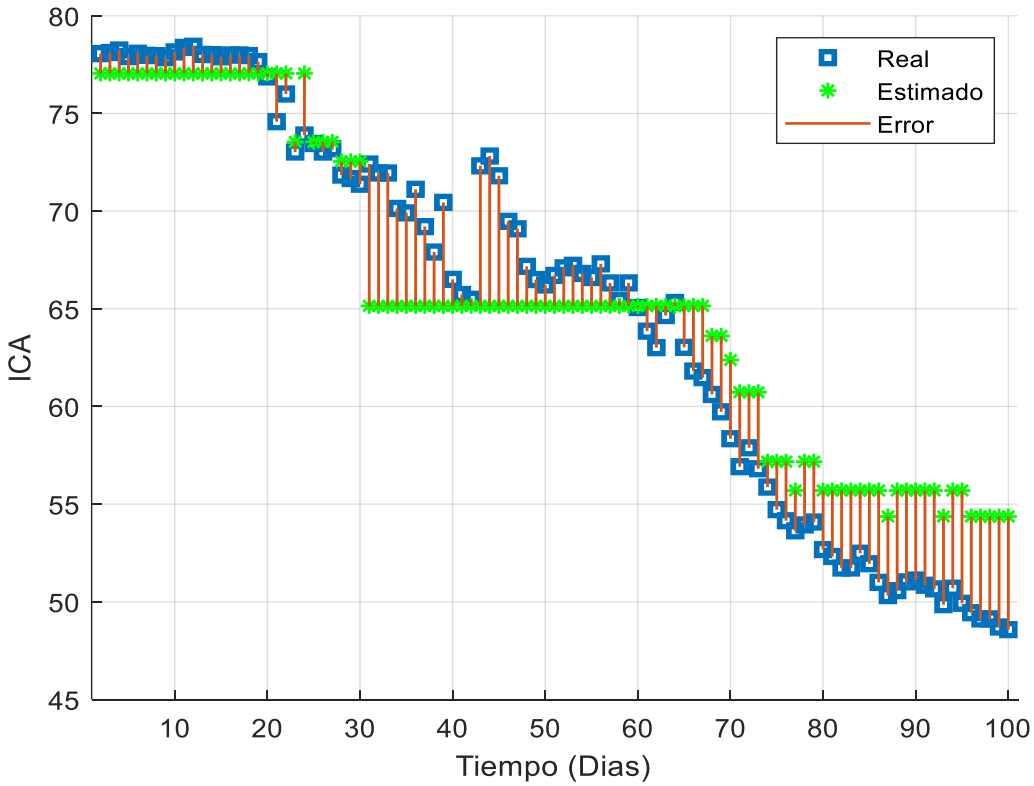


Figura 4.18. Validación ICA para muestra AE-4.

Tabla 4.12.Resultados Validación del ICA - muestra AE-4.

Modelo	**AR-AA1**
MAE	2.7476
RMSE	3.3950
R^2	0.8837

La validación del modelo **AR-AA1** en la estimación del ICA sobre la muestra de agua envasada, tiene resultados de 2.7476 para MAE, 3.3950 para RMSE y un R^2 de 0.8837, que equivale a un buen ajuste de las estimaciones del 88% sobre los valores del ICA reales.

Los resultados anteriores muestran la capacidad del modelo para estimar un índice de calidad de agua con una aproximación a los valores reales (R^2) superior al 82% aun en predicciones de clases de agua con características fisicoquímicas diferentes. Se estimaron 100 valores correspondientes a una evolución temporal dada en días, sin embargo, el modelo es capaz de predecir un número menor o mayor a 100 datos.

4.5. Discusión de Resultados

En este trabajo se estimaron parámetros y un índice de calidad de agua a partir de datos de monitoreo diario de parámetros fisicoquímicos como el pH, TDS y temperatura mediante algoritmos de regresión. Las muestras utilizadas provienen de fuentes de agua lluvia, acueducto y agua envasada comercial. Al igual que en otros estudios como en (Azad *et al.*, 2019), (Zhu y Heddam, 2019) y (Khaled *et al.*, 2018) del Capítulo 2, la correlación de parámetros se usó para determinar el conjunto de variables de entrada y el tipo de relación entre ellas. En este experimento se pudo evidenciar una fuerte correlación positiva entre el pH y el TDS y una débil correlación de estos parámetros con la temperatura en las diferentes muestras almacenadas, por lo cual, se calculó el factor de inflación de la varianza (FIV) para determinar si existen problemas de multicolinealidad entre las variables, dicha prueba se aplicó sobre los datos de las muestras AL-3, AA-1 y AE-2 utilizadas para calcular el índice de calidad de agua ICA. De acuerdo a los valores obtenidos para FIV, se observa que la muestra AE-2 presenta problemas de multicolinealidad con valores de FIV superiores a 10 por lo que sería importante aplicar algunas estrategias para resolver el problema de colinealidad (Guerrero y Melo, 2017) como trabajo futuro.

