UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA EN GESTIÓN AMBIENTAL

Técnicas de Machine Learning para la predicción del caudal efluente de la represa Condoroma

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para obtener el título profesional de Ingeniero en Gestión Ambiental

AUTORES

Encina Dávila, Astrid Floria Milagritos

Pacheco Hinojoza, Mirella Alejandra

Vargas Martell, Vannia Giovana

ASESOR

Fabian Arteaga, Junior John

ORCID Nº 0000-0001-9804-7795

Marzo, 2023

Informe Final TSP

INFORME DE ORIGINALIDAD

22%	21%	4%	9%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.esan.edu.pe Fuente de Internet	4%
2	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Trabajo del estudiante	2%
3	hdl.handle.net Fuente de Internet	1%
4	frogames.es Fuente de Internet	1%
5	docplayer.es Fuente de Internet	1%
6	pirhua.udep.edu.pe Fuente de Internet	1%
7	www.juntadeandalucia.es Fuente de Internet	1%
8	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	1%

RESUMEN

Distintos estudios están empleando técnicas de Machine Learning para el análisis de datos para hallar comportamientos que posibiliten crear modelos matemáticos predictivos y pronosticar diversas variables de salida. En este sentido, el presente trabajo de investigación se enfoca en los esfuerzos realizados para predecir el caudal efluente en la represa Condoroma, perteneciente a la Autoridad Autónoma de Majes (Autodema), donde se incluye el uso de técnicas de aprendizaje supervisado. Para ello, se utiliza una base de datos abiertos de dos plataformas de Autodema: Movimiento Hídrico Sistema Colca y Meteorología Represas. Estos datos históricos son resultados de mediciones mensuales del sistema de monitoreo del recurso hídrico. Además, se manejan para entrenar los modelos Regresión Lineal, Regresión de Vectores de Soporte (SVR) y ARIMA; asimismo, se utilizaron métricas como el MAE, MSE, RMSE y R^2 para medir el modelo con el mejor rendimiento. Con base en los resultados obtenidos, se determinó que para predecir el caudal efluente de la represa Condoroma la mejor técnica fue la de SVR que obtuvo un MAE de 5.536, un MSE de 83.701, un RMSE de 9.145 y una varianza igual a 0.427.

Palabras clave: Machine Learning, aprendizaje supervisado, regresión lineal, SVR, ARIMA

ABSTRACT

Different studies are using Machine Learning techniques for data analysis to find behaviors that make it possible to create predictive mathematical models and forecast divergent output variables. In this sense, the present research work focuses on the efforts made to predict the effluent flow in the Condoroma Dam, belonging to the Autonomous Authority of Majes (Autodema). In this case we are going to use supervised learning techniques. For this purpose, a set of open data from two Autodema platforms is used; Colca System Water Movement and Meteorology Dams. These historical data are the results of monthly measurements of the water resource monitoring system. Furthermore, Linear Regression, Support Vector Regression (SVR) and ARIMA models are used to train, and metrics such as MAE, MSE, RMSE and R^2 were used to measure the model with the best performance. Based on the results obtained, it was determined that the advantageous technique for predicting the effluent flow of the Condoroma Dam was the SVR technique, which obtained an MAE of 5.536, an MSE of 83.701, an RMSE of 9.145 and a variance equal to 0.427.

Key words: Machine learning, supervised learning, linear regression, SVR, ARIMA.

ÍNDICE DE CONTENIDOS

RESUMEN 2
ABSTRACT 3
ÍNDICE DE CONTENIDOS 4
INTRODUCCIÓN 10
Capítulo I: Planteamiento del problema 11
1.1 Descripción de la Realidad Problemática 11
1.2 Justificación de la Investigación 15
1.2.1 Justificación Teórica 15
1.2.2 Justificación Práctica 15
1.2.3 Justificación Metodológica 15
1.3 Delimitación de la Investigación 16
1.3.1 Espacial 16
1.3.2 Temporal 16
1.3.3 Conceptual 16
Capítulo II: Marco Teórico 17
2.1 Antecedentes de la Investigación 17
2.2 Bases Teóricas 32
2.2.1 Machine Learning 32
2.2.2 Aprendizaje Supervisado 33
2.2.2.1 Algoritmos de aprendizaje supervisado 34
2.2.4 Represas 44
Capítulo III: Entorno Empresarial 49
3.1 Descripción de la empresa 49
3.1.1 Reseña histórica y actividad económica 49
3.1.2 Descripción de la organización 49
3.1.3 Datos generales estratégicos de la empresa 52
3.2 Modelo de negocio actual (CANVAS) 57
3.3 Mapa de procesos actual 58
Capítulo IV: Metodología de la Investigación 59
4.1 Diseño de la Investigación 59
4.1.1 Enfoque de la investigación 59

4.1.2 Alcance de la investigación 59
4.1.3 Tipo de investigación 59
4.1.4 Población y muestra 60
4.2 Metodología de implementación de la solución 61
4.2.1 Adquisición de datos 62
4.2.2 Preparación 62
4.2.3 Análisis 62
4.2.4 Reporte 62
4.3 Metodología para la medición de resultados de la implementación 62
4.4 Cronograma de actividades y presupuesto 64
4.4.1 Cronograma 64
4.4.2 Presupuesto 65
Capítulo V: Desarrollo de la Solución 66
5.1 Propuesta solución 66
5.1.1 Planteamiento y descripción de Actividades 66
5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución 66
5.2 Medición de la solución 79
5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo. 80
5.2.2 Simulación de solución. Aplicación de Software 86
Capítulo VI: Conclusiones y recomendaciones 107
6.1 Conclusiones 107
6.2 Recomendaciones 108
Referencias bibliográficas 109

ÍNDICE DE TABLAS

Tabla 1: Mejores resultados de los modelos SVM y ANN para la predicción del flujo de entrada del embalse de la represa de Zayandehroud....24
Tabla 2: Casos especiales de los modelos ARIMA....42
Tabla 3: Matriz de Efectos Internos (EFI)....54
Tabla 4: Matriz de Efectos Externos (EFE)....54
Tabla 5: Presupuesto de Actividades....65
Tabla 6: Pruebas aplicadas para Regresión Lineal y SVR....78
Tabla 7: Métricas estadísticas para Regresión Lineal....80
Tabla 8: Métricas estadísticas para Regresión de Vectores de Soporte....82
Tabla 9: Métricas estadísticas - Técnica ARIMA....84
Tabla 10: Resumen de métricas estadísticas para el pronóstico del caudal efluente....86

