UNIVERSIDAD AUTÓNOMA DE ZACATECAS

"Francisco García Salinas"

"PREDICCIÓN DE LA DEMANDA BIOQUÍMICA DE OXÍGENO EN AGUAS SUPERFICIALES MEXICANAS USANDO APRENDIZAJE MÁQUINA"

Tesis para obtener el grado de:

Maestro en Ciencias del Procesamiento de la Información

Presenta

Maximiliano Guzmán Fernández

Director:

Dr. Héctor Antonio Duran Muñoz

Co-Directores:

Dra. Claudia Sifuentes Gallardo

Dr. José Ismael de la Rosa Vargas

Asesores:

Dra. Ireri Aydée Sustaita Torres

Dr. Pedro Daniel Alaniz Lumbreras

Zacatecas, Zac., 8 de agosto de 2022.

Zacatecas, Zac., 08 de agosto de 2022.

C. Maximiliano Guzmán Fernández
Estudiante de la MCPI
PRESENTE

Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis titulada "**Predicción de la Demanda Bioquímica de Oxígeno en Aguas Superficiales Mexicanas Usando Aprendizaje Máquina**", presentada por el estudiante Ing. Maximiliano Guzmán Fernández y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciban un cordial saludo.

COMITÉ TUTORIAL
PROCESAMIENTO Y ANÁLISIS DE DATOS

Dr. Héctor Antonio Duran Muñoz

Dra. Claudia Sifuentes Gallardo

Dr. Pedro Daniel Alaniz Lumbreras

Dr. José Ismael de la Rosa Vargas

Dra. Ireri Aydée Sustaita Torres

c.c.p. Interesado.

c.c.p. Responsable de la Maestría en Ciencias del Procesamiento de la Información.

Carta de similitud núm. 285/ IyP
Zacatecas, Zacatecas 11/agosto/2022

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

"Predicción de la demanda bioquímica de oxígeno en aguas superficiales mexicanas usando aprendizaje máquina" de Maximiliano Guzmán Fernández

Fue analizado con el software Copyleaks, con la intención de detectar similitudes; el resultado en cuestión fue

10 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente
"Somos Arte, Ciencia y Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas

Zacatecas, Zac., 08 de agosto de 2022
Carta Cesión de Derechos

A QUIEN CORRESPONDA

El que suscribe C. Maximiliano Guzmán Fernández alumno del Programa Maestría en Ciencias del Procesamiento de la Información con numero de matrícula 29105262, adscrito a la Unidad Académica de ingeniería Eléctrica de la Universidad Autónoma de Zacatecas, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. Héctor Antonio Duran Muñoz y Dra. Claudia Sifuentes Gallardo y cede los derechos del trabajo titulado: **"Predicción de la Demanda Bioquímica de Oxígeno en Aguas Superficiales Mexicanas Usando Aprendizaje Máquina"** a la Universidad Autónoma de Zacatecas para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o directores del trabajo. Este puede ser obtenido escribiendo al correo electrónico maxguzman@uaz.edu.mx o estableciendo contacto con el responsable de Maestría quien turnará la solicitud al autor y directores del trabajo de investigación. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo. Agradezco de antemano su atención a la presente, reciba un cordial saludo.

ATENATMENTE

MAXIMILIANO GUZMÁN FERNÁNDEZ

Agradecimientos

Agradezco a la Universidad Autónoma de Zacatecas por la oportunidad de formar parte de ella y obtener los conocimientos para mi crecimiento académico. A mi familia, por su apoyo, confianza y consejos en todos los momentos que los necesité, guiándome en decisiones y situaciones que se han presentado en esta etapa de mi vida.

A la Dra. Claudia Sifuentes Gallardo y al Dr. Héctor Antonio Durán Muñoz, por su tiempo y ayuda en todo el proceso de este trabajo. Además de su profesionalismo para aportar a mi formación académica y de su confianza por integrarme en su equipo de trabajo.

A mis compañeros, amigos y maestros del programa de la maestría en ciencias del procesamiento de la información por contribuir con ideas y experiencias para afrontar nuevos retos.

Dedicatoria

A mis abuelos …

Resumen

El monitoreo de la calidad del agua superficial es insuficiente en México debido a las limitadas estaciones de monitoreo del agua. El principal parámetro de monitoreo para evaluar la calidad del agua superficial es la demanda bioquímica de oxígeno. Este parámetro estima la materia orgánica biodegradable presente en el agua. Concentraciones superiores a 30 mg/l indican un alto nivel de contaminación por residuos domésticos e industriales. Por lo tanto, el objetivo de este trabajo es proporcionar una referencia al proceso convencional de determinación de la demanda bioquímica de oxígeno utilizando el aprendizaje máquina y un dispositivo electrónico de medición con sensores de bajo costo. La base de datos utilizada fue recopilada por la Comisión Nacional del Agua (CONAGUA). Se aplicaron las técnicas de correlación de Pearson y Forward Selection para identificar los parámetros con mayor contribución a la predicción de la demanda bioquímica de oxígeno. Se formaron tres grupos y se utilizaron como entrada a cuatro algoritmos de aprendizaje máquina. El algoritmo Random Forest obtuvo el mejor rendimiento. Los grupos A, B y C de parámetros obtuvieron un coeficiente de determinación de 0.76, 0.75 y 0.46 respectivamente. Esto permite elegir un grupo adecuado de parámetros que se pueden determinar con los instrumentos de análisis químicos disponibles en la zona de estudio.

Palabras Clave: aprendizaje máquina, Demanda bioquímica de oxígeno, aguas superficiales mexicanas.

Abstract

The monitoring of surface water quality is insufficient in Mexico due to the limited water monitoring stations. The main monitoring parameter to evaluate surface water quality is the biochemical oxygen demand. This parameter estimates the biodegradable organic matter present in the water. Concentrations above 30 mg/l indicates a high level of contamination by domestic and industrial waste. Therefore, the aim of this work to provide a reference to the conventional process of determining biochemical oxygen demand using machine learning and an electronic measuring device with low-cost sensors. The database used was collected by the National Water Commission (CONAGUA). Pearson's correlation and Forward Selection techniques were applied to identify the parameters with the most important contribution to prediction of biochemical oxygen demand. Three groups were formed and used as input to four machine learning algorithms. Random forest algorithm obtained the best performance. Group A, B and C of parameters obtained a coefficient of determination of 0.76, 0.75 and 0.46 respectively. This allows choosing an adequate group of parameters that can be determined with the chemical analysis instruments available in the study area.
Keywords: Machine Learning, Biochemical Oxygen Demand, Mexican Surface Waters.

Contenido General

Índice de Figuras

Índice de Tablas

Capítulo 1. Introducción

En este capítulo se plantea la importancia de la calidad del agua superficial y la relación que se le atribuye en México con el parámetro de la demanda bioquímica de oxígeno (DBO) presente en el agua después de 5 días. Después se presentan algunos de los problemas que surgen en el monitoreo de la calidad del agua y la determinación de la DBO después de 5 días. Continuando con el capítulo, se plantea la justificación, preguntas, objetivos e hipótesis de investigación y se finaliza con la descripción del trabajo a realizar y la estructura de este documento de tesis.

1.1 Antecedentes

La preservación, el tratamiento, el acceso y el uso eficiente del agua es fundamental para la humanidad, y varios países la consideran un recurso de seguridad nacional. Debido a la importancia de este recurso, el concepto de seguridad del agua ha sido discutido y definido por varias instituciones, como el Grupo de Agua de las Naciones Unidas (UN-WATER, por sus siglas en inglés) (UNU-INWEH, 2013), la Comisión Económica para América Latina y el Caribe (CEPAL) (Peña, 2016), entre otras. El uso del agua de ríos, pozos y lagunas es de gran importancia, ya que son las principales fuentes de abastecimiento de agua en los municipios y estados para diversos usos. Sin embargo, actividades como la minería, la ganadería, la agricultura y la demanda industrial generan explotación y contaminación del agua (Raynal-Gutierrez, 2020).

La calidad del agua es un factor importante por considerar, ya sea por las necesidades del ecosistema o por los niveles de contaminación que impactan directamente en la alimentación, higiene, salud y economía (Fonseca-Ortiz et al., 2020). Para garantizar el uso seguro del agua, se realiza un monitoreo continuo de los parámetros de calidad del agua en las fuentes de abastecimiento y en las zonas de descarga.

En México el organismo encargado de administrar, regular, controlar y proteger las aguas nacionales del país es la Comisión Nacional del Agua (CONAGUA). La CONAGUA lleva a cabo, a través de la Red Nacional de Medición de Calidad del Agua, el monitoreo de los

principales cuerpos de agua del país, tanto superficiales como subterráneos (CONAGUA, 2020).

Los parámetros principales que consideran para determinar la calidad del agua superficial se muestran en la Tabla 1.1. Estos y otros indicadores se comparten en los diferentes países del mundo dependiendo del establecimiento de su índice de calidad de agua.

Tabla 1.1 Parámetros principales considerados para evaluar la calidad del agua.

Parámetro	Descripción
Demanda bioquímica de oxígeno	Estima la cantidad de oxígeno que requiere un conjunto microbiano heterogénea para oxidar la materia orgánica en una muestra de agua por 5 días (Norma mexicana NMX-AA-028-SCFI-2001 Análisis de agua - determinación de la demanda bioquímica de oxígeno en aguas naturales, residuales (DBO5) y residuales tratadas - método de prueba, 2001).
Demanda química de oxígeno	Es la cantidad de oxígeno consumida en la oxidación química total de constituyentes orgánicos a productos inorgánicos finales (Norma mexicana NMX-AA-030/1-SCFI-2012 Análisis de agua - medición de la demanda química de oxígeno en aguas naturales, residuales y residuales tratadas. - método de prueba - parte 1 - método de reflujo abierto -, 2012).
Sólidos suspendidos totales	Es el material compuesto por Sólidos sedimentables, los sólidos suspendidos y coloidales que son retenidos por un filtro de fibra de vidrio (Norma mexicana NMX-AA-034-SCFI-2015 Análisis de agua - medición de sólidos y sales disueltas en aguas naturales, residuales y residuales tratadas – método de prueba, 2015).
Coliformes fecales	Organismos aerobios o anaerobios facultativos capaces de crecer a 35°C en un medio líquido de lactosa, con producción de ácido y gas en un periodo de 24 horas a 44.5°C (Norma mexicana NMX-AA-102-SCFI-2006 Calidad del agua – detección y enumeración de organismos coliformes, organismos coliformes termo tolerantes y escherichia coli prnmx-aa-102-scfi-2006 calidad del agua – detección y enumeración de organismos coliformes, 2006).
Escherichia coli	Organismos coliformes fecales los cuales producen indol a partir de triptófano en un lapso de 24 horas a 44.5°C (Norma mexicana NMX-AA-102-SCFI-2006 Calidad del agua – detección y enumeración de organismos coliformes, organismos coliformes termo tolerantes y escherichia coli prnmx-aa-102-scfi-2006 calidad del agua – detección y enumeración de organismos coliformes, 2006).
Enterococos	Son indicadores de contaminación fecal, ya que su hábitat es el intestino humano y animal. Son resistentes a condiciones extremas, por lo que su presencia indica una contaminación no reciente (Norma mexicana NMX-AA-102-SCFI-2006 Calidad del agua – detección y enumeración de organismos coliformes, organismos coliformes termo tolerantes y escherichia coli prnmx-aa-102-scfi-2006 calidad del agua – detección y enumeración de organismos coliformes, 2006).

Porcentaje de oxígeno disuelto	Representa la cantidad de oxígeno gaseoso que esta disuelta en el agua (Norma mexicana NMX-AA-012-SCFI-2001 Análisis de agua - determinación de oxígeno disuelto en aguas naturales, residuales y residuales tratadas - método de prueba, 2001).
Toxicidad	Capacidad de una sustancia de causar efectos adversos en organismos vivos (por ejemplo, toxicidad aguda Daphina magna) (Norma mexicana NMX-AA-087-SCFI-2010 Análisis de agua - evaluación de toxicidad aguda con daphnia magna, straus (crustácea - cladócera) - método de prueba, 2010).

En 2019 la Red Nacional de Medición de Calidad del Agua contó con 2764 sitios de monitoreo de agua superficial. Para evaluar de forma general la calidad de agua en sitios de monitoreo superficiales se planteó un semáforo de contaminación con las categorías de verde, amarillo y rojo. Para determinar el semáforo de contaminación se tomaron los 8 parámetros de la calidad del agua previamente mostrados en la Tabla 1.1. Los resultados de la evaluación mostraron que el 33.2% de los sitios se catalogaron con color verde, cumpliendo con los límites aceptables de la calidad de agua. El 31% de los sitios obtuvieron la categoría de amarillo al presentar incumplimiento en uno o varios de los parámetros de Escherichia coli, coliformes fecales, sólidos suspendidos totales y el porcentaje de oxígeno disuelto. El 35.8% de los sitios se catalogaron con color rojo presentando incumplimiento en uno o varios de los parámetros de demanda bioquímica de oxígeno, demanda química de oxígeno y enterococos (CONAGUA, 2020).

A partir de esto, la demanda bioquímica de oxígeno (DBO) es uno de los principales parámetros a la hora de evaluar la calidad del agua superficial en los sitios de monitoreo en México. Este parámetro indica la materia orgánica biodegradable presente en la muestra de los cuerpos de agua superficiales después de 5 días. Con base en el nivel de contaminación establecido por la norma mexicana NOM-001-SEMARNAT-1996 y CONAGUA, si la DBO es superior a 30 mg/l, el agua se considera contaminada para el uso público humano (Tabla-Vázquez et al., 2020).

Para determinar la DBO existen diversos métodos estandarizados con muestras diluidas (ISO 5815-1:2019) y no diluidas (ISO 5815-2:2003), entre otros. En la Figura 1.1 se presenta algunosde ellos y en la Figura 1.2 el tiempo que requieren para el análisis.

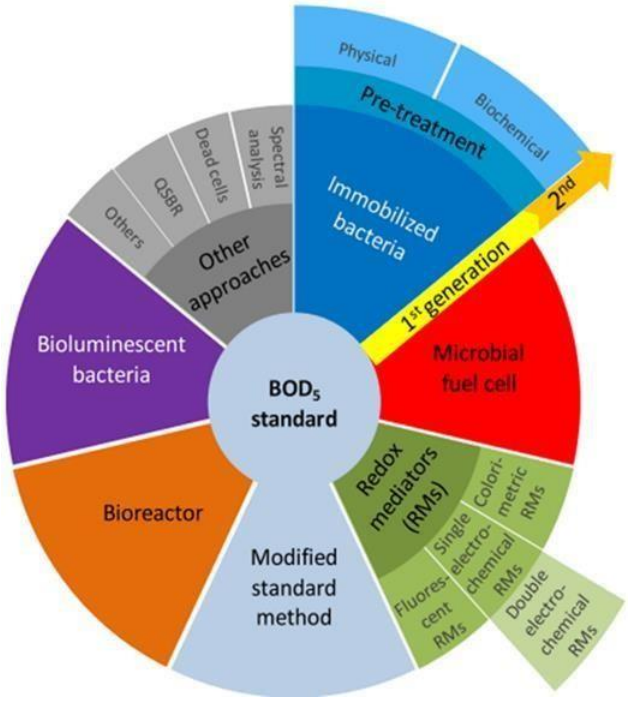


Figura 1.1 Métodos y estrategias para la estimación de DBO a 5 días. Jouanneau et al. (2014).

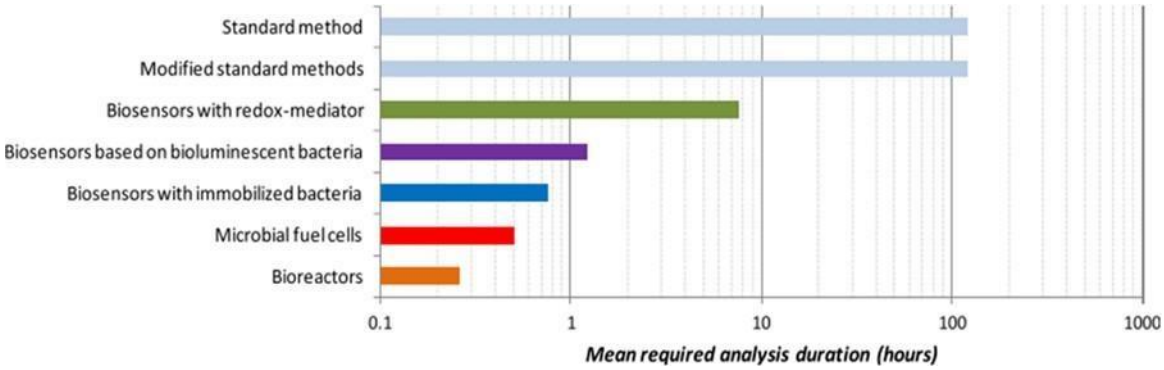


Figura 1.2 Tiempo requerido para el análisis. Jouanneau et al. (2014).

Debido al tiempo, materiales y conocimiento técnico para aplicar estos métodos en la determinación de la demanda bioquímica de oxígeno a 5 días (DBO5), el monitoreo en tiempo real de este parámetro puede presentar retrasos y complicaciones. A partir de lo anterior, la estimación de la DBO por medio de técnicas de inteligencia artificial es importante como herramienta de apoyo y referencia (Arreguin-Cortes & Cervantes-Jaimes, 2020). Por otro lado, identificar grupos de parámetros del agua superficial que muestren relación con la DBO facilitaría su determinación (M. Najafzadeh et al., 2019), esto al utilizar o implementar los instrumentos de medición de los parámetros identificados con los que cuente el especialista, en lugar de los métodos que comúnmente se aplican para determinar la DBO5. Por lo tanto, se tendría una alternativa más rápida para el diagnóstico y monitoreo de la calidad de agua superficial.

A partir de lo anterior, este trabajo presenta la identificación de parámetros del agua significativos con la DBO5 y la predicción de la DBO en aguas superficiales por medio de técnicas de inteligencia artificial. Además, se presenta el diseño, calibración y desarrollo de un dispositivo electrónico de medición que utilice tecnología de sensores para medir parámetros del agua.

1.2 Planteamiento del problema de investigación

La demanda bioquímica de oxígeno (DBO) se obtiene normalmente mediante la toma de muestras en la zona de estudio y su traslado a un laboratorio para el posterior análisis de la muestra. El análisis de las muestras se realiza mediante diversos métodos, en los que intervienen instrumentos y reactivos especializados. Este proceso convencional de recogida de muestras en la zona de estudio y su análisis en el laboratorio requiere mucho tiempo y trabajo. Como consecuencia, no es posible tener un monitoreo en tiempo real de la calidad del agua. Además, el diagnóstico de la contaminación se reduce a identificar y analizar con mayor frecuencia los puntos de control de las aguas superficiales que se encuentran cerca de la infraestructura de los laboratorios certificados.

Para ayudar el proceso de estudio realizado por los especialistas, es posible realizar análisis estadísticos y generar modelos predictivos mediante inteligencia artificial, a partir de los datos de medición obtenidos previamente por los especialistas. Se ha reportado la implementación de inteligencia artificial a través de aprendizaje máquina y minería de datos mediante algoritmos para la predicción de parámetros de calidad del agua en diferentes zonas de monitoreo (Di et al., 2019) (Melesse et al., 2020) (Mohammad Najafzadeh & Ghaemi, 2019) (Samsudin et al., 2019) (Abobakr Yahya et al., 2019) (Wang et al., 2017). Se caracterizan por diferentes etapas como el preprocesamiento, la normalización y la evaluación de los algoritmos de aprendizaje supervisado con estadísticas de bondad de ajuste (Ahmed et al., 2019).

De la misma manera, se han producido avances en los dispositivos electrónicos de monitoreo de la calidad del agua que utilizan la tecnología de sensores para ayudar a reducir el tiempo de recogida y análisis de las muestras de agua (Sagan et al., 2020) (Pujar et al., 2020). Además, para adquirir los valores de los parámetros se proponen diseños y desarrollos de dispositivos electrónicos utilizando sensores, actuadores, y microcontroladores (Rozario & Devarajan, 2020) (Gutiérrez et al., 2018) (Méndez-Barroso et al., 2020). Por ejemplo, para el análisis en ríos, la calidad del agua se clasificó por los parámetros de temperatura, pH, turbiedad y sólidos disueltos totales. Los datos se recogen en un río por medio de sensores, un microcontrolador como placa principal y la calidad del agua se clasifica mediante los algoritmos de K-Nearest Neighbors, máquina de vectores de apoyo, clasificador bayesiano y árboles de decisión. El rendimiento de los algoritmos se evalúa mediante la sensibilidad, la especificidad, la exactitud y la precisión (Rosero et al., 2020). Esto indica la posibilidad de involucrar parámetros fácilmente medibles en el área de estudio para diagnosticar la contaminación.

Del mismo modo, la aplicación del algoritmo de la máquina de vectores de apoyo para la predicción de la calidad del agua obtuvo un coeficiente de correlación de 0.97 y un error cuadrático medio de 0.058. Los datos se obtuvieron del Departamento de Medio Ambiente de Malasia. Los parámetros utilizados como entrada para los algoritmos fueron el pH, el oxígeno disuelto, la demanda bioquímica de oxígeno, la demanda química de oxígeno y el amoníaco-nitrógeno (Abobakr Yahya et al., 2019). Estos resultados muestran que los algoritmos de aprendizaje máquina pueden ser implementados para la predicción de parámetros particulares según las condiciones locales.

Para la predicción de la DBO se han implementado diferentes algoritmos de aprendizaje máquina. Por ejemplo, los algoritmos de regresión polinómica, el árbol modelo y la programación de la expresión génica. Los parámetros de entrada fueron Ca2+, Na+, Mg2+, NO-2, NO-3, PO3-4, conductividad eléctrica, pH y turbiedad. El algoritmo de programación de la expresión génica tuvo un rendimiento aceptable con un error cuadrático medio de 5.388 y un coeficiente de correlación de 0.86 (Najafzadeh et al., 2019).

Sin embargo, los parámetros que se utilizan como entrada de los algoritmos y en los que están en función la DBO5 dependen de las bases de datos y lo que contengan. Esto puede dificultar la selección de parámetros del agua que sean fáciles de medir o recolectar. Por lo que identificar alternativas de parámetros del agua de los cuales puede estar en función la DBO5 facilitaría su predicción. Además, contribuiría al monitoreo de la calidad del agua reduciendo el tiempo de análisis de las muestras de agua y permitiría utilizar y diseñar los instrumentos o medidores con los que se cuenten en los laboratorios.

1.3 Justificación del problema de investigación

La determinación de la demanda bioquímica de oxígeno a 5 días (DBO5) contribuye al establecimiento de la calidad del agua en cuerpos superficiales de México (CONAGUA, 2020). Debido a la importancia de este parámetro, la estimación de la DBO5 por medio de inteligencia artificial facilitaría y reduciría el tiempo que comúnmente los métodos estandarizados y variantes implican (Najafzadeh et al., 2019). Así mismo, la identificación de parámetros que se relacionan con la DBO5 permitirá establecer 3 grupos de estos y utilizarlos para la predicción en situaciones en las que se cuenten con los instrumentos, medidores y recursos para la medición de los parámetros. A partir de esto se podrá utilizar como herramienta de apoyo para el monitoreo continuo de este parámetro previo a obtener la medición de 5 días, alertar de altos niveles de contaminación y efectuar acciones que lo reduzcan.

1.4 Preguntas de investigación

Para esta investigación se presentan las siguientes preguntas.

1. ¿Qué relación existe entre parámetros del agua biológicos, químicos y físicos con la demanda bioquímica de oxígeno a 5 días (DBO5)?

2. ¿Cuáles parámetros del agua que presentan relación con la DBO5 son los relevantes para formar grupos de estos y adaptarse a los instrumentos de medición, disponibles en un laboratorio y zona de estudio?

3. ¿Qué algoritmo de aprendizaje máquina permite explicar el comportamiento de la DBO5 utilizando como entrada los grupos de parámetros previamente identificados?

4. ¿Qué algoritmo de aprendizaje máquina muestra un desempeño óptimo para la predicción de la DBO5 utilizando como entrada los grupos de parámetros previamente identificados?

5. ¿Qué componentes electrónicos permiten una replicación fácil para diseñar y desarrollar un dispositivo electrónico de medición de parámetros del agua utilizando tecnología de sensores?

6. ¿Qué sensores con precisión media facilitan la medición de parámetros del agua?

1.5 Objetivo general

Predecir la demanda bioquímica de oxígeno a 5 días (DBO5) en aguas superficiales de México por medio de técnicas de inteligencia artificial, minería, análisis de datos utilizando como entrada de los algoritmos tres grupos de parámetros del agua y un dispositivo electrónico de medición con sensores de bajo costo.

1.6 Objetivos específicos

1. Identificar parámetros que presenten relación con la demanda bioquímica de oxígeno a 5 días en aguas superficiales.

2. Formar 3 grupos de parámetros del agua que permitan predecir la DBO5 en aguas superficiales con las siguientes características: Grupo A de parámetros que se determinen en un laboratorio más rápido que la DBO5 por medio de métodos estandarizados. Grupo B de parámetros que se puedan determinar en la zona de estudio. Grupo C de parámetros que se puedan medir por medio de la tecnología de sensores.

3. Implementar los algoritmos de aprendizaje máquina, regresión lineal múltiple, regresión de cresta, bosques aleatorios y red elástica para identificar su comportamiento en la predicción de la DBO5 usando como entrada los grupos de parámetros A, B y C por separado.

4. Evaluar los algoritmos de aprendizaje máquina para la predicción de la DBO5 e identificar el que presente mayor desempeño al evaluarlos por estadísticos de bondad de ajuste de la raíz del error cuadrático medio, error absoluto medio y el coeficiente de determinación.

5. Diseñar, desarrollar y calibrar un dispositivo electrónico para medir el grupo de parámetros C que se identificaron por medio de la tecnología de sensores con precisión media.

1.7 Hipótesis

El desempeño del algoritmo de aprendizaje máquina bosques aleatorios al predecir la demanda bioquímica de oxígeno a 5 días (DBO5) en aguas superficiales es similar al utilizar como entrada 3 grupos con las siguientes características: grupo de parámetros que se determinan en un laboratorio más rápido que la DBO a 5 días por medio de métodos estandarizados, grupo de parámetros que se puedan determinar en la zona de estudio y grupo de parámetros que se puedan medir por medio de la tecnología de sensores.

1.8 Estructura de la tesis

En el capítulo dos se presenta el marco teórico con un contexto general sobre los parámetros de la calidad del agua y técnicas de inteligencia artificial que comúnmente se utilizan para analizar, recolectar e implementar la predicción de parámetros del agua como la demanda bioquímica de oxígeno a 5 días (DBO5).

En el capítulo tres se describe la metodología que se implementó en este trabajo para la identificación de los grupos de parámetros A, B y C propuestos anteriormente y las etapas para el entrenamiento y prueba de los algoritmos utilizados para la predicción de la DBO5. También se muestra el procedimiento de diseño, desarrollo y calibración del dispositivo electrónico de medición utilizando tecnología de sensores.

En el capítulo cuatro primero se presentan los resultados de la relación de los parámetros del agua con la DBO5 y la implementación y evaluación de los algoritmos de aprendizaje máquina utilizados para la predicción de la DBO5 utilizando como entrada los 3 grupos de parámetros.

En el capítulo cinco, se presentan las conclusiones, las contribuciones, y el trabajo futuro de esta investigación.

Capítulo 2. Marco Teórico

2.1 Calidad del agua

En México a partir de la década de 1970 se estableció un sistema de la calidad del agua con mediciones y escalas estandarizadas para representar la contaminación del agua. A ese sistema se le llamo índice da la calidad del agua (ICA) y permitía comparar los niveles de contaminación en diferentes zonas de la república. Con esto, si el agua estaba contaminada tendría un ICA cercano a 0% y si no estaba contaminada estaría cercano a 100% (Secretaria de Medio Ambiente y Recursos Naturales, 2013). Para la obtención del ICA se consideraban 18 parámetros con distinto peso para cada uno como se muestra en la Tabla 2.1.

Tabla 2.1 Parámetros con pesos relativos tomados por el ICA. SEMARNAT. (2013).

Parámetro	Peso (w)	Parámetro	Peso (w)
Demanda bioquímica de oxígeno	5.0	Nitrógeno en nitratos	2.0
Oxígeno disuelto	5.0	Alcalinidad	1.0
Coliformes fecales	4.0	Color	1.0
Coliformes totales	3.0	Dureza total	1.0
Sustancias activas al azul de metileno	3.0	Potencial de hidrogeno (pH)	1.0
Conductividad eléctrica	2.0	Sólidos suspendidos	1.0
Fosfatos totales	2.0	Cloruros	0.5
Gases y aceites	2.0	Sólidos disueltos	0.5
Nitrógeno amoniacal	2.0	Turbiedad	0.5

Después de tiempo, se sustituyó el ICA, debido al incremento de descargas en aguas nacionales por la población y la industria, ya que el ICA no consideraba algunos parámetros como metales pesados y compuestos orgánicos. Dentro de los nuevos parámetros que se consideraron para sustituir el ICA se decidió utilizar los que muestran una influencia humana ya sea por la presencia de centros urbanos o industriales.

Para el año 1996 se propuso la NOM-001-SEMARNAT-1996, que establece los límites máximos permisibles en aguas del territorio nacional (Norma mexicana NOM-001-

SEMARNAT-1996). Esta norma ya tomo en cuenta otro tipo de parámetros del agua y sus niveles máximos en diferentes cuerpos de agua, como mantos acuíferos, zonas de abastecimiento para uso público, agrícola entre otros. Se presentan en la Tabla 2.2 algunos de los parámetros más relevantes y sus límites máximos. A continuación, se describen una parte de los principales parámetros.

Tabla 2.2 Unidades y límites permisibles de los parámetros principales considerados para evaluar la calidad del agua. Norma mexicana NOM-001-SEMARNAT-1996.

Parámetros	Unidades	Límite máximo permisible
Color Verdadero	Unidades Pt/Co	15
Conductividad eléctrica	µS/cm	1- 200
pH	Unidades de pH	6,5-8,5
Turbiedad	NTU	3,0
Temperatura	°C	Limite a partir del parámetro a determinar
Demanda bioquímica de oxígeno a 5 días	mg/l	30<DBO<=120
Demanda química de oxígeno	mg/l	40<DQO>=200
Porcentaje de saturación de oxígeno disuelto	%	10<OD<=30 y 130<OD<=150
Eschericha coli	NMP/100	<1,1 NMP/100 ml
Sólidos suspendidos totales	mg/l	150<SST<=400
Coliformes fecales	NMP/100	<1,1 NMP/100 ml

2.1.1. Turbiedad

"La turbiedad en agua se debe a la presencia de partículas suspendidas y disueltas. Materia en suspensión como arcilla, cieno o materia orgánica e inorgánica finamente dividida, así como compuestos solubles coloridos, plancton y diversos microorganismos". (Norma mexicana NMX-AA-038-SCFI-2001. análisis de agua - determinación de turbiedad en aguas naturales, residuales y residuales tratadas - método de prueba). El método consiste en la comparación de intensidad de la luz dispersada por la muestra de agua bajo condiciones específicas y la intensidad de la luz dispersada por una referencia en las mismas condiciones, por lo que a mayor dispersión de luz se tiene mayor turbiedad.

2.1.2. pH

Se utiliza mayormente para evaluar las propiedades corrosivas del ambiente, o muestra. Los métodos electrométricos están basados en medir la diferencia de potencial de una celda

electroquímica, formada por dos medias celdas. La primera es un electrodo de medición y la segunda un electrodo de referencia. El potencial del electrodo de medición representa la función de la actividad del ion hidrogeno de la disolución de medición. (Norma mexicana NMX-AA-008-SCFI-2016. análisis de agua. - medición del pH en aguas naturales, residuales y residuales tratadas. - método de prueba).

2.1.3. Conductividad eléctrica

La conductividad eléctrica permite conocer el grado de mineralización de una muestra. Este parámetro muestra la capacidad de una solución para transportar una corriente eléctrica, que depende de la presencia de iones, concentración y de la temperatura. Este método se basa en la propiedad que adquiere el agua de conducir la corriente eléctrica cuando tiene iones disueltos (Norma mexicana NMX-AA-093-SCFI-2000. análisis de agua. -determinación de la conductividad electrolítica - método de prueba).

2.1.4. Sólidos suspendidos y disueltos totales

Todas las aguas contienen sustancias disueltas en cantidades variables que dependen de su origen. El agua puede contener varios tipos de sólidos, entre ellos, sólidos disueltos y los sólidos suspendidos. El principio de este método se basa en la medición cuantitativa de los sólidos y sólidos disueltos, así como la cantidad de materia orgánica contenidos en agua mediante la evaporación y calcinación de la muestra filtrada o no, en su caso, a temperaturas específicas, en donde los residuos son pesados y sirven de base para el cálculo de estos (Norma mexicana NMX-AA-034-SCFI-2015. análisis de agua. -medición de sólidos y sales disueltas en aguas naturales, residuales y residuales tratadas – método de prueba).

2.1.5. Color verdadero

El color en el agua puede deberse a la presencia del contenido natural de metales o iones metálicos en disolución, humus o residuos orgánicos, plancton o desechos industriales. El método se basa en la medición de color verdadero y/o aparente en una muestra de agua por medio de una comparación visual con una escala estandarizada de platino-cobalto, donde el analista realiza una comparación con una escala (Norma mexicana NMX-AA-045-SCFI-2001. análisis de agua. -determinación de color platino cobalto en aguas naturales, residuales y residuales tratadas - método de prueba).

2.1.6. Oxígeno disuelto

En el método electrométrico los electrodos de membrana sensible al oxígeno ya sean galvánicos o polarizados están constituidos por dos electrodos de metal en contacto con un electrolito soporte, separado de la disolución de muestra por medio de una membrana selectiva. En el cátodo, que usualmente es oro o platino, ocurre la reducción del oxígeno mientras que en el ánodo ocurre la oxidación del metal (plata o plomo). La diferencia básica entre el sistema galvánico y el polarizado es que en el primero la reacción en el electrodo ocurre espontáneamente, mientras que en el segundo es necesario aplicar un potencial externo para polarizar el electrodo indicador (Norma mexicana NMX-AA-012-SCFI-2001. análisis de agua. -determinación de oxígeno disuelto en aguas naturales, residuales y residuales tratadas - método de prueba).

2.1.7. Demanda química de oxígeno

La demanda química de oxígeno (DQO) del agua, medida a través del método del dicromato, puede ser considerada como una medida aproximada de la demanda teórica de oxígeno. Este parámetro es “la concentración de la masa de oxígeno equivalente a la cantidad de dicromato consumida por la materia disuelta y suspendida cuando una muestra de agua se trata con este oxidante bajo condiciones definidas”. (Norma mexicana NMX-AA-030/1-SCFI-2012. análisis de agua. -medición de la demanda química de oxígeno en aguas naturales, residuales y residuales tratadas. - método de prueba - parte 1 - método de reflujo abierto).

2.1.8. Demanda bioquímica de oxígeno

El método utilizado por la norma mexicana se basa en medir la cantidad de oxígeno que requieren los microorganismos para efectuar la oxidación de la materia orgánica presente en aguas naturales y residuales y se determina por la diferencia entre el oxígeno disuelto inicial y el oxígeno disuelto después de cinco días de incubación a 20°C (Norma mexicana NMX-AA-028-SCFI-2001 Análisis de agua - determinación de la demanda bioquímica de oxígeno en aguas naturales, residuales (DBO5) y residuales tratadas - método de prueba, 2001) . Además, existen varios métodos y propuestas que se han implementado. Estos métodos utilizan diferentes instrumentos y requieren diferente tiempo. En la Tabla 2.3 se presentan algunos de los más usados.

Tabla 2.3 Métodos para determinar la demanda bioquímica de oxígeno (DBO). Jouanneau et al. (2014).

Tecnología	Marcador de biodegradación	Transductor	Tiempo requerido
Método de referencia	O_2 Disuelto	Dosificación yodo métrica o Sonda electroquímica	5 días
Método de referencia modificado	O_2 Disuelto	Sonda óptica	5 días
Método fotométrico	O_2 Disuelto	Espectrofotómetro	5 días
Método manométrico	Presión	Manómetro	5 días
Biosensor basado en bacterias bioluminiscente	Actividad bioluminiscente	Iluminómetro	72 minutos
Células de combustible microbianas	Potencial eléctrico	Amperímetro	315 minutos
Mediador REDOX	Mediador-REDOX	Amperímetro, iluminómetro, fluorímetro o espectrofotómetro	15 minutos
Biosensor con bacterias atrapadas	O_2 Disuelto	Electroquímico o sonda óptica	10 minutos
Biorreactor	O_2 Disuelto	Electroquímico o sonda óptica	20 minutos

Como se mencionó previamente la CONAGUA a través de la Red Nacional de Medición de Calidad del Agua se basa en la NOM-001-SEMARNAT-1996 para identificar la calidad de cuerpos de agua superficial y subterránea. Los principales parámetros que establecen una calidad de agua contaminada para uso público son la demanda bioquímica de oxígeno a 5 días (DBO5), demanda química de oxígeno, enterococos y toxicidad. Si se incumple el límite máximo permisible en al menos unos de esto parámetros el agua se considera contaminada. Por lo que elegir la DBO5 como parámetro de referencia de la calidad del agua superficial permite realizar un mapeo de los cuerpos de agua del país, con el inconveniente del tiempo y condiciones que requiere el análisis y la muestra por medio de métodos estandarizados.
Algunos de los métodos de determinación de la DBO son cortos en tiempo, pero implican contar con equipos especializados y procedimientos con cierto nivel de dificultad, por lo que determinar la DBO por medio de técnicas de inteligencia artificial ayudaría a los especialistas al tener una referencia o estimado del nivel de DBO en poco tiempo.