Se estimó el ICA sin incluir la temperatura como parámetro de entrada, Anexo D, debido a la baja correlación de este parámetro con el pH y TDS. Los resultados fueron muy similares a los obtenidos cuando se incluye la temperatura en el modelo, por lo que podría pensarse en reducir el número de variables independientes y generar así una estructura más simple. Sin embrago dada la importancia de este parámetro en estudios de calidad de agua como se aprecia en (Seo, Yun y Choi, 2016), en donde se considera como variable de salida, podría plantearse cambiar las condiciones del experimento de tal manera que se puedan generar cambios en la temperatura y evidenciar así su relación con otros parámetros fisicoquímicos, además, es posible que la débil incidencia de este parámetro en el pH, TDS y ICA se deban a las condiciones de almacenamiento propias del experimento.

La estimación de parámetros en este campo se ha enfocado en monitorear fuentes de agua continuas y superficiales como ríos, quebradas, lagunas, agua de mar principalmente. En (Al-Musawi y Al-Rubaie, 2017) en donde estima el ICA en un rio, se obtienen un de R^2 de 0.93, de manera similar en (Hussein Ewaid, Ali Abed y Kadhum, 2018), (Wang, Zhang y Ding, 2017) con resultados de 0.97 y 0.92 para R^2. Al comparar los resultados, se observan desempeños similares a los presentados en este documento cuyos valores de ajuste en la estimación del ICA se encuentran entre 0.94 a 0.98 para R^2. Sin embargo, los datos recolectados en los trabajos mencionados presentan una variación espacial y temporal asociada a los puntos de muestreo y fechas de toma de datos, lo que representan cambios en las propiedades fisicoquímicas del agua de un lugar a otro (muestras en diferentes puntos a lo largo de un rio, por ejemplo), a diferencia de los datos recolectados en este trabajo que solo evidencian una variación temporal en un periodo de tres meses.

En otros estudios, también se estimaron parámetros como la temperatura (Seo, Yun y Choi, 2016) con resultados de ajuste de 0.99, conductividad en (Azad et al., 2019) con un valor de R^2 de 0.91 y solidos disueltos totales (TSS) en (Granata et al., 2017) con 0.97 de ajuste en la técnica de máquinas de vectores de soporte y R^2 de 0.90 para la técnica de árboles de regresión (Tabla 2.3). Resultados similares se presentaron en este trabajo al estimar el parámetro de pH y TDS con valores entre 0.85 y 0.97 para R^2. Los estudios relacionados en este campo, se centran en identificar cambios en las concentraciones de parámetros que evidencien la presencia de contaminantes que deterioran la calidad del líquido, utilizando incluso variables categóricas en relación con la interacción de los factores que rodean la fuente de agua. Son escasos los trabajos enfocados al estudio del agua para consumo humano en reposo o en el comportamiento de agua que ha sufrido alguna clase de tratamiento y/o adición de sustancias químicas para su limpieza, como es el caso del agua de acueductos o envasada industrial, por lo que, no sería posible comparar de manera directa los resultados aquí encontrados.

Entre las limitaciones de este trabajo se deben destacar los problemas de correlación que reflejan la homogeneidad de los datos de los parámetros fisicoquímicos, la clasificación de la calidad de agua en las muestras al inicio del experimento, refiere a un agua poco contaminada cuyas propiedades de origen orgánico, bajo estas condiciones, tienen normalmente poca variación. Los modelos construidos en este trabajo están concebidos para agua de consumo humano, de bajas concentraciones de sales muy típicas en ríos y arroyos o aguas provenientes de procesos de tratamiento, purificación y limpieza y almacenadas bajo condiciones físicas específicas, sin embargo, podría evaluarse su desempeño en agua salobre o fuentes subterráneas.

Las muestras de agua lluvia tienen una calidad aceptable, lo cual puede asociarse por ejemplo a que la zona que se encuentra alejada de la alta contaminación industrial o grandes centros urbanos (factores de contaminación externa). En el caso de las muestras de agua de acueducto, también podría considerarse recolectar las muestras de acueductos ubicados en diferentes regiones, lo que adicionaría más variabilidad a los modelos de regresión. En este trabajo tampoco se considerar el error asociado a los instrumentos de medición por lo que podría esperarse que el implementar estrategias para reducir este porcentaje en trabajos futuros pueda contribuir en la confiabilidad de los resultados de los modelos de regresión.