ÍNDICE DE FIGURAS

Figura 1: Demanda Hídrica, Escenario 2020-2050.....11
Figura 2: Crecimiento mundial de la población y el volumen de almacenamiento de los embalses.....12
Figura 3: Precipitación en la represa Condoroma (2009-2022).....14
Figura 4: Caudal de la represa Condoroma (2009-2022).....14
Figura 5: Precipitación anual y caudal medio anual de entrada de la represa durante 1980-2019 de la represa del río Soyang.....20
Figura 6: Variación de la serie de tiempo de la entrada de la represa y datos meteorológicos.....20
Figura 7: Descripción de los modelos de aprendizaje automático.....21
Figura 8: Precipitaciones mensuales de las estaciones hidrométricas de Ghaleh-Shahrokh durante un período de 44 años (1971 a 2015).....22
Figura 9: Entrada mensual en el embalse de la represa de Zayandehroud durante un período de 44 años (1971 a 2015).....23
Figura 10: Caudal diario de la Estación Ardilla periodo de 66 años (1951 a 2017).....25
Figura 11: Diagrama de flujo de la Metodología.....29
Figura 12: Aprendizaje supervisado para un algoritmo de regresión y un algoritmo de clasificación.....34
Figura 13: Algoritmo KNN.....35
Figura 14: Algoritmo de regresión.....36
Figura 15: Representación gráfica de modelos lineales de ε--SVR.....38
Figura 16: Representación gráfica de modelos no lineales SVR.....39
Figura 17: Desarrollo de un algoritmo de árbol de decisión.....43
Figura 18: Estructura del modelo de red neuronal.....44
Figura 19: Modelo de pérdida de evaporación - alta temperatura.....45
Figura 20: Modelo de pérdida de evaporación - baja temperatura.....46
Figura 21: Organigrama Autoridad Autónoma de Majes (Autodema).....50
Figura 22: Cadena de Suministro de Autodema.....51
Figura 23: Situación interna y externa de Autodema.....55
Figura 24: CANVAS Autodema.....57
Figura 25: Mapa de Procesos de Autodema.....58
Figura 26: Diseño de la implementación.....61
Figura 27: Cronograma de actividades.....64
Figura 28: Plataforma Movimiento Hídrico Cuenca Regulado.....66
Figura 29: Plataforma Meteorología Represas.....67
Figura 30: Recopilación de datos descargados en una hoja de cálculo.....67
Figura 31: Análisis de datos faltantes.....68
Figura 32: Caudal afluente diario (m^3/s).....69
Figura 33: Caudal afluente promedio mensual (m^3/s).....69
Figura 34: Nivel de embalse diario (m.s.n.m).....70
Figura 35: Nivel de embalse promedio mensual (m.s.n.m).....70
Figura 36: Volumen útil diario (m^3).....71

Figura 37: Volumen útil promedio mensual (m^3) 71
Figura 38: Pérdidas por evaporación diario (m^3) 72
Figura 39: Pérdidas por evaporación promedio mensual (m^3) 72
Figura 40: Evaporación diaria (m.m) 73
Figura 41: Evaporación promedio mensual (m.m) 73
Figura 42: Precipitación diaria (m.m) 74
Figura 43: Precipitación acumulada mensual (m.m) 74
Figura 44: Temperatura mínima diaria (°C) 75
Figura 45: Temperatura mínima promedio mensual (°C) 75
Figura 46: Temperatura máxima diaria (°C) 76
Figura 47: Temperatura máxima promedio mensual (°C) 76
Figura 48: Plataforma Anaconda Navigator 87
Figura 49: Interfaz de Jupyter Notebook 6.4.12 87
Figura 50: Importación de Librerías pandas y matplotlib.pyplot 87
Figura 51: Lectura del dataset - Regresión Lineal y SVR 88
Figura 52: Identificación de variables numéricas 88
Figura 53: Descripción estadística y verificación de data 89
Figura 54: Designación de columna Fecha como índice 89
Figura 55: Transformación de datos diarios a mensuales 89
Figura 56: Función para graficar variables 89
Figura 57: Separación de variables - Regresión Lineal No Normalizada 90
Figura 58: Asignación de variables X - Regresión Lineal No Normalizada 90
Figura 59: Asignación de variable Y - Regresión Lineal No Normalizada 91
Figura 60: Datos train y test - Regresión Lineal No Normalizada 91
Figura 61: Modelamiento de la Regresión Lineal No Normalizada 91
Figura 62: Métricas - Regresión Lineal No Normalizada 92
Figura 63: Coeficientes - Regresión Lineal No Normalizada 92
Figura 64: Función StandarScarler - Regresión lineal 92
Figura 65: Asignación de variables X - Regresión Lineal Normalizada 93
Figura 66: Normalización de datos - Regresión Lineal 93
Figura 67: Datos train y test - Regresión Lineal Normalizada 93
Figura 68: Modelamiento de la Regresión Lineal Normalizada 93
Figura 69: Métricas - Regresión Lineal Normalizada 94
Figura 70: Separación de variables - SVR No Normalizada 94
Figura 71: Asignación de variables X - SVR No Normalizada 95
Figura 72: Asignación de variable Y - SVR No Normalizada 95
Figura 73: Datos train y test - SVR No Normalizada 95
Figura 74: Modelamiento de SVR No Normalizada 96
Figura 75: Métricas - SVR No Normalizada 96
Figura 76: Función StandarScarler - SVR 97
Figura 77: Asignación de variables X - SVR Normalizada 97
Figura 78: Normalización de datos - SVR 97
Figura 79: Datos train y test - SVR Normalizada 98
Figura 80: Modelamiento SVR Normalizada 98

Figura 81: Métricas - SVR Normalizada 98
Figura 82: Instalación de librería pmdarima 99
Figura 83: Librerías - Técnica ARIMA 99
Figura 84: Lectura del dataset - Técnica ARIMA 99
Figura 85: Fecha como índice 100
Figura 86: Gráfico de Caudal Efluente 100
Figura 87: Descomposición aditiva 101
Figura 88: Prueba de Dickey-Fuller 101
Figura 89: División de datos en entrenamiento y prueba 102
Figura 90: Muestra gráfica de los datos de entrenamiento y prueba 102
Figura 91: ARIMA (1,1,1) 103
Figura 92: ARIMA (2,0,2) 103
Figura 93: Auto ARIMA 104
Figura 94: ARIMA (3,1,2) 104
Figura 95: Métricas - ARIMA (1,1,1) 105
Figura 96: Métricas - ARIMA (2,0,2) 105
Figura 97: Métricas - ARIMA (3,1,2) 106

INTRODUCCIÓN

El presente Trabajo de Suficiencia Profesional “Técnicas de Machine Learning para la predicción del caudal efluente de la represa Condoroma” tiene como finalidad proponer un modelo de predicción de caudales efluentes para la represa Condoroma que permita dar soporte en la gestión de riesgos de Autodema y las demás instituciones involucradas.