2.2 Técnicas de inteligencia artificial

La inteligencia artificial se compone de distintas técnicas con el propósito de replicar o simular el comportamiento de la inteligencia del humano por medio de una máquina. El aprendizaje máquina es una de las técnicas que se encarga de aprender e identificar patrones a partir de datos. Este conocimiento sirve para después proporcionarle nuevos datos y obtener el comportamiento del fenómeno a partir de la entrada (Flach, P. 2012). El proceso de enseñanza se caracteriza por diferentes etapas como el preprocesamiento, la normalización y la evaluación de los algoritmos de aprendizaje supervisado con estadísticas de bondad de ajuste. A continuación, se describen las técnicas principales del proceso.

2.2.1 Descripción y exploración de base de datos

En esta etapa se identifica las variables de la base de datos con la que se trabaja, así como análisis de estadística básica para conocer el comportamiento de los datos. Además, se eliminan errores de captura por posibles ausencias del valor (Rodríguez-Ruiz et al., 2020).

2.2.2 Detección de valores atípicos

Para la detección de valores atípicos, el diagrama de caja, estadística descriptiva e histogramas de frecuencia permiten visualizar e identificar posibles valores que por errores de captura de la base de datos o anomalías de la variable se encuentren en la base de datos (Ahmed et al., 2019).

2.2.3 Normalización Z-Score

Z-Score es un método de normalización y estandarización que representa el número de desviaciones estándar y permite saber cuán lejos se está de la media para cada punto o parámetro bruto (Al-Ghamdi et al., 2021). La Ecuación 2.1 muestra la expresión de normalización z-score aplicada a cada parámetro, donde x representa el valor del parámetro, μ es la media del parámetro y σ es la desviación estándar:

$$z - score = (x - \mu)/\sigma \tag{2.1}$$

2.2.4 Coeficiente de correlación de Pearson

El método de Pearson permite extraer los parámetros que tienen mayor relación. Con el fin de encontrar las variables dependientes e independientes que tienen un comportamiento lineal (Melesse et al., 2020). La Ecuación 2.2 presenta la correlación de Pearson entre los valores de dos vectores con valores continuos, X_i y Y_i, donde $\bar{x}$ es la media del vector X_i, $\overline{y}$ es la media del vector Y_i y n el número de valores totales de la muestra.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2} \tag{2.2}$$

2.2.5 Forward Selection

Esta técnica evalúa primero la contribución individual de cada característica a la predicción de la variable dependiente, esto utilizando como única entrada del algoritmo una característica individual. A continuación, las características con la mayor contribución individual se clasifican en orden descendente y se agrupan. Se crea un conjunto de bases de datos añadiendo una característica cada vez. Este conjunto de características se utiliza como entrada al algoritmo y se evalúa el desempeño con métricas al predecir la variable dependiente (Noori et al., 2011). La Figura 2.1 ejemplifica el proceso.

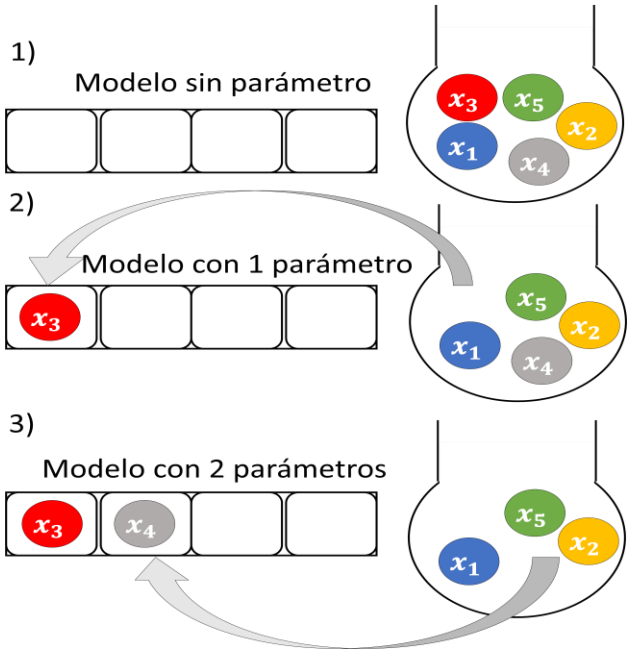


Figura 2.1 Proceso de Forward Selection. (1) Modelo sin parámetro, (2) Modelo con el primer parámetro con mayor contribución, (3) Modelo con el primer y segundo parámetro con mayor contribución ordenado de forma ascendente.

2.2.6 Curvas de aprendizaje

Las curvas de aprendizaje permiten identificar el número de datos de entrenamiento de los algoritmos de aprendizaje máquina necesarios para mejorar el desempeño. De igual manera que en Forward Selection en este proceso se evalúa el desempeño del algoritmo en el entrenamiento y prueba del algoritmo de aprendizaje máquina con estadísticos de bondad de ajuste. Al observar el comportamiento en entrenamiento y prueba del algoritmo se puede conocer si el modelo necesita más características y datos. El proceso inicia utilizando un ejemplo para el entrenamiento y se va aumentando sucesivamente este número hasta completar con los ejemplos que se establezcan de la base de datos para el entrenamiento (Raj, S., & Masood, S. 2020).

2.2.7 División de datos

La división de los datos para el entrenamiento y prueba de los algoritmos de aprendizaje máquina se enfoca en validar los algoritmos con un balance de las medidas. Una de las técnicas utilizadas es la validación cruzada, ya que esta técnica divide los datos en k subpartes e itera sobre todas las subpartes de la base de datos completa, teniendo para el entrenamiento k-1 subpartes y 1 subparte para la prueba (Chen et al., 2019). La Figura 2.2 muestra el proceso conk=3.

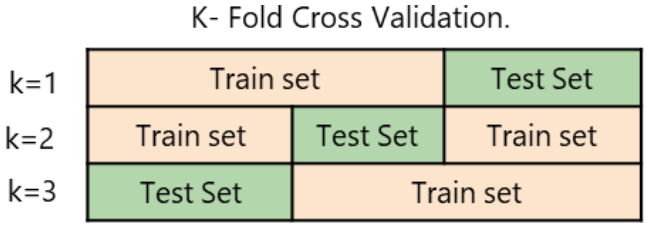


Figura 2.2 Proceso de división y validación cruzada K-Fold.

2.2.8 Algoritmos de aprendizaje máquina supervisado

2.2.8.1.1 Regresión lineal múltiple

La regresión lineal múltiple, generalmente se usa para la predicción de datos continuos. Es una forma de aplicar regresión lineal y se aplica cuando los datos de estudio tienen más de una variable de entrada en relación con la salida. Con este algoritmo, se puede estimar una Ecuación de la variable de salida en función de dos o más variables de entrada (El Bilali, A., & Taleb, A. 2020). La Figura 2.3 muestra una representación de una variable de salida *Y*en función de dos de entrada, X_1 y X_2.

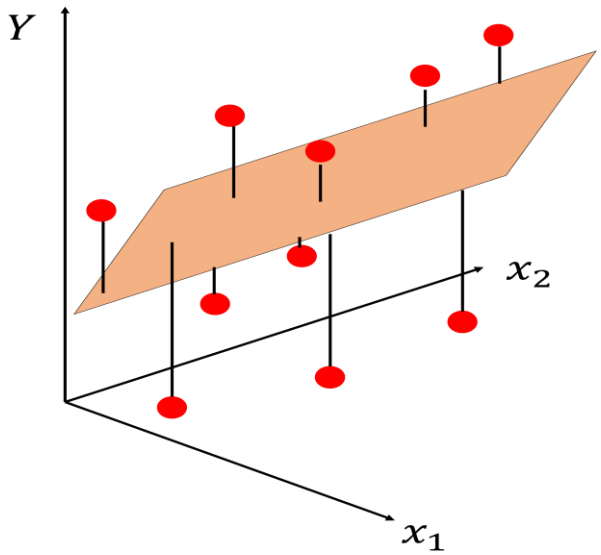


Figura 2.3 Representación visual de una regresión lineal múltiple con una variable de salida en función de dos variables de entrada.

2.2.8.1.2 Bosques aleatorios

Es un algoritmo que utiliza varios modelos de base por subgrupos de los datos de entrenamiento y toma decisiones a partir de todos los modelos generados. El modelo base es un árbol de decisión y se va generando varios modelos base dando buena eficiencia. Se puede utilizar para regresión y clasificación (Melesse et al., 2020). La Figura 2.4 muestra el proceso.

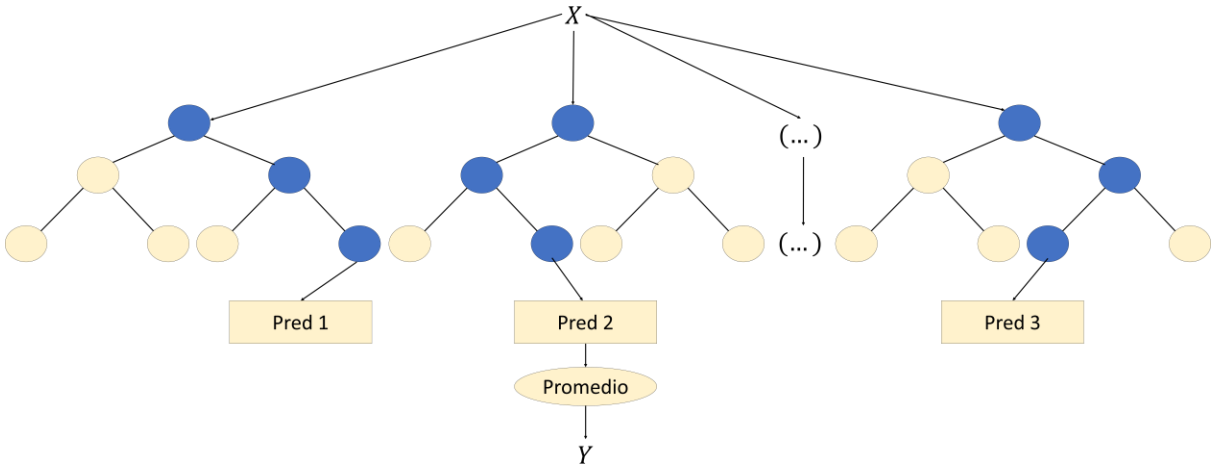


Figura 2.4 Representación algoritmo de bosques aleatorios.

2.2.8.1.3 Regresión de cresta

Trabaja por los mismos principios que una regresión lineal, y agrega un cierto sesgo para evitar el efecto de tener altas variaciones. Además, minimiza la suma de los residuos al cuadrado (Li et al., 2020).

2.2.8.1.4 Red elástica

Este algoritmo combina la eficiencia de regresión de cresta. Minimiza la función de costo al combinar los métodos de penalización de ambos algoritmos (Aheto et al., 2021).

2.2.9 Estadísticos de bondad de ajuste para evaluación de algoritmos de aprendizaje máquina

2.2.9.1.1 Coeficiente de determinación

El coeficiente de determinación representa la variación que existe entre las predicciones, los valores reales y la media de los valores. La Ecuación 2.3 presenta la expresión del coeficiente de determinación, donde y_i son los valores reales, y' son las predicciones, $\bar{y}$ representa la media de los valores y n el número de valores totales de la muestra:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2.3)$$

2.2.9.1.2 Raíz del error cuadrático medio

La raíz del error cuadrático medio va escalando los valores al rango de los valores del error cuadrático medio. La Ecuación 2.4 muestra la expresión, donde y_i son los valores reales, y' son las predicciones y n el número de valores totales de la muestra:

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

2.2.9.1.3 Error medio absoluto

El error absoluto medio representa la suma del valor absoluto del error, y se divide por el número total de valores de la muestra. La Ecuación 2.5 muestra la expresión, donde y_i son los valores reales, y' son las predicciones y n el número de valores totales de la muestra:

$$MAE = \frac{1}{n}\|y_i - y_i'\| \tag{2.5}$$

2.3 Adquisición de datos

La adquisición de datos es una herramienta que permite medir y censar parámetros físicos y recopilar información relevante de estos fenómenos. Por lo general el proceso de adquisición de datos se conforma de lo siguiente:

- Identificar la señal física de interés.
- Transformar la señal física a eléctrica por medio de un sensor.
- Acondicionar la señal eléctrica por medio de filtros y amplificación.
- Conversión de la señal analógica a digital.
- Procesamiento de la señal para su interpretación.

2.3.1 Tarjetas de adquisición de datos

Las tarjetas de adquisición de datos se utilizan para procesar y recopilar información acondicionada proveniente de sensores. Permiten comunicar y transferir información con una computadora, convertir señales analógicas a digitales entre otras aplicaciones. Se utilizan en diferentes áreas como investigación científica y procesos industriales.

2.3.2 Sensores

Los sensores convierten una señal proveniente de un fenómeno físico en otra señal por lo general de tipo eléctrica. Al diseñar sistemas que adquieren estas señales, hay aspectos acerca de los sensores que es necesario tener en cuenta:

- La naturaleza de la señal que el sensor genera, como pueden ser el voltaje, rango de amplitud y respuesta en frecuencia, entre otros. Y determina la precisión necesaria, el tipo de acondicionamiento de señal, convertidor analógico digital y cualquier otro hardware a utilizar
- La calibración del sensor para que describa el fenómeno físico y su comportamiento ya sea lineal o de otro tipo.
- La elección de los componentes y el diseño del sistema de adquisición de datos para que el sensor y el convertidor analógico digital tenga el mismo desempeño.
- La precisión para que en repetidas mediciones se obtenga el mismo valor en condiciones idénticas.
- El tiempo de respuesta al cambiar el valor de la medición de forma repentina.

Para este trabajo se analizan los sensores de turbiedad, conductividad eléctrica, pH, temperatura y temperatura ambiente.

2.4 Principales estudios relacionados

Zare Abyaneh (2014), utiliza los algoritmos de regresión lineal múltiple y redes neuronales artificiales para la predicción de la demanda bioquímica de oxígeno a 5 días y a demanda química de oxígeno. Los estadísticos de bondad de ajuste que evaluó fueron el coeficiente de correlación, la raíz del error cuadrático medio. Temperatura, pH, sólidos suspendidos totales como entrada de la red neuronal artificial obtuvo 25.1mg/l de RMSE, 0.83 de coeficiente de correlación y para la predicción de DBO, y para la predicción de DQO obtuvo 49.4 mg/l de RMSE y 0.81 de coeficiente de correlación. Este algoritmo obtuvo los mejores resultados al predecir la DBO.

De forma similar, Verma y Singh (2013) utilizan los parámetros de temperatura, pH, sólidos disueltos totales, sólidos suspendidos totales, oxígeno disuelto y aceites para estimar la DBO y DQO utilizando redes neuronales artificiales. Los datos los recolectaron de una zona de descarga de mina en India. Como resultado de evaluación del algoritmo se obtuvo de RMSE 0.114 y 0.98% para DBO y DQO respectivamente. Se destaca por utilizar parámetros que se pueden medir en el área de estudio.

Granata et al. (2017) propone la modelación de la DBO5 por medio de los algoritmos de máquina de vectores de apoyo (SVM) y árboles de regresión. Utilizaron una base de datos recolectada por la Universidad de Alabama en el periodo de 1992 a 2002. El coeficiente de determinación y la raíz del error cuadrático medio (RMSE) fueron los estadísticos de bondad de ajuste con los que evaluaron los algoritmos. Los algoritmos mostraron resultados similares, con 103 de RMSE y 0.8871 de coeficiente de determinación para arboles de regresión y 104 de

RMSE y 0.87 de coeficiente de determinación para SVM. Este trabajo utiliza el mismo parámetro de DBO5 como entrada y salida.

2.5 Comparación entre los trabajos relacionados y la propuesta de investigación

Para la comparación de los trabajos previos y el de esta investigación se presenta la Tabla 2.4. Con una revisión de literatura preliminar se identifica el uso de parámetros fáciles de medir en la zona de estudio y con la posibilidad de utilizar los instrumentos y medidores con los que se cuente en el laboratorio. Además, parámetros que podrían determinarse más rápido que la demanda bioquímica a 5 días en un laboratorio.

Tabla 2.4 Comparación trabajos relacionados

Autor	**Parámetros**	**Algoritmos**	**Parámetros fáciles de medir**
Najafzadeh et al. (2019)	Ca2+, Na+, Mg2+, NO-2, NO-3, PO3-4, conductividad eléctrica, pH y turbiedad	Regresión polinómica, árbol modelo, programación de expresión génica	3
Zare Abyaneh. (2014)	Temperatura, pH, sólidos suspendidos totales	Regresión lineal múltiple y redes neuronales artificiales	3
Verma and Singh. (2013)	Temperatura, pH, sólidos disueltos totales, sólidos suspendidos totales, oxígeno disuelto y aceites	Redes neuronales artificiales	5
Granata et al. (2017)	DBO5	Máquina de vectores de apoyo, árboles de regresión	0
Najafzadeh and Ghaemi. (2019)	Ca2+, Na+, Mg2+, NO-2, NO-3, PO3-4, conductividad eléctrica, pH y turbiedad	Spline de regresión adaptativa multivariante, máquina de vectores de apoyo de mínimos cuadrados, red neuronal artificial y sistema de inferencia neuro-fuzzy adaptativo.	3
Este trabajo	Demanda química de oxígeno, nitrógeno amoniacal, nitrógeno Kjeldahl, fósforo, color verdadero, absorción UV, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, pH, temperatura y temperatura del agua.	Regresión lineal múltiple, regresión de cresta, arboles aleatorios, red elástica.	***8***

2.6 Modelo o esquema general de investigación

A partir de lo planteado por Sampieri et al. (2017) en donde presenta el método científico, este trabajo seguirá un estudio correlacional y explicativo. Correlacional para identificar la relación que existe entre los parámetros del agua y la demanda bioquímica de oxígeno a 5 días (DBO5) para después predecir este parámetro por medio de técnicas de inteligencia artificial. Explicativo para tratar de representar el comportamiento de la DBO5 en aguas superficiales de México por medio de un grupo de parámetros del agua. Además, esta investigación se plantea como no experimental longitudinal ya que no se manipularán los parámetros de la calidad del agua utilizados de la base de datos de CONAGUA y solo se analizarán los parámetros del agua y la DBO5 para después predecir este parámetro. Se utilizará como muestras las regiones hidrológicas de México que están recolectadas en la base de datos. Esto en un periodo de tiempo del 2012 al 2019.

Se formarán 3 grupos de parámetros que cumplan con lo siguiente: Grupo de parámetros que se determinen en un laboratorio más rápido que la DBO a 5 días por medio de métodos estandarizados, grupo de parámetros que se puedan determinar en la zona de estudio y un grupo de parámetros que se puedan medir por medio de la tecnología de sensores. Continuando, se utilizarán estos grupos de parámetros para predecir la DBO5 por medio de algoritmos de aprendizaje máquina e identificar el que obtenga mayor desempeño.

Después, se plantea un cuasi experimento, al diseñar, calibrar y desarrollar un dispositivo electrónico con tecnología de sensores para medir los parámetros del grupo que se puedan obtener por medio de tecnología de sensores.

Capítulo 3. Método y propuesta de investigación

3.1Modelo de investigación

La metodología utilizada en este trabajo consiste en 2 etapas. La primera consiste en el procesamiento de la información para la predicción de la demanda bioquímica de oxígeno. La segunda fue el diseño e implementación de un dispositivo electrónico de medición. La Figura 3.1 muestra la metodología utilizada.

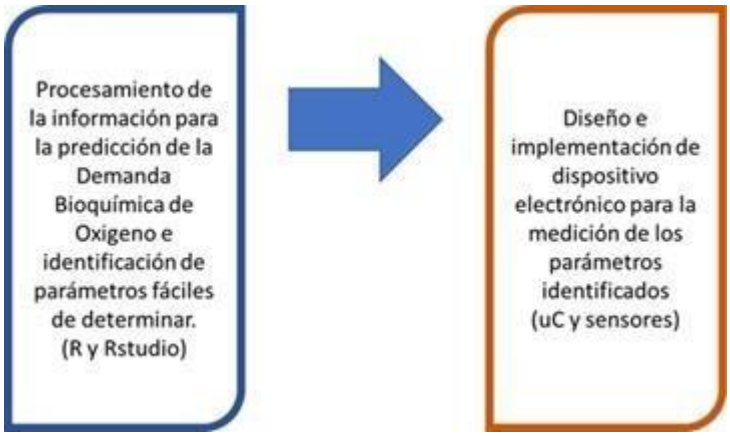


Figura 3.1 Etapas principales para determinar la demanda bioquímica de oxígeno a 5 días.

Se utilizó la base de datos recopilada por la CONAGUA, que contiene ubicación de sitios de monitoreo y sus parámetros del agua desde 2012 hasta 2019. El tipo de datos es cuantitativo con variables nominales y de medición de razón. Para el análisis se utilizó el software de Rstudio.

Se realizó un análisis descriptivo por parámetro del agua en los sitios de agua superficial de México. Por medio de diagramas de caja y estadística descriptiva se analizaron los parámetros del agua. Para conocer la relación significativa de parámetros del agua con la demanda bioquímica de oxígeno a 5 días (DBO5) se utilizó el coeficiente de correlación de Pearson y Forward Selection (FS). Después, las curvas de aprendizaje se propusieron para conocer el

número mínimo de muestras que se necesitan para entrenar los algoritmos de aprendizaje máquina.

A partir de lo anterior se seleccionaron los parámetros del agua que presentan relación significativa con la DBO5 y se formaron 3 grupos de parámetros para utilizarse como entrada de los algoritmos de aprendizaje máquina. Estos 3 grupos de parámetros se proponen que cumplan con lo siguiente:

- Grupo A de parámetros que se determinen en un laboratorio más rápido que la DBO5 por medio de métodos estandarizados.
- Grupo B de parámetros que se puedan determinar en la zona de estudio.
- Grupo C de parámetros que se puedan medir por medio de la tecnología de sensores.

Al seleccionar los grupos de parámetros se probó y entrenó algoritmos de aprendizaje máquina que recientemente se han reportado en la literatura con buen desempeño y fácil implementación para la estimación de la DBO utilizando como entrada todos los parámetros y los 3 grupos de parámetros seleccionados del agua que contiene la base de datos. El algoritmo que mostró mejor desempeño en coeficiente de determinación, el error absoluto medio y la raíz del error cuadrático medio se seleccionó para utilizarse y diseñar un dispositivo electrónico de medición con sensores que midan los parámetros del grupo C.

Finalmente, el dispositivo electrónico se diseñó con componentes y sensores comerciales de bajo costo que presentan una fácil reproducción y mantenimiento.

3.2 Preprocesamiento de la información para la predicción de la demanda bioquímica de oxígeno a 5 días

En esta sección se describe la primera etapa principal del procesamiento de la información. Consta de tres etapas secundarias para obtener la predicción de la demanda bioquímica de oxígeno. La primera etapa fue el preprocesamiento de los datos. La segunda etapa consistió en el análisis de los datos para la predicción de la demanda bioquímica de

oxígeno. En la última etapa se validan los algoritmos de aprendizaje máquina. En la Figura 3.2 se muestra el proceso.

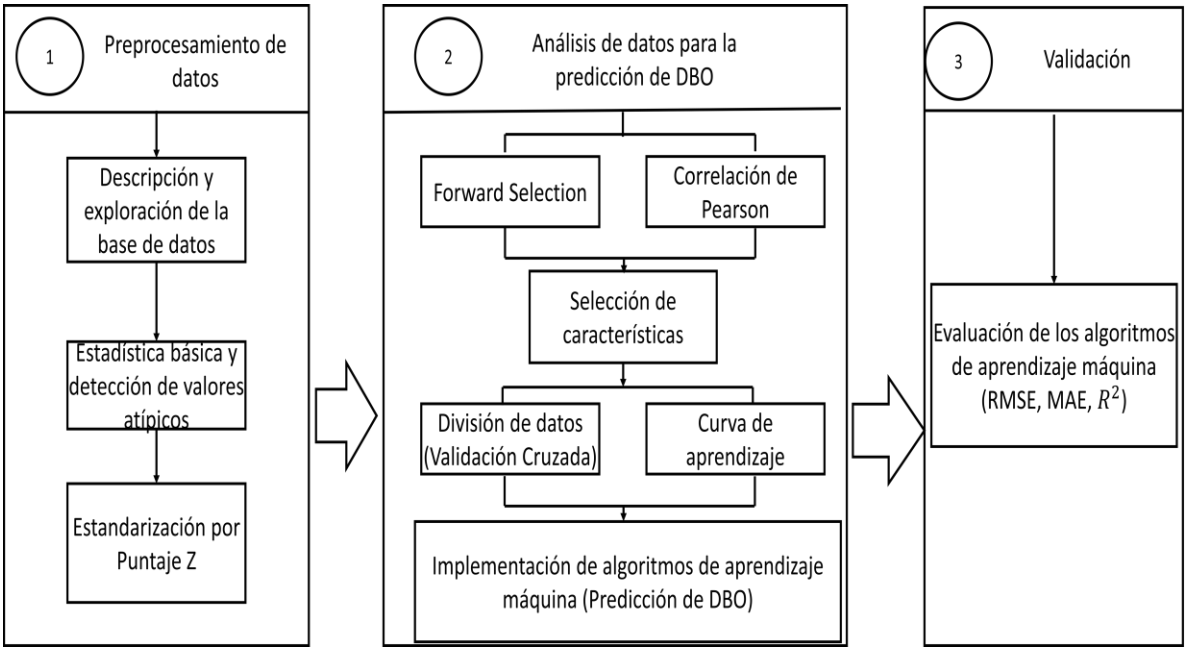


Figura 3.2 Etapas del procesamiento de la información. (1) Preprocesamiento de datos, (2) Análisis de datos para la predicción de la demanda bioquímica de oxígeno a 5 días y (3) Validación.

3.2.1 Preprocesamiento de datos

En esta sección se hace la descripción y exploración, detección de valores atípicos y normalización de base de datos. Se utilizó la base de datos recopilada por la CONAGUA, que contiene indicadores de sitios de monitoreo y parámetros del agua desde 2012 hasta 2019 (CONAGUA, 2020). La base de datos está compuesta por 177 parámetros químicos y biológicos del agua superficial en México, con un total de 110827 muestras. Se eliminaron los errores de captura y se obtuvieron 31 parámetros químicos y biológicos con un total de 59129 muestras. Los parámetros utilizados y sus estadísticas básicas se presentan en la Tabla 3.1.

Tabla 3.1 Estadística básica de los parámetros de la base de datos procesada. CONAGUA, (2020)

Parámetro (Unidad)	**Min**	**Media**	**Max**	**Parámetro (Unidad)**	**Min**	**Media**	**Max**
Coliforme fecal (NMP/100mL)	1	55772	24196000	Sólidos Suspendidos Totales (mg/L)	0.1	105	20812
Escherichia Coli (NMP/100mL)	1	46459	24196000	Turbiedad (NTU)	0.01	75	21500
Demanda bioquímica de oxígeno a 5 días(mg/L)	0.1	23.2	7667	Arsénico (mg/L)	0.0001	0.006	1
Demanda Química de Oxígeno (mg/L)	0.9	77.7	14489	Cadmio (mg/L)	0.00002	0.0002	0.1
Fósforo total(mg/L)	0.001	1.3	95.2	Cromo (mg/L)	0.0002	0.01	76.5
Nitrógeno Orgánico (mg/L)	0	2.5	827.8	Mercurio (mg/L)	0.00001	0.0003	0.5
Color Verdadero (U Pt/Co)	2.5	55.2	8000	Níquel (mg/L)	0	0.005	7.3
Absorción UV (U Abs/cm)	0.002	0.17	17	Plomo (mg/L)	0.001	0.003	1.8
Sólidos Disueltos Totales (mg/L)	2.4	354.5	159520	Dureza(mg/L)	3.8	295.2	37965
Conductividad Eléctrica (µS/cm)	3.8	1056	199400	Temperatura (°C)	-6	27.6	51
pH (UpH)	2.9	7.8	11.8	Temperatura Agua(°C)	4	24.9	62
Porcentaje de Oxígeno Disuelto (% Saturación)	0.6	73.2	1113.3	Carbono Orgánico Total (mg/L)	0.06	12.8	2490
Oxígeno Disuelto (mg/L)	0.05	5.7	762	Nitrógeno (mg/L)	0.008	7.4	1244.1
Nitrógeno Amoniacal (mg/L)	0.003	3.7	497	Nitrógeno Kjeldahl (mg/L)	0.003	6.34	1239.8
Dióxido de Nitrógeno (mg/L)	0.0005	0.1	21.84	Ortofosfato (mg/L)	0.0005	0.87	144.4
Nitrato (mg/L)	0.0004	1	336.2				

3.2.2 Detección valores atípicos

Para detectar los valores atípicos se eligió el diagrama de caja y la estadística básica. La mayoría de los parámetros variaron sus valores máximos debido a errores de medición o anomalías en la recolección. En la Figura 3.3 se muestra el código implementado para obtener el diagrama de caja de cada parámetro. Se utilizó la librería ggplot2.

```
diagrama_de_caja <- ggplot(datos, aes(x=AÑO, y=PARAMETRO,group=AÑO,color=AÑO)) +
  geom_boxplot()+ stat_summary(fun=mean, colour="black", geom="point",shape=18, size=2)+
  labs(title = "PARAMETRO", "\n boxplot",x="Year",y="UNIDAD")+
  theme(plot.title = element_text(color="blue",size=18,hjust = 0.5),
        axis.text.y = element_text(color="blue",size=12,hjust=1),
        axis.text.x = element_text(color="darkred",size=12,hjust=.5,vjust=.5),
        axis.title.x = element_text(color="blue", size=14),
        axis.title.y = element_text(size=14))+scale_color_gradientn(colours = rainbow(5))
```

Figura 3.3 Código implementado para obtener el diagrama de caja de los parámetros.

Implementar el código permite analizar los parámetros por año y se observó el comportamiento en todas las estaciones de monitoreo. En la Figura 3.4 se muestra el diagrama de caja como ejemplo de la demanda bioquímica de oxígeno.

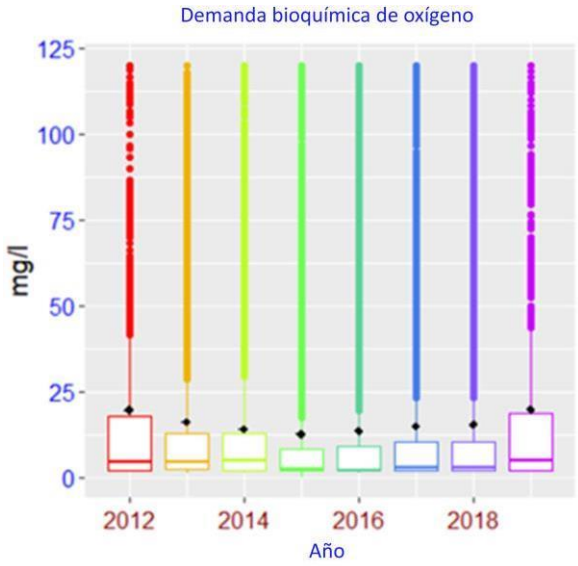


Figura 3.4 Diagrama de caja del parámetro demanda bioquímica de oxígeno a 5 días del año 2012 al 2019.

De forma similar en la Figura 3.5 se presenta el código implementado para obtener la estadística descriptiva de cada parámetro. Con este código se obtiene los valores mínimos, máximos, media y desviación estándar de cada parámetro.

```
mean(datos$PARAMETRO)
max(datos$PARAMETRO)
min(datos$PARAMETRO)
sd(datos$PARAMETRO)
```

Figura 3.5 Código implementado para obtener la estadística descriptiva de los parámetros.

3.2.3 Normalización

Para finalizar la primera etapa, se normalizaron los parámetros para establecer los valores de los parámetros en una escala común utilizando Z-Score. La Ecuación 3.1 muestra la expresión de normalización z-score aplicada a cada parámetro, donde x representa el valor del parámetro, μ es la media del parámetro y σ es la desviación estándar:

$$z - score = (x - \mu)/\sigma \tag{3.1}$$

La Figura 3.6 muestra el código implementado para realizar la normalización de cada parámetro.

```
dato$PARAMETRO<-((dato$PARAMETRO)-mean(dato$PARAMETRO))/sd(dato$PARAMETRO)
```

Figura 3.6 Código implementado para realizar la normalización de los parámetros.

3.3 Análisis de los datos para la predicción de la demanda bioquímica de oxígeno a 5 días

En esta sección, por un lado, se lleva a cabo el análisis de correlación de Pearson entre los parámetros de la base de datos, la implementación de forward selection y la selección de características para formar los 3 grupos de parámetros establecidos por este trabajo. Por otro lado, se realizan curvas de aprendizaje para los grupos de parámetros seleccionados, la división de datos para el entrenamiento y prueba de los algoritmos de aprendizaje máquina y por último la implementación de los algoritmos de aprendizaje máquina.

3.3.1 Coeficiente de correlación de Pearson

Para comenzar la etapa 2, se realizó un análisis de correlación mediante el método de Pearson. La Ecuación 3.2 presenta la expresión aplicada de la correlación de Pearson entre los valores de dos parámetros X_i y Y_i, donde $\bar{x}$ es la media del parámetro X_i, y $\bar{y}$ es la media del parámetro Y_i y n el número de valores totales de la muestra. Además, para observar la matriz de los coeficientes de correlación de Pearson entre los parámetros se generó un mapa de calor utilizando la librería corrplot. La Figura 3.7 muestra el código implementado para generar la matriz de coeficientes de correlación y el mapa de calor.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2} \tag{3.2}$$

```
matriz_coeficientes_correlacion_pearson<-cor(datos,use="pairwise.complete.obs", method = "pearson")
mapa_de_calor<-corrplot(matriz_coeficientes_correlacion_pearson, method = "ellipse",tl.srt =90)
```

Figura 3.7 Código implementado para generar la matriz de coeficientes de correlación de Pearson y el mapa de calor de los parámetros.

3.3.2 Forward Selection

Continuando con la etapa 2, se aplicó la técnica Forward Selection (FS) para apoyar la selección de grupos de parámetros. Este proceso se realiza con el 90% de las mediciones para el entrenamiento del algoritmo y el 10% para la prueba del mismo. El algoritmo utilizado para aplicar Forward Selection fue regresión lineal múltiple. El estadístico de bondad de ajuste utilizado para evaluar la contribución individual y conjunta de los parámetros fue el coeficiente de determinación R^2. La Figura 3.8 muestra el código implementado para seleccionar los datosde prueba y entrenamiento.

```
datos_para_utilizar<-datos[,1:ncol(datos)]
indices<-createDataPartition(datos$PARAMETRO_DBO,p=0.9, list=FALSE)
datos_Entrenamiento<-datos_para_utilizar[indices,]
datos_Prueba<-datos_para_utilizar[-indices,]
```

Figura 3.8 Código implementado para seleccionar las observaciones para entrenamiento y prueba.

La Figura 3.9 muestra el código implementado para dividir la base de datos para que solo contenga en esta etapa la demanda bioquímica de oxígeno a 5 días y otro parámetro.

```
for(i in 1:31){
  max_indice<-i
  particiones_para_entrenamiento[i]<-list(datos_Entrenamiento[,c(PARAMETRO_DBO,max_indice)])
}
for(i in 1:31){
  max_indice<-i
  particiones_para_prueba[i]<-list(datos_Prueba[,c(PARAMETRO_DBO,max_indice)])
}
```

Figura 3.9 Código implementado para dividir la base de datos con la demanda bioquímica de oxígeno a 5 días y otro parámetro individual.

Esto permite formar una base de datos por cada parámetro, por ejemplo, la primera base de datos contiene los parámetros demanda bioquímica de oxígeno a 5 días y coliformes fecales. La segunda base de datos contiene los parámetros demanda bioquímica de oxígeno a 5 días y Escherichia coli y así sucesivamente hasta generar las 30 bases de datos.
La Figura 3.10 muestra el código que se implementó para entrenar y probar el algoritmo de regresión lineal en cada base de datos que se obtuvo de dividir la base de datos original.

```
for(j in 1:31){
  datos_test<-data.frame(particiones_para_entrenamiento[j])
  datos_train<-data.frame(particiones_para_prueba[j])
  modelo<-lm(DBO_TOT ~.,data =  datos_train)
  predicciones_train = predict(modelo, newdata = datos_train)
  predicciones_test = predict(modelo, newdata = datos_test)
  #Coeficiente de determinacion en entrenamiento
  datosAComparar_train<-data.frame(Reales=datos_train$DBO_TOT, Predicciones= predicciones_train)
  R2_train<-R2(datosAComparar_train$Reales,datosAComparar_train$Predicciones)
  R2_train_1[j]<-R2_train
  #Coeficiente de determinacion en prueba
  datosAComparar_test<-data.frame(Reales=datos_test$DBO_TOT, Predicciones=predicciones_test)
  R2_test<-R2(datosAComparar_test$Reales,datosAComparar_test$Predicciones)
  R2_test_1[j]<-R2_test
}
```

Figura 3.10 Código implementado para el entrenamiento y prueba de algoritmo de regresión lineal con cada base de datos obtenida.