Conclusiones

La construcción de los modelos de regresión mediante las técnicas de aprendizaje automático como la regresión lineal, arboles de regresión y máquinas de vectores de soporte, permiten estimar parámetros fisicoquímicos e índices de calidad, para agua almacenada de diferentes fuentes como el agua lluvia, agua de acueducto y agua envasada a partir de datos de monitoreo diario.

Los modelos construidos para la estimación de parámetros como el pH y TDS a partir de datos de pH, TDS y temperatura, mostraron buenos resultados de desempeño con valores de ajuste de las estimaciones a los datos reales entre el 80% y 97%. En los modelos para la estimación del ICA los resultados también fueron positivos con valores de 97% y 98% de ajuste y en la validación resultados superiores al 82% de aproximación a los valores reales.

El procedimiento realizado en este trabajo permitió construir y validar un modelo (AR-AA-1) capaz de estimar el comportamiento de la calidad del agua a partir de los índices de calidad (ICA) en un periodo de tiempo (días), utilizando los datos de monitoreo de parámetros como pH, TDS y temperatura, sin que se requiera ingresar al modelo ecuaciones o procedimientos matemáticos adicionales. Además, este modelo permite estimar el ICA en muestras de otra clase de agua almacenada como el agua lluvia y el agua envasada.

La calidad de las muestras de agua almacenada, tienen un comportamiento decreciente pasando de una clasificación *Buena* a *Media* alrededor del día 45 y una clasificación *Mala* hacia el día 85. Es importante mencionar que las muestras de agua lluvia tienen un proceso de descomposición más lenta que las muestras de agua de acueducto y envasada, lo que podría asociarse a los procesos de tratamiento que sufren antes de su distribución.

Al comparar los modelos de regresión lineal arboles de regresión y máquinas de vectores de soporte, el modelo de árboles de regresión presento un desempeño superior seguido por los modelos de regresión lineal, los resultados de AR pueden atribuirse a la etapa de poda durante la fase de entrenamiento del modelo la cual elimina ramificaciones redundantes para optimizar su estructura. Dada la tendencia lineal de los datos en los casos de estudio, estas técnicas suelen acoplarse mejor, para datos de mayor varianza es posible que la técnica MVS u otras de mayor complejidad tengan mejores resultados.

Los parámetros pH, TDS y temperatura son de medición directa y no requieren de procesamientos posteriores en laboratorio ni equipo muy complejo y costoso para su monitoreo, lo que podría representar una ventaja si se desea utilizar el modelo en fuentes de agua en lugares remotos. Como trabajo futuro se plantea considerar el error de medición como variable dentro del experimento y modelos de recesión

El análisis de correlación permitió establecer asociaciones importantes entre los parámetros y de utilidad a la hora de determinar las entradas para la construcción de los modelos, también fue de utilizar para determinar la existencia de problemas de multicolinealidad entre los parámetros en cuyo caso podrían aplicarse estrategias para resolver este problema como trabajo futuro.

Si bien los modelos construidos en este trabajo están diseñados para condiciones específicas como muestras almacenadas y clases de agua como agua lluvia, de acueducto y envasada, sería interesante evaluar su desempeño bajo otros escenarios. Como trabajos posteriores, también podría considerarse la implementación de los modelos en dispositivos móviles que puedan facilitar las labores de control y monitoreo de agua en lugares de difícil acceso, adicionar otros parámetros fisicoquímicos y biológicos o la estimación de otros parámetros que no se pueden obtener fácilmente por medición directa.

Producto de este trabajo de investigación se elaboraron los artículos “A*prendizaje automático para la predicción de calidad de agua potable”* publicado en la revista *Ingeniare - Universidad Libre* y el articulo “*Multivariate Prediction of Nitrogen Concentration in a Stream Using Regression Models”* en proceso de revisión en una revista internacional.

Referencias

Al-Musawi, N. O. y Al-Rubaie, F. M. (2017) “Prediction and Assessment of Water Quality Index using Neural Network Model and GIS Case Study: Tigris River in Baghdad City”, *Applied Research Journal*, 3(11), pp. 343–353.

Azad, A. *et al.* (2019) “Modeling river water quality parameters using modified adaptive neuro fuzzy inference system”, *Water Science and Engineering*, 12(1), pp. 45–54. doi: 10.1016/j.wse.2018.11.001.