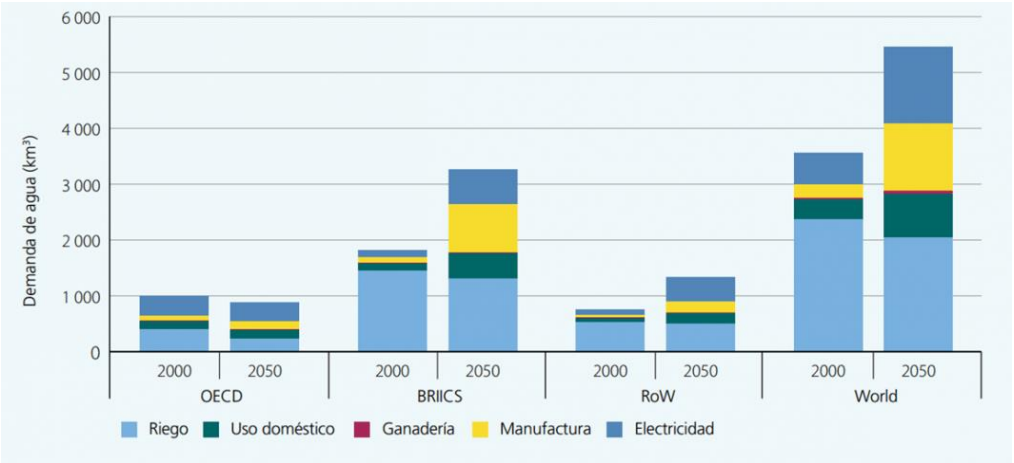
El trabajo se divide en seis capítulos. Inicia con el Capítulo I en el que se detalla la realidad problemática, puesto que se sabe que el agua es fundamental para la vida y para cubrir las necesidades del ser humano; sin embargo, es uno de los recursos cuya disponibilidad viene siendo amenazada, por lo que su gestión resulta importante y necesaria. En comparación con otras décadas, hoy encontramos cada vez menos disponibilidad hídrica en la superficie. Por tanto, las represas simbolizan un sistema que nos permite gestionar mejor este recurso; por ello este capítulo muestra datos que aportan a la problemática; además, contiene la justificación y la delimitación de la investigación. El Capítulo II, presenta los antecedentes de investigación y las bases teóricas en relación al tema de estudio; por un lado, conceptos hidrológicos, como precipitación, represas, entre otros y, por otro lado, conceptos de Machine Learning como técnicas de aprendizaje supervisado y no supervisado. Por otra parte, el Capítulo III describe el entorno empresarial, el cual comprende la descripción de la empresa y su ubicación, modelo de negocio, cadena de suministro, entre otros. Luego el Capítulo IV, explica y desarrolla la metodología empleada, la cual incluye la determinación del diseño, alcance y enfoque de la investigación. Posteriormente el Capítulo V detalla la propuesta y medición de la solución mediante la aplicación de técnicas de Machine Learning. Finalmente, el Capítulo VI, precisa las conclusiones alcanzadas en base a los hallazgos y también presenta las recomendaciones que pueden ser consideradas en futuras líneas de investigación.

Capítulo I: Planteamiento del problema

1.1 Descripción de la Realidad Problemática

Una de las principales necesidades humanas es el acceso al agua, debido a su impacto en la salud, educación y equidad de género; la gestión de recursos hídricos es clave para lograr un desarrollo sostenible. Sin embargo, la disponibilidad hídrica es limitada por naturaleza, porque si bien en todo el mundo hay agua, en muchas ocasiones, el momento, el lugar y la calidad de agua no siempre es la adecuada (Lozano-Parra, 2018). En la Figura 1 se muestra la imagen que ilustra el escenario de demanda hídrica entre el año 2000-2050 a nivel mundial.

Figura 1*: Demanda Hídrica, Escenario 2020-2050*



Nota. Obtenido de iAgua (2017)

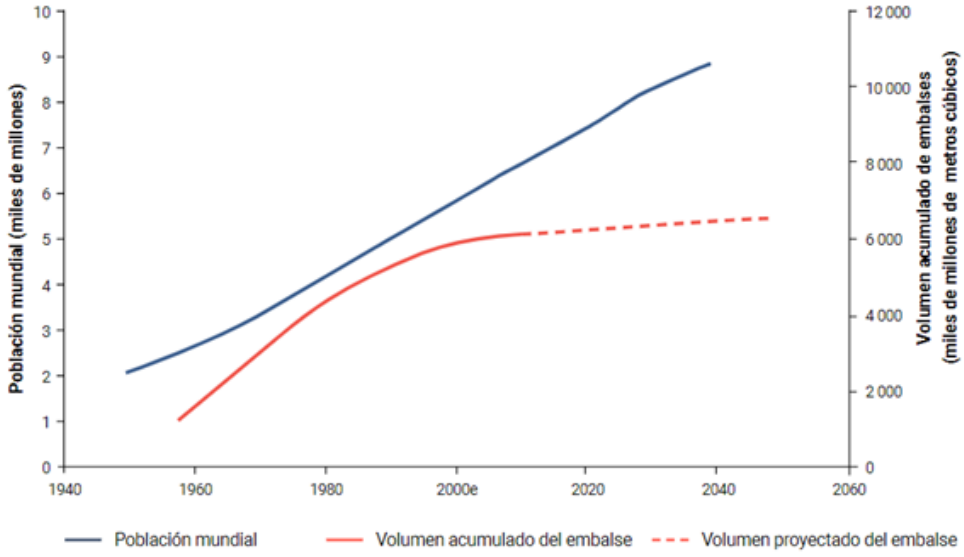
Además, las presiones antropogénicas actuales, especialmente ligadas a sectores económicos, ejercen un gran impacto sobre el agua, disminuyendo la oferta hídrica necesaria para cubrir las necesidades humanas y asegurar los caudales suficientes para mantener un ambiente equilibrado (Banco Mundial, 2014).

En ese sentido, es importante la implementación de medidas de adaptación ante este problema como, por ejemplo, la ejecución de sistemas de almacenamiento, como represas, que garanticen reservas suficientes de agua (Banco Mundial, 2014). En todo el mundo, los océanos, los suelos, los lagos, los acuíferos y la atmósfera funcionan como embalses, al igual que los que construye el hombre, como las represas. Es así que, el almacenamiento de recursos hídricos es un objetivo relevante que recae en la infraestructura hídrica, ya que su principal finalidad es hacer frente a las variaciones de disponibilidad, suministro y demanda de agua, almacenando

o reteniendo agua para ser aprovechada en diversos ámbitos como el consumo humano o el riego (WWAP UNESCO, 2021).