Después, se ordenaron de manera descendente los parámetros con mayor coeficiente de determinación. La Figura 3.11 muestra el código implementado que se utilizó para ordenar y generar las bases de datos los parámetros que obtuvieron mayor coeficiente de determinación.

```
R2_train_2<-R2_train_1
R2_test_2<-R2_test_1
indices_ordenados_train<-sort(R2_train_2, decreasing=TRUE,index.return = TRUE)
indices_ordenados_test<-sort(R2_test_2, decreasing=TRUE,index.return = TRUE)
particiones_ordenadas_para_entrenamiento<-list()
particiones_ordenadas_para_prueba<-list()
for(i in 2:31){
  particiones_ordenadas_para_entrenamiento[i]<-list(datos_Entrenamiento[, c(indices_ordenados_train$ix[PARAMETRO_DBO:i])])
}
for(i in 2:31){
  particiones_ordenadas_para_prueba[i]<-list(datos_Prueba[, c(indices_ordenados_test$ix[PARAMETRO_DBO:i])])
}
```

Figura 3.11 Código implementado para ordenar y generar las bases de datos de los parámetros que obtuvieron mayor coeficiente de determinación.

La Figura 3.12 muestra el código que se implementó para entrenar y probar el algoritmo de regresión lineal múltiple en cada base de datos que se obtuvo al ordenar los parámetros que obtuvieron mayor coeficiente de determinación.

```
R2_train_3<-c()
R2_test_3<-c()
for(j in 2:31){
  datos_test1<-data.frame(particiones_ordenadas_para_prueba[j])
  datos_train1<-data.frame(particiones_ordenadas_para_entrenamiento[j])
  modelo1<-lm(DBO_TOT~.,data =  datos_train1)
  predicciones_train1 = predict(modelo1, newdata = datos_train1)
  predicciones_test1 = predict(modelo1, newdata = datos_test1)
  #Coeficiente de determinacion en entrenamiento
  datosAComparar_train1<-data.frame(Reales=datos_train1$DBO_TOT, Predicciones= predicciones_train1)
  R2_train11<-R2(datosAComparar_train1$Reales,datosAComparar_train1$Predicciones)
  R2_train_3[j]<-R2_train11
  #Coeficiente de determinacion en prueba
  datosAComparar_test1<-data.frame(Reales=datos_test1$DBO_TOT, Predicciones=predicciones_test1)
  R2_test11<-R2(datosAComparar_test1$Reales,datosAComparar_test1$Predicciones)
  R2_test_3[j]<-R2_test11
}
```

Figura 3.12 Código implementado para el entrenamiento y prueba de algoritmo de regresión lineal múltiple con cada base de datos obtenida con los parámetros ordenados.

3.3.3 Selección de características

Después de aplicar Forward Selection, la selección de características consiste en formar 3 grupos de parámetros que cumplieran con los objetivos de este trabajo utilizando el análisis de correlación de Pearson y Forward Selection. Los grupos buscados son los siguientes:

A. Grupo de parámetros que se determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio.
B. Grupo de parámetros que se puedan determinar en la zona de estudio.
C. Grupo de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores.

En la Figura 3.13 se muestran los grupos de parámetros que se propusieron.

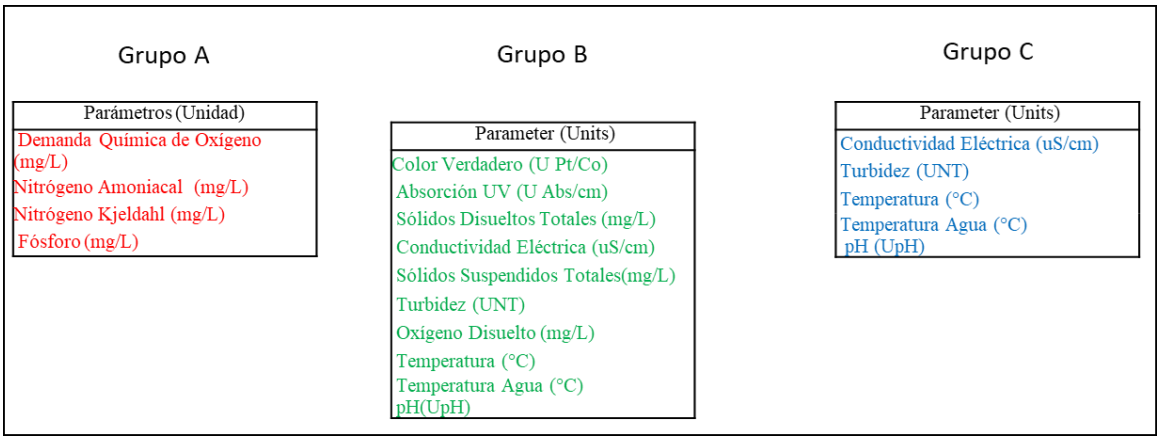


Figura 3.13 Grupos de parámetros propuestos.

3.3.4 Curvas de aprendizaje

Al seleccionar los grupos de parámetros, se implementaron curvas de aprendizaje para cada grupo y lograr identificar el número de datos de entrenamiento de los algoritmos de aprendizaje máquina necesarios para mejorar el desempeño.
De igual manera que en Forward Selection en este proceso regresión lineal múltiple fue el algoritmo utilizado y como estadístico de bondad de ajuste el coeficiente de determinación, además se realiza con el 90% de los ejemplos para el entrenamiento del algoritmo y el 10% para la prueba del mismo. La Figura 3.14 muestra el código que se implementó para entrenar y probar el algoritmo de regresión lineal múltiple aumentando en cada iteración el número de datos para entrenamiento.

```
Max_indice<-0
R2_train_4<-c()
R2_test_4<-c()
for(j in 3:dim(datos_Entrenamiento)[1]){
  Max_indice<-j
  datos_train5<-(datos_Entrenamiento[(1) : (Max_indice-1),])
  datos_test5<-datos_Prueba
  modelo3<-lm(DBO_TOT ~.,data =  datos_train5)
  predicciones_train4 = predict(modelo3, newdata = datos_train5)
  predicciones_test4 = predict(modelo3, newdata = datos_test5)
  #Coeficiente de determinacion en entrenamiento
  datosAComparar_train5<-data.frame(Reales=datos_train5$DBO_TOT, Predicciones= predicciones_train4)
  R2_train7<-R2(datosAComparar_train5$Reales,datosAComparar_train5$Predicciones)
  R2_train_4[j]<-R2_train7
  #Coeficiente de determinacion en prueba
  datosAComparar_test5<-data.frame(Reales=datos_test5$DBO_TOT, Predicciones=predicciones_test4)
  R2_test9<-R2(datosAComparar_test5$Reales,datosAComparar_test5$Predicciones)
  R2_test_4[j]<-R2_test9
}
```

Figura 3.14 Código implementado para el entrenamiento y prueba de algoritmo de regresión lineal múltiple utilizando como entrada los tres grupos de parámetros y aumentando en cada iteración el número de datos para entrenamiento.

3.3.5 División de datos

Después de conocer el número de datos necesarios para el entrenamiento para cada grupo, se llevó a cabo la división de los datos para el entrenamiento y prueba de los algoritmos de aprendizaje máquina. La técnica utilizada fue la validación cruzada y para cada grupo de parámetros se utilizó k=3. De las 59129 mediciones de la base de datos para entrenamiento se seleccionaron 53218 y para prueba 5911. El número de mediciones para entrenamiento se tomaron de los 53218 y en cada doblez se entrenó con 35479 mediciones. En la Figura 3.15 semuestra el código implementado y la estructura del entrenamiento.

```
tipoEntrenamiento<-trainControl(
  method = "cv",
  number = 3,
  savePredictions = "final",
  classProbs = FALSE
)
```

```
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 35479, 35479, 35478
```

Figura 3.15 Código implementado para establecer la estructura de entrenamiento de los algoritmos.

3.4 Predicción de la demanda bioquímica de oxígeno a 5 días por medio de algoritmos de aprendizaje máquina supervisado

Para finalizar la segunda etapa, se implementaron cuatro algoritmos de aprendizaje máquina utilizando la librería mlr3verse para predecir la demanda bioquímica de oxígeno a 5 días en aguas superficiales utilizando los 3 grupos de parámetros previamente seleccionados.

3.4.1 Regresión lineal múltiple

El primer algoritmo utilizado fue la regresión lineal múltiple. La Figura 3.16 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio. La Ecuación 3.3 muestra los coeficientes al utilizar como entrada del algoritmo el grupo A.

```
predictores<-c("N_TOTK","DQO_TOT","P_TOT", "N_NH3")
Salida<-"DBO_TOT"
modelo_lm<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "lm",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_lm<-predict(object = modelo_lm,
                                      datos_Prueba[,predictores])
```

Figura 3.16 Código implementado del algoritmo de regresión lineal múltiple entrada con grupo A.

$$DBO = -3.58 + 0.28 DemandaQuímicaOxígeno - 0.12 ANitrógenoAmoniacal + 0.35\ NitrógenoKjeldahl\ + 1.01 Fósforo \tag{3.3}$$

La Figura 3.17 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se puedan determinar en la zona de estudio. La Ecuación 3.4 muestra los coeficientes al utilizar como entrada del algoritmo el grupo B.

```
predictores<-c("COLOR_VER" ,"ABS_UV","OD_mg.L","SDT","CONDUC_CAMPO", "SST",
               "TURBIEDAD","pH_CAMPO", "TEMP_AMB", "TEMP_AGUA" )

Salida<-"DBO_TOT"
modelo_lm<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "lm",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_lm<-predict(object = modelo_lm,
                                      datos_Prueba[,predictores])
```

Figura 3.17 Código implementado del algoritmo de regresión lineal múltiple entrada con grupo B.

$$DBO = 12.7 + 0.09 ColorVerdadero + 34.4 AbsorciónUV + 0.025 SólidosDisueltosTotales - 0.002 ConductividadEléctrica + 0.044 SólidosSupendidosTotales - 0.03 Turbiedad, -2.8 OxígenoDisuelto - 0.09 Temperatura - 0.48 TemperaturaAgua + 1.51 pH \tag{3.4}$$

La Figura 3.18 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores. La Ecuación 3.5 muestra los coeficientes al utilizar como entrada del algoritmo el grupo C.

```
predictores<-c( "CONDUC_CAMPO",
                "TURBIEDAD", "TEMP_AMB",
                "TEMP_AGUA", "pH_CAMPO")

Salida<-"DBO_TOT"
modelo_lm<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "lm",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_lm<-predict(object = modelo_lm,
                                      datos_Prueba[,predictores])
```

Figura 3.18 Código implementado del algoritmo de regresión lineal múltiple entrada con grupo C.

$$DBO = 77 + 0.007ConductividadEléctrica + 0.05Turbidad - 0.47Temperatura \quad (3.5)$$
$$-0.8TemperaturaAgua - 4.9pH$$

Las ecuaciones 3.3, 3.4 y 3.5 generadas por el algoritmo de regresión lineal múltiple se utilizaron para la predicción de la predicción de la demanda bioquímica de oxígeno a 5 días a partir de los grupos de parámetros A, B y C.

3.4.2 Bosques aleatorios

El segundo algoritmo utilizado fue el de bosques aleatorios. La Figura 3.19 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio. En la Tabla 3.2 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("N_TOTK","DQO_TOT","P_TOT", "N_NH3")
Salida<-"DBO_TOT"
modelo_rf<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ranger",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rf<-predict(object = modelo_rf,
                                      datos_Prueba[,predictores])
```

Figura 3.19 Código implementado del algoritmo bosques aleatorios entrada con grupo A.

Tabla 3.2 Condiciones de operación del algoritmo de bosques aleatorios entrada con grupo A.

Tipo	Regresión
Número de árboles	500
Tamaño Muestra	41392
Número de variables independientes	4
Mtry	2
Tamaño del nodo de destino	5
Modo de importancia variable	Ninguno
Splitrule	Extratrees

La Figura 3.20 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se puedan determinar en la zona de estudio. En la Tabla 3.3 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("COLOR_VER" ,"ABS_UV","OD_mg.L","SDT","CONDUC_CAMPO", "SST",
               "TURBIEDAD","pH_CAMPO", "TEMP_AMB", "TEMP_AGUA" )

Salida<-"DBO_TOT"
modelo_rf<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ranger",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rf<-predict(object = modelo_rf,
                                      datos_Prueba[,predictores])
```

Figura 3.20 Código implementado del algoritmo bosques aleatorios entrada con grupo B.

Tabla 3.3 Condiciones de operación del algoritmo de bosques aleatorios entrada con grupo B.

Tipo	Regresión
Número de árboles	500
Tamaño Muestra	41392
Número de variables independientes	10
Mtry	2
Tamaño del nodo de destino	5
Modo de importancia variable	Ninguno
Splitrule	Variance

La Figura 3.21 muestra el código implementado al utilizar como entrada del grupo de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores. En la Tabla 3.4 se muestran las condiciones de operación del algoritmo.

```
predictores<-c( "CONDUC_CAMPO",
                "TURBIEDAD", "TEMP_AMB",
                "TEMP_AGUA", "pH_CAMPO")

Salida<-"DBO_TOT"
modelo_rf<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ranger",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rf<-predict(object = modelo_rf,
                                      datos_Prueba[,predictores])
```

Figura 3.21 Código implementado del algoritmo bosques aleatorios entrada con grupo C.

Tabla 3.4 Condiciones de operación del algoritmo de bosques aleatorios entrada con grupo C.

Tipo	Regresión
Número de árboles	500
Tamaño Muestra	41392
Número de variables independientes	5
Mtry	3
Tamaño del nodo de destino	5
Modo de importancia variable	Ninguno
Splitrule	Extratrees

Las condiciones de operación mostradas en las Tablas 3.2, 3.3 y 3.4 generadas por el algoritmo de bosques aleatorios se utilizaron para la predicción de la predicción de la demanda bioquímicade oxígeno a 5 días a partir de los grupos de parámetros A, B y C.

3.4.3 Regresión de cresta

El tercer algoritmo implementado fue regresión de cresta. La Figura 3.22 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio. En la Figura 3.23 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("N_TOTK","DQO_TOT","P_TOT", "N_NH3")
Salida<-"DBO_TOT"
modelo_rg<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ridge",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rg<-predict(object = modelo_rg,
                                      datos_Prueba[,predictores])
```

Figura 3.22 Código implementado del algoritmo regresión de cresta entrada con grupo A.

```
.
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Sequence of  moves:
     DQO_TOT N_TOTK P_TOT N_NH3
Var        2      1     3     4 5
Step       1      2     3     4 5
```

Figura 3.23 Condiciones de operación del algoritmo regresión de cresta entrada con grupo A.

La Figura 3.24 muestra el código implementado al utilizar como entrada del algoritmo el grupo de parámetros que se puedan determinar en la zona de estudio. En la Figura 3.25 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("COLOR_VER" ,"ABS_UV","OD_mg.L","SDT","CONDUC_CAMPO", "SST",
               "TURBIEDAD","pH_CAMPO", "TEMP_AMB", "TEMP_AGUA" )

Salida<-"DBO_TOT"
modelo_rg<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ridge",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rg<-predict(object = modelo_rg,
                                      datos_Prueba[,predictores])
```

Figura 3.24 Código implementado del algoritmo regresión de cresta entrada con grupo B.

```
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Sequence of  moves:
     ABS_UV OD_mg.L COLOR_VER SDT TEMP_AGUA SST TEMP_AMB pH_CAMPO TURBIEDAD CONDUC_CAMPO
Var       2       3         1   4        10   6        9        8         7            5 11
Step      1       2         3   4         5   6        7        8         9           10 11
```

Figura 3.25 Condiciones de operación del algoritmo regresión de cresta entrada con grupo B.

La Figura 3.26 muestra el código implementado al utilizar como entrada el grupo de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores. En la Figura 3.27 se muestran las condiciones de operación del algoritmo.

```
predictores<-c( "CONDUC_CAMPO",
                "TURBIEDAD", "TEMP_AMB",
                "TEMP_AGUA", "pH_CAMPO")

Salida<-"DBO_TOT"
modelo_rg<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "ridge",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_rg<-predict(object = modelo_rg,
                                      datos_Prueba[,predictores])
```

Figura 3.26 Código implementado del algoritmo regresión de cresta entrada con grupo C.

```
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Sequence of  moves:
     CONDUC_CAMPO TURBIEDAD TEMP_AGUA TEMP_AMB pH_CAMPO
Var             1         2         4        3        5 6
Step            1         2         3        4        5 6
```

Figura 3.27 Condiciones de operación del algoritmo regresión de cresta entrada con grupo C.

Las condiciones de operación mostradas en las Figuras 3.23, 3.25 y 3.27 generadas por el algoritmo de regresión de cresta se utilizaron para la predicción de la predicción de la demanda bioquímica de oxígeno a 5 días a partir de los grupos de parámetros A, B y C.

3.4.4 Red elástica

Por último, en este trabajo se utilizó el algoritmo de red elástica. En la Figura 3.28 se muestra el código implementado utilizando como entrada del algoritmo el grupo de parámetros que se determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio. En la Figura 3.29 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("N_TOTK","DQO_TOT","P_TOT", "N_NH3")
Salida<-"DBO_TOT"
modelo_el<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "enet",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_el<-predict(object = modelo_el,
                                      datos_Prueba[,predictores])
```

Figura 3.28 Código implementado del algoritmo red elástica entrada con grupo A.

```
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Sequence of  moves:
     DQO_TOT N_TOTK P_TOT N_NH3
Var        2      1     3     4 5
Step       1      2     3     4 5
```

Figura 3.29 Condiciones de operación del algoritmo red elástica entrada con grupo A.

En la Figura 3.30 se muestra el código implementado utilizando como entrada del algoritmo el grupo de parámetros que se puedan determinar en la zona de estudio. En la Figura 3.31 se muestran las condiciones de operación del algoritmo.

```
predictores<-c("COLOR_VER" ,"ABS_UV","OD_mg.L","SDT","CONDUC_CAMPO", "SST",
               "TURBIEDAD","pH_CAMPO", "TEMP_AMB", "TEMP_AGUA" )

Salida<-"DBO_TOT"
modelo_el<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "enet",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_el<-predict(object = modelo_el,
                                      datos_Prueba[,predictores])
```

Figura 3.30 Código implementado del algoritmo red elástica entrada con grupo B.

```
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Cp statistics of the Lasso fit
Cp: 36660.056 25447.219 18822.606 13037.526  3948.301  2638.518  1212.151   717.844   680.486   483.965    11.000
DF:  1  2  3  4  5  6  7  8  9 10 11
Sequence of  moves:
     ABS_UV OD_mg.L COLOR_VER SDT TEMP_AGUA SST TEMP_AMB pH_CAMPO TURBIEDAD CONDUC_CAMPO
Var       2       3         1   4        10   6        9        8         7           5 11
Step      1       2         3   4         5   6        7        8         9          10 11
```

Figura 3.31 Condiciones de operación del algoritmo red elástica entrada con grupo B.

En la Figura 3.32 se muestra el código implementado utilizando el grupo de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores. En la Figura 3.33 se muestran las condiciones de operación del algoritmo.

```
predictores<-c( "CONDUC_CAMPO",
                "TURBIEDAD", "TEMP_AMB",
                "TEMP_AGUA", "pH_CAMPO")

Salida<-"DBO_TOT"
modelo_el<-train(datos_Entrenamiento[,predictores], datos_Entrenamiento[,Salida], method = "enet",
                 trControl = tipoEntrenamiento,
                 tuneLength = 3)
datos_Prueba$predicciones_el<-predict(object = modelo_el,
                                      datos_Prueba[,predictores])
```

Figura 3.32 Código implementado del algoritmo red elástica entrada con grupo C.

```
Call:
elasticnet::enet(x = as.matrix(x), y = y, lambda = param$lambda)
Sequence of  moves:
     CONDUC_CAMPO TURBIEDAD TEMP_AGUA TEMP_AMB pH_CAMPO
Var             1         2         4        3        5 6
Step            1         2         3        4        5 6
```

Figura 3.33 Condiciones de operación del algoritmo red elástica entrada con grupo C.

Las condiciones de operación mostradas en las Figuras 3.29, 3.31 y 3.33 generadas por el algoritmo de red elástica se utilizaron para la predicción de la predicción de la demanda bioquímica de oxígeno a 5 días a partir de los grupos de parámetros A, B y C.
A continuación, se describe la evaluación de los algoritmos de aprendizaje máquina.

3.4.5 Evaluación de algoritmos de aprendizaje máquina

En esta sección se describen los estadísticos de bondad de ajuste utilizados para evaluar los algoritmos de aprendizaje máquina. En la etapa 3, los algoritmos se evaluaron en el entrenamiento y en la prueba. La evaluación se realizó mediante los estadísticos de bondad de ajuste de la raíz del error cuadrático medio (RMSE), el error medio absoluto (MAE) y el coeficiente de determinación (R^2). En la Figura 3.34 se muestra el código implementado para calcular los estadísticos de bondad de ajuste en la etapa de entrenamiento y prueba de los algoritmos.