Brentan, B. M. *et al.* (2017) “Hybrid regression model for near real-time urban water demand forecasting”, *Journal of Computational and Applied Mathematics*, 309, pp. 532–541. doi: 10.1016/j.cam.2016.02.009.

Carrasquilla-Batista, A. *et al.* (2016) “Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal”, *Revista Tecnología en Marcha*, 29(8), p. 33. doi: 10.18845/tm.v29i8.2983.

Castro, M. *et al.* (2015) “Indicadores de la calidad del agua: evolución y tendencias a nivel global”, *Ingeniería solidaria*, 9(17). doi: 10.16925/in.v9i17.811.

Chen, Z., Ye, X. y Huang, P. (2018) “Estimating Carbon Dioxide (CO2) Emissions from Reservoirs Using Artificial Neural Networks”, *Water*. MDPI AG, 10(1), p. 26. doi: 10.3390/w10010026.

Chou, J.-S., Ho, C.-C. y Hoang, H.-S. (2018) “Determining quality of water in reservoir using machine learning”, *Ecological Informatics*, 44, pp. 57–75. doi: 10.1016/j.ecoinf.2018.01.005.

Chow S, Shao J, Wang H. 2008. "Sample Size Calculations in Clinical Research". 2nd Ed. Chapman & Hall/CRC Biostatistics Series. pag. 58.

Cruz, G. R., Alonso, L. M. y Franco, Á. A. (2017) “Hybrid Predictive Model and Recommendations with Techniques of Data Mining and Artificial Intelligence”, *Programación Matemática y Software*, 9(3), pp. 18–24.

Duran, E. P. *et al.* (2010) “Captación de agua de lluvia, alternativa sustentable”, en *Congreso Nacional del Medio Ambiente*.

Espino, T. C. (2017) *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso*.

Fan, J. *et al.* (2017) “Predicting bio-indicators of aquatic ecosystems using the support vector machine model in the Taizi River, China”, *Sustainability (Switzerland)*, 9(6), pp. 1–11. doi: 10.3390/su9060892.

García-García, J. A., Reding-Bernal, A., y López-Alvarenga, J. C. (2013). "Cálculo del tamaño de la muestra en investigación en educación médica". *Investigación en educación*

médica, *2*(8), 217-224.

Gómez-Gutiérrez, A. *et al.* (2016) “La calidad sanitaria del agua de consumo”, *Gaceta Sanitaria*. Ediciones Doyma, S.L., pp. 63–68. doi: 10.1016/j.gaceta.2016.04.012.

Gorde, S. P. y Jadhav, M. V (2013) “Assessment of Water Quality Parameters : A Review”, *International Journal of Engineering Research and Applications*, 3(6), pp. 2029–2035.

Granata, F. *et al.* (2017) “Machine learning algorithms for the forecasting of wastewater quality indicators”, *Water (Switzerland)*, 9(2), pp. 1–12. doi: 10.3390/w9020105.

Guerrero, S. C., y Melo, O. O. (2017). “Una metodología para el tratamiento de la multicolinealidad a través del escalamiento multidimensional”. *Ciencia en Desarrollo*, *8*(2), 9-24. (FIV)

Günnemann, S. (2017) “Machine Learning Meets Databases”, *Datenbank Spektrum*, 17, pp. 77–83. doi: 10.1007/s13222-017-0247-8.

Haghiabi, A. H., Nasrolahi, A. H. y Parsaie, A. (2018) “Water quality prediction using machine learning methods”, *Water Quality Research Journal of Canada*. IWA Publishing, 53(1), pp. 3–13. doi: 10.2166/wqrj.2018.025.

Harrington, P. (2012) *Machine Learning in Action*. Manning Publications Co.

Hussein Ewaid, S., Ali Abed, S. y Kadhum, S. A. (2018) “Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis”, *Environmental Technology & Innovation*, 11, pp. 390–398. doi: 10.1016/j.eti.2018.06.013.

Janet Gil, M. *et al.* (2012) “Contaminantes emergentes en aguas, efectos y posibles tratamientos”, *Producción+Limpia*, 7(2).

Kang, G. K., Gao, J. Z. y Xie, G. (2017) “Data-driven Water Quality Analysis and Prediction: A Survey”. doi: 10.1109/BigDataService.2017.40.

Khaled, B. *et al.* (2018) “Modelling of biochemical oxygen demand from limited water quality variable by anfis using two partition methods”, *Water Quality Research Journal of Canada*. IWA Publishing, 53(1), pp. 24–40. doi: 10.2166/wqrj.2017.015.