La capacidad de reserva de recursos hídricos per cápita de los embalses está disminuyendo a nivel mundial, debido a que el crecimiento demográfico ha avanzado más que la ampliación de la capacidad de reserva hídrica (ver Figura 2); a este hecho se suma, la disminución de capacidad de almacenamiento de los embalses existentes debido, principalmente, a la sedimentación (WWAP UNESCO, 2021).

Figura 2: *Crecimiento mundial de la población y el volumen de almacenamiento de los embalses*



Nota. Obtenido de Annandale et al. (2016)

Según el Inventario de Represas elaborado por la Autoridad Nacional del Agua (2015), hay cerca de 743 represas en Perú con condiciones adecuadas, de las cuales la mayor parte (442) tienen fines de riego; seguido de las represas de relave que suman 113. El presente inventario resalta la importancia de implantar sistemas de control y monitoreo en las represas, con el objetivo de gestionar riesgos y disminuir la probabilidad de ocurrencia.

Respecto a los riesgos asociados a la operación y mantenimiento de represas, la manera en la que este tipo de infraestructura almacena y libera el agua, tiene importantes valores. Por un lado, liberar demasiada agua puede poner en riesgo suministros de uso directo y puede tener costos a futuro. Por el contrario, no liberar la cantidad necesaria de agua puede crear pérdidas

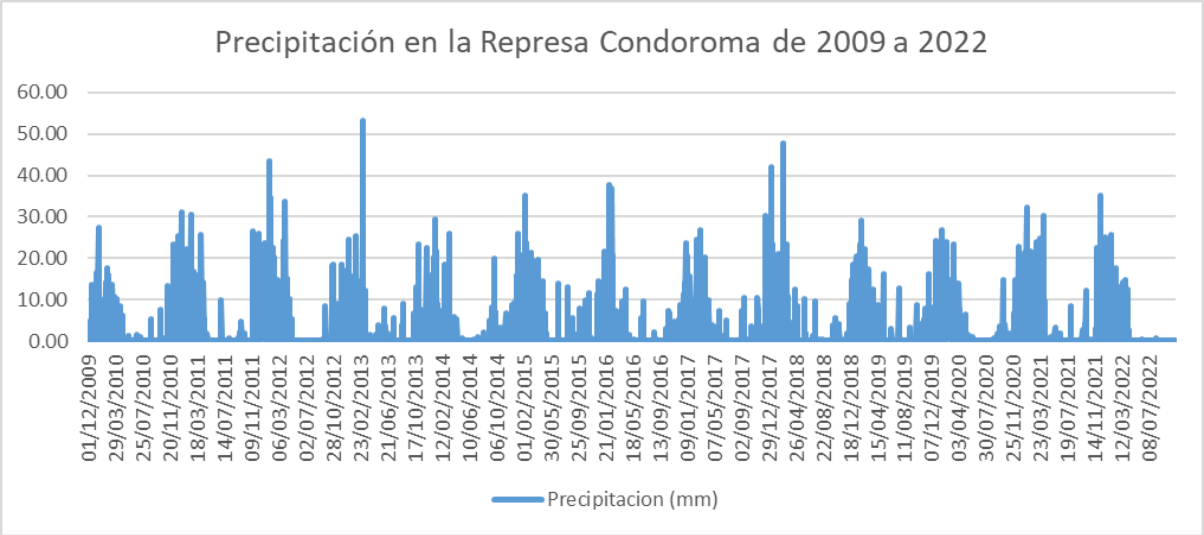
ambientales, económicas y sociales de forma inmediata. La liberación de agua de las represas tiene grandes impactos en los actores involucrados alrededor de esta (WWAP UNESCO, 2021).

Ante los inminentes riesgos de desbordamiento de represas, es importante que los involucrados en la operación y mantenimiento de estas infraestructuras hidráulicas, dispongan de datos históricos y de un registro de afecciones ocurridas, con el propósito de anticipar un nuevo episodio y sus posibles consecuencias; así como, ponerse en un contexto ya conocido. Es así que, la predicción de variables hidrológicas, por más que lleve añadida cierto nivel de incertidumbre, es una herramienta relevante en la gestión de represas (Confederación Hidrográfica del Ebro, 2021).

En ese sentido, la represa Condoroma, principal componente de la infraestructura hidráulica mayor del Sistema Colca, la cual es operada por la Autoridad Autónoma de Majes (Autodema), afronta cada año problemas en temporada de precipitaciones altas, ya que la recaudación de agua aumenta debido a la intensidad de lluvias, lo cual ocasiona desbordamientos y genera que incrementen las descargas. Es así que, en marzo de 2020, la represa Condoroma descargó agua a un 90% de su capacidad, 234.36 hectómetros cúbicos (hm^3), lo cual causó inundaciones en el distrito de Coporaque, ubicado en la provincia de Caylloma. (Orihuela, 2020). Además, en febrero de 2019, la represa Condoroma alcanzó un almacenamiento del 87% de su capacidad, es decir, 203 hm^3; ante esto el Servicio Nacional de Meteorología e Hidrología (SENAMHI) alertó que las precipitaciones iban a continuar en las partes altas y que podrían originar que la represa se llene a un 100% de su capacidad (Hancco, 2019). Ante estos escenarios Autodema realiza descargas de agua en las cuencas del Chili y Colca; sin embargo, esta práctica es efectuada cuando las represas están al borde de alcanzar su capacidad máxima, lo cual origina que el cauce de los ríos aguas abajo aumente de forma intempestiva, ocasionando inundaciones que impactan a los pobladores y cultivos de la zona.