```
R2_entrenamiento<-modelo$results[9,4]
MAE_entrenamiento<-modelo$results[9,5]
RMSE_entrenamiento<-modelo$results[9,3]
R2_prueba<-R2(datos_prueba$DBO_TOT,datos_prueba$predicciones_modelo)
MAE_prueba<-MAE(datos_prueba$DBO_TOT,datos_prueba$predicciones_modelo)
RMSE_prueba<-RMSE(datos_prueba$DBO_TOT,datos_prueba$predicciones_modelo)
```

Figura 3.34 Código implementado para el cálculo de los estadísticos de bondad de ajuste.

3.5 Diseño e implementación de dispositivo electrónico de medición

En esta sección se presenta la segunda etapa principal. Se describe el procedimiento que se realizó para el diseño y desarrollo del dispositivo electrónico de medición utilizando sensores que midan los parámetros del grupo C previamente identificados en la selección de características por las técnicas de coeficiente de correlación de Pearson y Forward Selection (FS). Los parámetros fueron turbiedad, conductividad eléctrica, pH, temperatura del agua y temperatura ambiente. En la Figura 3.35 se muestran las etapas.

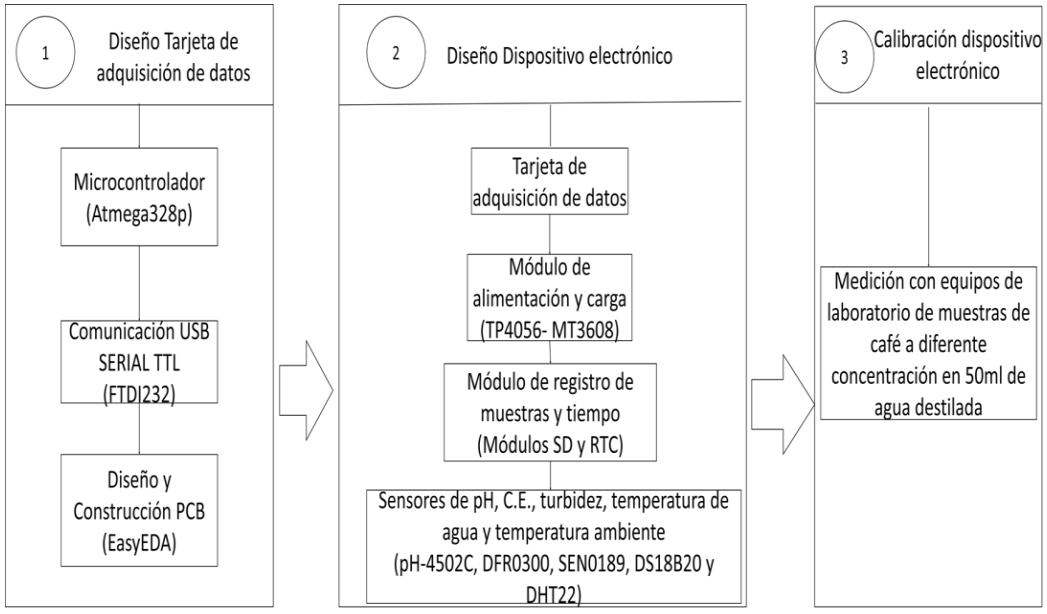


Figura 3.35 Etapas secundarias del diseño e implementación del dispositivo electrónico para la medición de parámetros identificados. (1) Diseño de tarjeta de adquisición de datos. (2) Diseño de dispositivo electrónico. (3) Calibración.

3.5.1 Tarjeta de adquisición de datos

Para el control del dispositivo se diseñó y construyó una tarjeta de adquisición de datos utilizando como componentes principales el microcontrolador atmega328p y un dispositivo FTDI FT232 para la comunicación serial y programación con una computadora. En la Figura 3.36 se muestra el esquema de la tarjeta de adquisición de datos.

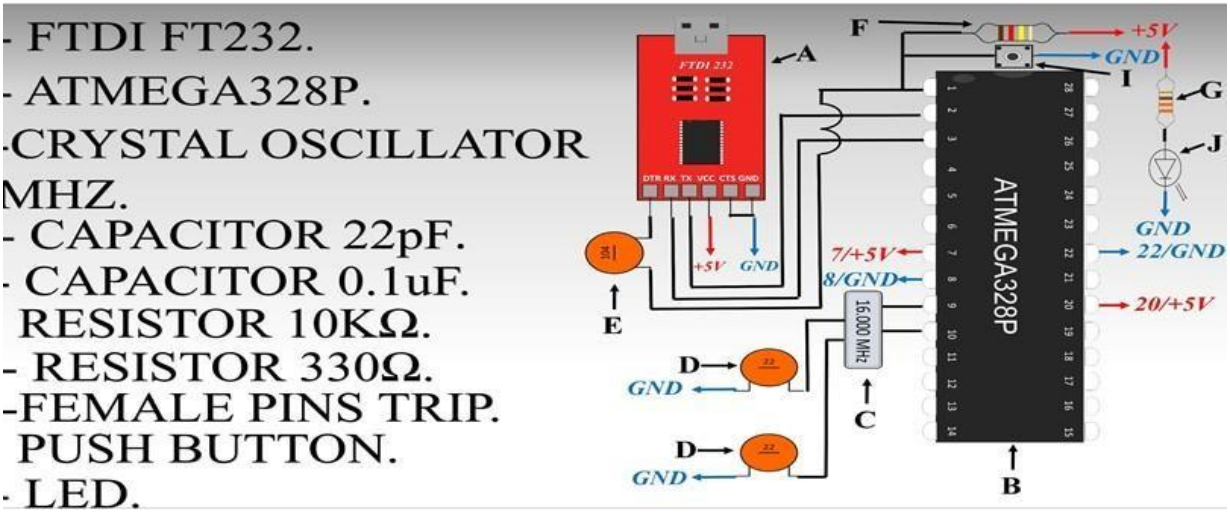


Figura 3.36 Esquema de tarjeta de adquisición de datos con microcontrolador atmega328p.

Se diseñó el circuito de la tarjeta de adquisición de datos para que su construcción fuera sencilla y rápida de realizar. La placa base se obtuvo por un método casero al transferir la tinta del circuito en impresión láser a una placa de cobre por medio de calor, para después introducirla en ácido y eliminar el cobre excedente. Después se realizaron perforaciones y se soldaron los componentes. La Figura 3.37 muestra la tarjeta de adquisición de datos construida y los componentes de presentan en la Tabla 3.5.

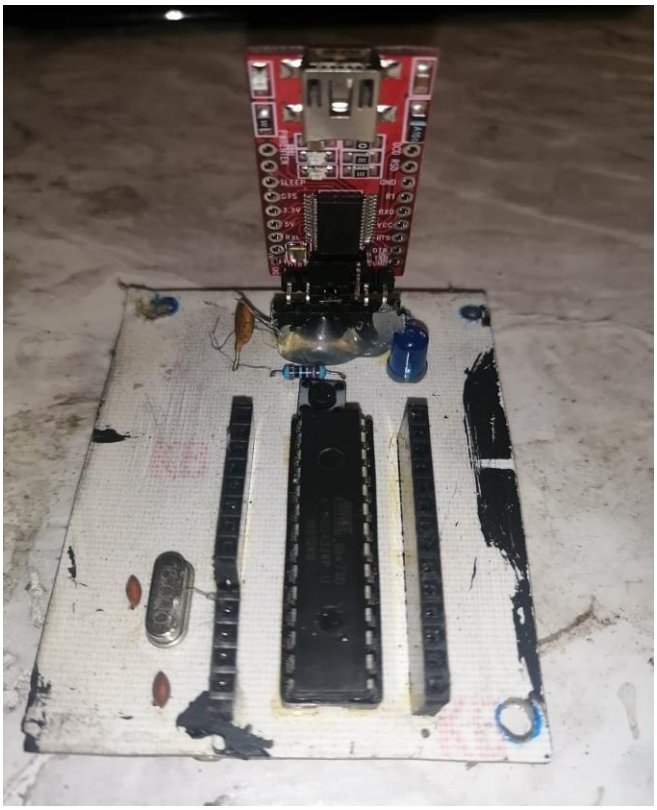


Figura 3.37 Tarjeta de adquisición de datos con microcontrolador atmega328p.

Tabla 3.5 Componentes de la tarjeta de adquisición de datos.

Componente	Cantidad	Costo (Pesos mexicanos)
Resistencias de 1000 Ω y 330 Ω (1/4W), 1 de 100 Ω y 220 Ω (1/4).	1	0.30
Diodo emisor de luz (LED).	1	0.50
Condensador cerámico 0.1uF.	1	1.30
Condensadores cerámicos de 22pF.	2	1.40
Cristal oscilador de 16MHz.	1	8.64
FTDI FT232.	1	38.53
Microcontrolador ATmega328P-PU.	1	46.00
Placa fenólica.	1	20.00
Conexión de cable USB-Mini B.	1	15.00
	Total	133.07

3.5.2 Sensor de turbiedad

Este sensor mide la turbiedad del agua utilizando luz para detectar las partículas suspendidas en el agua esto al medir la transmitancia de la muestra. El módulo que se utilizó fue el modelo SEN0189 del fabricante DFRobot (Sensor Turbiedad DFRobot, 2020). Este módulo cuenta con un sensor que se introduce en la muestra de agua y un circuito de acondicionamiento y amplificación de señal. El circuito del sensor está dentro de una estructura de plástico. En la Figura 3.38 se muestra la estructura y el circuito que se utiliza para detectar el cambio de intensidad de luz.

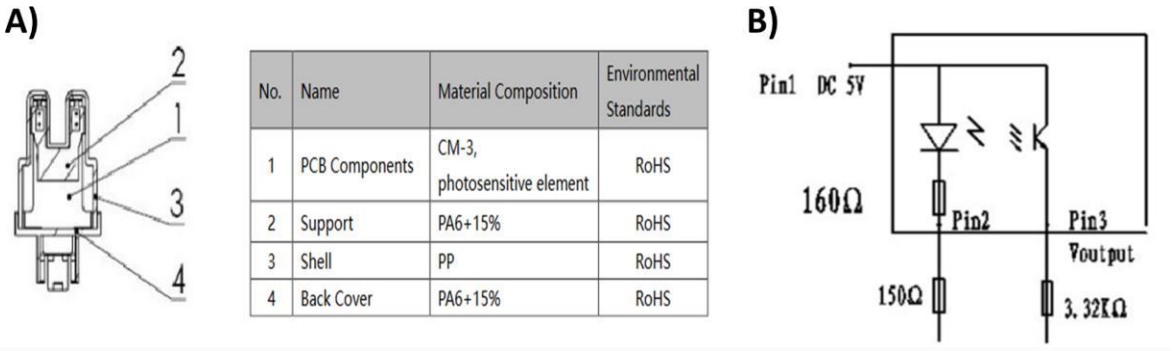


Figura 3.38 A) Estructura del sensor, B) Circuito del sensor (Sensor Turbiedad DFRobot, 2020).

El voltaje de operación del módulo es de +5V DC. Además, tiene un rango de detección de unidades nefelométricas de turbiedad (NTU) aproximado de 0 NTU a 3000 NTU. En la Figura 3.39 se muestra el sensor y el circuito de acondicionamiento de señal.

Figura 3.39 Circuito de acondicionamiento de señal y sensor de turbiedad (Sensor Turbiedad DFRobot, 2020).

El módulo puede entregar la señal de forma analógica o digital. Se eligió adquirir la señal analógica y la conexión con la tarjeta de adquisición de datos es por medio de la entrada analógica del pin 23. La conexión se muestra en la Figura 3.40.

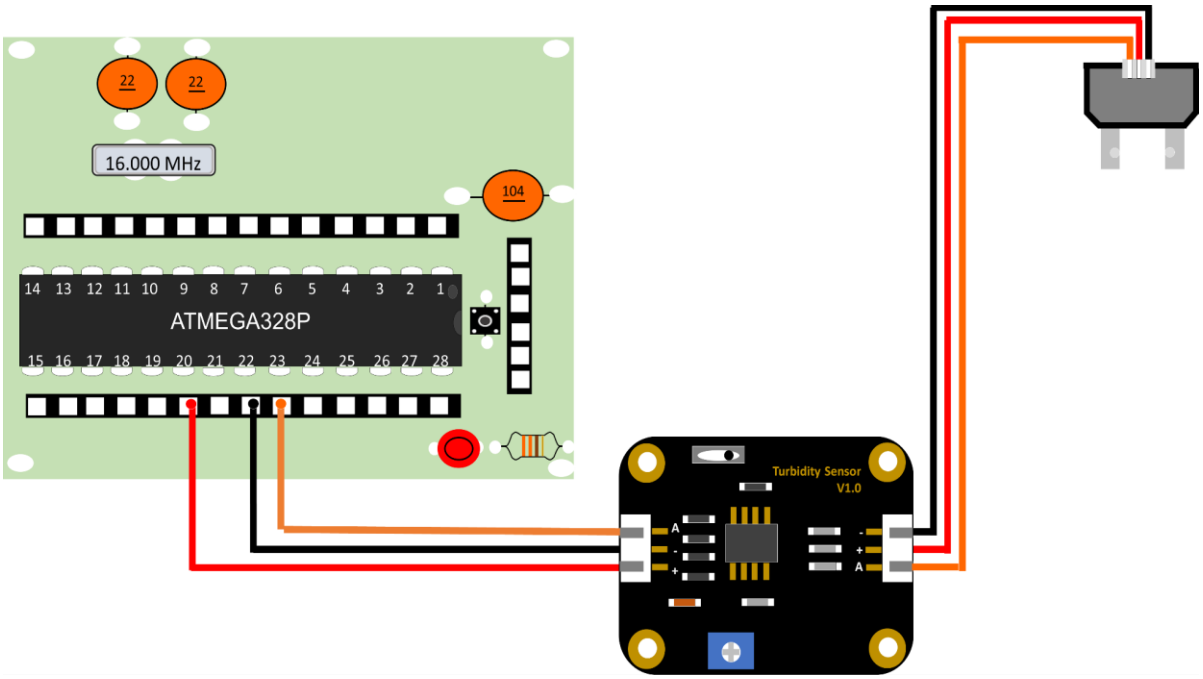


Figura 3.40 Esquema de tarjeta de adquisición de datos y módulo de turbiedad.

3.5.3 Sensor de conductividad eléctrica

Se utilizó el módulo de conductividad eléctrica modelo SKU: DFR0300-H del fabricante DFRobot (Sensor analógico conductividad eléctrica DFRobot, 2020). El módulo se compone de un sensor con dos electrodos en forma de sonda y un circuito de acondicionamiento y amplificación de señal. Se incluye una solución de referencia para calibrar el módulo. El kit completo se muestra en la Figura 3.41.

Figura 3.41 Módulo y sensor de conductividad eléctrica (Sensor analógico conductividad eléctrica DFRobot, 2020).

El módulo puede utilizar un voltaje de alimentación entre 3.3V y 5V, obteniendo mediciones de hasta 100 unidades de conductividad eléctrica en ms/cm. La conexión con la tarjeta de adquisición de datos es por medio de la entrada analógica del pin 24. La conexión se muestra en la Figura 3.42.

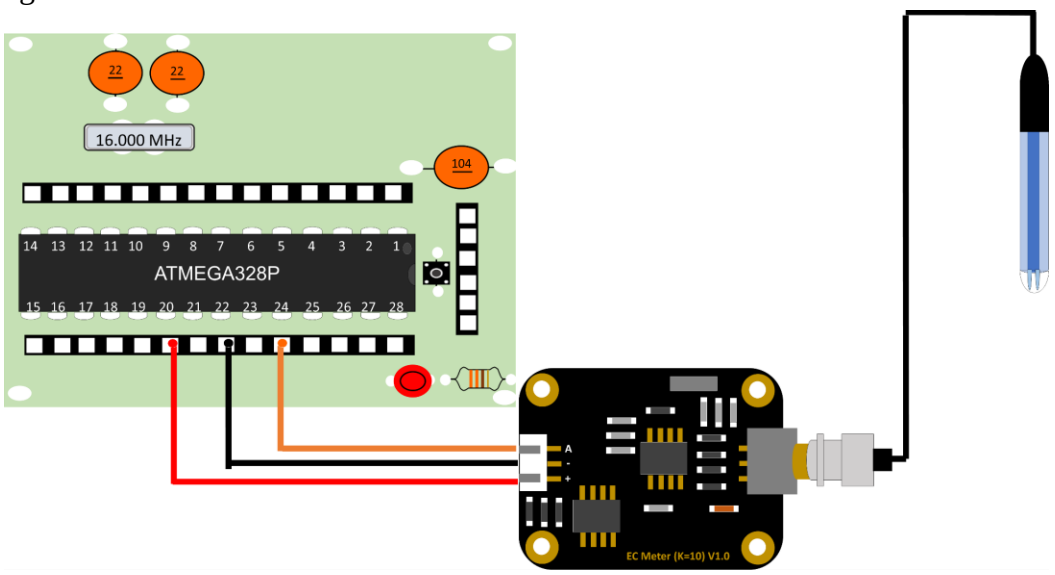


Figura 3.42 Esquema de tarjeta de adquisición de datos y módulo de conductividad eléctrica.

3.5.4 Sensor pH

El sensor de pH que se utilizó fue el modelo pH-4502c (Sensor pH-4502c, 2020) que cuenta con un circuito de acondicionamiento y amplificación de señal y una sonda compuesta de electrodos. Tiene un rango de detección de 0 unidades de pH a 14 unidades de pH con un voltaje de alimentación de +5V CD. En la Figura 3.43 se muestra el circuito de acondicionamiento de señal y el sensor.

Figura 3.43 Módulo de pH (Sensor pH-4502c, 2020).

El módulo entrega la señal de forma digital y analógica, este último se eligió para conectarse al pin número 25 de la tarjeta de adquisición de datos. En la Figura 3.44 se muestra la conexión.

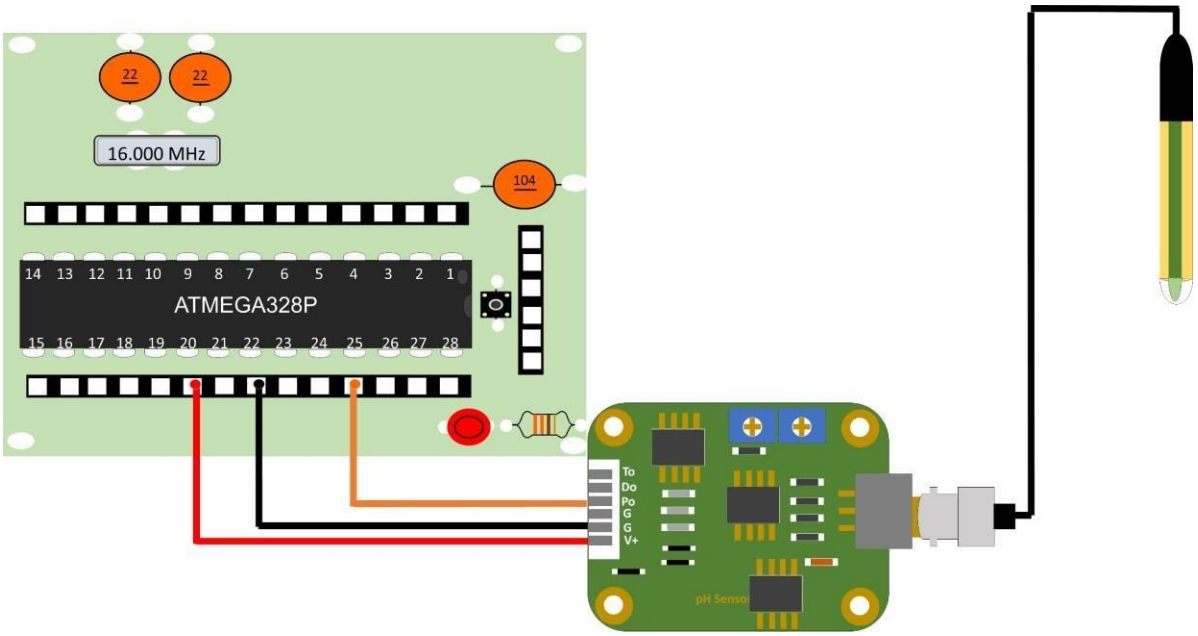


Figura 3.44 Esquema de tarjeta de adquisición de datos y módulo de pH.

3.5.5 Sensor de temperatura del agua

El sensor de temperatura modelo DS18B20 (Sensor DS18B20 Dallas Semiconductor, 2020) se utilizó para la medición de la temperatura del agua. Tiene un rango de medición de temperatura de -55°C a 125°C y un voltaje de alimentación de +3V y +5V. En la Figura 3.45 se muestra el sensor DS18B20. Entrega una señal digital con una resolución configurable de 9, 10, 11 y 12 bits. La conexión con la tarjeta de adquisición de datos fue en el pin 15 como se muestra en la Figura 3.46.

Figura 3.45 Sensor de temperatura del agua (Sensor DS18B20 Dallas Semiconductor, 2020).

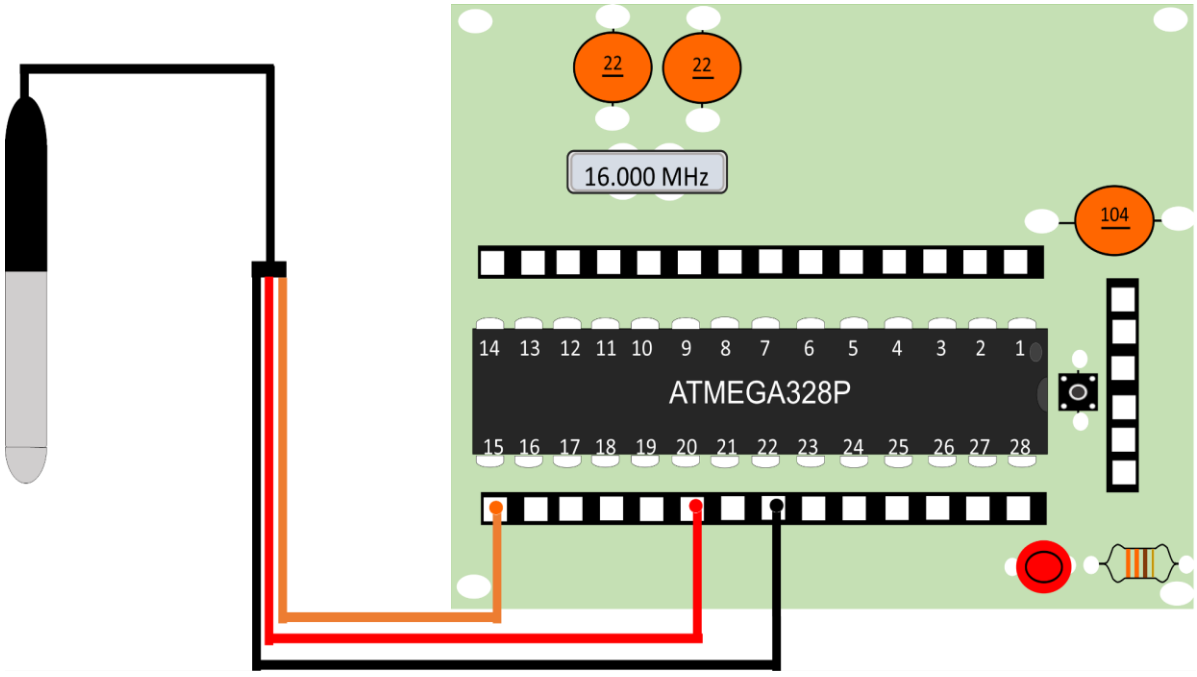


Figura 3.46 Esquema de tarjeta de adquisición de datos y sensor de temperatura del agua.

3.5.6 Sensor de temperatura ambiente

El sensor de temperatura y humedad modelo DHT22 (Sensor DHT22,2020) se utilizó para medir la temperatura ambiente. Es un sensor capacitivo con un voltaje de alimentación de +3V y +5V, y un rango de medición de temperatura de -40°C a 80°C. En la Figura 3.47 se muestra el sensor DHT22. La transferencia de datos con la tarjeta de adquisición de datos es por el protocolo de comunicación de un bus de datos conectado al pin 14. En la Figura 3.48 se muestrala conexión.

Figura 3.47 Sensor de temperatura ambiente (Sensor DHT22,2020).

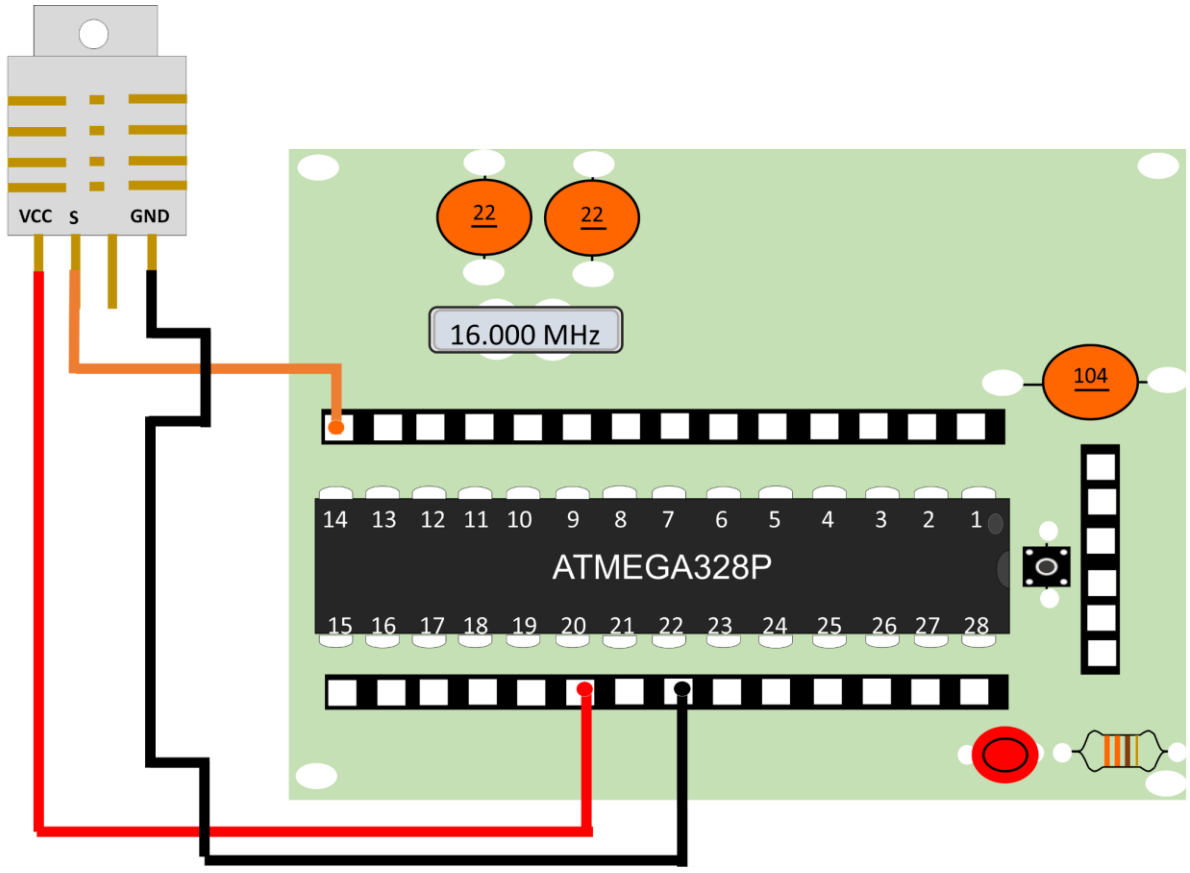


Figura 3.48 Esquema de tarjeta de adquisición de datos y sensor de temperatura ambiente.

3.5.7 Fuente de energía y carga

El dispositivo se compone de un módulo de energía y carga para alimentar los sensores y componentes. Se utilizó el módulo TP4056 (TP4056 cargador lineal, 2020) para cargar dos baterías 18650 de 3.7V y 6000mA conectadas en paralelo. El voltaje de +3.7V que proporcionan las baterías se eleva a +5V utilizando el módulo MT3608 (MT3608 convertidor de voltaje,2020). En la Figura 3.49 se muestra la conexión entre estos componentes.

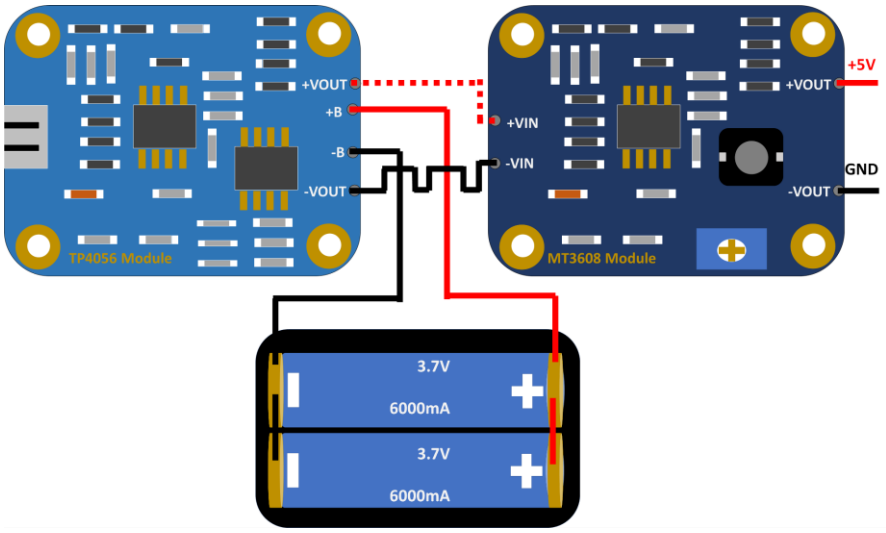


Figura 3.49 Esquema de módulos de carga de baterías y convertidor boost.

3.5.8 Módulo de almacenamiento y tiempo

Para el almacenamiento de las mediciones y registro de tiempo se utilizaron los módulos de tarjeta SD (SD module, 2020) y reloj RTC DS3231 (DS3231,2020). El módulo SD permite guardar información como archivo de texto .txt en una memoria SD. La comunicación con los microcontroladores es por medio de SPI. El reloj RTC registra la fecha y hora de la medición y se comunica con los microcontroladores por medio del bus I2C. Este módulo de almacenamiento y tiempo facilita el análisis y portabilidad de las mediciones. Cada uno de ellos trabaja con un voltaje de alimentación de +5V. En la Figura 3.50 se muestra la conexión entre estos componentes.

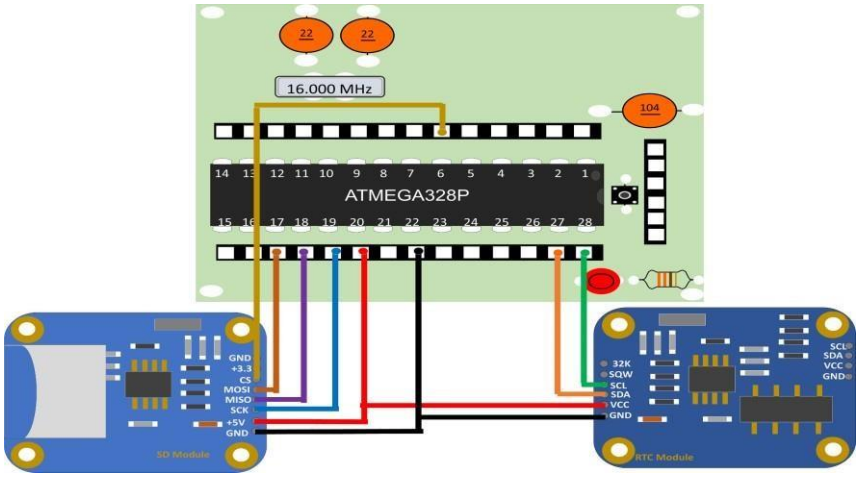


Figura 3.50 Esquema de módulos de almacenamiento en tarjeta SD y reloj RTC.

3.5.9 Diseño de estructura del dispositivo

La Figura 3.51 muestra un esquema de la integración de la tarjeta de adquisición de datos,los sensores, el módulo de energía y carga, y el módulo de almacenamiento y tiempo.

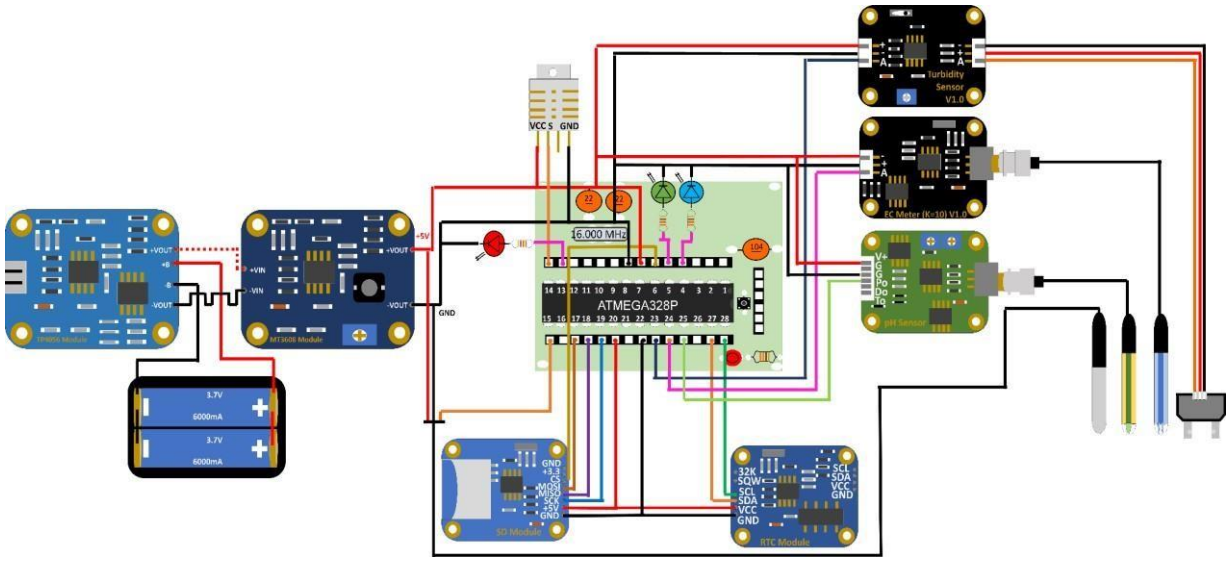


Figura 3.51 Esquema de módulos del dispositivo.

Como primer prototipo la estructura del dispositivo electrónico de medición se propuso que fuera de madera con dimensiones de 10cm de ancho, 10cm de alto y 35cm de largo. Al interior de la estructura se posicionaron todos los componentes. Además, se agregaron 5 interruptores para controlar el encendido del dispositivo y los sensores. En las Figuras 3.52 y 3.53 se muestra el diseño de la estructura y el acomodo de los componentes. En la Figura 3.54 se muestra el prototipo del dispositivo y en la Tabla 3.6 el costo de los componentes.

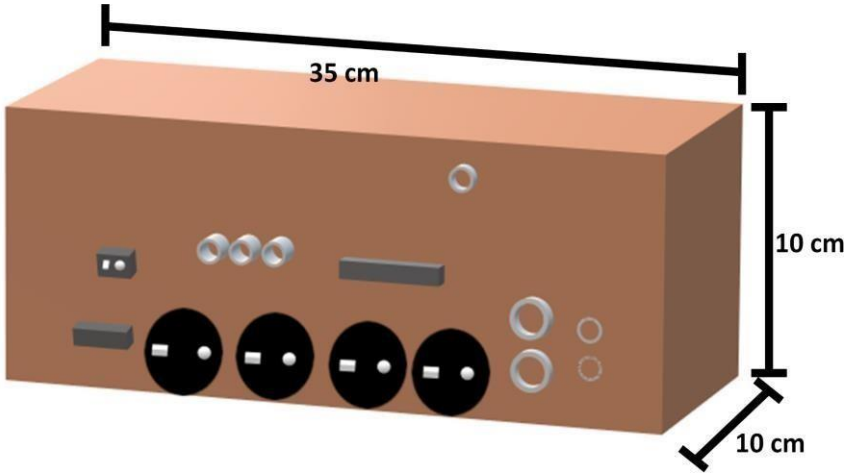


Figura 3.52 Estructura del dispositivo.

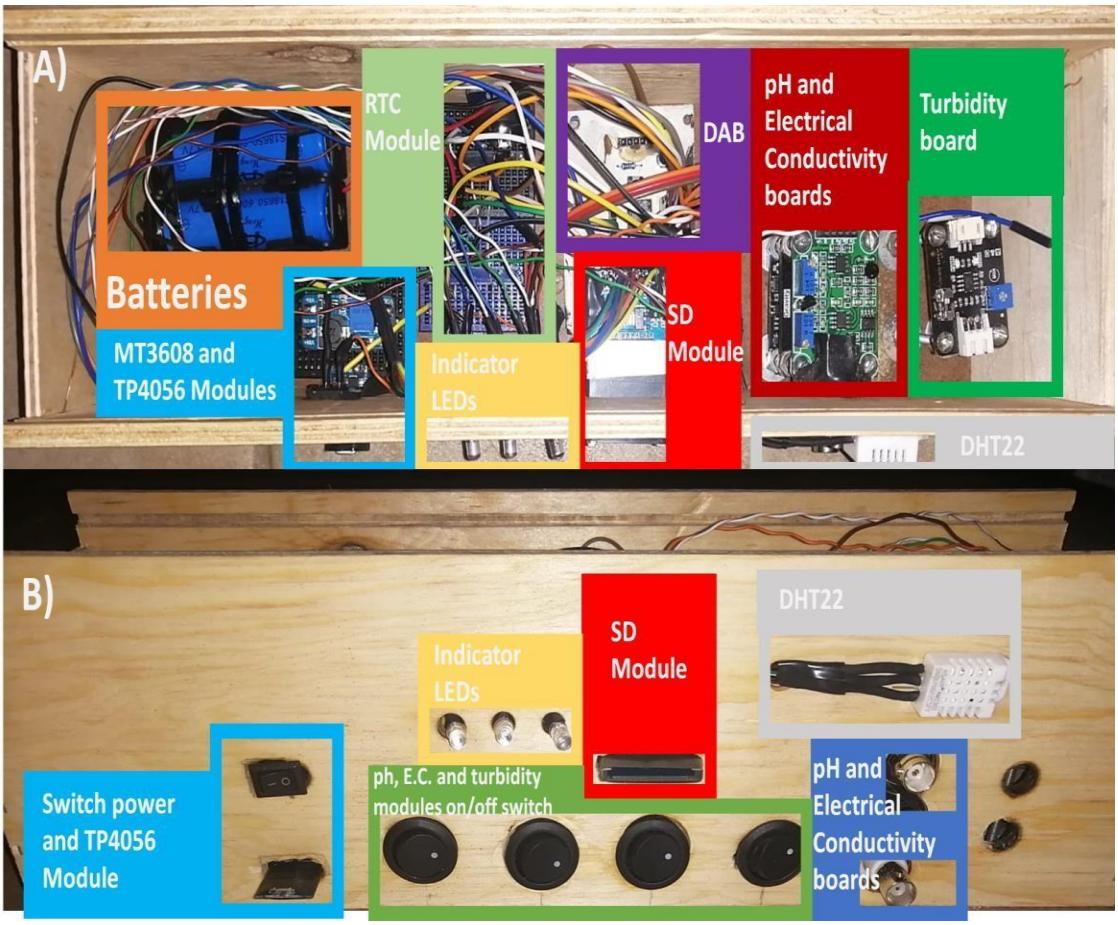


Figura 3.53 Dispositivo armado. A) Vista del interior. B) Vista frontal.

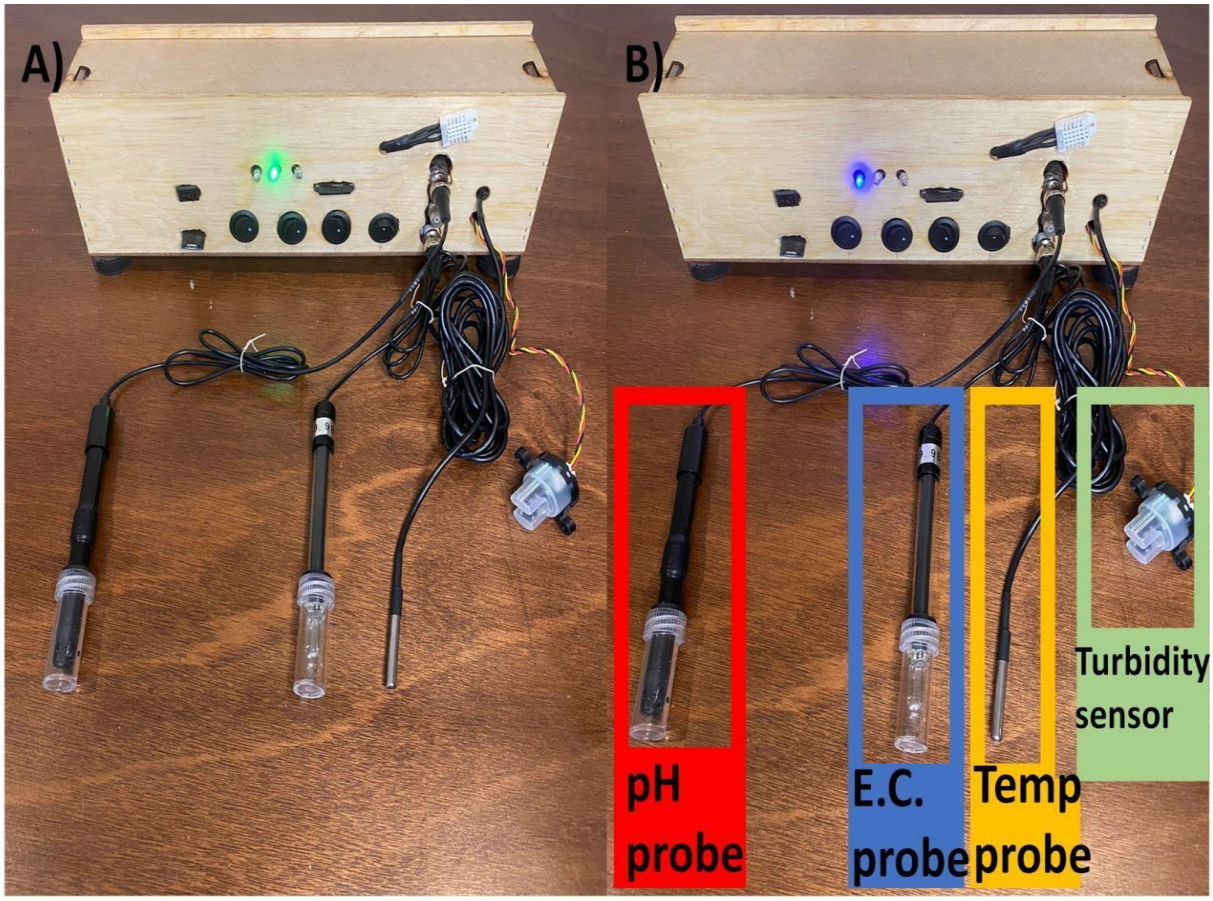


Figura 3.54 Dispositivo armado. A) Función encendido. B) Función guardando mediciones.

Tabla 3.6 Componentes del dispositivo electrónico de medición.

Componente	**Cantidad**	**Costo (Pesos mexicanos)**
Sensor DTH22- humedad y temperatura ambiente	1	218
Sensor DS18B20 -temperatura agua	1	168
Sensor SEN0189- turbiedad	1	429
Módulo DS3231 RTC	1	91
Sensor DFR0300- conductividad eléctrica	1	3880
Módulo memoria SD	1	67
Memoria SD 2GB	1	40

Tarjeta DABOP con atmega328p	1	133.07
Baterías 18650 3.7V 6000mAh	2	150
Porta baterías	2	17
Módulo convertidor DC-DC-MT3608	1	80
Módulo cargador de baterías - TP4056	1	45
Switch	4	52
Led ultrabrillante	3	4.5
Protoboard Mini	3	35
Estructura	1	100
Recipiente de toma de muestras	1	40
Cables de conexión (10 piezas)	4	60
Total		5609.6

3.5.10 Calibración

Después de la construcción del dispositivo electrónico de medición se calibraron los sensores de conductividad eléctrica, pH y turbiedad con el siguiente procedimiento. Se prepararon 7 muestras de café soluble (marca" NESCAFE") con diferente masa. En la Figura 3.55 se observa el procedimiento para medir 0.5004 gr, 0.2006 gr, 0.1007 gr, 0.0804gr, 0.0603 gr, 0.0401 gr y 0.0201 gr del café utilizando una báscula modelo OHRUS. Después se procedió a mezclar cada una con 20 ml de agua destilada previamente filtrada. En la Figura 3.56 se observan las 7 muestras.

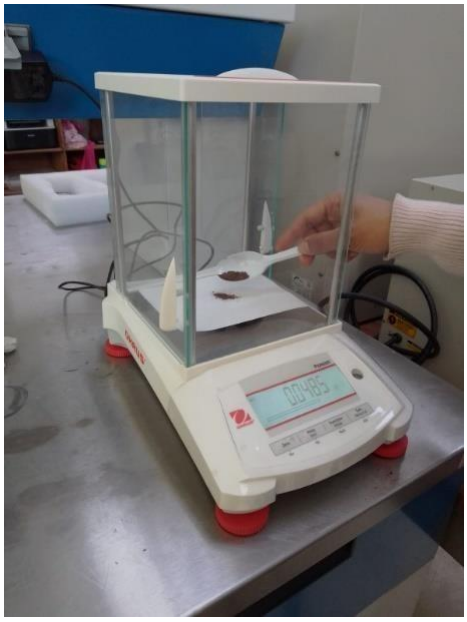


Figura 3.55 Medición de gramos de café.

Figura 3.56 Muestras de café soluble con diferente concentración en 20 ml de agua destilada.

Se utilizaron equipos de laboratorio para caracterizar las muestras y los sensores. La Tabla 3.7 muestra las mediciones de conductividad eléctrica, pH y turbiedad que se obtuvieron con los equipos de laboratorios al aumentar la concentración de café en el agua destilada. La Figura 3.57 muestra el equipo HACH HQ40D para medir pH en unidades de pH (UpH), el equipo HACH H170 para medir conductividad eléctrica en micro Siemens/cm (µS/cm) y el equipo HACH DR900 para medir turbiedad en unidades de atenuación de formacina (FAU).

Tabla 3.7 Mediciones de conductividad eléctrica, pH y turbiedad con equipos de laboratorios.

Café (gr)	Conductividad eléctrica equipo HACH H170 (µS/cm)	pH equipo HACH HQ40D (UpH)	Turbiedad equipo HACH DR900(FAU)
0.50	940	5.15	Fuera de rango del equipo
0.2014	419	5.4	1017
0.1005	221	5.68	408
0.08	169.5	5.83	268
0.0608	144.7	6	239
0.0401	102.6	6.39	155
0.0204	65.4	7.01	80
Blanco (0)	8.61	7.27	0

Figura 3.57 Equipos de laboratorio. A) pH-HACH HQ40D,

B) Turbiedad-HACH DR900, C) C.E.-HACH H170.

La Tabla 3.8 muestra las mediciones de conductividad eléctrica que se obtuvieron con el equipo HACH H170 y el voltaje del sensor DRF0300 al aumentar la concentración de café en el aguadestilada.

Tabla 3.8 Mediciones de conductividad eléctrica con equipo HACH H170 y voltaje del sensor DFR0300.

Café (gr)	Conductividad eléctrica equipo HACH H170 (mS/cm)	Voltaje sensor DFR0300 (mV)
0.50	940	16
0.2014	419	12.3
0.1005	221	11.5
0.08	169.5	11.4
0.0608	144.7	11.3
0.0401	102.6	11.2
0.0204	65.4	11.1
Blanco (0)	8.61	11

La Figura 3.58 muestra el comportamiento de conductividad eléctrica medida con el equipoHACH H170 al aumentar la concentración de café en el agua destilada.

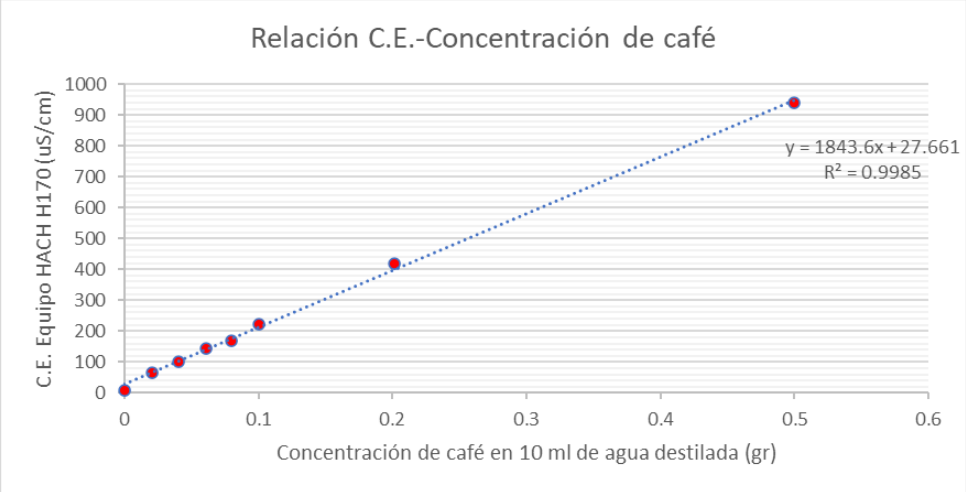


Figura 3.58 Relación conductividad eléctrica- concentración de café.

La Figura 3.59 muestra la relación de conductividad eléctrica medida con el equipo HACHH170 y el voltaje del sensor DFR0300.

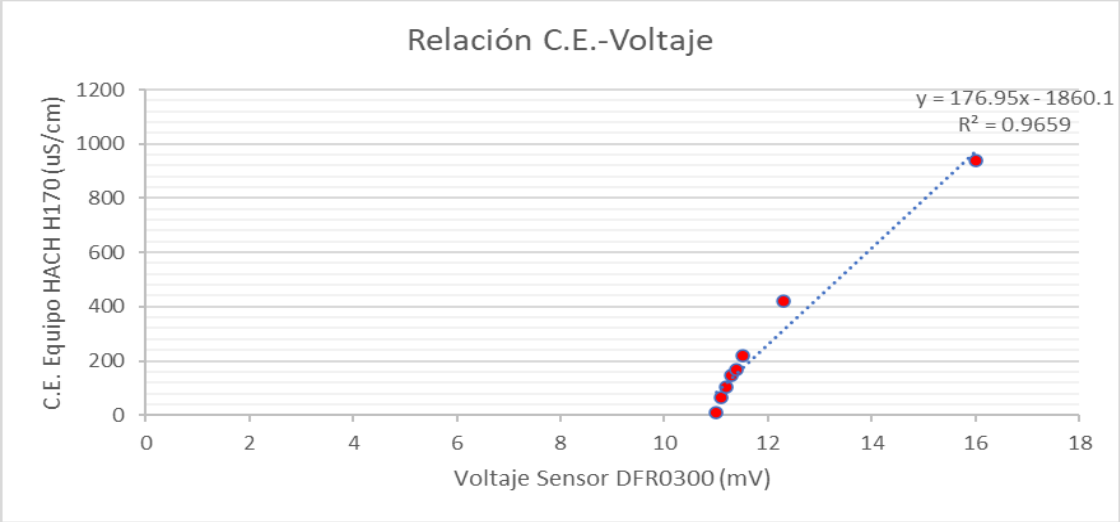


Figura 3.59 Relación conductividad eléctrica – voltaje.

Para obtener la relación conductividad eléctrica-voltaje se aplicó una regresión lineal para facilitar su programación e implementación en el microcontrolador y se obtuvo la Ecuación 3.6 que se muestra a continuación:

$$C.E = 176.95(mV) - 1860.1 \qquad (3.6)$$

Por último, se programó esta Ecuación en la tarjeta de adquisición de datos y se obtuvo la conductividad eléctrica que proporciona el dispositivo electrónico de medición. En la Tabla 3.9 se presenta la comparación de las mediciones de conductividad eléctrica de las muestras realizadas.

Tabla 3.9 Verificación de mediciones conductividad eléctrica con equipo HACH H170 y voltaje del sensor DFR0300.

Café (gr)	Conductividad eléctrica equipo HACH H170 (mS/cm)	Voltaje sensor DFR0300 (V)	Conductividad eléctrica dispositivo electrónico(mS/cm)
0.50	940	16	971.1
0.2014	419	12.3	316.3
0.1005	221	11.5	174.8
0.08	169.5	11.4	157.1
0.0608	144.7	11.3	139.4
0.0401	102.6	11.2	121.7
0.0204	65.4	11.1	104
Blanco (0)	8.61	11	86.3

La Tabla 3.10 muestra las mediciones de turbiedad que se obtuvieron con el equipo HACH DR900 y el voltaje del sensor SEN0189 al aumentar la concentración de café en el agua destilada.

Tabla 3.10 Mediciones de turbiedad con equipo HACH DR900 y voltaje del sensor SEN0189.

Café (gr)	Turbiedad HACH DR900 (FAU)	Voltaje Sensor SEN0189 (V)
0.50	Fuera de rango del equipo	2.17
0.2014	1017	2.69
0.1005	408	2.97
0.08	268	3.04
0.0608	239	3.14
0.0401	155	3.27
0.0204	80	3.38
Blanco (0)	0	3.48

La Figura 3.60 muestra el comportamiento de turbiedad medida con el equipo HACH DR900 al aumentar la concentración de café en el agua destilada.

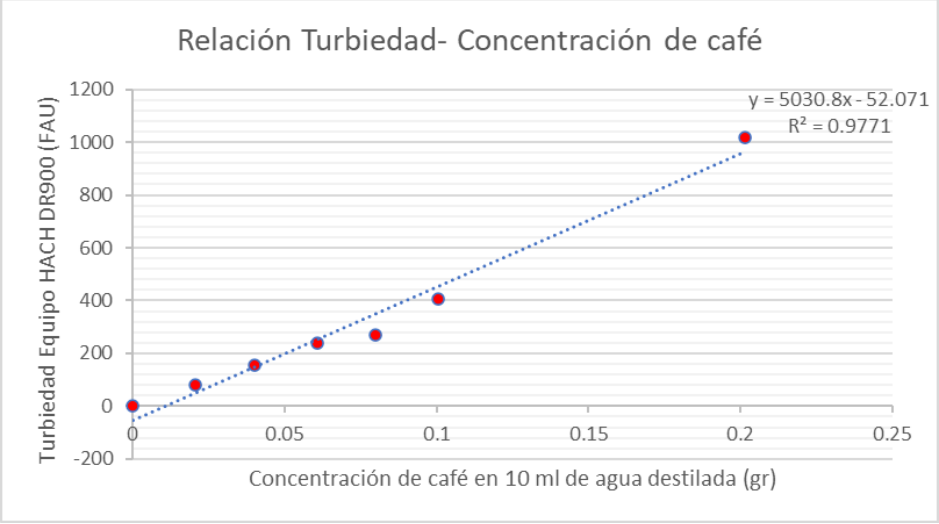


Figura 3.60 Relación Turbiedad- concentración de café.

La Figura 3.61 muestra la relación de turbiedad medida con el equipo HACH DR900 y el voltajedel sensor SEN0189.

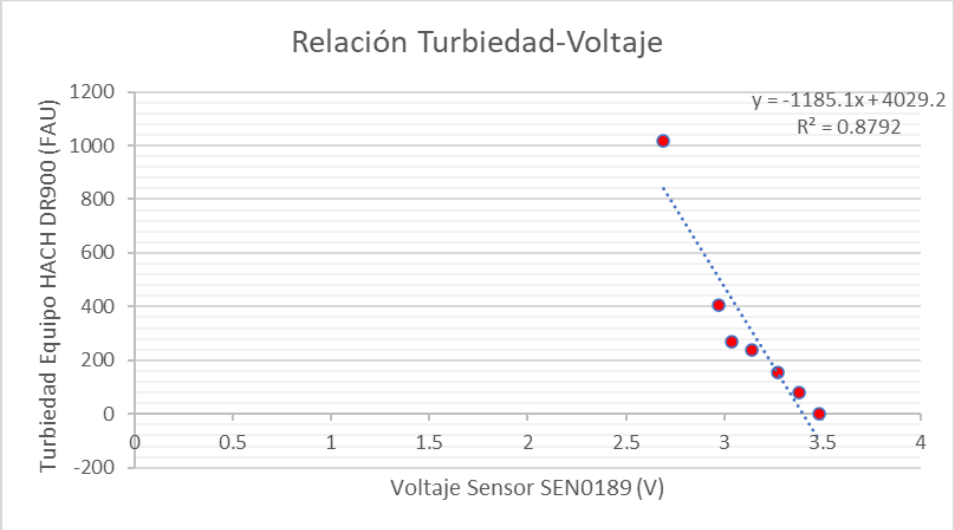


Figura 3.61 Relación Turbiedad – voltaje.

Para obtener la relación turbiedad-voltaje se aplicó una regresión lineal para facilitar su programación e implementación en el microcontrolador y se obtuvo la Ecuación 3.7 que se muestra a continuación:

$$Turbiedad = -1185.1(V) + 4029.2 \quad (3.7)$$

Por último, se programó esta Ecuación en la tarjeta de adquisición de datos y se obtuvo la turbiedad que proporciona el dispositivo electrónico de medición. En la Tabla 3.11 se presentala comparación de las mediciones de turbiedad de las muestras realizadas.

Tabla 3.11 Verificación de mediciones de turbiedad con equipo HACH DR900 y voltaje del sensor SEN0189.

Café (gr)	Turbiedad HACH DR900(FAU)	Voltaje Sensor SEN0189 (V)	Turbiedad dispositivo electrónico (FAU)
0.50	Fuera de rango del equipo	2.17	1457.5
0.2014	1017	2.69	841.2
0.1005	408	2.97	509.4
0.08	268	3.04	426.4
0.0608	239	3.14	307.9
0.0401	155	3.27	153.9
0.0204	80	3.38	23.5
Blanco (0)	0	3.48	-94.9

La Tabla 3.12 muestra las mediciones de pH que se obtuvieron con el equipo HACH HQ40D y el voltaje del sensor pH-4502c al aumentar la concentración de café en el agua destilada.

Tabla 3.12 Mediciones de pH con equipo HACH HQ40D y voltaje del sensor pH-4502c.

Café (gr)	pH HACH HQ40D(UpH)	Voltaje Sensor pH-4502c (V)
0.50	5.15	2.88
0.2014	5.4	2.86
0.1005	5.68	2.85
0.08	5.83	2.84
0.0608	6	2.83
0.0401	6.39	2.82
0.0204	7.01	2.77
Blanco (0)	7.27	2.49

La Figura 3.62 muestra el comportamiento de pH medida con el equipo HACH HQ40D alaumentar la concentración de café en el agua destilada.

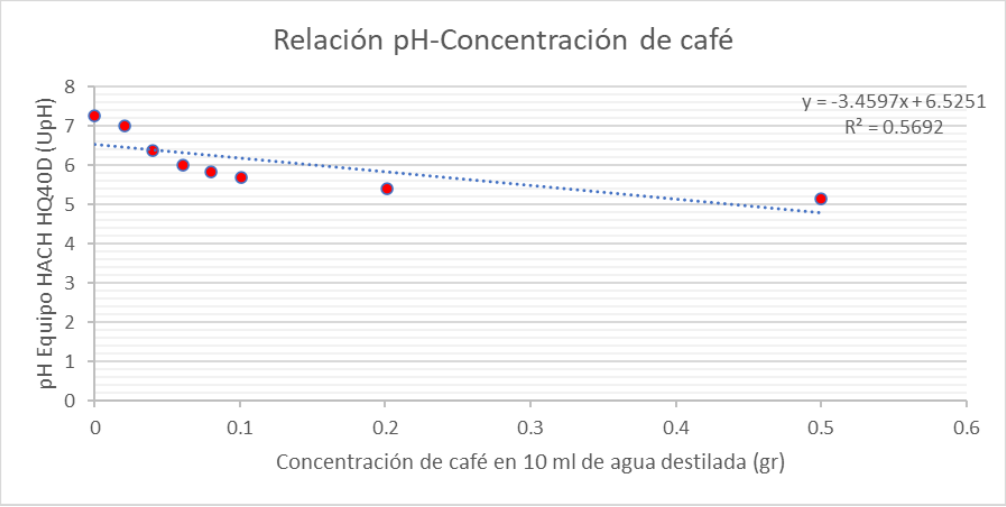


Figura 3.62 Relación pH- concentración de café.

La Figura 3.63 muestra la relación de pH medida con el equipo HACH HQ40D y el voltaje delsensor pH-4502c.

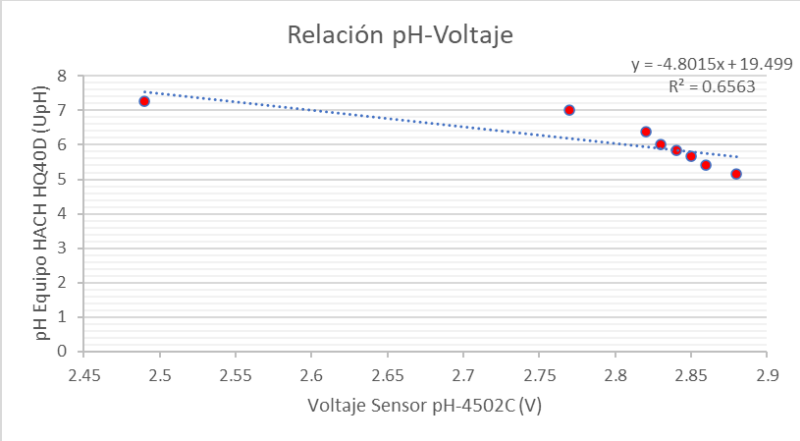


Figura 3.63 Relación pH – voltaje.

Para obtener la relación pH-voltaje se aplicó una regresión lineal para facilitar su programación e implementación en el microcontrolador y se obtuvo la Ecuación 3.8 que se muestra a continuación:

$$pH = -4.8015(V) + 19.499 \quad (3.8)$$

Por último, se programó esta Ecuación en la tarjeta de adquisición de datos y se obtuvo el pH que proporciona el dispositivo electrónico de medición. En la Tabla 3.13 se presenta la comparación de las mediciones de pH de las muestras realizadas.

Tabla 3.13 Verificación de mediciones de pH con equipo HACH HQ40D y voltaje del sensor pH-4502c.

Café (gr)	pH HACH HQ40D (UpH)	Voltaje Sensor pH-4502c (V)	pH dispositivo electrónico (UpH)
0.50	5.15	2.88	5.67
0.2014	5.4	2.86	5.76
0.1005	5.68	2.85	5.81
0.08	5.83	2.84	5.86
0.0608	6	2.83	5.91
0.0401	6.39	2.82	5.95
0.0204	7.01	2.77	6.19
Blanco (0)	7.27	2.49	7.54

Capítulo 4. Resultados y Limitaciones

4.1 Preprocesamiento de la información para la predicción de la demanda bioquímica de oxígeno a 5 días

4.1.1. Detección de valores atípicos

Cada parámetro del agua se analizó por separado del año 2012 al 2019 y se observó el comportamiento en todas las estaciones de monitoreo. Las figuras que se muestran a continuación contienen las mediciones de los parámetros después de eliminar valores por error de medición o por no incorporar el valor en alguna casilla de la base de datos. La Figura 4.1 muestra el diagrama de caja de la demanda bioquímica de oxígeno a 5 días con valores máximos fuera de la caja de 125 mg/l.

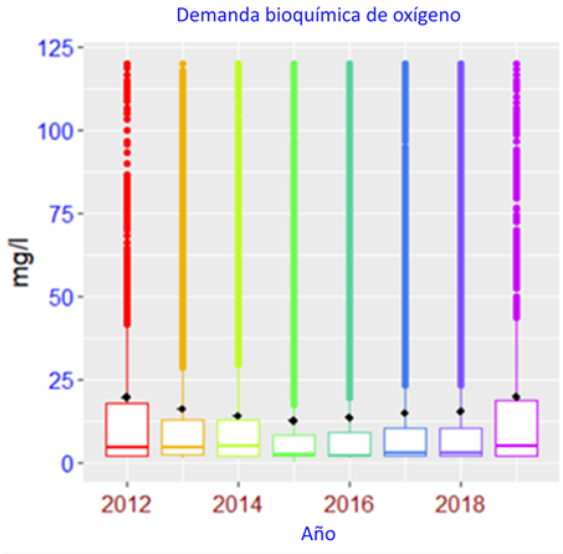


Figura 4.1 Diagrama de caja del parámetro demanda bioquímica de oxígeno a 5 días del año 2012 al 2019.

Del año 2012 al 2015 la demanda bioquímica de oxígeno a 5 días presentó una disminución en el número de estaciones con valores menores a 25 mg/l. Para los años 2016 al 2019 aumento hasta llegar a el número de estaciones similares que se tenían con valores mínimos de 25 mg/l en el año 2012. Estas variaciones están dentro de la categoría de cumplimiento, sin embargo, en el periodo de tiempo de 2012 al 2019 se han presentado valores mayores a 30 mg/l en las estaciones que se catalogaron como contaminadas.

Solo para la demanda bioquímica de oxígeno a 5 días se observó el número de muestras registradas del año 2012 al 2019 por su nivel de concentración en mg/l y su categorización con alrededor de 28111 de muestras en excelente (DBO<3), 9428 en buena calidad (3<DBO<=6), 14378 en aceptable (6<DBO<=30), 4803 en contaminada (30<DBO<=120) y 2409 en fuerte contaminada (DBO>120). En la Figura 4.2 se observan los niveles de contaminación de la demanda bioquímica de oxígeno.

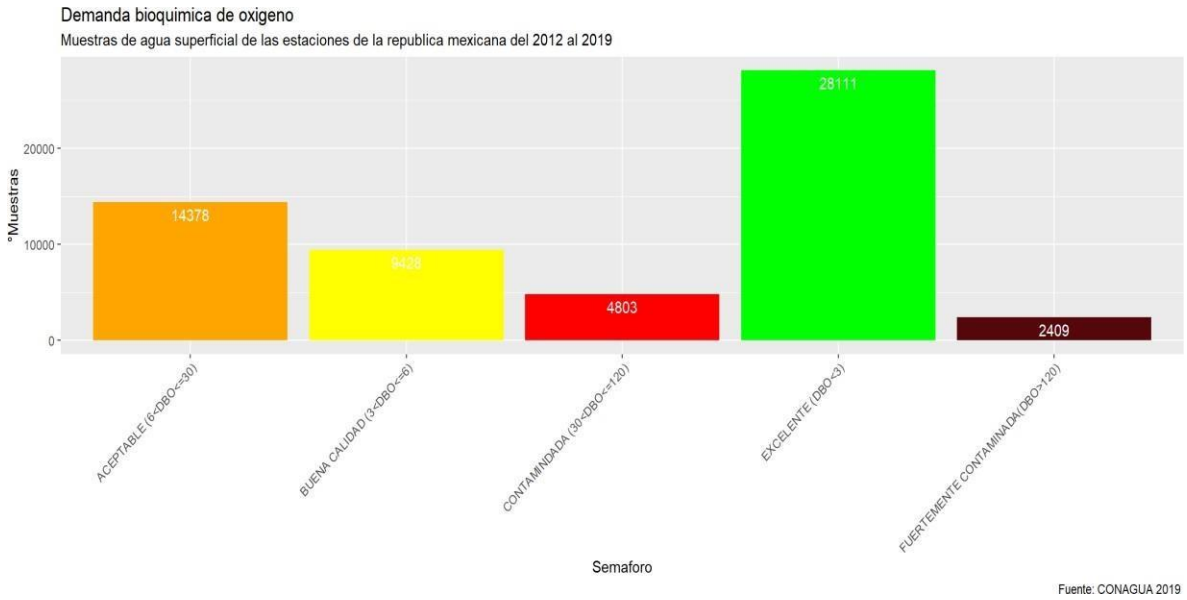


Figura 4.2 Nivel de DBO5 en Estados de la república mexicana del 2012 al 2019**.**

La Figura 4.3 muestra el diagrama de caja de la demanda química de oxígeno, fósforo total, nitrógeno Kjeldahl y nitrógeno amoniacal con valores máximos fuera de la caja de 250 mg/L, 20 mg/L, 400mg/L y 200mg/L respectivamente.

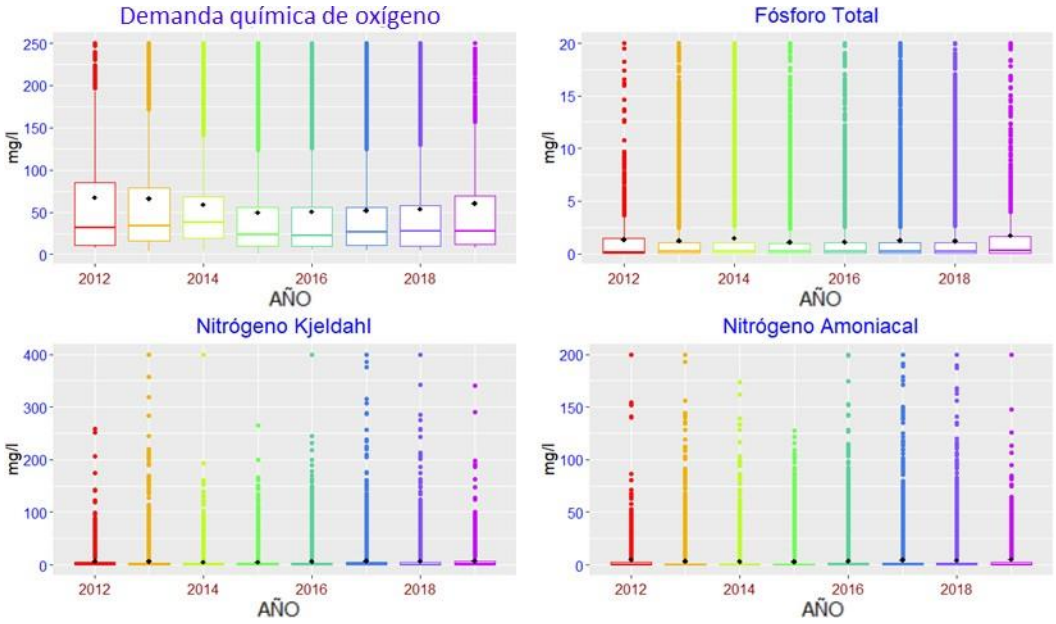


Figura 4.3 Diagrama de caja de los parámetros demanda química de oxígeno, fósforo total, nitrógeno Kjeldahl y nitrógeno amoniacal del año 2012 al 2019.

La demanda química de oxígeno presentó una media de 70 mg/l a 50mg/l, con valores mínimos en los años 2015 y 2016. Sin embargo, se observó que más de la mitad de las estaciones de monitoreo tuvieron valores mayores a 40 mg/l que los catalogan como contaminadas. El fósforo

total, nitrógeno amoniacal y nitrógeno Kjeldahl presentaron valores medios de 1.3mg/l, 3.7 mg/l y 6.34 mg/l, catalogándose como no contaminadas a partir de los límites máximos permisibles de 20 mg/l y 40 mg/l respectivamente.

La Figura 4.4 muestra el diagrama de caja de color verdadero, absorción UV, sólidos disueltos totales y conductividad eléctrica con valores máximos fuera de la caja de 200 Pt/Co, 2 Abs/cm,1000 mg/L y 5000 uS/cm respectivamente.

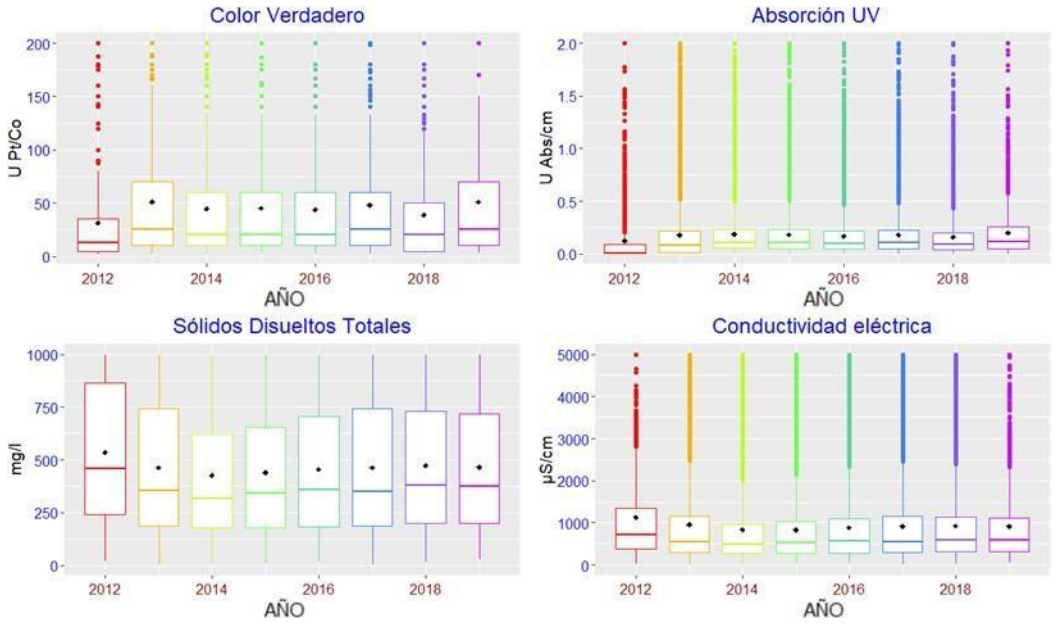


Figura 4.4 Diagrama de caja de los parámetros color verdadero, absorción UV, sólidos disueltos totales y conductividad eléctrica del año 2012 al 2019.

Los parámetros solidos disueltos totales y absorción UV presentaron una media de 354.5 mg/l, y 0.17 Abs/cm respectivamente. Para estos parámetros la NOM-001-SEMARNAT-1996 no establece límites máximos permisibles. Sin embargo, para conductividad eléctrica y color verdadero los límites máximos son 200 uS/cm y 15 Pt/Co. Estos parámetros presentaron una media de 1056 uS/cm y 55.2 Pt/Co respectivamente.

La Figura 4.5 muestra el diagrama de caja de pH, oxígeno disuelto, sólidos suspendidos totales y turbiedad con valores máximos fuera de la caja de 11.8 UpH, 40 mg/L, 400 mg/L y 500 NTUrespectivamente.

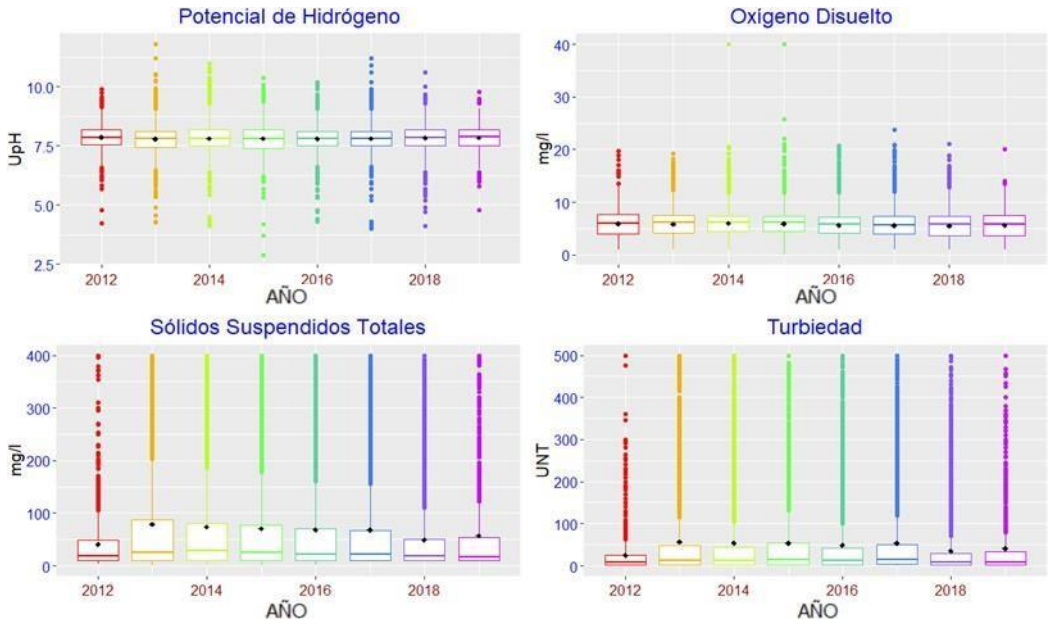


Figura 4.5 Diagrama de caja de los parámetros pH, oxígeno disuelto, sólidos suspendidos totales y turbiedad del año 2012 al 2019.

Los parámetros pH, oxígeno disuelto, sólidos suspendidos totales y turbiedad presentaron una media de 7.8 UpH, 5.7 mg/l, 105 mg/l y 75 NTU respectivamente. Los límites máximos permisibles catalogados como no contaminados son 8.5UpH, 150 mg/l para sólidos suspendidos totales y 3 NTU para turbiedad. Para el parámetro de oxígeno disuelto el límite máximo permisible está representado por el porcentaje de saturación de oxígeno disuelto mayor a 30% y menor a 50% (30%<OD<=50% y), y mayor a 120% y menor a 130% (120%<OD<=130%).

La Figura 4.6 muestra el diagrama de caja de temperatura ambiente y temperatura del agua con valores máximos fuera de la caja de 51 °C y 62 °C respectivamente. La temperatura ambiente y la temperatura del agua presentaron una media de 27.6 °C y 24.9 °C respectivamente, dondeel límite máximo permisible para la temperatura del agua es de 40 °C.

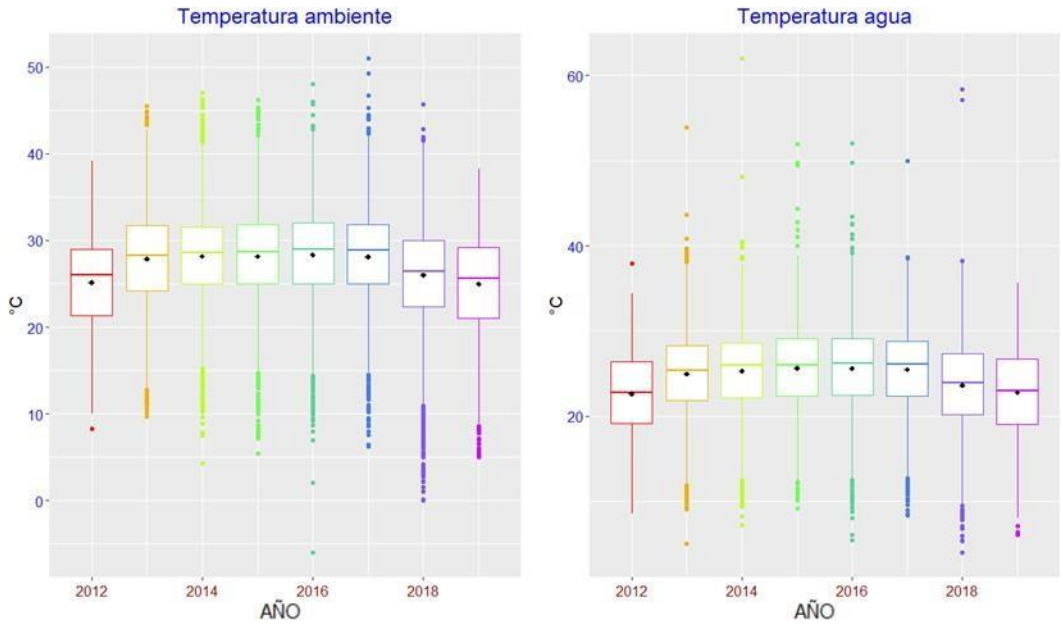


Figura 4.6 Diagrama de caja de los parámetros temperatura ambiente y temperatura agua del año 2012 al 2019.

A partir de la NOM-001-SEMARNAT-1996 los parámetros de temperatura agua, pH, porcentaje de saturación de oxígeno disuelto, turbiedad, sólidos suspendidos totales, conductividad eléctrica, color verdadero, demanda química de oxígeno, fósforo total, nitrógeno Kjeldahl y demanda bioquímica de oxígeno a 5 días presentaron mediciones fuera de los límites máximos permisibles al analizarlos por separado del año 2012 al 2019, sin embargo para identificar las zonas con estas características es necesario realizar un análisis delimitado por estado, región hidrológica y parámetro. Dentro de los parámetros que se destaca el número de estaciones con valores mayores a los límites máximos permisibles son la demanda química de oxígeno, conductividad eléctrica y color verdadero, por lo que es importante considerarlos para la determinación de la calidad del agua.

Estos diagramas de caja permitieron identificar de forma visual el valor máximo y medio de cada parámetro del agua en las estaciones de monitoreo del año 2012 al 2019. Los valores que se muestran fuera de la caja no se eliminaron ya que fueron mediciones con valores que comúnmente se pueden presentar en muestras de agua.

4.2 Análisis de datos para la predicción de la demanda bioquímica de oxígeno a 5 días

En esta sección se presentan los coeficientes de correlación de Pearson obtenidos entre los parámetros de la base de datos con correlación baja, media y alta. Después, los coeficientes de determinación al implementar Forward Selection para la predicción de la demanda bioquímica de oxígeno a 5 días en función de varias combinaciones de parámetros del agua.

Los coeficientes de determinación utilizando como entrada del algoritmo los parámetros de forma independiente mostraron resultados mínimos de 0.001 y máximos de 0.66 para varios parámetros que a continuación se presentan.
Al analizar lo anterior se seleccionaron las características para formar los 3 grupos de parámetros establecidos por este trabajo para poder elegir el grupo de parámetros que faciliten la determinación de la demanda bioquímica de oxígeno a 5 días a partir de los equipos de medición, las condiciones de la zona de estudio y laboratorio.
Por otro lado, las curvas de aprendizaje mostraron el número de ejemplos necesarios para el entrenamiento y prueba de los algoritmos de aprendizaje máquina alrededor de 40000 muestras.
Por último, se obtuvo el desempeño de implementar los algoritmos de aprendizaje máquina utilizando como entrada los 3 grupos de parámetros para la predicción de la demanda bioquímica de oxígeno.

4.2.1. Coeficiente de correlación de Pearson

La Figura 4.7 muestra el mapa de calor representando el coeficiente de correlación dePearson entre los parámetros previamente listados en la Tabla 3.1.

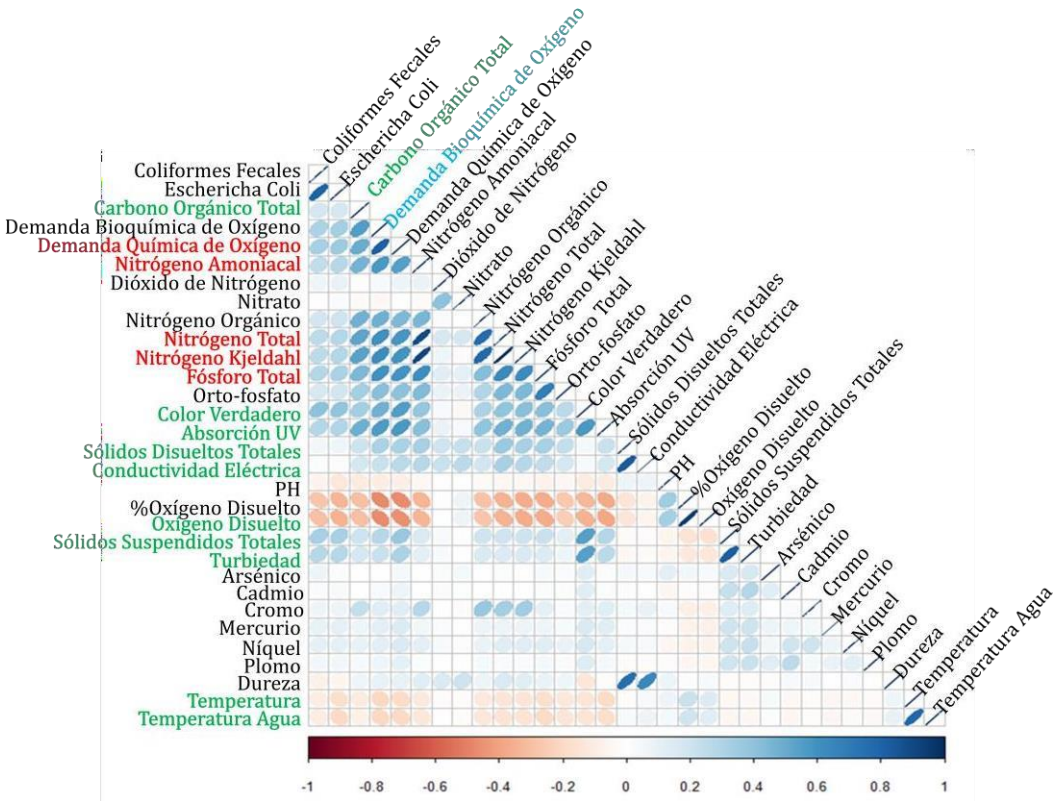


Figura 4.7 Mapa de calor representando la matriz de coeficiente de correlación.

El mapa de calor indica lo siguiente: Los parámetros que mostraron correlación positiva superior a 0.5 con la demanda bioquímica a 5 días fueron carbono orgánico total, demanda química de oxígeno, nitrógeno amoniacal, nitrógeno, nitrógeno Kjeldahl, fósforo total y absorción UV. Los parámetros que mostraron correlación positiva entre 0.2 y 0.5 con la demanda bioquímica a 5 días fueron coliformes fecales, Escherichia coli, nitrógeno orgánico, orto-fosfato, color verdadero, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales y turbiedad. Los parámetros que mostraron correlación negativa entre -0.2 y -0.5 con la demanda bioquímica a 5 días fueron Porcentaje de Saturación de Oxígeno disuelto, Oxígeno disuelto, temperatura y temperatura del agua.
La demanda química de oxígeno mostró una alta correlación ($|r| > 0.7$). Este parámetro implica la Demanda Bioquímica de Oxígeno, midiendo la oxidación completa de la muestra, tanto orgánica, biodegradable como no biodegradable (r=0.81). Carbono orgánico total, nitrógeno amoniacal, nitrógeno, nitrógeno Kjeldahl, fósforo, absorción UV, coliforme fecal, Escherichia coli, nitrógeno orgánico, ortofosfato, color verdadero, sólidos disueltos totales y oxígeno disuelto, mostraron una correlación moderada ($0.3 < |r| < 0.7$). Esto puede estar relacionado con el método y la técnica de determinación de parámetros. Dióxido de nitrógeno, nitrato de nitrógeno, conductividad eléctrica, PH, sólidos suspendidos totales, turbiedad, arsénico, cadmio, cromo, mercurio, níquel, plomo, dureza, temperatura y temperatura del agua mostraron una correlación débil ($0 < |r| < 0.3$). La conductividad eléctrica proporciona información general sobre la concentración de sales de iones, por lo que muestra una correlación débil con la demanda bioquímica de oxígeno a 5 días ($r = 0.21$) y una alta correlación con los sólidos disueltos totales ($r = 0.83$).

4.2.2. Forward Selection

En la Tabla 4.1 se presenta el coeficiente de determinación al utilizar cada parámetro individual como entrada del algoritmo de regresión lineal para la predicción de la demanda bioquímica de oxígeno a 5 días. Se muestra el comportamiento para la etapa de entrenamiento y de prueba al aplicar Forward Selection.

Tabla 4.1 Identificación de parámetros individuales con mayor coeficiente de determinación al aplicar Forward Selection.

	Entrenamiento	**Prueba**
Parámetro	Coeficiente de Determinación	Coeficiente de Determinación
Coliformes Fecales	0.09	0.08
Escherichia Coli	0.11	0.09

Demanda Bioquímica de Oxígeno a 5 días	1	1
Demanda Química de oxigeno	0.66	0.66
Sólidos Suspendidos Totales	0.07	0.06
Solidos Disueltos Totales	0.1	0.1
Fósforo Total	0.36	0.32
Color Verdadero	0.22	0.21
Absorción UV	0.31	0.3
Conductividad Eléctrica	0.04	0.04
pH	0.008	0.008
Porcentaje de saturación de Oxígeno Disuelto	0.22	0.23
Oxígeno Disuelto	0.21	0.22
Turbiedad	0.04	0.04
Arsénico	0.00001	0.00003
Cadmio	0.001	0.002
Cromo	0.016	0.019
Mercurio	0.01	0.009
Níquel	0.013	0.025
Plomo	0.004	0.001
Dureza	0.01	0.01
Temperatura	0.04	0.04
Temperatura Agua	0.04	0.04
Carbono Orgánico Total	0.3	0.2
Nitrógeno Amoniacal	0.3	0.29
Dióxido de Nitrógeno	0.001	0.003
Nitrato	0.001	0.001
Nitrógeno Orgánico	0.24	0.19
Nitrógeno	0.37	0.33

Nitrógeno Kjeldahl	0.39	0.34
Orto-fosfato	0.17	0.17

Los parámetros que mostraron un coeficiente de determinación mayor a 0.3 al aplicar Forward Selection fueron: carbono orgánico total, demanda química de oxígeno, nitrógeno amoniacal, nitrógeno, Kjeldahl nitrógeno, fósforo total y absorción UV. Los parámetros más significativos que mostraron un coeficiente de determinación menor a 0.3 al aplicar Forward Selection fueron: Escherichia coli, nitrógeno orgánico, orto-fosfato, color verdadero, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, temperatura, pH y temperatura del agua.

En la Tabla 4.2 se muestra el aumento del coeficiente de determinación al agrupar los parámetros en conjuntos. A partir de agrupar los 4 parámetros más eficientes se obtuvo 0.70 de coeficiente de determinación. Por lo que al agregar más parámetros no mostró aumentos considerables y se probaron otros conjuntos de parámetros.

Tabla 4.2 Comportamiento de agrupar parámetros con coeficiente de determinación mayor a 0.3 y de agrupar los parámetros más significativos con coeficiente de determinación menor a 0.3.

	Entrenamiento	**Prueba**
Parámetro	Coeficiente de Determinación	Coeficiente de Determinación
Demanda Química de Oxígeno	0.66	0.66
Demanda Química de Oxígeno, Nitrógeno Amoniacal	0.69	0.69
Demanda Química de Oxígeno, Nitrógeno Amoniacal, Nitrógeno Kjeldahl	0.70	0.70
(A) Demanda Química de Oxígeno, Nitrógeno Amoniacal, Nitrógeno Kjeldahl, Fósforo Total	0.70	0.70
(B) Color Verdadero, Absorción UV, Sólidos Disueltos Totales, Conductividad Eléctrica, Sólidos Suspendidos Totales, Turbiedad, Oxígeno Disuelto, Temperatura, Temperatura Agua	0.42	0.41
(C) Conductividad Eléctrica, Turbiedad, Temperatura, Temperatura Agua, pH.	0.29	0.28

4.2.3. Selección de características

A partir del análisis de correlación de Pearson y de Forward Selection (FS) se realizó la selección de parámetros para formar los grupos. Para (A) grupo de parámetros que se

determinan más rápido que la demanda bioquímica de oxígeno a 5 días en un laboratorio fueron: demanda química de oxígeno, nitrógeno amoniacal, nitrógeno Kjeldahl y fósforo. Estos parámetros fueron los que mostraron correlación mayor a 0.5 con el parámetro de la demanda bioquímica de oxígeno a 5 días, 0.81, 0.59, 0.62 y 0.60 respectivamente. El coeficiente de determinación que se obtuvo de Forward Selection para estos parámetros de forma individual fue 0.66, 0.328, 0.39 y 0.36 respectivamente.

De igual manera a partir del análisis de correlación de Pearson y de Forward Selection se realizó la selección de parámetros para el grupo B. (B) Grupo de parámetros que se puedan determinar en la zona de estudio fueron: Color verdadero, absorción UV, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, temperatura y temperatura del agua. Estos parámetros mostraron correlación positiva y negativa entre 0.2 y 0.5 con la demanda bioquímica a 5 días. El coeficiente de determinación que se obtuvo de Forward Selection para estos parámetros de forma individual fue 0.22, 0.31, 0.1, 0.04, 0.07, 0.04, 0.21, 0.04 y 0.04 respectivamente.

Para el grupo (C) de parámetros que se puedan medir en la zona de estudio por medio de tecnología de sensores se seleccionaron los siguientes: conductividad eléctrica, turbiedad, pH, temperatura y temperatura del agua. La correlación de estos parámetros con la demanda bioquímica de oxígeno a 5 días fue positiva y negativa entre 0.2 y 0.5. El coeficiente de determinación que obtuvo de Forward Selection de forma individual fue, 0.04, 0.04, 0.008, 0.04 y 0.04 respectivamente. En la Figura 4.8 se muestran los grupos de parámetros.

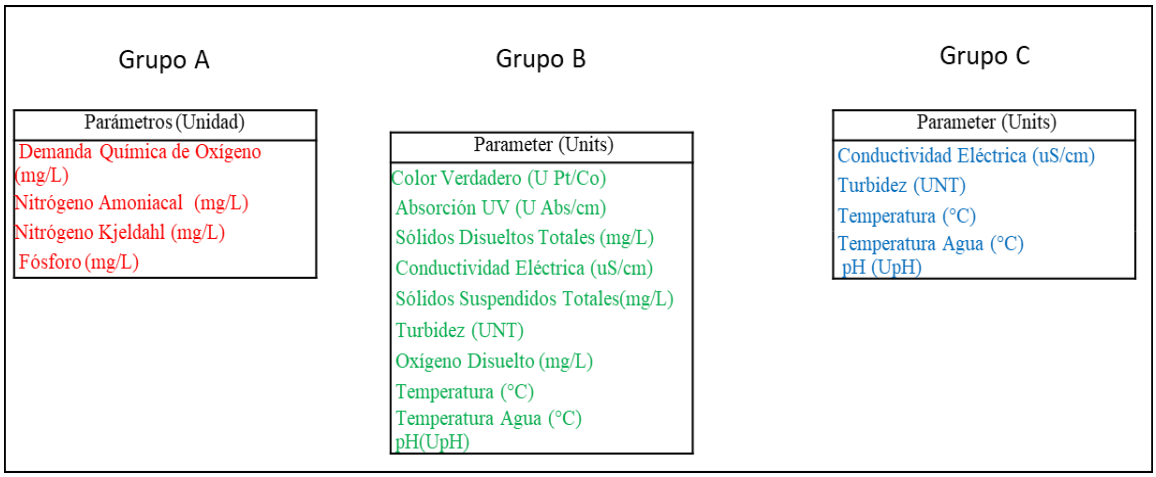


Figura 4.8 Grupos de parámetros A, B y C.

A partir de la metodología empleada en este trabajo, el FS confirmó el comportamiento y relación del análisis de correlación, al mostrar los mismos parámetros que cumplieron con los arreglos establecidos. De igual manera, FS permitió conocer de forma previa a aplicar los algoritmos de aprendizaje máquina, el rendimiento de utilizar un algoritmo de regresión lineal múltiple con todos los parámetros disponibles en la base de datos procesada y tener una referencia del rendimiento máximo de ese algoritmo.

4.2.4. Curvas de aprendizaje

Las Figuras 4.9, 4.10 y 4.11 muestran las curvas de aprendizaje al aumentar el número de ejemplos en el entrenamiento. Se observó que el número de ejemplos que se necesita para no sobre ajustar ni sub ajustar los algoritmos fue de alrededor de 40000 ejemplos para cada grupo de parámetros. El coeficiente de determinación (R^2) en la etapa de prueba de los algoritmos para el grupo A, B y C fue de 0.70, 0.41 y 0.29 respectivamente. La Figura 4.9 muestra el rendimiento del algoritmo de regresión lineal múltiple al aumentar el número de ejemplos, utilizando como entrada el grupo A de parámetros.