Kutner, M. H., Nachtsheim, C. J., Neter, J. (2004). "Applied Linear Regression Models" (4th edición). McGraw-Hill Irwin.

Lisseth Guzmán, B., Nava, G. y Díaz, P. (2015) “Quality of water for human consumption and its association with morbimortality in Colombia”, *Biomédica*, 35(2), pp. 177–90. doi: 10.7705/biomedica.v35i0.2511.

Matlab (2020a) “Growing Decision Trees”. Disponible en: https://la.mathworks.com/help/stats/growing-decision-trees.html.

Matlab (2020b) “Least-Squares Fitting”. Disponible en: https://la.mathworks.com/help/curvefit/least-squares-fitting.html?lang=en.

Matlab (2020c) “Understanding Support Vector Machine Regression”. Disponible en: https://la.mathworks.com/help/stats/understanding-support-vector-machine-regression.html?lang=en.

Mohammed, H., Longva, A. y Seidu, R. (2018) “Predictive analysis of microbial water quality using machine-learning algorithms”, *Environmental Research, Engineering and Management*, 74(1), pp. 7–20. doi: 10.5755/j01.erem.74.1.20083.

MPS (2007) “Ministerio de la Protección Social. Ministerio de Ambiente, Vivienda y Desarrollo Territorial. Resolución 2115 de 22 junio de 2007”.

Nazemi, A. y Madani, K. (2018) “Urban water security: Emerging discussion and remaining challenges”, *Sustainable Cities and Society*. Elsevier Ltd, pp. 925–928. doi: 10.1016/j.scs.2017.09.011.

Palacio Salazar, J. E., y Castaño Vélez, E. A. (2016). “Detección de datos extremos y de multicolinealidad en modelos no lineales: una interfaz gráfica en R”.

RadFard, M. *et al.* (2019) “Protocol for the estimation of drinking water quality index (DWQI) in water resources: Artificial neural network (ANFIS) and Arc-Gis”, *MethodsX*, 6, pp. 1021–1029. doi: 10.1016/j.mex.2019.04.027.

Rial, E. P. (2014) *Análisis estadístico multivariante de un conjunto de datos biológicos experimentales*.

Rivas-Ruiz, R., Moreno-Palacios, J., & Talaveraa, J. O. (2013). "Clinical research XVI. Differences between medians with Mann-Whitney U test". *Revista médica del Instituto Mexicano del Seguro Social*, *51*(4), 414-419.

Seo, I. won, Yun, S. H. y Choi, S. Y. (2016) “Forecasting Water Quality Parameters by ANN Model Using Pre-processing Technique at the Downstream of Cheongpyeong Dam”, *Procedia Engineering*, 154, pp. 1110–1115. doi: 10.1016/j.proeng.2016.07.519.

Soltani, F., Kerachian, R. y Shirangi, E. (2010) “Developing operating rules for reservoirs considering the water quality issues: Application of ANFIS-based surrogate models”, *Expert Systems with Applications*. Elsevier Ltd, 37(9), pp. 6639–6645. doi: 10.1016/j.eswa.2010.03.057.

Sotiropoulos, D. N. y Tsihrintzis, G. A. (2016) *Machine Learning Paradigms. Artificial Immune Systems and Their Applications in Software Personalization*. doi: 10.1007/978-3-319-47194-5.

Tejedor, F.J. y Etxeberria, J. (2006). "Análisis inferencial de datos en educación". Madrid. La Muralla.

Wang, X., Zhang, F. y Ding, J. (2017) “Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China”, *Scientific Reports*. Nature Publishing Group, 7(1). doi: 10.1038/s41598-017-12853-y.

WHO (2019) *Safer water, better health*. World Health Organization.

Wills, M. y Irvine, K. (1996) “Application of the National Sanitation Foundation Water Quality Index in Cazeno- via Creek, NY, Pilot Watershed Management Project”, *Middle States Geographer*, pp. 95–104.

Wu, E. M.-Y. *et al.* (2014) “The Application of Water Quality Monitoring Data in a Reservoir Watershed Using AMOS Confirmatory Factor Analyses”, *Environmental Modeling & Assessment*. Springer International Publishing, 19(4), pp. 325–333. doi: 10.1007/s10666-014-9407-5.

Zhu, S. y Heddam, S. (2019) “Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN)”, *Water Quality Research Journal*. IWA Publishing. doi: 10.2166/wqrj.2019.053.