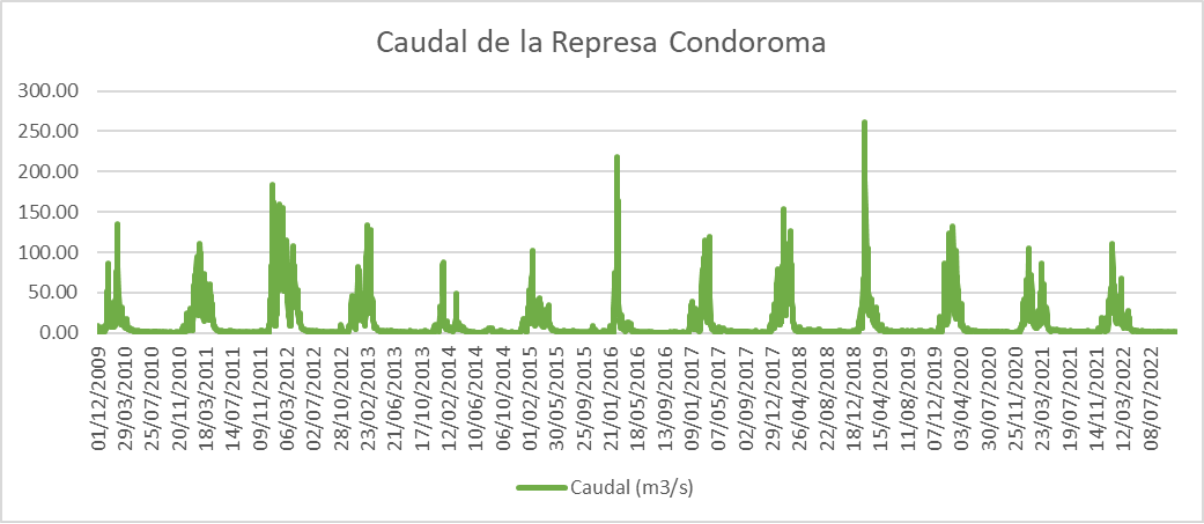
Como se muestra en la Figura 3, los meses en los que incrementan las precipitaciones en la represa Condoroma son de diciembre a marzo, lo cual coincide con la temporada de lluvias altas en la región; es decir, cada año se observa un comportamiento similar en este parámetro que tiene un impacto directo con el nivel de caudal de la represa Condoroma, el cual también incrementa en los mismos periodos (Ver Figura 4).

Figura 3: *Precipitación en la represa Condoroma (2009-2022)*



Nota. Elaboración propia utilizando los datos disponibles de la plataforma Movimiento Hídrico Sistema Colca de la Autoridad Autónoma de Majes (2023)

Figura 4: *Caudal de la represa Condoroma (2009-2022)*



Nota. Elaboración propia utilizando los datos disponibles de la plataforma Movimiento Hídrico Sistema Colca de la Autoridad Autónoma de Majes (2023)

Según la Confederación Hidrográfica del Ebro (2021), las causas del incremento del caudal en una cuenca son variadas. Por un lado, puede depender de las precipitaciones registradas en una cuenca; sin embargo, no siempre que llueve los ríos crecen, lo cual representa una incertidumbre para la ejecución de predicciones hidrológicas. Por otro lado, el caudal en una cuenca se puede incrementar debido al deshielo de la nieve; en ese sentido, las

causas mencionadas pueden ser predecidas mediante modelos hidrometeorológicos que aporten datos a futuro de variables como escorrentía, temperatura, precipitación, radiación solar, humedad relativa, entre otros.

1.2 Justificación de la Investigación

1.2.1 Justificación Teórica

El presente trabajo de investigación se realizará con el fin generar un modelo predictivo de caudales efluentes en la represa Condoroma y así aportar al conocimiento existente acerca de la aplicación de Machine Learning como herramienta que puede ser utilizada para la predicción de caudales.

1.2.2 Justificación Práctica

El objetivo del presente trabajo es proponer un modelo de predicción de caudales efluentes para la represa Condoroma que asista en la gestión de riesgos de Autodema y demás organizaciones involucradas, ya que cuando se demuestre la confiabilidad de los modelos de predicción, podrán ser aplicados como mecanismos que aporten en la toma de decisiones respecto a la gestión, operación y mantenimiento de la Represa y, a su vez, podrá ser aprovechado como guía para la predicción de caudales del resto de represas operadas por Autodema o podrá ser de utilidad para futuros trabajos que sigan esta línea de investigación.

1.2.3 Justificación Metodológica

El trabajo se iniciará obteniendo datos de la Plataforma Movimiento Hídrico Cuenca Colca Regulado de Autodema. Se trabajará con los parámetros disponibles para la represa Condoroma: Nivel de Embalse (m.s.n.m), Volumen Útil (m^3), Caudal Afluente (m^3/s), Pérdidas evaporadas (m^3), Evaporación (m.m.) y Precipitación (m.m.), Temperatura Máxima (°C) y Temperatura Mínima (°C) como variables independientes y Caudal Efluente (m^3/s) como variable dependiente. Posteriormente, se identificará el problema a resolver y se evaluará la técnica de Machine Learning con una metodología de enfoque predictivo para que se obtenga un modelo adecuado para solucionar el problema planteado.

1.3 Delimitación de la Investigación

1.3.1 Espacial

La presente investigación se realizará en la represa Condoroma, ubicada sobre el río Colca a 4158 m.s.n.m. en la localidad Callalli, distrito de Callalli, provincia de Caylloma, departamento de Arequipa. La represa Condoroma forma parte del sistema hidráulico mayor del Sistema Regulado Colca Siguas-Chivay y es gestionado por Autodema.

1.3.2 Temporal

La investigación aplicará Machine Learning del tipo de Aprendizaje Supervisado, para predecir el caudal efluente de la represa Condoroma, utilizando data histórica de mediciones diarias realizadas a parámetros hídricos y climáticos, desde diciembre de 2009 hasta febrero de 2023.

1.3.3 Conceptual

La presente investigación se llevará a cabo mediante la aplicación y análisis de Machine Learning del tipo de Aprendizaje Supervisado (Regresión) para desarrollar un modelo predictivo del caudal efluente de la represa Condoroma. Para ello, se cuenta con datos históricos del caudal efluente (variable a modelar); así como, datos de distintos parámetros medidos en la Represa que pueden influir en el comportamiento del caudal efluente (variables independientes).

Capítulo II: Marco Teórico

2.1 Antecedentes de la Investigación

Artículos relacionados

Esayase, T. (2022) Predicting the Peak Flow and Assessing the Hydrologic Hazard of Kessem Dam, Ethiopia using Machine Learning and RMC-RFA Software.

Problema

Diversas fallas en el funcionamiento de la represa de Terraplén provocado por las inundaciones debido al desbordamiento de agua durante el caudal máximo, genera la necesidad de realizar predicciones de inundaciones apoyándose en la utilización de modelos de aprendizaje automático.

Objetivo

El objetivo del artículo es realizar modelos mediante técnicas de Machine Learning que puedan predecir las inundaciones en la cuenca hidrográfica del río Kessem y en la cuenca Awash en el valle del Rift de la región de Afar en Etiopía.