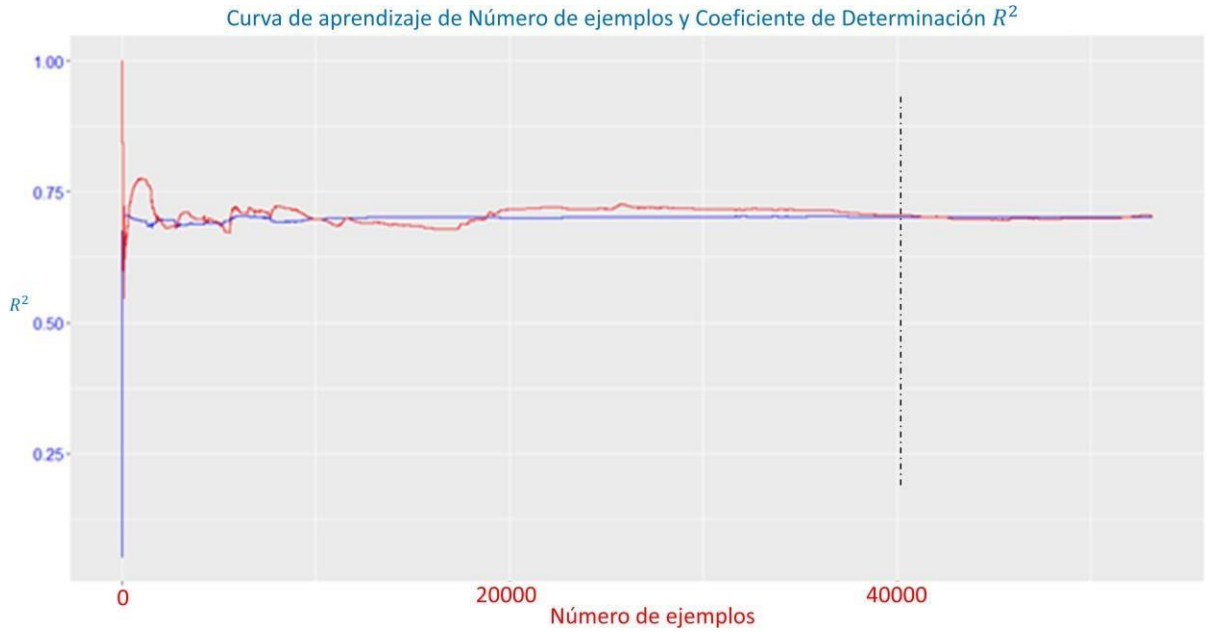


Figura 4.9 Curva de aprendizaje para grupo A (Línea roja entrenamiento y línea azul prueba).

La Figura 4.10 muestra el rendimiento del algoritmo de regresión lineal múltiple al aumentarel número de ejemplos, utilizando como entrada el grupo B de parámetros.

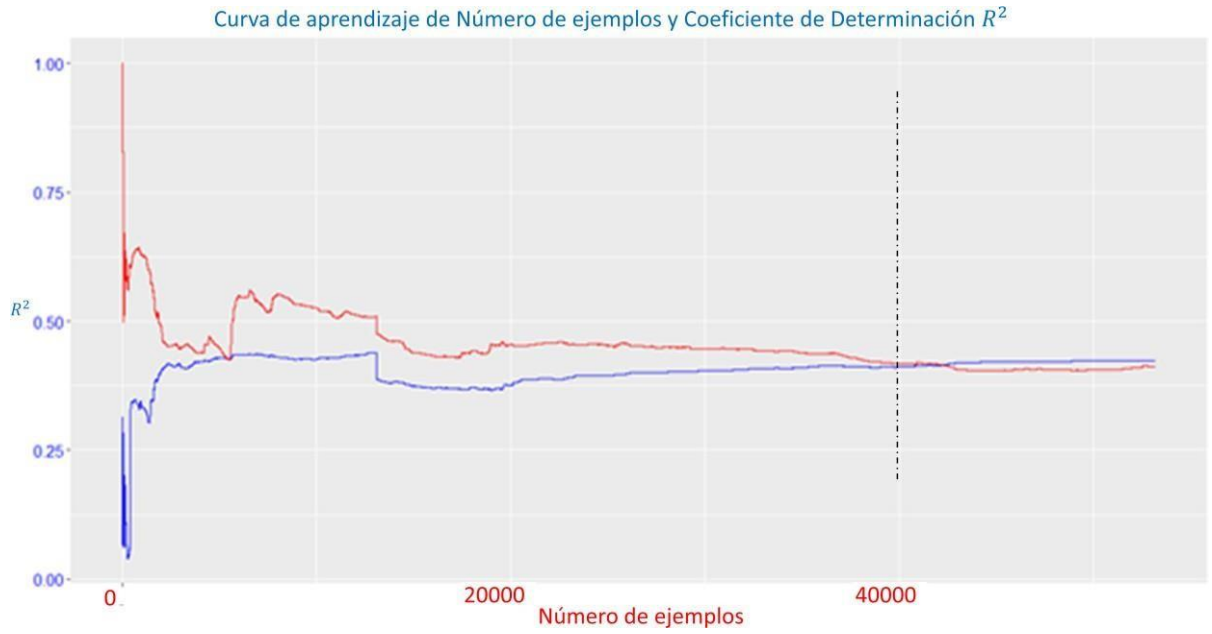


Figura 4.10 Curva de aprendizaje para grupo B (Línea roja entrenamiento y línea azul prueba).

La Figura 4.11 muestra el rendimiento del algoritmo de regresión lineal múltiple al aumentarel número de ejemplos, utilizando como entrada el grupo C de parámetros.

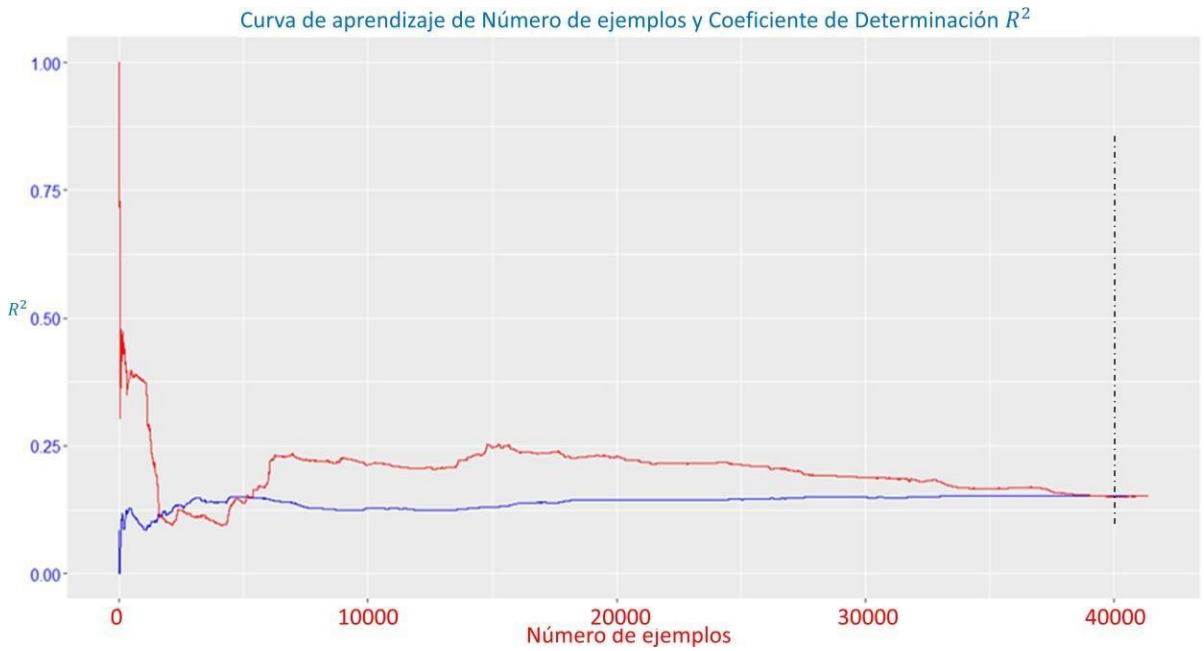


Figura 4.11 Curva de aprendizaje para grupo C (Línea roja entrenamiento y línea azul prueba).

4.3 Evaluación de algoritmos de aprendizaje máquina

Los algoritmos utilizados para la predicción de la demanda bioquímica a 5 días fueron regresión lineal múltiple, regresión de cresta, bosques aleatorios y red elástica. Estos algoritmos se entrenaron y probaron utilizando por separado los grupos de parámetros. La evaluación de los algoritmos fue por medio de los estadísticos de bondad de ajuste raíz del error cuadrático medio (RMSE), error absoluto medio (MAE) y el coeficiente de determinación R^2.
El grupo A se conformó por los parámetros demanda química de oxígeno, nitrógeno amoniacal, nitrógeno Kjeldahl y fósforo total. La Tabla 4.3 presenta los resultados en la etapa de prueba delos algoritmos utilizando como entrada los parámetros del grupo A.

Tabla 4.3 Resultados en la etapa de prueba usando los parámetros del grupo A.

Algoritmo	Estadísticos de bondad de ajuste		
	Raíz del error cuadrático medio (RMSE)	Coeficiente de determinación (R^2)	Error absoluto medio (MAE)
Regresión lineal múltiple	0.53	0.7	0.30
Regresión de cresta	0.53	0.7	0.30
Bosques aleatorios	**0.48**	**0.76**	**0.23**
Red elástica	0.53	0.7	0.30

El grupo B se conformó por los parámetros color verdadero, absorción UV, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, temperatura y temperatura del agua. La Tabla 4.4 presenta los resultados en la etapa de prueba de los algoritmos utilizando como entrada los parámetros del grupo B.

Tabla 4.4 Resultados en la etapa de prueba usando los parámetros del grupo B.

Algoritmo	Estadísticos de bondad de ajuste		
	Raíz del error cuadrático medio (RMSE)	Coeficiente de determinación (R^2)	Error absoluto medio (MAE)
Regresión lineal múltiple	0.67	0.52	0.42
Regresión de cresta	0.67	0.52	0.42
Bosques aleatorios	**0.48**	**0.75**	**0.24**
Red elástica	0.67	0.52	0.42

El grupo C se conformó por los parámetros conductividad eléctrica, turbiedad, pH, temperatura y temperatura del agua. La Tabla 4.5 presenta los resultados en la etapa de prueba de los algoritmos utilizando como entrada los parámetros del grupo C.

Tabla 4.5 Resultados en la etapa de prueba usando los parámetros del grupo C.

Algoritmo	Estadísticos de bondad de ajuste		
	Raíz del error cuadrático medio (RMSE)	Coeficiente de determinación (R^2)	Error absoluto medio (MAE)
Regresión lineal múltiple	0.92	0.11	0.55
Regresión de cresta	0.92	0.11	0.55
Bosques aleatorios	**0.73**	**0.45**	**0.4**
Red elástica	0.92	0.11	0.55

El algoritmo de bosques aleatorios obtuvo resultados óptimos al utilizar los tres grupos de parámetros como entrada. En la etapa de prueba, se obtuvieron 0.48 de RMSE, 0,76 de R^2 y 0.23 de MAE cuando se utilizó el grupo A. Estos parámetros de calidad del agua se determinan en laboratorios con base en las Normas Mexicanas. La selección de estos parámetros para el grupo A puede acelerar la determinación de la demanda bioquímica de oxígeno. Estos parámetros no requieren un largo tiempo de análisis en el laboratorio.

Del mismo modo, el algoritmo de bosques aleatorios obtuvo 0.48 de RMSE, 0,75 de R^2 y 0.24 de MAE utilizando el grupo B. La selección de estos parámetros para el grupo B permite ampliar considerablemente el número de sitios de supervisión. Como estos parámetros se pueden determinar con instrumentos en el área de estudio, se facilita el diagnóstico de la contaminación del agua mediante la predicción de la demanda bioquímica de oxígeno a 5 días con un desempeño similar al del grupo A.

Utilizando como entrada el grupo C, el algoritmo de bosques aleatorios obtuvo 0.72 de RMSE, 0,45 de R^2 y 0.4 de MAE. Al medir estos parámetros con tecnología de sensores para obtener una aproximación de la demanda bioquímica de oxígeno a 5 días se reduce el transporte de muestras y el tiempo de análisis. Además, ofrece la posibilidad de analizar el agua superficial requerida independientemente de su ubicación, proximidad a laboratorios químicos y disposición de instrumentos especializados al diseñar prototipos electrónicos utilizando sensores que midan conductividad eléctrica, turbiedad, pH, temperatura y temperatura del agua. Para aumentar el desempeño de la predicción de la demanda bioquímica de oxígeno a 5 días utilizando los grupos de parámetros se podrían aplicar diferentes técnicas de entrenamiento de algoritmos como el aprendizaje de conjuntos y algoritmos genéticos cumpliendo con los requerimientos computacionales que necesitan estas técnicas.

Capítulo 5. Conclusiones

5.1 Objetivos alcanzados

De los objetivos propuestos por este trabajo se cumplieron los siguientes:

1. Identificar parámetros que presenten relación con la demanda bioquímica de oxígeno a 5 días (DBO5) en aguas superficiales.
2. Establecer 3 grupos de parámetros que permitan predecir la DBO5 en aguas superficiales con las siguientes características: Grupo A de parámetros que se determinan en un laboratorio más rápido que la DBO5 por medio de métodos estandarizados. Grupo B de parámetros que se puedan determinar en la zona de estudio. Grupo C de parámetros que se puedan medir por medio de la tecnología de sensores.

Los grupos de parámetros identificados se seleccionaron a partir de implementar las técnicas de coeficiente de correlación y Forward Selection (FS). Los grupos fueron: Grupo A, Demanda química de oxígeno, nitrógeno amoniacal, nitrógeno Kjeldahl y fósforo. Grupo B, color verdadero, absorción UV, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, temperatura y temperatura del agua. Grupo C, Turbiedad, conductividad eléctrica, temperatura, temperatura del agua y pH.

Estos grupos de parámetros del agua dan la posibilidad al especialista elegir los instrumentos y medidores con los que cuente en el laboratorio y zona de estudio. Además, se facilita y acelera la determinación de la demanda bioquímica de oxígeno a 5 días ya que el medir los parámetros seleccionados en los grupos no se requiere esperar los 5 días para obtener las mediciones.

3. Implementar los algoritmos de aprendizaje máquina, regresión lineal múltiple, regresión de cresta, bosques aleatorios y red elástica para identificar su comportamiento en la predicción de la DBO5 usando como entrada los grupos de parámetros A, B y C por separado.

Estos algoritmos de aprendizaje máquina son fáciles de implementar al utilizar las librerías que incluye el software Rstudio. Además, el ajuste de parámetros para su operación y conceptualización de funcionamiento es sencillo.

4. Evaluar los algoritmos de aprendizaje máquina para la predicción de la DBO5 e identificar el que presente mayor desempeño al evaluarlos por los estadísticos de bondad de

ajuste raíz del error cuadrático medio (RMSE), error absoluto medio (MAE) y el coeficiente de determinación (R^2).

El algoritmo de bosques aleatorios fue el que presentó mayor desempeño al obtener lo siguiente para los grupos de parámetros: Grupo A 0.48 de RMSE, 0.76 de R^2 y 0.23 de MSE. Grupo B 0.48 de RMSE, 0.75 de R^2 y 0.24 de MSE. Grupo C 0.73 de RMSE, 0.45 de R^2 y 0.4 de MAE.

5. Diseñar, desarrollar y calibrar un dispositivo electrónico para medir el grupo de parámetros C que se identificaron por medio de la tecnología de sensores con precisión media. A partir del rendimiento de los algoritmos de aprendizaje máquina utilizando como entrada los parámetros del grupo C turbiedad, conductividad eléctrica, temperatura, temperatura del agua y pH permitieron conducir el diseño, desarrollo y calibración del dispositivo electrónico de medición. Este dispositivo se compone de una tarjeta de adquisición de datos basada en el microcontrolador atmega328p, módulos de alimentación, registro, y los sensores de turbiedad, conductividad eléctrica, pH temperatura ambiente y temperatura del agua.

El uso de estos componentes de bajo costo permite construir el dispositivo electrónico con pocos conocimientos en electrónica, programación y experiencia en mantenimiento, esto al contar con una amplia documentación para el manejo de los componentes. Además, existe una diferencia de costo al comparase con equipos comerciales similares de alrededor de $1500 dólares contra $350 dólares del dispositivo electrónico presentado en este trabajo. Sin embargo, la calibración debe realizarse continuamente con soluciones utilizadas como patrón de trabajo o con equipos de medición de laboratorio previamente calibrados.

5.2 Hipótesis / Proposiciones demostradas

Se puede concluir que el algoritmo de aprendizaje máquina bosques aleatorios obtiene desempeños similares de predecir la demanda bioquímica de oxígeno a 5 días en aguas superficiales, al utilizar como entrada los grupos A y B con las siguientes características:

- Grupo A de parámetros que se determinan en un laboratorio más rápido que la DBO a 5 días por medio de métodos estandarizados (Demanda química de oxígeno, nitrógeno amoniacal, nitrógeno Kjeldahl y fósforo).
- Grupo B de parámetros que se puedan determinar en la zona de estudio (color verdadero, absorción UV, sólidos disueltos totales, conductividad eléctrica, sólidos suspendidos totales, turbiedad, oxígeno disuelto, temperatura y temperatura del agua.).

Por otro lado, utilizar el grupo C (Turbiedad, conductividad eléctrica, temperatura, temperatura del agua y pH) disminuye el rendimiento de la predicción, sin embargo, los parámetros a medir son comunes en calidad del agua, facilitando su medición con tecnología de sensores, equipos de medición en el área de estudio y el dispositivo electrónico de medición presentado en este trabajo.

5.3 Contribuciones de la investigación

1. La identificación de parámetros y su significancia con la demanda bioquímica de oxígeno a 5 días permite utilizar otros parámetros como entrada para la predicción de la DBO5.
2. Los algoritmos implementados variaron su desempeño de la etapa de entrenamiento y prueba. Sin embargo, resaltó el algoritmo de bosques aleatorios, mostrando resultados similares utilizando dos grupos de parámetros como entrada. Esto permite tener una alternativa y referencia a los métodos especializados que comúnmente se utilizan en los laboratorios químicos.
3. El diseño, desarrollo y calibración de un dispositivo electrónico de medición por medio de sensores de turbiedad, conductividad eléctrica, temperatura, temperatura del agua y pH con precisión media. Las lecturas de este equipo en combinación con el algoritmo de bosques aleatorios ofrecen una alternativa a la determinación de la demanda bioquímica de oxígeno a 5 días.

Al analizar y comparar este trabajo con los trabajos de Najafzadeh et al. (2019), Reza Golabi. et al. (2020), Alsulaili and Refaie. (2021) y Najafzadeh and Ghaemi. (2019) resaltan cuatro aspectos, el tipo y número de parámetros que utilizan para predecir la DBO5, la estadística básica de los parámetros, los algoritmos que implementan y el desempeño obtenido en la predicción.

Najafzadeh et al. (2019), utiliza 9 parámetros, pH, turbiedad, conductividad eléctrica, sodio, calcio, magnesio, dióxido de nitrógeno, nitrato y ortofosfato. La estadística básica de la DBO fue de 3.7 mg/l de mínimo, 19.21 mg/l de media y 40.6 mg/l de valor máximo. El algoritmo que obtuvo mejor desempeño para la predicción de la DBO fue programación de expresión génica (GEP) con 0.86 de correlación y 5.388 de RMSE.

Najafzadeh and Ghaemi. (2019) solo implementan diferentes algoritmos a los de Najafzadeh et al. (2019), obteniendo el mejor desempeño el algoritmo de máquina de vectores de apoyo de mínimos cuadrados con kernel polinomial (LS-SVM-Poly) con 0.85 de coeficiente de correlación y 5.46 de RMSE.

Reza Golabi. et al. (2020) utiliza el mismo parámetro de DBO y transformaciones de esta. La estadística básica de la DBO fue de 0.47 mg/l de mínimo, 3.27 mg/l de media y 6.22 mg/l de valor máximo. El algoritmo que obtuvo mejor desempeño para la predicción de la DBO fue bosques aleatorios con transformación de ondas y seleccionando los parámetros por optimización de colonia de hormigas (WRF-PCA) con 0.92 de coeficiente de correlación y 0.0241 de RMSE.

Alsulaili and Refaie. (2021) utiliza 5 parámetros, pH, temperatura, conductividad eléctrica, sólidos suspendidos totales y demanda química de oxígeno. La estadística básica de la DBO fue de 15 mg/l de mínimo, 279.6 mg/l de media y 541 mg/l de valor máximo. El algoritmo que

obtuvo mejor desempeño para la predicción de la DBO fue redes neuronales artificiales con 5 capas ocultas y 9 neuronas por capa con 0.75 de coeficiente de correlación.
En este trabajo el número de parámetros utilizados fueron 4 para el grupo A, 10 para el grupo B y 5 para el grupo C sin utilizar el mismo parámetro de la DBO en estos grupos. Los valores de DBO con la que se contó fueron de 0.1 mg/l de mínimo, 23.2 de media y 125 mg/l de valor máximo. Estos valores de la DBO muestran el rango con el que se entrenó para la predicción de la DBO, por lo que el reducir el rango de valores de DBO para el entrenamiento podría mejorar el rendimiento de los algoritmos. El algoritmo de bosques aleatorios es sencillo de implementar e interpretar, y el desempeño que obtuvo utilizando como entrada los parámetros del grupo A fue de 0.76 de coeficiente de determinación y 0.48 de RMSE. De igual manera, el algoritmo de bosque aleatorios utilizando como entrada los parámetros del grupo B y C fue de 0.75 de coeficiente de determinación, 0.48 de RMSE y 0.45 de coeficiente de determinación y 0.73 de RMSE respectivamente. Por lo que al comparar los desempeños con los trabajos anteriormente presentados son similares, destacando el uso de 4 parámetros del grupo A y 5 del grupo C fáciles de medir y que permitieron construir un dispositivo electrónico de medición en este trabajo.

5.4 Trabajos publicados

El trabajo titulado “Prediction of Biochemical Oxygen Demand in Mexican Surface Waters using Machine Learning” participó en la 5ta Conferencia internacional en computación, matemáticas y estadística con sede en la Universidad Tecnológica Mara de Kedah Branch Malasia. Se obtuvieron resultados preliminares de los algoritmos de aprendizaje máquina usando como entrada dos propuestas de grupos de parámetros del agua.

El trabajo titulado “Arduino: a Novel Solution in the Problem of High-Cost Experimental Equipment in Higher Education” se publicó en la revista Experimental Techniques. Se describió el procedimiento para la construcción de la tarjeta de adquisición de datos basado en el microcontrolador atmega328p y sus aplicaciones en equipo experimental de laboratorio.

Se participó en el décimo cuarto congreso latinoamericano de apicultura Filapi con sede en Chile. Se propuso la construcción de un sistema de monitoreo utilizando como placa principal la tarjeta de adquisición de datos basado en el microcontrolador atmega328p.

Referencias

Abdalrahman Alsulaili., Abdelrahman Refaie. (2021) Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. Water Supply, 21 (5): 1861–1877. doi: https://doi.org/10.2166/ws.2020.199

Abobakr Yahya, A. S., Ahmed, A. N., Binti Othman, F., Ibrahim, R. K., Afan, H. A., El-Shafie, A., Fai, C. M., Hossain, M. S., Ehteram, M., & Elshafie, A. (2019). Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. Water, 11(6), 1231. https://doi.org/10.3390/w11061231

Aheto, J. M. K., Duah, H. O., Agbadi, P., & Nakua, E. K. (2021). A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, ¿Ridge and Elastic net regression approaches compare? Preventive Medicine Reports, 23, 101475. https://doi.org/10.1016/j.pmedr.2021.101475

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water, 11(11), 2210. https://doi.org/10.3390/w11112210

Al-Ghamdi, A. B., Kamel, S., & Khayyat, M. (2021). Evaluation of Artificial Neural Networks Performance Using Various Normalization Methods for Water Demand Forecasting. Proceedings - 2021 IEEE 4th National Computing Colleges Conference, NCCC 2021, 1–6. https://doi.org/10.1109/NCCC49330.2021.9428856

Arreguin-Cortes, F. I., & Cervantes-Jaimes, C. E. (2020). Water Security and Sustainability in Mexico. In J. A. Raynal-Villasenor (Ed.), Journal of the American Water Resources Association (Vol. 6, Issue 1, pp. 177–195). Springer International Publishing. https://doi.org/10.1007/978-3-030-40686-8_10

Conagua. (2020). CONAGUA. https://www.gob.mx/conagua/articulos/calidad-del-agua

Chen, N., Xiong, C., Du, W., Wang, C., Lin, X., & Chen, Z. (2019). An Improved Genetic Algorithm Coupling a Back-Propagation Neural Network Model (IGA-BPNN) for Water-Level Predictions. Water, 11(9), 1795. https://doi.org/10.3390/w1109179

Di, Z., Chang, M., & Guo, P. (2019). Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales. Water, 11(2). https://doi.org/10.3390/w11020339

DS3231 RTC. (2020). Extremely Accurate I2C-Integrated RTC/TCXO/Crystal. https://datasheets.maximintegrated.com/en/ds/DS3231.pdf

El Bilali, A., & Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. Journal of the Saudi Society of Agricultural Sciences, 19(7), 439–451. https://doi.org/10.1016/j.jssas.2020.08.001

Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press. https://doi.org/10.1017/CBO9780511973000

Fonseca-Ortiz, C. R., Mastachi-Loza, C. A., Díaz-Delgado, C., & Esteller-Alberich, M. V. (2020). The Water–Energy–Food Nexus in. In J. A. Raynal-Villasenor (Ed.), Journal of the American Water Resources Association (Vol. 6, Issue 1, pp. 65–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-40686-8_4

Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; De Marinis, G. (2017). Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. Water, 9(2), 105. https://doi.org/10.3390/w9020105

Golabi, M.R., Farzi, S., Khodabakhshi, F. (2020). Biochemical oxygen demand prediction: development of hybrid wavelet-random forest and M5 model tree approach using feature selection algorithms. Environ Sci Pollut Res 27, 34322–34336.https://doi.org/10.1007/s11356-020-09457-x

Gutiérrez, M., Ramírez, J., Gutiérrez, A., García, N., & Villalobos, J. (2018). Prototipo para el monitoreo automatizado de parámetros de calidad del agua en una granja de camarón. Científica, 22, 87–95. https://www.redalyc.org/jatsRepo/614/61458109001/html/index.html#:~:text=Resumen%3A En la acuicultura uno, desarrollo de los organismos cultivados.) http://aquaticcommons.org/16644/1/86. Various Institutions. MBP 2010%5B1%5D.pdf

Hernández-Sampieri, R., Fernández-Collado, C., Baptista-Lucio, P, (2017). Metodología de la investigación.

Jouanneau, S., Recoules, L., Durand, M. J., Boukabache, A., Picot, V., Primault, Y., Lakel, A., Sengelin, M., Barillon, B., & Thouand, G. (2014). Methods for assessing biochemical oxygen demand (BOD): A review. Water Research, 49(1), 62–82. https://doi.org/10.1016/j.watres.2013.10.066

Li, Y., Wang, X., Zhao, Z., Han, S., & Liu, Z. (2020). Lagoon water quality monitoring based on digital image analysis and machine learning estimators. Water, 172, 115471. https://doi.org/10.1016/j.watres.2020.115471

Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River Water Salinity Prediction Using Hybrid Machine Learning Models. Water, 12(10), 2951. https://doi.org/10.3390/w12102951

Méndez-Barroso, L. A., Rivas-Márquez, J. A., Sosa-Tinoco, I., & Robles-Morúa, A. (2020). Design and implementation of a low-cost multiparameter probe to evaluate the temporal variations of water quality conditions on an estuarine lagoon system. Environmental Monitoring and Assessment, 192(11), 710. https://doi.org/10.1007/s10661-020-08677-5

MT3608 convertidor de voltaje. (2020). High Efficiency 1.2MHz 2A Step Up Converter. https://www.olimex.com/Products/Breadboarding/BB-PWR-3608/resources/MT3608.pdf

Najafzadeh, M., Ghaemi, A., & Emamgholizadeh, S. (2019). Prediction of water quality parameters using evolutionary computing-based formulations. International Journal of Environmental

Science and Technology, 16(10), 6377–6396. https://doi.org/10.1007/s13762-018-2049-4

Najafzadeh, M, & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. Environmental Monitoring and Assessment, 191(6), 380. https://doi.org/10.1007/s10661-019-7446-8

NORMA MEXICANA NOM-001-SEMARNAT-1996 LIMITES MAXIMOS PERMISIBLES DE CONTAMINANTES EN LAS DESCARGAS DE AGUAS RESIDUALES EN AGUAS Y BIENES NACIONALES, Pub. L. No. NOM-001-SEMARNAT-1996, Diario Oficial de la Federación 6 (1997).

NORMA MEXICANA NMX-AA-012-SCFI-2001 ANÁLISIS DE AGUA - DETERMINACIÓN DE OXÍGENO DISUELTO EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-012-SCFI-2001, Diario Oficial de la Federación 6 (2001).

NORMA MEXICANA NMX-AA-038-SCFI-2001 ANÁLISIS DE AGUA - DETERMINACIÓN DE TURBIEDAD EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-038-SCFI-2001, Diario Oficial de la Federación (2001).

NORMA MEXICANA NMX-AA-008-SCFI-2016 ANÁLISIS DE AGUA. - MEDICIÓN DEL pH EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS. - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-008-SCFI-2016, Diario Oficial de la Federación (2016).

NORMA MEXICANA NMX-AA-093-SCFI-2000 ANÁLISIS DE AGUA. - DETERMINACIÓN DE LA CONDUCTIVIDAD ELECTROLÍTICA - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-093-SCFI-2000, Diario Oficial de la Federación (2000).

NORMA MEXICANA NMX-AA-045-SCFI-2001 ANÁLISIS DE AGUA. - DETERMINACIÓN DE COLOR PLATINO COBALTO EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-045-SCFI-2001, Diario Oficial de la Federación (2001).

NORMA MEXICANA NMX-AA-028-SCFI-2001 ANÁLISIS DE AGUA - DETERMINACIÓN DE LA DEMANDA BIOQUÍMICA DE OXÍGENO EN AGUAS NATURALES, RESIDUALES (DBO5) Y RESIDUALES TRATADAS - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-028-SCFI-2001, Diario Oficial de la Federación 16 (2001). http://www.conagua.gob.mx/CONAGUA07/Noticias/NMX-AA-029-SCFI-2001.pdf

NORMA MEXICANA NMX-AA-102-SCFI-2006 CALIDAD DEL AGUA – DETECCIÓN Y ENUMERACIÓN DE ORGANISMOS COLIFORMES, ORGANISMOS COLIFORMES TERMOTOLERANTES Y Escherichia coli PRNMX-AA-102-SCFI-2006 CALIDAD DEL AGUA – DETECCIÓN Y ENUMERACIÓN DE ORGANISMOS COLIFORMES, Pub. L. No. NMX-AA-102-SCFI-2006 CALIDAD, 1 (2006).

NORMA MEXICANA NMX-AA-087-SCFI-2010 ANÁLISIS DE AGUA - EVALUACIÓN DE

TOXICIDAD AGUDA CON Daphnia magna, Straus (Crustacea - Cladocera) - MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-087-SCFI-2010 ANÁLISIS, Diario Oficial de la Federación 1 (2010).

NORMA MEXICANA NMX-AA-030/1-SCFI-2012 ANÁLISIS DE AGUA - MEDICIÓN DE LA DEMANDA QUÍMICA DE OXÍGENO EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS. - MÉTODO DE PRUEBA - PARTE 1 - MÉTODO DE REFLUJO ABIERTO -, Pub. L. No. NMX-AA-030/1-SCFI-2012 ANÁLISIS, 1 Diario Oficial de la Federación 1 (2012). http://www.gob.mx/cms/uploads/attachment/file/166774/NMX-AA-030-1-SCFI-2012.pdf

NORMA MEXICANA NMX-AA-034-SCFI-2015 ANÁLISIS DE AGUA - MEDICIÓN DE SÓLIDOS Y SALES DISUELTAS EN AGUAS NATURALES, RESIDUALES Y RESIDUALES TRATADAS – MÉTODO DE PRUEBA, Pub. L. No. NMX-AA-034-SCFI-2015 ANÁLISIS, Diario Oficial de la Federación 16 (2015). https://www.gob.mx/cms/uploads/attachment/file/166146/nmx-aa-034-scfi-2015.pdf

Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A., & Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology, 401(3–4), 177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021

Peña, H. (2016). Recursos Naturales e infraestructura. Desafíos de la seguridad hídrica en América Latina y el Caribe. Naciones Unidas. https://repositorio.cepal.org/bitstream/handle/11362/40074/S1600566_es.pdf?sequence=1&isAllowed=y

Pujar, P. M., Kenchannavar, H. H., Kulkarni, R. M., & Kulkarni, U. P. (2020). Real-time water quality monitoring through Internet of Things and ANOVA-based analysis: a case study on river Krishna. Applied Water Science, 10(1), 22. https://doi.org/10.1007/s13201-019-1111-9

Raynal-Gutierrez, M. E. (2020). Water Use and Consumption: Industrial and Domestic. In J. A. Raynal-Villasenor (Ed.), Journal of the American Water Resources Association (Vol. 6, Issue 1, pp. 103–116). Springer International Publishing. https://doi.org/10.1007/978-3-030-40686-8_6

Raj, S., & Masood, S. (2020). Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. Procedia Computer Science, 167, 994–1004. https://doi.org/10.1016/j.procs.2020.03.399

Rodríguez-Ruiz, J. G., Galván-Tejada, C. E., Zanella-Calzada, L. A., Celaya-Padilla, J. M., Galván-Tejada, J. I., Gamboa-Rosales, H., Luna-García, H., Magallanes-Quintanar, R., & Soto-Murillo, M. A. (2020). Comparison of Night, Day and 24 h Motor Activity Data for the Classification of Depressive Episodes. Diagnostics, 10(3), 162. https://doi.org/10.3390/diagnostics10030162

Rosero-Montalvo, P.D., López-Batista, V.F., Riascos, J.A., Peluffo-Ordóñez, D.H. (2020). Intelligent WSN System for Water Quality Analysis Using Machine Learning Algorithms: A Case Study

(Tahuando River from Ecuador). Remote Sens, 12(12), 1988. https://doi.org/10.3390/rs12121988

Rozario, A. P. R., & Devarajan, N. (2020). Monitoring the quality of water in shrimp ponds and forecasting of dissolved oxygen using Fuzzy C means clustering based radial basis function neural networks. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020-01900-8

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., Maalouf, S., & Adams, C. (2020). Monitoring inland water quality using remote sensing: potential and limitations of spectral Indices, bio-optical simulations, machine learning, and cloud computing. Earth-Science Reviews, 205(April), 103187. https://doi.org/10.1016/j.earscirev.2020.103187

SD module. (2020). Micro SD Card Micro SDHC Mini TF Card Adapter Reader Module for Arduino. http://datalogger.pbworks.com/w/file/fetch/89507207/Datalogger%20%20SD%20Memory%20Reader%20Datasheet.pdf

Semarnat. (2013). Compendio de estadística ambientales edición 2013. https://apps1.semarnat.gob.mx:8443/dgeia/compendio_2013/dgeiawf.semarnat.gob.mx_8080/ibi_apps/WFServlet28b9.html

Samsudin, M. S., Azid, A., Khalit, S. I., Sani, M. S. A., & Lananan, F. (2019). Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones. Marine Pollution Bulletin, 141(February), 472–481. https://doi.org/10.1016/j.marpolbul.2019.02.045

Sensor analógico conductividad eléctrica DFRobot. (2020). Gravity: Analog electrical conductivity sensor/meter(K=10). https://wiki.dfrobot.com/Gravity_Analog_Electrical_Conductivity_Sensor_Meter_K=10_SKU_DFR0300-H

Sensor DHT22, DFRobot. (2020). DHT22 Temperature and humidity module. https://wiki.dfrobot.com/DHT22_Temperature_and_humidity_module_SKU_SEN0137

Sensor DS18B20 Dallas Semiconductor. Dallas Semiconductor (2020). DS18B20 Programmable resolution 1-wire digital thermometer https://cdn.sparkfun.com/datasheets/Sensors/Temp/DS18B20.pdf

Sensor pH-4502c. (2020). PH-4502C Sensor de pH liquido con electrodo E201-BNC. https://uelectronics.com/producto/sensor-de-pH-liquido/

Sensor Turbiedad DFRobot. (2020). Gravity: Arduino Turbidity sensor. https://wiki.dfrobot.com/Turbidity_sensor_SKU_SEN0189

Tabla-Vázquez, C. G., Chávez-Mejía, A. C., Orta Ledesma, M. T., & Ramírez-Zamora, R. M. (2020). Wastewater Treatment in Mexico. In J. A. Raynal-Villasenor (Ed.), Journal of the American Water Resources Association (Vol. 6, Issue 1, pp. 133–155). Springer International Publishing. https://doi.org/10.1007/978-3-030-40686-8_8

TP4056 cargador lineal. (2020). TP4056 1A Standalone Linear Li-lon Battery Charger with Thermal Regulation in SOP-8. https://dlnmh9ip6v2uc.cloudfront.net/datasheets/Prototyping/TP4056.pdf

UNU-INWEH. (2013). Water Security & the Global Water Agenda. The UN-Water analytical brief. In E. & H. (UNU-I. Institute for Water (Ed.), Journal of Chemical Information and Modeling (United Nat, Vol. 53, Issue 9). United Nations University. https://www.unwater.org/publications/water-security-global-water-agenda/

Verma, A.K., Singh, T.N. (2013). Prediction of water quality from simple field parameters. Environ Earth Sci, 69, 821-829. https://doi.org/10.1007/s12665-012-1967-6

Wang, X., Zhang, F., & 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, 7(1), 12858. https://doi.org/10.1038/s41598-017-12853-y

Water quality — Determination of biochemical oxygen demand after n days (BODn) — Part 1: Dilution and seeding method with allylthiourea addition from (ISO 5815-1), (2019).

Water quality — Determination of biochemical oxygen demand after n days (BODn) — Part 2: Method for undiluted samples from (ISO 5815-2), (2003).

Zare Abyaneh, H. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J Environ Health Sci Engineer, 12(40). https://doi.org/10.1186/2052-336X-12-40

Anexos

UNIVERSITI TEKNOLOGI MARA | Cawangan Kedah Kampus Sungai Petani

2021 ICMS

INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

CERTIFICATE OF *Participation*

THIS IS TO CERTIFY THAT

MAXIMILIANO GUZMAN FERNANDEZ

has presented the paper entitled

PREDICTION OF BIOCHEMICAL OXYGEN DEMAND IN MEXICAN SURFACE WATERS USING MACHINE LEARNING

in **The 5th International Conference on Computing, Mathematics and Statistics 2021**

virtually on
4-5 AUGUST 2021

Organized by
FACULTY OF COMPUTER & MATHEMATICAL SCIENCES
UiTM KEDAH BRANCH

PROF. DR. MOHAMAD ABDULLAH HEMDI
RECTOR,
UNIVERSITI TEKNOLOGI MARA (UiTM) KEDAH BRANCH

STRATEGIC PARTNER

SPONSORED BY

FILAPI2020
XIV CONGRESO LATINOAMERICANO DE APICULTURA

CERTIFICADO

XIV CONGRESO LATINOAMERICANO DE APICULTURA

Maximiliano Guzmán-

Asistió al Congreso Latinoamericano de Apicultura Filapi 2020 que se desarrolló en formato virtual desde el 03 y hasta el 07 de noviembre de 2020.

Lucas Martinez
Presidente
Federación Latinoamericana de Apicultura FILAPI

RED NACIONAL APICOLA - CHILE
FILAPI
UNIVERSIDAD DE LA FRONTERA
ProChile
PATROCINA
RED NACIONAL APICOLA - CHILE

UNIVERSITI TEKNOLOGI MARA | Cawangan Kedah Kampus Sungai Petani

2021 iCMS INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e-PROCEEDINGS

of The 5th International Conference on Computing, Mathematics and Statistics (iCMS2021)

4-5 August 2021

Driving Research Towards Excellence

e-Proceedings of the 5th International Conference on Computing, Mathematics and Statistics (iCMS 2021)

Driving Research Towards Excellence

e-ISBN: 978-967-2948-12-4
DOI

Library of Congress Control Number:

Publication by
Department of Mathematical Sciences
Faculty of Computer & Mathematical Sciences
UiTM Kedah

PREDICTION OF BIOCHEMICAL OXYGEN DEMAND IN MEXICAN SURFACE WATERS USING MACHINE LEARNING

Maximiliano Guzmán-Fernández[1], Misael Zambrano-de la Torre[2], Claudia Sifuentes-Gallardo[3], Oscar Cruz-Dominguez[4], Carlos Bautista-Capetillo[5], Juan Badillo-de Loera[6], Efrén González Ramírez[7], Héctor Durán-Muñoz[8]

[1,2,3,6,7,8] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, México,
[4] Universidad Politécnica de Zacatecas, [5] Doctorado en Ciencias de la Ingeniería, Universidad Autónoma de Zacatecas, México.
([1] maxguzman1@hotmail.com, [2] misaelzambrano1997@gmail.com, [3] clauger17@gmail.com,
[4] racso_zurc@hotmail.com, [5] baucap@uaz.edu.mx, [6] l_badillo@uaz.edu.mx,
[7] gonzalezefren@uaz.edu.mx, [8] hectorduranm@hotmail.com)

The monitoring of surface water quality is insufficient in Mexico due to the limited water monitoring stations. The main monitoring parameter to evaluate surface water quality is the biochemical oxygen demand. This parameter estimates the biodegradable organic matter present in the water. Concentrations above 30 mg/l indicates a high level of contamination by domestic and industrial waste. Therefore, the aim of this work to provide a reference to the conventional process of determining biochemical oxygen demand using machine learning. The database used was collected by the National Water Commission (CONAGUA). Pearson's correlation and Forward Selection techniques were applied to identify the parameters with the most important contribution to prediction of biochemical oxygen demand. Two groups were formed and used as input to four machine learning algorithms. Random forest algorithm obtained the best performance. Group 1 and 2 of parameters obtained a 0.76 and 0.75 coefficient of determination respectively. This allows choosing an adequate group of parameters that can be determined with the chemical analysis instruments available in the study area.

***Keywords*:** Machine Learning, Biochemical Oxygen Demand, Mexican Surface Waters.

1. Introduction

The preservation, treatment, access, and efficient use of water is fundamental for humanity, and several countries consider it as a national security resource. Due to the importance of this resource, the concept of water security has been discussed and defined by several institutions, such as the United Nations Water Group (UN-Water), the Economic Commission for Latin America and the Caribbean (CEPAL), among others. The use of water from rivers, wells and lagoons is of great importance, as they are the main sources of water supply in the municipalities and states for various uses. However, activities such as mining, livestock, agriculture and industrial demand generate water exploitation and contamination (Raynal, 2020). Water quality is an important factor to consider, whether for ecosystem needs or for contamination levels that directly impact food, hygiene, health, and economy. To ensure the safe use of water, continuous monitoring of water quality parameters in the water supply sources and discharge areas should be carried out.