Base de datos

Para el artículo se utilizaron datos hidrometeorológicos, modelo de elevación digital (DEM), uso de la tierra, cobertura de la tierra (LULC) y datos de mapas de suelo para predecir la cantidad de inundación y la evaluación del riesgo en la seguridad de la represa. Los datos de precipitación diaria se recopilaron de la Agencia Meteorológica de Etiopía (EMA) desde 1988 a 2018 (30 años), en total se utilizaron 15 estaciones en la cuenca de Kessem y sus alrededores, los datos de flujo de corriente diarios para el río Kessem en las estaciones de represa Aware Melka y Kessem, se recopilaron del departamento de hidrología del Ministerio de Agua y Energía (MoWE).

Metodología

Durante el desarrollo de la metodología, primero se usaron datos recopilados de estaciones de monitoreo, los cuales se procesaron mediante la herramienta ArcGis y, con la función Arc Hydro Tools, se delimitaron las cuencas y se estimaron sus características. Posteriormente, se analizaron los datos climáticos mediante el modelo del sistema terrestre canadiense (Canadian Earth System Model conocido por sus siglas en inglés CanESM2) para

el área de estudio, se redujeron a una resolución espacial más fina en el nivel de la cuenca mediante la corrección del sesgo a través del enfoque estadístico del modelo estadístico de reducción de escala (Statistical DownScaling Model, conocido por sus siglas en inglés SDSM) y mediante el uso de los predictores potenciales seleccionados se realizaron proyecciones del clima a futuro. Los resultados se usaron como entrada del modelo Machine Learning para predecir el flujo en la salida de la cuenca hidrográfica de la represa Kessem y los eventos de inundación estimados dentro de futuros horizontes de tres tiempos. Por último, los resultados se importaron al software RMC-RFA para evaluar el riesgo hidrológico futuro de los eventos de inundación.

Técnica de Machine Learning

Se utilizó el modelo climático CanESM2, el cual es un modelo climático global acoplado de cuarta generación desarrollado por el Centro Canadiense de Modelado y Análisis del Clima (CCCma) de Medio Ambiente y Cambio Climático de Canadá. Además, el estudio de detección de cambios de uso de la tierra/cobertura de la tierra (LULC) se realizó mediante el método de clasificación supervisada. También para predecir el flujo del río Kessem en la represa de Kessem dentro de un horizonte de tres tiempos, se identificó que el modelo bidireccional de memoria a corto y largo plazo (Bi-LSTM) supera al modelo memoria a corto y largo plazo (LSTM) y al modelo unidad recurrente cerrada (GRU) y es el ideal para predecir datos de flujo de cuerpos de agua.

Resultados

Los resultados preliminares de la caracterización de las curvas de riesgo hidrológico y los hidrogramas de inundación indican que la represa de Kesem puede ser afectada por inundaciones en un período de retorno entre 2076 y 2100. Sin embargo, la represa requiere más estudios de análisis de riesgos y modificaciones de seguridad para controlar este modo probable de falla.

Hong, et al. (2020). Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow.

Problema

El calentamiento global ha llevado a la preocupación por el cambio climático y ha provocado variaciones en el ciclo hidrológico, lo que se traduce en una mayor incertidumbre en la gestión de los recursos hídricos. Por ello, es importante establecer un plan para la gestión de los recursos hídricos a través de una gestión eficiente del agua mediante la mejora de la operación de estructuras hidráulicas; sin embargo, los cambios en los patrones de caudal de las represas, provocados por el cambio climático, están dificultando el uso de recursos hídricos estables y el establecimiento de planes de abastecimiento.

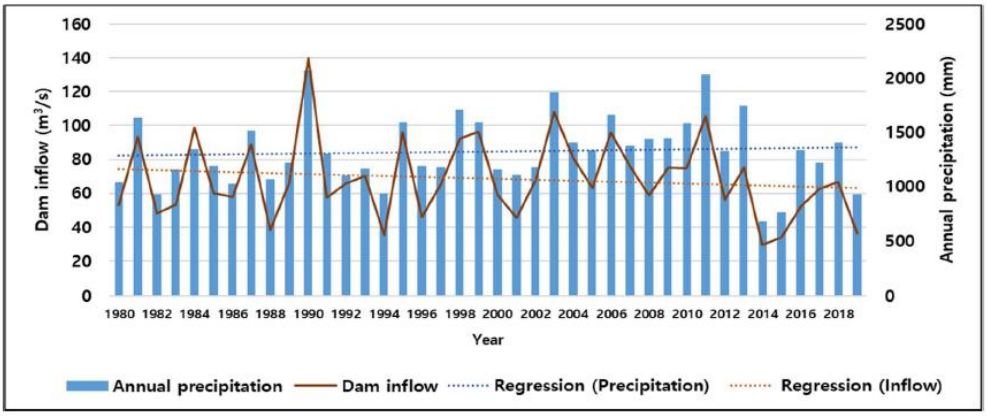
Objetivo

Evaluar el rendimiento del algoritmo para predecir la cantidad de afluencia de la represa del río Soyang. Desarrollar y evaluar los algoritmos de aprendizaje automático combinados, teniendo en cuenta la duración del flujo.

Base de datos

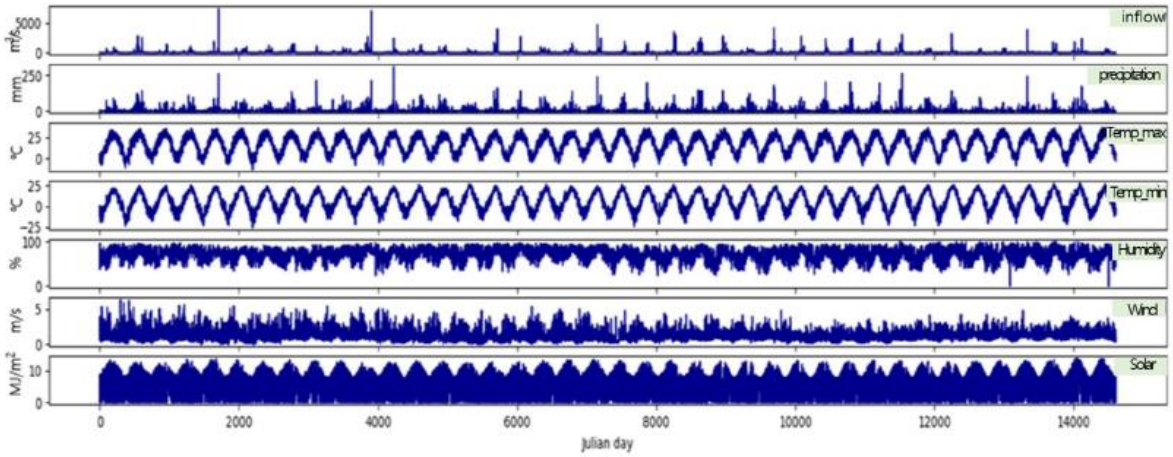
Se obtuvieron datos de precipitación proporcionados por la estación meteorológica de Chuncheon, esta estación facilita información respecto a la precipitación anual de un periodo de 40 años, desde 1980 hasta 2019. Además, se obtuvieron datos de la serie temporal del clima (precipitación, temperatura máxima, temperatura mínima, humedad, velocidad del viento y radiación solar) proporcionados por la Administración Meteorológica de Corea. También, se recopilaron datos de entrada de la represa del río Soyang otorgados por la Corporación de Recursos Hídricos de Corea (K-water).

Figura 5: *Precipitación anual y caudal medio anual de entrada de la represa durante 1980-2019 de la represa del río Soyang.*



Nota. Obtenido de Hong et al., (2020)

Figura 6: *Variación de la serie de tiempo de la entrada de la represa y datos meteorológicos.*



Nota. Obtenido de Hong et al., (2020)

Metodología

El método utilizado para el preprocesamiento es escalado y métodos de estandarización, incluido el escalado de formas, la normalización y la estandarización mediante la función 'StandardScaler'. También, para predecir el caudal de la represa del río Soyang se utilizó el aprendizaje supervisado. Este tipo de aprendizaje se aplica para el entrenamiento de datos e infiere una función a partir de estos. Se utilizaron un total de seis métodos: árbol de decisión (DT), perceptrón multicapa (MLP), bosque aleatorio (RF), aumento de gradiente (GB), red neuronal recurrente: memoria a corto y largo plazo (RNN-LSTM), y red neuronal convolucional–LSTM (CNN–LSTM), para construir modelos que estimen la cantidad de agua que ingresa a la represa.

Figura 7: *Descripción de los modelos de aprendizaje automático.*

Modelos de aprendizaje automático	Módulo	Función	Notación
Árbol de decisión	sklearn.tree	DecisionTreeRegressor	DT
Perceptrón multicapa	sklearn.neural_network	MLPRegresor	MLP
Bosque aleatorio	sklearn.ensemble	RandomForestRegresor	RF
Aumento de gradiente	sklearn.ensemble	GradientBoostingRegressor	ES
RNN-LSTM	keras.models.Sequential	LSTM, denso, abandono	LSTM
CNN-LSTM	keras.models.Sequential	LSTM, denso, abandono, Conv1D, MaxPooling1D	CNN-LSTM

Nota. Obtenido de Hong et al. (2020)

Técnicas de Machine Learning

Una de las técnicas empleadas es el árbol de decisión, un modelo ampliamente utilizado para la clasificación y la regresión. Este método aprende a medida que continúa haciendo preguntas de sí o no para llegar a una decisión. La siguiente técnica utilizada es el perceptrón multicapa, una de las estructuras de redes neuronales de avance (FFNN), que consta de un total de tres capas: capa de entrada, capa oculta y capa de salida. También utiliza la técnica del bosque aleatorio, que combina el algoritmo de embolsado, el método de aprendizaje por conjuntos y el algoritmo del árbol de clasificación y registro y el aumento de gradiente un modelo de conjunto que aprende el algoritmo de impulso mediante el aprendizaje de conjunto para el árbol de decisión. A su vez, la investigación aplica cuatro técnicas adicionales: MLP, GB, RNN-LSTM y CNN-LSTM.

Resultados

Se demostró que Perceptrón Multicapa (MLP) es el mejor algoritmo para la predicción de caudales; sin embargo, a pesar de que es el algoritmo ideal para la predicción de flujo, en términos de evaluación del rendimiento del modelo, hubo un límite de predicción de flujo en toda la duración, lo que significa que un solo uso del algoritmo no podría considerar perfectamente los regímenes de flujo. Para mejorar esto, se desarrollaron algoritmos de aprendizaje automático de combustión (CombML) y los resultados muestran que es posible predecir el flujo de entrada mediante el aprendizaje del flujo de entrada, considerando las características del flujo, como el flujo en Corea.

Babaei et al. (2019). Artificial Neural Network and Support Vector Machine Models for Inflow Prediction of Dam Reservoir (Case Study: Zayandehroud Dam Reservoir)

Problema

Determinar la cantidad real de liberación de agua de los embalses aporta información valiosa que ayuda a los tomadores de decisiones a gestionar y asignar recursos hídricos de forma óptima. La liberación de agua se ve influenciada por diversos parámetros, especialmente por el caudal de entrada y caudal de salida; sin embargo, debido a la incertidumbre de los valores de entrada a un embalse es importante predecir las descargas de entrada a los embalses.

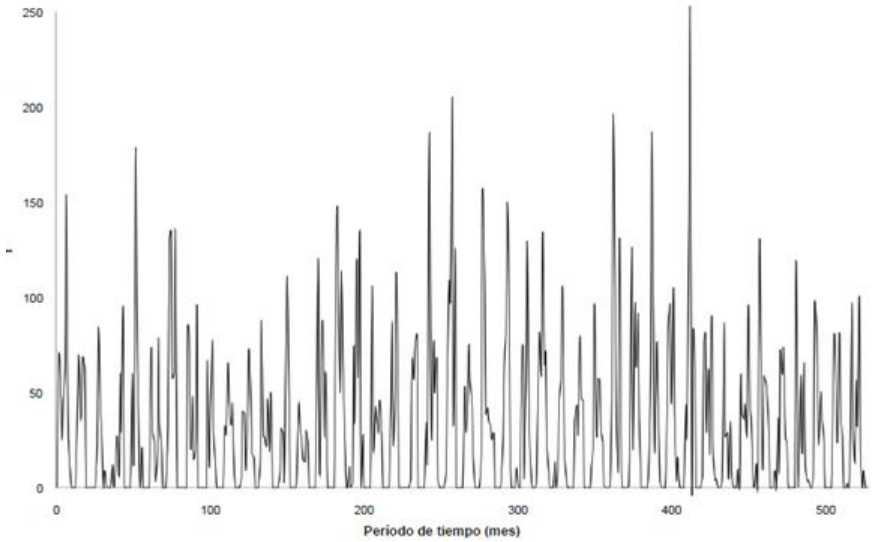
Objetivo

Predecir el caudal entrante en el embalse de la represa Zayandehroud, ubicada en la meseta central de Irán, mediante los modelos de redes neuronales artificiales y máquinas de vectores de soporte.