In Mexico, the agency in charge of managing, regulating, controlling and protecting the country's national waters is the National Water Commission (CONAGUA). CONAGUA performs the monitoring of the main water bodies in the country, both surface and groundwater. Biochemical oxygen demand is one of the main parameters when evaluating surface water quality at monitoring sites in Mexico. This parameter indicates the biodegradable organic material present in the sample of surface water bodies after 5 days. Based on the contamination level established by CONAGUA, if the biochemical oxygen demand is above 30 mg/l, the water is considered contaminated. This parameter is normally obtained by taking samples in the study area and then transferring them to a laboratory for subsequent analysis of the sample. Sample analysis is through biochemical and manometric methods, involving specialized

instruments and reagents. This conventional process of collecting samples from the study area and analyzing them in the laboratory requires considerable time and labor. As a consequence, it is not possible to have real-time monitoring of water quality. In addition, the diagnosis of contamination is reduced to identifying and analyzing more frequently surface water monitoring sites that are located near certified laboratory infrastructure.

To assist the study process performed by the specialists, it is possible to perform statistical analyses and generate predictive models using artificial intelligence, based on measurement data previously obtained by the specialists. The implementation of artificial intelligence through machine learning and data mining using algorithms for the prediction of water quality parameters in different monitoring zones has been reported in the literature. They are characterized by different stages such as preprocessing, normalization and evaluation of the supervised learning algorithms with goodness- of-fit statistics. For the analysis in rivers, water quality was classified by temperature, pH, turbidity and total dissolved solids. Data are collected in a river and water quality is classified using K-Nearest Neighbors, support vector machine, Bayesian classifier and decision trees. The performance of the algorithms is evaluated by sensitivity, specificity, accuracy and precision (Rosero et al., 2020). This indicates the possibility of involving easily determinable parameters in the study area to diagnose contamination. Similarly, the implementation of the support vector machine algorithm for water quality prediction obtained a correlation coefficient of 0.97 and 0.058 mean square error. The data was collected from the Malaysian Department of Environment. The parameters used as input to the algorithms were ph, dissolved oxygen, biochemical oxygen demand, chemical oxygen demand and ammonia-nitrogen (Abobakr Yahya et al., 2019). These results show that machine learning algorithms can be implemented for the prediction of particular parameters according to local conditions. Different machine learning algorithms such as polynomial regression, model tree and gene expression programming have been implemented for the prediction of biochemical oxygen demand. Ca2+, Na+, Mg2+, NO-2, NO-3, PO3-4, electrical conductivity, pH and turbidity were the input parameters. The gene expression programming algorithm performed acceptably with 5.388 root mean square error and 0.86 correlation coefficient (Najafzadeh et al., 2019). However, most of the parameter groups used as input to algorithms for prediction of biochemical oxygen demand are chosen depending on laboratory capabilities and no matter the determination time. Therefore, choosing parameter groups that are obtained quicker than determining biochemical oxygen demand would save analysis time and allow more study areas to be diagnosed.

The aim of this work is to predict the biochemical oxygen demand in surface waters of Mexico using machine learning algorithms. This study is presented using measurements of water quality parameters from the 2764 surface water monitoring sites in Mexico, acquired by CONAGUA from 2012 to 2019. Pearson's correlation and Forward Selection techniques were applied to select two groups of parameters as input to the multiple linear regression, ridge regression, random forest and elastic net algorithms. The groups of parameters were: (1) if it is possible to transfer the sample to a laboratory, a group of parameters that are obtained quicker than determining biochemical oxygen demand. (2) a group of parameters that can be measured in the study area. The database was split by cross-validation for training and testing of the algorithms. The performance of the algorithms was evaluated by goodness-of-fit statistics. Random forest obtained the best prediction. Similar results were obtained when using both groups of parameters as input. Therefore, this work provides a group of parameters that can be measured in the study area and a group of parameters that can be quickly determined in a laboratory.

2. Materials and Methods

The methodology used in this work, to obtain the prediction of biochemical oxygen demand in surface waters of Mexico, consists of three stages. The first stage was data preprocessing. The second stage consisted of data analysis for the prediction of biochemical oxygen demand.

In the last stage, the machine learning algorithms are validated. The methodology is shown in Figure 1 and was implemented with R studio software version 4.0.2. The following sections describe each stage.

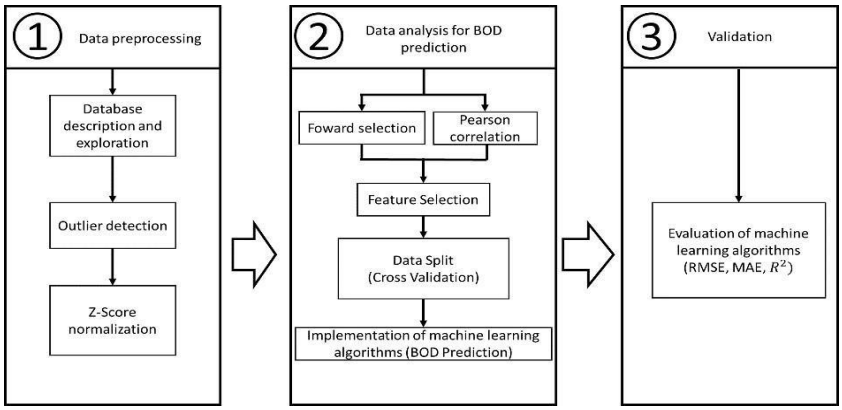


Figure 1: Stages of the methodology: (1) Data preprocessing, (2) Data analysis for prediction of biochemical oxygen demand and (3) Validation

2.1 Data preprocessing

In this first stage, the database collected by CONAGUA was used, which contains indicators of monitoring sites and water parameters from 2012 to 2019. The database is composed of 177 chemical and biological parameters of surface water in Mexico, with a total of 110827 samples. Errors of capture were eliminated and 31 chemical and biological parameters were obtained with a total of 58824 samples. The parameters used and their basic statistics are presented in Table 1.

Table 1: Basic statistics of database parameters.

Parameter (Units)	Min	Mean	Max	Parameter (Units)	Min	Mean	Max
Fecal Coliform (NMP/100mL)	1	55772	24196000	Total Suspended Solid (mg/L)	0.1	105	20812
Escherichia Coli (NMP/100mL)	1	46459	24196000	Turbidity (UNT)	0.01	75	21500
Biochemical Oxygen Demand (mg/L)	0.1	23.2	7667	Arsenic (mg/L)	0.0001	0.006	1
Chemical Oxygen Demand (mg/L)	0.9	77.7	14489	Cadmium (mg/L)	0.00002	0.0002	0.1
Phosphorus (mg/L)	0.001	1.3	95.2	Chromium (mg/L)	0.0002	0.01	76.5
Organic Nitrogen (mg/L)	0	2.5	827.8	Mercury (mg/L)	0.00001	0.0003	0.5
True Color (U Pt/Co)	2.5	55.2	8000	Nickel (mg/L)	0	0.005	7.3
UV Absorbance (U Abs/cm)	0.002	0.17	17	Lead (mg/L)	0.001	0.003	1.8
Total Dissolved Solids (mg/L)	2.4	354.5	159520	Hardness (mg/L)	3.8	295.2	37965
Electrical Conductivity(uS/cm)	3.8	1056	199400	Temperature (°C)	-6	27.6	51
PH (UpH)	2.9	7.8	11.8	Water Temperature (°C)	4	24.9	62
% Dissolved Oxygen (% Saturation)	0.6	73.2	1113.3	Total Organic Carbon (mg/L)	0.06	12.8	2490
Dissolved Oxygen (mg/L)	0.05	5.7	762	Nitrogen (mg/L)	0.008	7.4	1244.1
Ammoniacal Nitrogen (mg/L)	0.003	3.7	497	Kjeldahl Nitrogen (mg/L)	0.003	6.34	1239.8
Nitrogen Dioxide (mg/L)	0.0005	0.1	21.84	Orto-Phosphate (mg/L)	0.0005	0.87	144.4
Nitrate Nitrogen (mg/L)	0.0004	1	336.2				

Box plot and basic statistics were chosen to detect outliers. Most of the parameters varied their maximum values due to measurement errors or collection anomalies. Each parameter was analyzed separately per year and outliers were adjusted to a threshold value. This value was considered as a maximum due to the rest of the measurement values. E.g. in 2012, the Chemical Oxygen Demand had outliers of 14489 mg/L, so the limit value of 250 mg/L was determined and all values exceeding the limit value were assigned to 250 mg/L (Ahmed et al., 2019). To finalize stage one, the parameters were normalized in order to stablish the parameter values on a common scale. The z-score is a method for normalization and standardization that represents the number of standard deviations and allows one to know how far away one is from the mean for each point or raw parameter. Equation (1) shows the z-score normalization expression applied to each parameter, where x represents the parameter value, μ is the mean of the parameter and σ is the standard deviation:

$$z - score = (x - \mu)/\sigma \tag{1}$$

2.2 Data analysis for prediction of biochemical oxygen demand

To begin stage 2, a correlation analysis was performed using Pearson's method. In order to find the dependent and independent variables that have a linear behavior. This method allows us to extract the parameters that have the highest relationship. Equation (2) presents the Pearson correlation between the values of two vector Xi and Yi, where $\bar{x}$ is the mean of the vector x_i $\bar{y}$ is the mean of the vector y_i and n the number of total values in the sample.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2} \tag{2}$$

Continuing with stage 2, the Forward Selection technique was applied to support the selection of parameter groups. This technique first evaluates the individual contribution of each parameter to the prediction of biochemical oxygen demand. Then, the parameters with the highest individual contribution are sorted in descending order and grouped together. Creating database sets by adding one parameter at a time. This set of parameters is used as input to the algorithm and the coefficient of determination is evaluated when predicting biochemical oxygen demand. This process is performed with 70% of measurements for algorithm training and 30% for algorithm testing (Melesse et al., 2020). The algorithm used to apply Forward Selection was multiple linear regression. The goodness-of-fit statistic used to evaluate the individual and joint contribution of the parameters was the coefficient of determination R^2.

After applying the forward selection technique, data split was carried out. The purpose of this division is to validate the algorithms with a balance of the measurements. Cross-validation was the technique used, as this technique divides the data into k subparts and iterates on all subparts of the entire database, having for training k-1 subparts and 1 subpart for testing. In this work was used k=3 since the database consists of 58824 measurements and allows us to use a large balanced number of data for training and testing. 41176 was the number of measurements used for training and 17648 was the number of measurements used for testing.

To finalize stage two, four machine learning algorithms were implemented to predict the biochemical oxygen demand in surface water. The first algorithm used was multiple linear regression, with this algorithm it was possible to obtain an equation of the output variable as a function of the input variables. The second algorithm used was Random Forest, which is based on a decision tree and generates several base models giving good efficiency, it can be used for regression and classification. Also, in this work the Ridge Regression algorithm was used, which uses the same principles as a linear regression, and adds some bias to avoid the effect of having high variances. It also minimizes the sum of the squared residuals. Finally, the Elastic net algorithm, which combines the efficiency of ridge regression, was used in this work. It minimizes the cost function by combining the penalty methods of both algorithms.

2.3 Validation

In stage 3, the algorithms were evaluated in training and testing. The evaluation was through the goodness- of-fit statistics of the root mean square error, mean absolute error and the coefficient of determination. The coefficient of determination was used. It determines the variation that exists between the predictions, the true values and the mean of the values. Equation (3) presents the expression of the coefficient of determination, where y_i are the actual values, y'_i are the predictions, $\bar{y}$ represents the mean of the values and n the number of total values in the sample:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

It was also necessary to use the root mean square error, by scaling the values to the range of the mean square error values. Equation (4) shows the expression, where y_i are the actual values, y'_i are the predictions and n the number of total values in the sample:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2} \quad (4)$$

In addition, the mean absolute error, which represents the sum of the absolute value of the error, was taken and then divided by the total number of values in the sample. Equation (5) shows the expression, where y_i are the actual values, y'_i are the predictions and n the number of total values in the sample:

$$MAE = \frac{1}{n}\|y_i - y'_i\| \quad (5)$$

3. Results and Discussion

3.1 Outlier detection

The box-plot analysis and basic statistics showed that most of the parameters had outliers, with maximum values significantly off the mean, so these values are replaced by the measurement limits for each parameter. Table 2 presents the parameters used and the basic statistics after removing the outliers. By changing the values for each parameter, the data used for training and testing the algorithms were free of bias. Only by modifying the values that seemed to be out of the limits.

Table 2: Basic statistics of the parameters after assigning a limit value.

Parameter (Units)	Min	Mea	Max	Parameter (Units)	Min	Mean	Max
Biochemical Oxygen Demand (mg/L)	0.1	14.7	120	Total Suspended Solids (mg/L)	0.1	66.8	400
Chemical Oxygen Demand (mg/L)	0.9	55.2	250	Phosphorus (mg/L)	0.001	1.2	20
Dissolved Oxygen (mg/L)	0.05	5.7	40	Temperature (°C)	-6	27.6	51
True Color (U Pt/Co)	2.5	45.1	200	Turbidity (UNT)	0.01	49.2	500
UV Absorbance (U Abs/cm)	0.002	0.17	2	Water Temperature (°C)	4	24.9	62
Ammoniacal Nitrogen (mg/L)	0.003	3.7	200	Kjeldahl Nitrogen (mg/L	0.003	6.3	400
Electrical Conductivity(uS/cm)	3.8	900	5000	Total Dissolved Solids (mg/L)	2.4	455.4	1000
Total Organic Carbon (mg/L)	0.06	12.5	1000				

3.2 Features Selection

Based on Pearson's correlation and Forward Selection, parameters were selected for group 1 and group 2. Figure 2 shows the heat map representing the Pearson correlation between the parameters.

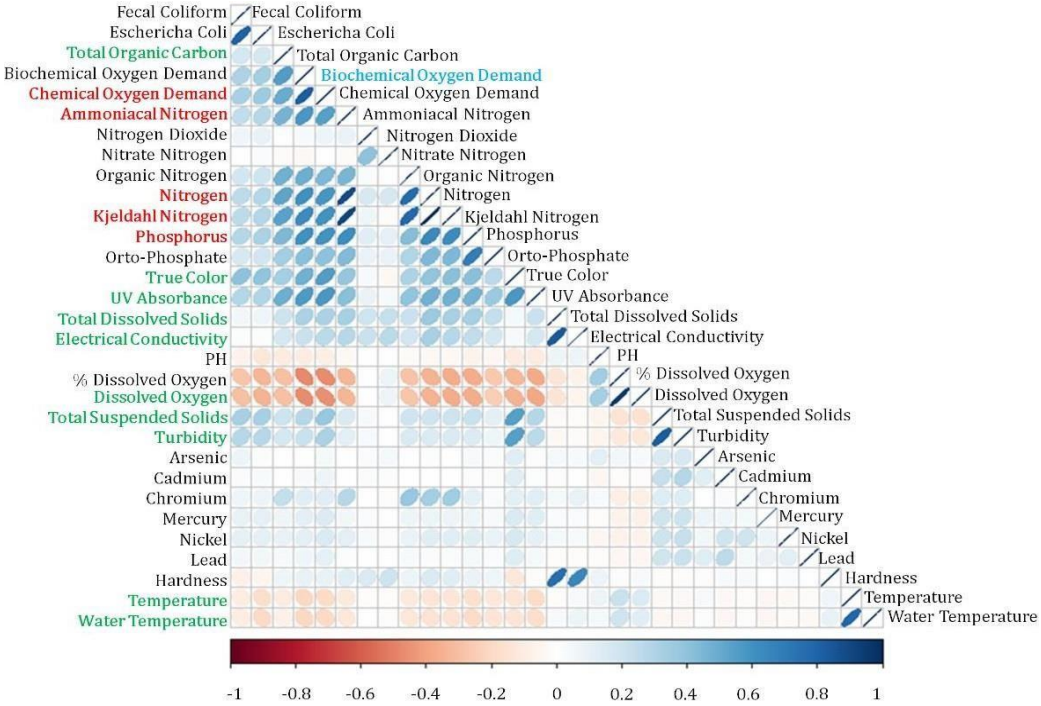


Figure 2: Heat map for Pearson's correlation for all parameters in the processed database

Chemical Oxygen Demand shown a high correlation ($|r| > 0.7$). This parameter involves Biochemical Oxygen Demand, by measuring the complete oxidation of the sample, both organic, biodegradable and non-biodegradable materia (r=0.81). Total Organic Carbon, Ammoniacal Nitrogen, Nitrogen, Kjeldahl Nitrogen, Phosphorus, UV Absorbance, Fecal Coliform, Escherichia Coli, Organic Nitrogen, Ortho-Phosphate, True Color, Total Dissolved Solids, and Dissolved Oxygen, shown a moderate correlation ($0.3 < |r| < 0.7$). This can be related to the method and technique of parameter determination. Nitrogen Dioxide, Nitrate Nitrogen, Electrical Conductivity, PH, Total Suspended Solids, Turbidity, Arsenic, Cadmium, Chromium, Mercury, Nickel, Lead, Hardness, Temperature and Water Temperature shown a weak correlation ($0 < |r| < 0.3$). Electrical Conductivity provides general information on the concentration of salts and ions, so it shows a weak correlation with the Biochemical Oxygen Demand (r=0.21) and high correlation with Total Dissolved Solids (r=0.83).

The results of applying Forward Selection were as follows. The coefficient of determination individually obtained for Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen and Phosphorus was 0.66, 0.328, 0.39 and 0.36, respectively. The coefficient of determination individually obtained for Total Organic Carbon, True Color, UV Absorption, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Water Temperature and Temperature were 0.31, 0.22, 0.31, 0.1, 0.04, 0.07, 0.04, 0.21, 0.04 and 0.04, respectively. After several tests using different combinations of parameters with coefficients of determination greater than 0.3 and less than 0.3 as input to the multiple linear regression algorithm, two sets were formed. This process also allowed us to determine the performance of using a multiple linear regression algorithm with all the parameters available in the processed database and get a reference of the maximum performance of that algorithm. Table 3 shows the increase in the coefficient of determination when grouping the parameters into sets.

Table 3: Coefficient of determination when combining the parameters by forward selection into two sets.

Sets of parameters used as input to multiple linear regression algorithm	Coefficient of Determination [0-1]	
	Training	Testing
Chemical Oxygen Demand, Ammoniacal Nitrogen	0.69	0.69
(Set 1) Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen, Phosphorus.	0.70	0.70
Total Organic Carbon, True Color, UV Absorbance, Total Dissolved Solids, Electrical Conductivity.	0.48	0.46
(Set 2) Total Organic Carbon, True Color, UV Absorbance, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Temperature, Water Temperature.	0.53	0.51

The parameters selected for group 1 were: Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen and Phosphorus. The parameters selected for group 2 were: Total Organic Carbon, True Color, UV Absorption, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Water Temperature and Temperature. According to the methodology used in this work, Forward Selection confirmed the performance and relationship of the Pearson correlation, showing the same parameters that complied with the established groups.

3.3 Validation of machine learning algorithms

After implementing the multiple linear regression, ridge regression, random forest and elastic net algorithms, it was found that the best performing algorithm was random forest. The performance of the algorithms when using group 1 and 2 as input are shown in Table 4 and Table 5.

Table 4: Results in the testing stage of the algorithms using group 1 parameters as input.

Algorithm	Goodness of fit		
	Root Mean Square Error	Coefficient of Determination	Mean Absolute Error
Multiple Linear Regression	0.53	0.7	0.30
Ridge Regression	0.53	0.7	0.30
Random Forest	**0.48**	**0.76**	**0.23**
Elastic Net	0.53	0.7	0.30

Table 5: Results in the testing stage of the algorithms using group 2 parameters as input.

Algorithm	Goodness of fit		
	Root Mean Square Error	Coefficient of Determination	Mean Absolute Error
Multiple Linear Regression	0.67	0.52	0.42
Ridge Regression	0.67	0.52	0.42
Random Forest	**0.48**	**0.75**	**0.24**
Elastic Net	0.67	0.52	0.42

The random forest algorithm obtained optimal results when using the two groups of parameters as input. At the test step, 0.48 RMSE, 0.76 $R2$ and 0.23 MAE were obtained when using group 1. These water quality parameters are determined in laboratories based on Mexican Standards. Selecting these parameters for group 1 can accelerate the determination of biochemical oxygen demand. These parameters do not require a long analysis time in the laboratory.

Similarly, the random forest algorithm obtained 0.48 RMSE, 0.75 $R2$ and 0.24 MAE using group 2. Selecting these parameters for group 2 allows the number of monitoring sites to be greatly expanded. As these parameters can be determined with instruments or sensors in the study area, the diagnosis of water pollution by predicting biochemical oxygen demand is facilitated. This reduces sample transport and analysis time. It offers the possibility of analyzing the required surface water regardless of its location or proximity to chemical laboratories. Additionally, using the groups of parameters identified by this work, different algorithm training techniques could be applied to increase performance such as ensemble learning and genetic algorithms.

4. Conclusion

Water quality is essential for the human life development. Through the present work it was possible to identify the best algorithm that can predict the biochemical oxygen demand in surface waters of Mexico. Also, the parameters that have the most influence. Random forest showed flexibility when implemented in the prediction of biochemical oxygen demand by obtaining 0.48 RMSE, 0.76 R^2 and 0.23 MAE using the parameters Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen and Phosphorus. In addition, 0.48 RMSE, 0.75 R^2 and 0.24 MAE were obtained using the parameters Total Organic Carbon, True Color, UV Absorption, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Water Temperature and Temperature. This indicates that based on the local conditions and the study area, the biochemical oxygen demand can be obtained in a similar way and diagnose water contamination in Mexico in a relatively short time. As a future work, it is proposed to design and develop a real-time electronic monitoring device to measure the parameters of group 2 obtained in this work.

Acknowledgment

We would like to thank the Mexican National Council for Science and Technology (CONACYT) for their support in all activities.

References

M. E. Raynal Gutierrez. (2020). Water use and consumption: industrial and domestic in water resources of Mexico. Vol. 6, J. A. Raynal-Villasenor, ed. GEWERBESTRASSE 11, 6330 Cham, Switzerland: Springer Nature, 2020, pp. 103–116.

P. D. Rosero-Montalvo, V. F. López-Batista, J. A. Riascos, and D. H. Peluffo-Ordóñez. (2020). Intelligent WSN system for water quality analysis using machine learning algorithms: a case study (Tahuando river from Ecuador). *Remote Sens*, 12(12).

A. S. Abobakr Yahya., A. N. Ahmed, F. Binti Othman, R. K. Ibrahim, H. A. Afan, A. El-Shafie, C. Fai, M.S. Hossain, M. Ehteram and A. Elshafie. (2019). Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. *Water*, 11(6).

M. Najafzadeh, A. Ghaemi, and S. Emamgholizadeh. (2019). Prediction of water quality parameters using evolutionary computing-based formulations. *Int. J. Environ. Sci. Technol*, 16(10).

U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, 11(11).

A. M. Melesse, K. Khosravi, J.P. Tiefenbacher, S. Heddam, S. Kim, A. Mosavi, B. T. Pham. (2020). River Water Salinity Prediction Using Hybrid Machine Learning Models. *Water*, 12(10).

iCMS2021: 4 - 5 August

2021 ICMS

INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e ISBN 978-967-2948-12-4

9 7 8 9 6 7 2 9 4 8 1 2 4

Experimental Techniques
https://doi.org/10.1007/s40799-021-00449-1

RESEARCH PAPER

Arduino: a Novel Solution to the Problem of High-Cost Experimental Equipment in Higher Education

M. Guzmán-Fernández[1] • M. Zambrano de la Torre[1] • J. Ortega-Sigala[2] • C. Guzmán-Valdivia[3] • J. I. Galvan-Tejeda[1] • O. Crúz-Domínguez[3] • A. Ortiz-Hernández[3] • M. Fraire-Hernández[1] • C. Sifuentes-Gallardo[1] • H.A. Durán-Muñoz[1]

Received: 5 December 2019 / Accepted: 1 February 2021

Abstract

The acquisition of experimental equipment has become a problem, due to its high costs. To partially solve this problem, the scientific community has used the Arduino data acquisition board (DAB). This board has a low-cost and allows the automation of equipment in a simple and practical way. However, using only the Arduino board is no longer enough, it is now necessary to combine it with the implementation of an user-friendly interface. Through this combination, it is possible to develop sophisticated laboratory equipment. The aim of this work is to describe step by step and in a simple way the development and implementation of a low-cost experimental equipment based on Arduino for use in higher education. Also, this work tries to encourage users to develop their own laboratory equipment, and solve their equipment needs. The novelty of this work pretends to be that any user without a deep knowledge of electronics and programming can easily assemble their own lab equipment, avoiding high equipment fees. In this work, three simple examples are shown, and any teacher, researcher, or student can easily reproduce them. With the combination of three previous examples, it is possible to develop sophisticated laboratory equipment, for example a Ph Meter with temperature sensor and stirring setup. The commercial cost of this lab equipment is approximately 600 USD, but with this homemade setup, the total cost drops to 75 USD.

Keywords Arduino · Development and implementation of new technologies · Higher education

Introduction

The acquisition of experimental equipment has become a problem due to its high costs. To partially solve this problem, the scientific community has developed new low-cost technologies. These new low-cost technologies have been applied in different study fields [1, 2]. For example, the use of social networking apps in education, the study of physical phenomena, etc. In the case of laboratory equipment, there is the problem to perform chemistry practices, where it is needed to use high-cost lab equipment. Even, the majority of universities do not have the financial capacity to acquire such laboratory equipment. This problem is causing a growing need to develop additional new low-cost laboratory equipment [3, 4]. A powerful tool to partially solve this problem is the Arduino data acquisition board (DAB).

This board has a low-cost, and the user does not need to have deep knowledge on electronics and programming. Basically, this board is a unit of control and the heart of the entire automation [5]. In the literature, there are many articles using the Arduino board to solve particular problems. For example: systems to recognize crying infants [6], DNA recognition systems [7], low-cost cameras for the study of animal behavior [8], for monitoring temperature in greenhouses [9] and even systems for fluorescence detection [10], among others [11]. However, most of the articles do not detail the electronic diagrams used, nor the programming code of the Arduino board. Therefore, it is difficult for users who do not have deep knowledge on electronics and programming to reproduce such lab equipment. Also, a full real-time automatization of lab equipment is not possible using only the Arduino board. It is necessary to combine the board with a friendly interface that controls lab equipment in real-time, with this

* H.A. Durán-Muñoz
hectorduranm@hotmail.com

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zac. Zacatecas, Mexico

[2] Unidad Académica de Física, Universidad Autónoma de Zacatecas, Fresnillo, Mexico

[3] Carrera de Ingeniería Industrial, Universidad Politécnica de Zacatecas, Fresnillo, Zacatecas, Mexico

combination it is possible to develop sophisticated lab equipment and avoid high equipment fees.

To control the Arduino board in real-time, there are several kinds of software for interface design. However, most are paid with a high costs. An attractive alternative is Visual Basic software, with its latest version Visual Studio. There is also a wide use of Visual Studio for different applications, for example, the characterization of materials [12], statistical analysis [13, 14], and computational simulations for chemical processes [15]. The combination of Arduino and Visual Studio turns out to be an extremely powerful tool, developing new low-cost experimental equipment and solving the current problem to acquire high-cost equipment. Anothers advantage of the user who designs his own control interface is that user licenses do not have to be paid, it is possible to incorporate additional elements to carry out more complex experiments, technical support is carried out by the user, encourages technological development and creates the possibility of generating additional financial dividends for universities.

The aim of this work is to describe step by step and in a simple way the development and implementation of a low-cost experimental equipment based on Arduino for use in university laboratories. Also, this work tries to encourage users to develop their own laboratory equipment. So that in the future they can develop their own laboratory equipment, and solve their particular equipment needs. The novelty of this work pretends to be that any user without a deep knowledge of electronics and programming can easily assemble their own lab equipment, avoiding acquiring high-cost experimental equipment. This work presents a low-cost and multifunctional solution that can be implemented to replace high-cost experimental equipment. For this, the manufacturing process of a DAB, based on Arduino, is described step by step. Subsequently, the automation of the board will be carried out by designing a friendly interface of Visual Studio.

The development of the graphic interface is also carried out step by step and allows the user to have control over the actuators connected to theboard.Also, threesimpleexamplesare presentedinthis manuscript.Theseexamples canbe easily implementedand thatcan solve the needs ofany researcher, teacher orstudent to develop their own experimental equipment or a device complementary to experimental equipment. Finally, with the combination of three previous examples, itwas possible tomanufacture a Ph Meter with a temperature sensor and stirring setup. The commercial cost of this experimental equipment is approximately 600 USD, but with this homemade system, the total cost drops to 75 USD.

Materials and Methods

DAB Design and Control Software

The data acquisition process is the physical phenomenon that will be analyzed, then a transducer acts by transforming the received signal into an electrical signal, then digitizing it and achieving its processing and reading through the DAB, thus reaching the information to the computer to present it graphically with a binary system.

Data Acquisition Board (DAB)

The main feature of a data acquisition board (DAB) is to use the minimum of components, which is low cost and has an optimal operation to solve the necessary tasks of the user. The ATmega328P microcontroller was used to design the DAB, its main feature being low-cost and versatility. For the design and construction of the board, low cost and easily obtainable components were selected. The components are shown in Table 1.

The FTDI FT232 device (USB to TTL serial adapter) allows the communication between the microcontroller through peripherals and programming software. The electronic configuration of the board is shown in Fig. 1. Microcontroller programming can be done using the free-use Arduino IDE software.

For the design of the printed circuit board (PCB), the free-use EasyEDA online software was used, offering essential tools, easy to use in the design of electronic circuits and generation of printed circuits (Fig. 2). Considering each of the components of the previous design, the printed circuit template is generated, necessary for the creation of the PCB.

Table 1 Components used for the DAB

1000Ω and 330Ω (1/4 W) resistance.
Light emitting diode. 3 mm in diameter, 4.2 V maximum voltage and 20 mA. Visually indicates the power on the board.
Pushbutton with two terminals. It has the function of resetting the microcontroller.
0.1uF ceramic condenser. Necessary for the communication of the microcontroller with the FT232 module.
Two ceramic capacitors of 22 pF. Necessary in the connection of the external oscillator crystal of the microcontroller.
16 MHz oscillator crystal. Necessary for the operation of the microcontroller.
FTDI FT232. Component responsible for communication between the microcontroller and the PC.
ATmega328P-PU microcontroller.
Phenolic plate. DAB base, copper sheet with dimensions of 6×8 cm.
Ferric chloride. Chemical compound used for the construction of the PCB board.
USB-Mini B cable connection. Necessary component for communication with the PC.
Strips of female pins. Extension for easy connection management.

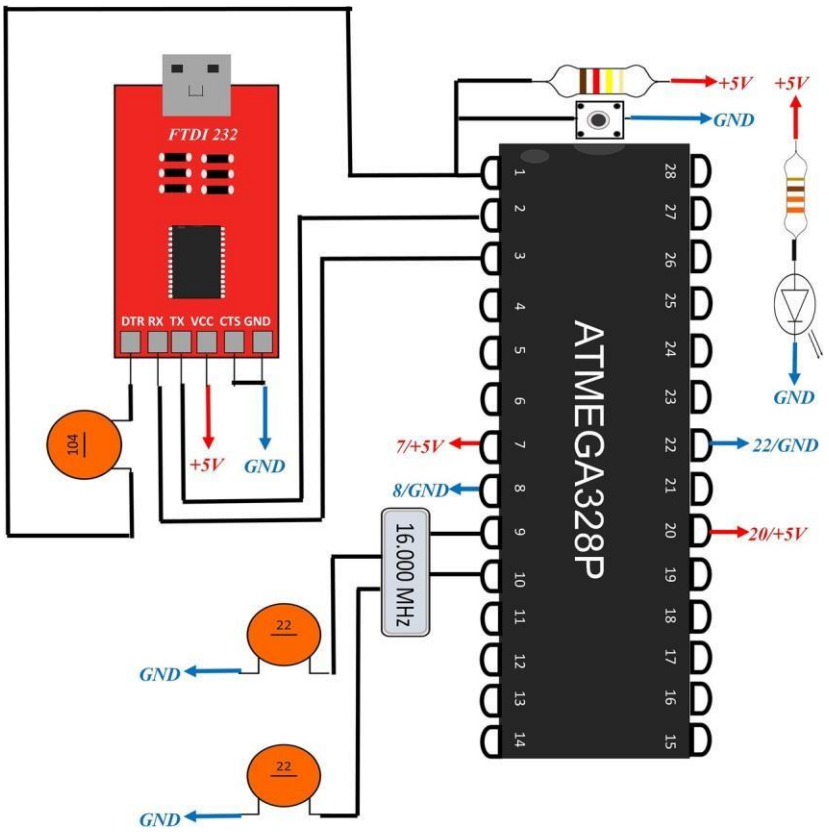


Fig. 1 Electronic diagram of the data acquisition board with ATmega328p microcontroller

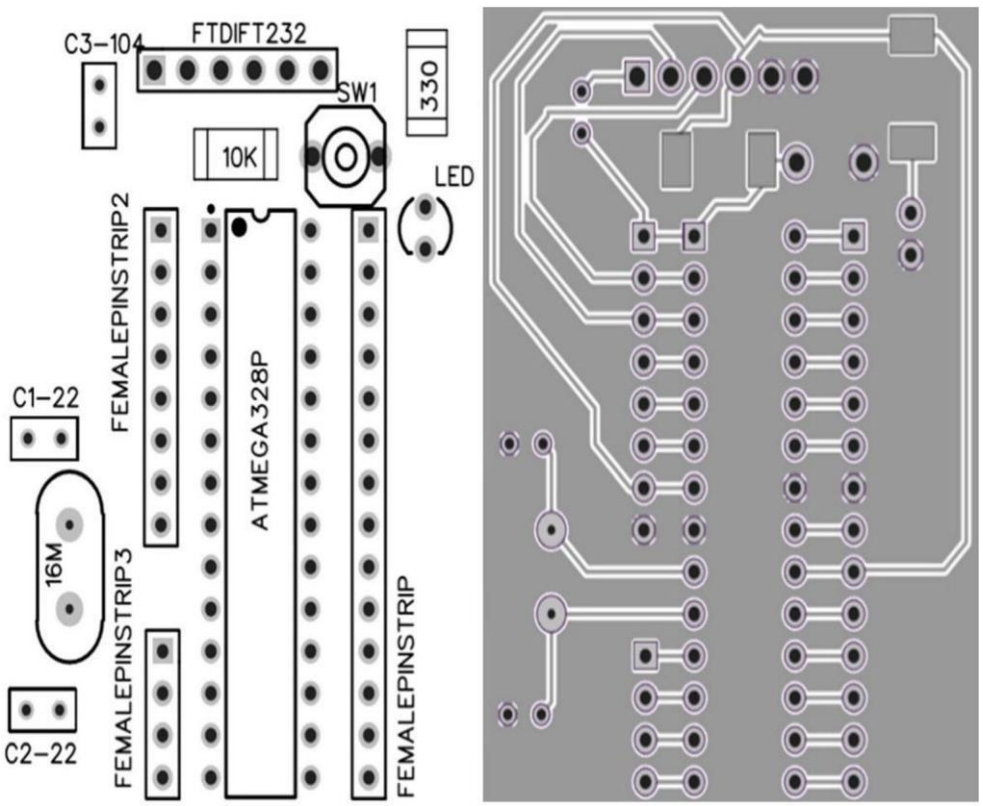


Fig. 2 DAB circuit designed with free-use EasyEDA online software, with ATmega328p microcontroller

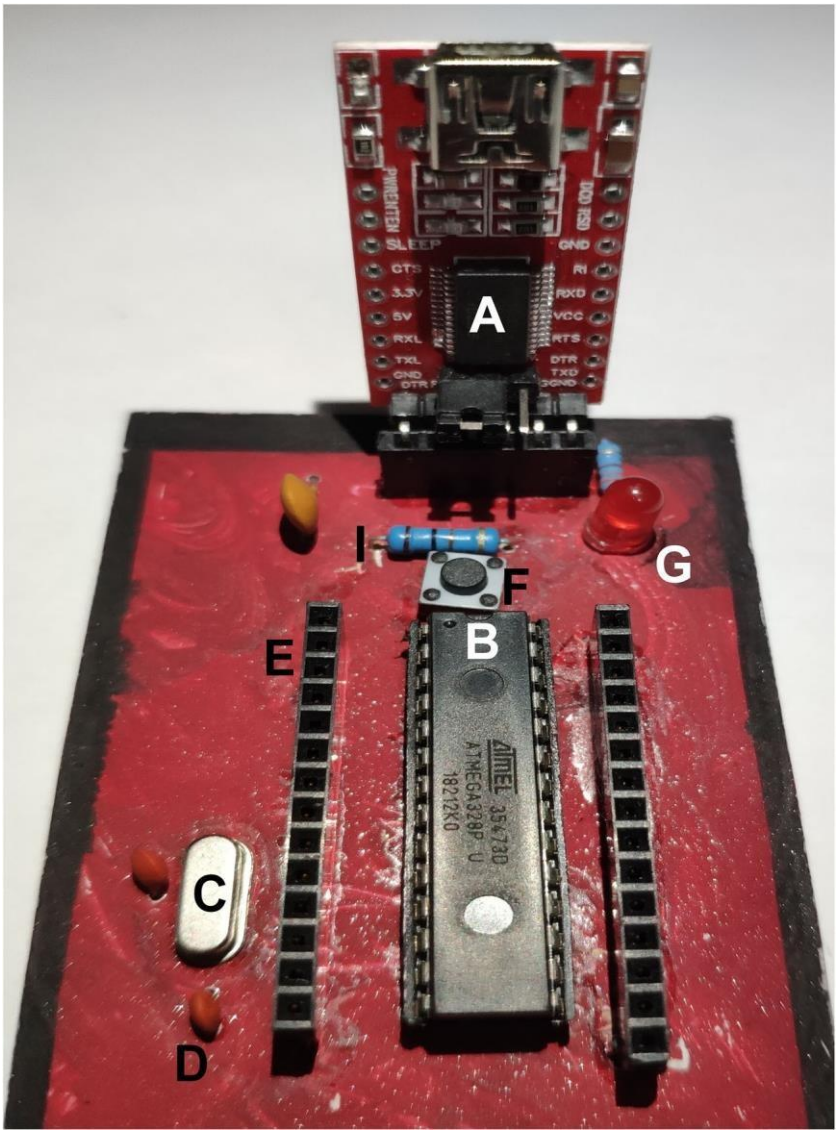


Fig. 3 Process of assembling components on the data acquisition board. (a) FTDI FT232, (b) ATmega328P-PU microcontroller, (c) 16 MHz Oscillator crystal, (d) Ceramic condenser, (e) Strips of female pins, (f) Pushbutton and (g) Light emitting diode

Construction

For the construction of the DAB (PCB board) a Phenolic plate is used. In it, the ferric chloride is poured to trace the circuit in copper. While the circuit design is printed on a coated type sheet. Then, the method of "ironing" is used; heat transfer to the design of the circuit of the coated type sheet is impregnated to the Phenolic plate. As a result of the "ironing" method, the Phenolic plate is obtained with the circuit impregnated in ink, from the coated type sheet.

In the next step, the phenolic plate is placed in a container to expose it to ferric chloride and trace the circuit on copper lines. The next step is to drill the PCB board to

Fig. 4 Creating a Project in Visual Studio 2019

Fig. 5 (a) Work area and (b) Elements to turn off /on LED

assemble the components, using as a tool a mini hand drill with a bit like the pins of the circuit components. In the assembly of components, tools such as 40w soldering iron, tin solder and paste are used. After component assembly, short-circuit and continuity tests are carried out before the first ignition. The final construction result of the first prototype of the data acquisition board is concluded with the first power-up (Fig. 3).

Control Software

The purpose of this section is to show the basic characteristics of the software used, in subsequent sections the procedure for developing a graphical interface implemented for three basic applications is shown in detail. The free software used is Visual Basic, hosted in Visual Studio 2019. This software is available in different operating

Fig. 6 (a) Initialization of the elements, (b) Status configuration

a)

```
private void TURNON_Click(object sender, EventArgs e)
{
    flag = 1;
    TurnON.Enabled = false;
    TurnOFF.Enabled = true;
    LED.Enabled = true;
}

0 references
private void TURNOFF_Click(object sender, EventArgs e)
{
    flag = 2;
    TurnON.Enabled = true;
    TurnOFF.Enabled = false;
    LED.Enabled = false;
    serialPort1.Write("0");
    LED.Checked = false;
    label7.Text = "";
```

b)

```
7   public partial class Form1 : Form
8   {
9       int flag = 0;
10
        1 reference
11      public Form1()
12      {
13          InitializeComponent();
14          TurnOFF.Enabled = false;
15          LED.Enabled = false;
16          try{ serialPort1.Open(); }
17
18          catch(Exception msg) { MessageBox.Show(msg.ToString()); }
19      }
```

systems, has a simple programming language, a high range of features in the design options and a friendly interface.

Another important advantage of this software is that it generates the programming code automatically by adding graphic elements in the user interface and finally offers multiple files and templates to create different projects or applications. For our case, we used the template Windows Form Application (Fig. 4), this template allows you to create an executable file (.exe).

The Arduino IDE code is for programming the Atmega328P microcontroller. The interface designed in Visual Studio is for the user to send a command to the microcontroller. The communication between both programs is the serial port, and both programs work together as follows. First, the user sends a command to the serial port (from the Visual Studio interface). Later, then the microcontroller takes the value in the serial port and executes the corresponding instruction.

Results

Implementation of DAB and Control Software in Three Different Applications

The following section deals with three different applications in detail: (1) Control of turning off/on an LED, (2) Temperature reading with plotted values and (3) Control of a servomotor. For each application, the necessary steps to develop the graphical interface, the programming code for the microcontroller and the associated electronic scheme are shown. Finally, the combination of three previous applications can produce more sophisticated devices. For example, a Ph

```
sketch_aug22a §
int led=10;
int dato;

voidsetup() {
pinMode(led,OUTPUT);
Serial.begin(9600);}

voidloop() {
dato=Serial.read();
if (dato == '1') {
digitalWrite(led, HIGH);}
if (dato == '0') {
digitalWrite(led, LOW);}
}
```

Fig. 7 Arduino code for LED control

Meter with temperature sensor and stirring setup. The main result of this work is to show in a simple and step-by-step way how the combination between a DAB and the user-friendly interface is possible to potentiate different applications to solve particular equipment needs.

In addition, this work seeks to be a basis for any user without deep knowledge of electronics and programming to assemble their own experimental equipment and solve their needs, avoiding acquiring high-cost experimental equipment.

LED Controlled by Visual Studio

The first step is to create the project in the work area (Fig. 5(a)), adding the toolbox (Fig. 5(b)). In this case, a graphical interface was made for serial communication with the DAB to control the on/off from a LED.

The first step will be to declare the variable “flag” (Fig. 6(a)), with the stages of the program. In addition, the attributes of each element of the graphical interface and the enablement of the serial port will be defined.

Then the conditions to be performed are programmed, depending on the state of the “flag” variable, with a value of 1 the “Turn ON” element will be disabled, and the other elements enabled. When the “flag” variable has a value of 2, the opposite will happen and a “0” will also be written on the serial port (Fig. 6(b)).

LED Controlled by Arduino

The Arduino IDE is a programming environment with the characteristics of code editor, compiler and debugger. To create the project in the Arduino IDE it is necessary to select the “File” tab and press the “New” option. The code used to control the LED (Fig. 7) is based on reading information from the serial port that is sent from the graphical interface.

Once you have the code in the IDE, you proceed to compile it by pressing the “verify” tab, then the COM connection port is checked in the “tools” and “port” tab, with the program loading as the last step to the data acquisition board in the “upload” tab.

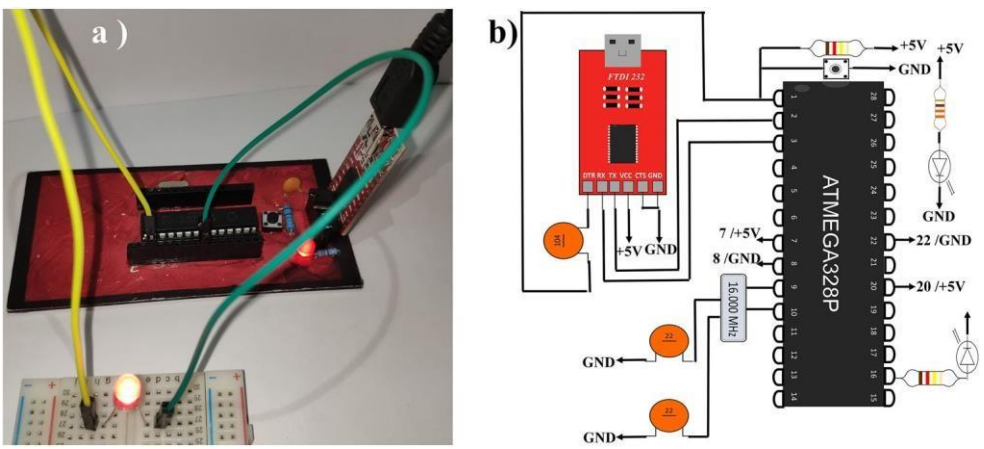


Fig. 8 (a) Electronic configuration to control a led. b) Schematic diagram

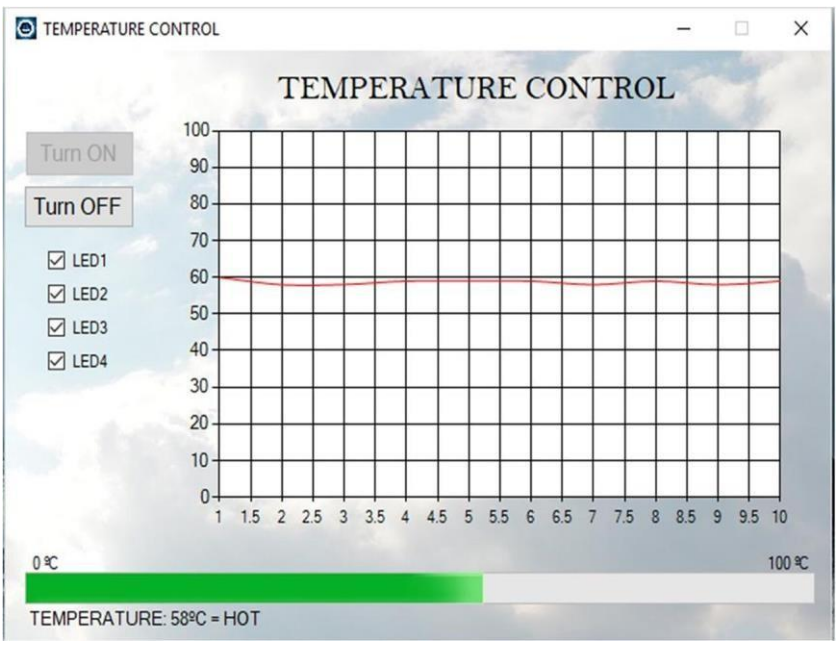
Fig. 9 (a) Programming of each "Tool" (LED), (b) Programming the "Timer" element (c) Programming of the "Progress Bar" element

a)
```
private void LED1_CheckedChanged(object sender, EventArgs e)
{
    if (flag == 1)
    {
        if (LED1.Checked == true)
        {
            serialPort1.Write("1");
        }
        else
        {
            serialPort1.Write("2");
        }
```

b)
```
private void Timer1_Tick(object sender, EventArgs e)
{
    if (flag == 1)
    {
        serialPort1.Write("9");
        time++;
        if (time == 10)
        {
            time = 0;
            chart1.Series[0].Points.Clear();
        }
        chart1.Series["TEMPERATURE"].Points.AddY(temperatureProBar);
    }
```

c)
```
private void ProgressBar1_Click(object sender, EventArgs e)
{
    progressBar1.Value = temperatureProBar;
    if (temperatureProBar <= 20)
    {
        label1.Text = "TEMPERATURE: "+ temperatureProBar.ToString()+ "ºC = COLD";
    }
    if (temperatureProBar > 20 && temperatureProBar < 50)
    {
        label1.Text = "TEMPERATURE: " + temperatureProBar.ToString() + "ºC = MILD";
    }
    if (temperatureProBar >= 50)
    {
        label1.Text = "TEMPERATURE: " + temperatureProBar.ToString() + "ºC = HOT";
    }
```

Fig. 10 Temperature recording interface running



```
int LM35; int dato; int led=9; int led1=10; int led2=11; int led3=12; float TEMP;
void setup() { Serial.begin(9600);}
void loop() {
dato=Serial.read();
if (dato == '1') {digitalWrite(led,HIGH);}
if (dato == '2') {digitalWrite(led,LOW);}
if (dato == '3') {digitalWrite(led1,HIGH);}
if (dato == '4') {digitalWrite(led1,LOW);}
if (dato == '5') {digitalWrite(led2,HIGH);}
if (dato == '6') {digitalWrite(led2,LOW);}
if (dato == '7') {digitalWrite(led3,HIGH);}
if (dato == '8') {digitalWrite(led3,LOW);}
if (dato == '9') {LM35 = analogRead(A0);
TEMP = ((LM35 * 5000.0) / 1023) / 10;
Serial.println(TEMP, 1);}}
```

Fig. 11 Arduino code to record the temperature

LED Electronic Configuration

Once the application is programmed, it only remains to execute and check any type of error. The final electronic configuration for the LED control is shown in Fig. 8(b).

Implementation of DAB and Temperature Logging Software

Temperature registration using visual studio

The interface design for temperature recording consists of four buttons, which are programmed to send signals to the board. To know if each button is activated, a LED is associated to it; the programming of each LED is the same and is shown in Fig. 9(a). The "timer" function was used to record the temperature values, which allows to send the values to the graphical interface. For example, if the "flag" variable is declared to be "1", the value of "9" must be written to the serial port. In the application presented, there is a counter called "time", which represents the time (X axis) on the graph. While the variable "temperatureProBar" is plotting the information coming from the response of the temperature value board (Fig. 9(b)). Finally, the conditions for the "Progress bar" are programmed, Fig.

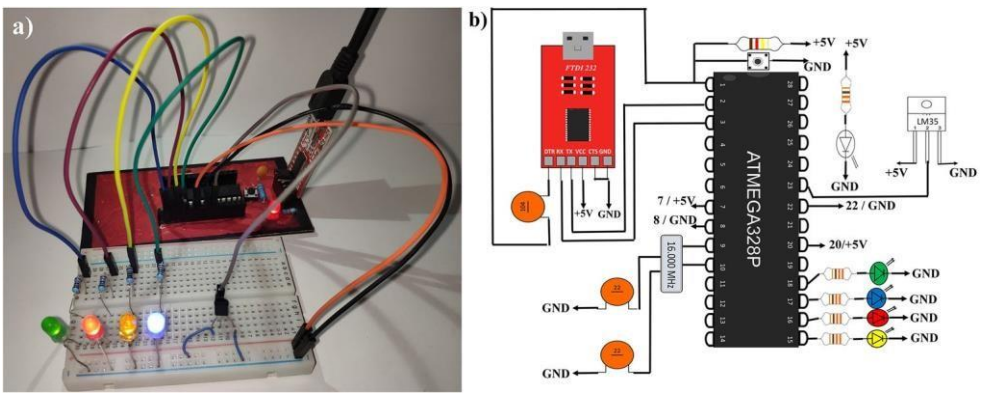


Fig. 12 (a) Electronic configuration to record the temperature. (b) Schematic diagram

Fig. 13 (a) Function to configure communication port, (b) Programming of the "TrackBar"

a)
```
private void PortProperties()
{

    Port = new SerialPort();
    Port.PortName = "COM5";
    Port.BaudRate = 9600;
    try
    {
        Port.Open();
    }
    catch(Exception e1)
    {
        MessageBox.Show(e1.Message);
    }
```

b)
```
private void TrackBar1_Scroll(object sender, EventArgs e)
{
    if(Port.IsOpen)
    {
        Port.WriteLine(Val_TrackBar.Value.ToString());
        label1.Text = "DEGREE = " + Val_TrackBar.Value.ToString();

    }
}
```

9(c), which controls the temperature increase. On the computer bar were the temperature value and an indication of "cold, warm or hot".

Finally, the software works by measuring the temperature shown in the "Graph" and "Programming bar" and has four "selection boxes", which can be assigned to different tasks

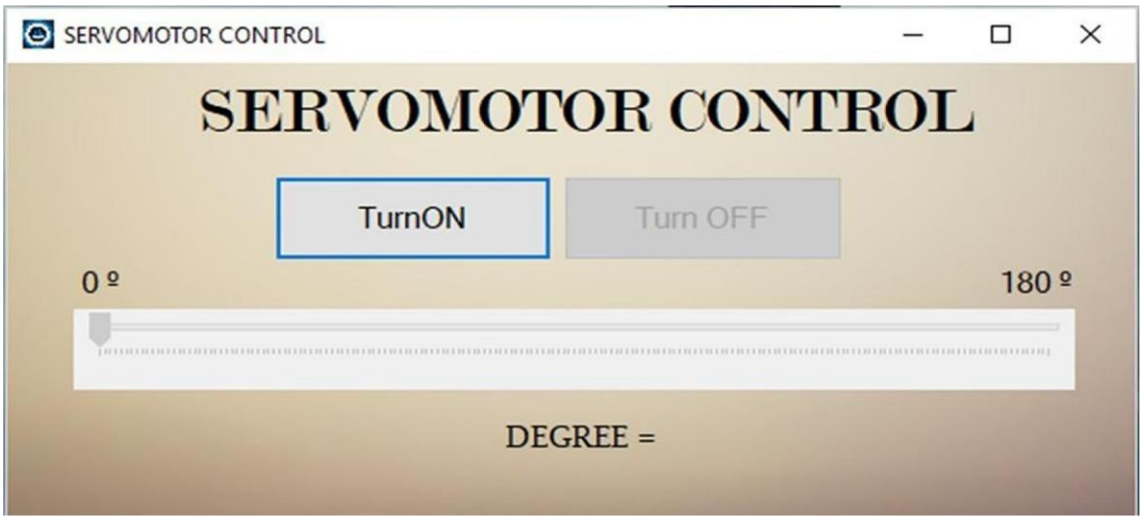


Fig. 14 Servomotor control interface

Fig. 15 Arduino programming code for servomotor control

```
Servomotor
#include <Servo.h>
 Servo myservo;
 int val;
void setup() {
 Serial.begin(9600);
myservo.attach(10);}
void loop() {}
void serialEvent() {
val = Serial.parseInt();
if (val!=0){ myservo.write(val);}}
```

that the user may need, in this case they represent with "LED" (Fig. 10).

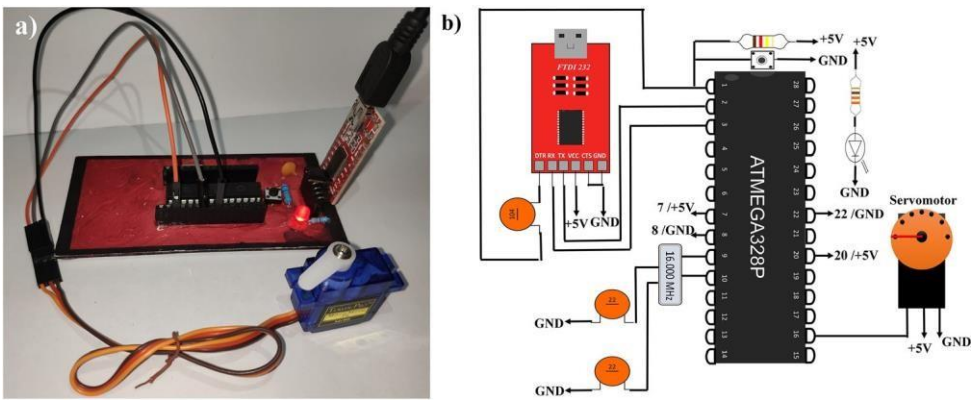
Temperature record using Arduino

The Arduino programming code is shown in Fig. 11. This consists of identifying the value of the serial port and assigning it to the variable "data". After identifying this value, the program carried out in Visual Studio registers it and graphs it in the interface.

Electronic temperature recording configuration

The conversion of the physical phenomenon of temperature to an electrical signal and then adapting it and reading it by means of the DAB is necessary to use the LM35 sensor.

Fig. 16 (a) Electronic configuration to control a servomotor. (b) Schematic diagram



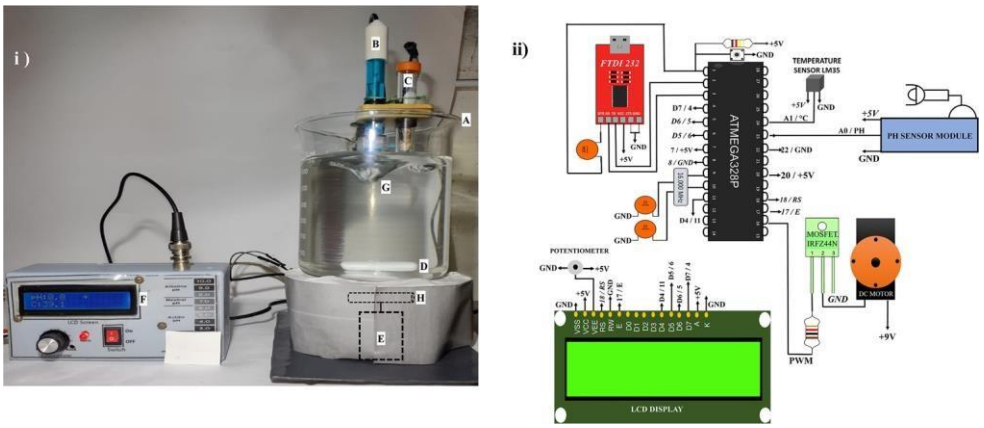


Fig. 17 Ph Meter with temperature sensor and stirring system: ii) Electronic diagram. i) Homemade system: (a) Beaker, (b) PH Sensor, (c) Temperature sensor, (d) Stir bar, (e) DC Motor and (f) LCD Display, (g) Aqueous solution, (h) Magnet

This sensor is inexpensive, accessible and uses a 5 V supply voltage. In addition, it is a 3-pin sensor, which provides an output of 10 mV for every °C. The electronic configuration is shown in Fig. 12b, and the configuration setup is shown in Fig. 12a.

Implementation of DAB and Servomotor Control Software

Servomotor controlled by visual studio

For this application, the position of a servomotor is required to be controlled. Basically the operation is through serial communication, being able to send the corresponding values to the position desired by the user. A servomotor is controlled by pulse width modulation (PWM), and depending on the voltage level, it moves position, commonly by 180° movements.

For the graphic interface design, first a function is created that will allow us to configure the serial port for communication, this time it will be done by configuring the desired port by programming (Fig. 13(a)). Once the communication is done, simply program the item called "TrackBar", which will work only if the port is "open", writing the value provided by the "TrackBar" and showing the value in the "Label" (Fig. 13(b)).

Once this is done, just run the software and move the pointer through the bar (Fig. 14)

Table 2 Cost of electronic components

Electronic component	Cost [US$]
Ph Sensor	54
Temperature sensor	1
Stir bar	3
DC Motor	3
LCD Display	4
1000Ω and 330Ω (1/4 W) resistance, 0.1uF ceramic condenser, Two ceramic capacitors of 22 pF, 16 MHz oscillator crystal and LED.	2
Atmega328P	3
Mosfet IRFZ44N	2
Potentiometer	1
FTDI 232	2
Total	75

Servomotor controlled by Arduino

The programming code for Arduino is shown in Fig. 15. In this code the library "<Servo.h > "is added.

Electronic configuration to control servomotor

Finally, the electronic configuration to control the position of a servomotor is shown in Fig. 16(b).

Implementation of a Ph Meter with Temperature Sensor and Stirring Setup

Finally, this last section presents a novel homemade Ph Meter with sensing temperature and stirring setup. This homemade setup has a low-cost, and it is very easy to make by teachers, researchers and students. Also, the above examples serve as a basis for teachers to design new effective didactic setups, and to avoid paying high costs for experimental equipment. The electronic diagram is presented in Fig. 17(a), and the final setup is presented in Fig. 17(b).

The cost of each electronic component is presented in Table 2. In addition, the total cost of this device is around 75 dollars, with a lower value than commercial equipment.

Finally, the measurement reliability of the proposed device was checked, its response was compared with a commercial pH meter (Table 3). Therefore, the performance of the device is adequate, with an error of less than 5%.

Table 3 pH response compared with a commercial pH meter

pH [a. u.]	pH Comercial	Error [%]	Sensor	Eror [%]
3	3.06	1.70	3.11	2.59
9	8.83	1.45	8.72	2.7

Conclusions

In this work, the design and implementation of a low-cost data acquisition board was managed in a clear and concrete way. Also, through three simple applications, the basic principles were made known so that someone without deep knowledge in electronics or programming can develop their own low-cost experimental equipment. In each application, the graphic interface design for the control of different actuators is approached in parallel. Finally, in this work was presented a novel homemade Ph Meter with sensing temperature and stirring system. This homemade experimental equipment has a low-cost and is very easy to make by teachers, researchers and students. Also, the above examples serve as a basis for teachers to design new effective didactic systems, and to avoid paying high costs for experimental equipment.

Acknowledgements The authors gratefully acknowledge the scholarships for this work from CONACYT.

Declarations

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Damcı E, Şekerci C (2019) Development of a low-cost single-axis shake table based on Arduino. Exp Tech 43:179–198. https://doi.org/10.1007/s40799-018-0287-5
2. Cave A, Roslyakov S, Iskander M, Bless S (2016) Design and performance of a laboratory pneumatic gun for soil ballistic applications. Exp Tech 40:541–553. https://doi.org/10.1007/s40799-016-0055-3
3. Hercog D, Gergič B (2014) A flexible microcontroller-based data acquisition device. Sensors. 14:9755–9775. https://doi.org/10.3390/s140609755
4. Oates M, Ruiz-Canales M, Ferrández-Villena M, Fernández-López A (2017) A low cost sunlight analyser and data logger measuring radiation. Comput Electron Agric 143:38–48. https://doi.org/10.1016/j.compag.2017.09.024
5. Bajer L, Krejcar O (2015) Design and realization of low cost control for green house environment with remote control. IFAC-Papers On Line 48(4):368–373. https://doi.org/10.1016/j.ifacol.2015.07.062
6. Schubert T, D'Ausilio A, Canto R (2013) Using Arduino microcontroller boards to measure response latencies. Behav Res Methods 45:1332–1346. https://doi.org/10.3758/s13428-013-0336-z
7. Kim K-W, Lee M-S, Ryu M-H, Kim J-W (2016) Arduino-based automation of a DNA extraction system. Technol Health Care 24:105–112. https://doi.org/10.3233/THC-151048
8. Devarakonda K, Nguyen KP, Kravitz AV (2016) ROBucket: a low cost operant chamber based on the Arduino microcontroller. Behav Res Methods 48:503–509. https://doi.org/10.3758/s13428-015-0603-2
9. Guzmán C, Carrera J, Durán H, Berumen J, Ortiz A, Guirette O, Arroyo A, Brizuela J (2019) Implementation of virtual sensors for monitoring temperature in greenhouses using CFD and control. Sensors 19(1):60. https://doi.org/10.3390/s19010060
10. Bueno-Hernández D, Rupesh K, Roberto M, Marty JL (2017) Low cost optical device for detection of fluorescence from Ochratoxin A using a CMOS sensor. Sensors Actuators B 246:606–614. https://doi.org/10.1016/j.snb.2017.02.097
11. Wen-Hsuan K, Chi-Hung T, Sufen C, Ching-Chang W (2016) Development of a computer-assisted instrumentation curriculum for physics students: using LabVIEW and Arduino Platform. J Sci Educ Technol 25:427. https://doi.org/10.1007/s10956-016-9603-y
12. Durán-Muñoz H, Hernández-Ortiz M, Sifuentes-Gallardo C, Galván-Tejeda I, Sánchez-Zeferino R, Castaño-Meneses V (2018) Comparative study of kinetic parameters induced by different excitation sources: using a novel and user-friendly glow curve deconvolution spreadsheet. J Mater Sci Mater Electron 29:15732–15740. https://doi.org/10.1007/s10854-018-9226-6
13. Zhou Y (2016) Mathematical modeling of chain drive geometries for a durability test rig. Exp Tech 40:1137–1146. https://doi.org/10.1007/s40799-016-0108-7
14. Poppe L, Eliason H, Hastings M (2004) A visual basic program to generate sediment grain-size statistics and to extrapolate particle distributions. Comput Geosci 30:791–795. https://doi.org/10.1016/j.cageo.2004.05.005
15. Irie M, Terada T, Katsura T, Matsuoka S, Inui K-i (2005) Computational modelling of H+-coupled peptide transport via human PEPT1. J Physiol 565:429–439. https://doi.org/10.1113/jphysiol.2005.084582