Base de datos

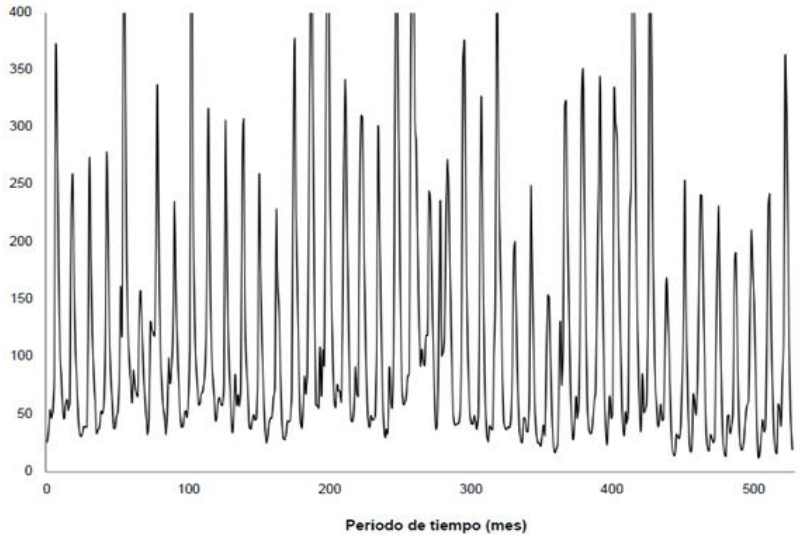
Se utilizó la variable precipitación mensual durante 44 años (1971 a 2015) de la estación hidrométrica Ghaleh Shahrokh (Ver Figura 8) y caudal de entrada mensual en el embalse de la represa Zayandehroud durante un período de 44 años (1971 a 2015) (Ver Figura 9).

Figura 8: *Precipitaciones mensuales de las estaciones hidrométricas de Ghaleh-Shahrokh durante un período de 44 años (1971 a 2015)*



Nota. Obtenido de Babaei et al. (2019)

Figura 9: *Entrada mensual en el embalse de la represa de Zayandehroud durante un período de 44 años (1971 a 2015)*



Nota. Obtenido de Babaei et al. (2019)

Metodología

Se utilizaron modelos de Machine Learning para estimar el flujo de entrada a un embalse. Posteriormente, se combinaron las bases de datos para proponer patrones de entrada para ingresarlos a los modelos en mención con diferentes tiempos de retraso y datos de precipitación.

Técnica de Machine Learning

Se utilizaron dos técnicas: redes neuronales artificiales y máquinas de vectores de soporte. El modelo de red neuronal artificial se basa en la red neuronal del cerebro humano. Según la revisión de literatura realizada por Babaei et al. (2019), el modelo de redes neuronales artificiales (ANN) ha sido utilizado para resolver muchos problemas de ingeniería de recursos hídricos debido a su capacidad para simular funciones no lineales con una precisión adecuada. Por otro lado, el modelo máquina de vectores (SVM) de soporte es un método caracterizado por simplificar procesos complejos, simular comportamientos no lineales y abarcar varios aspectos de la incertidumbre en un problema de predicción.

Resultados

Los mejores resultados para R y RMSE se obtuvieron con el noveno modelo (Ver Tabla 1), que utilizó como datos de entrada la precipitación mensual estación Ghaleh-Ghahrokh con

diferentes desfases temporales, a diferencia del resto de modelos para lo que no se consideró el desfase en mención.

La comparación de los resultados indica que el noveno modelo propuesto tiene el menor error para la predicción del flujo de entrada en el que los resultados del modelo SVM superan a los del modelo ANN.

Tabla 1: *Mejores resultados de los modelos SVM y ANN para la predicción del flujo de entrada del embalse de la represa de Zayandehroud*

Modelo	**Etapa de entrenamiento**		**Etapa de validación**		**Etapa de evaluación**	
	RMSE	**R**	**RMSE**	**R**	**RMSE**	**R**
SVM	47.9346	0.89620	42.69093	0.93030	23.56193	0.96220
ANN	48.5441	0.89269	43.74800	0.92983	28.51250	0.95333

Nota. Obtenido de Babaei et al. (2019)

Tesis relacionadas

Seminario Gastelo, J. (2021). Modelos de predicción para el caudal del río Chira en la estación Ardilla.

Problema

Las obras de ingeniería hidráulica del Proyecto Especial Chira Piura tienen dificultades para lograr sus objetivos debido a la recurrencia de los Fenómenos El Niño y a la creciente colmatación del Embalse Poechos, el cual es el componente más importante del conjunto de estructuras de ingeniería de esta técnica de riego.

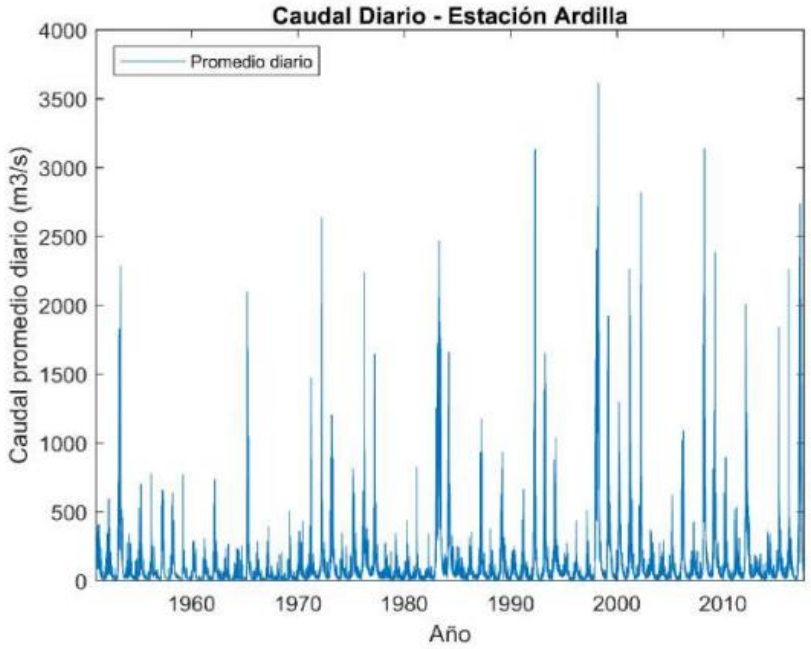
Objetivo

Identificar los modelos de pronóstico unidimensionales adecuados que pueden facilitar la gestión de los recursos hídricos en el embalse de Poechos.

Base de datos

Se utilizaron datos que fueron recolectados en la Estación Ardilla (un promedio de 66 años). Los datos obtenidos abarcan información desde 1951 hasta el 2017, además se utilizan datos de estaciones meteorológicas ubicadas en la cuenca del río Chira.

Figura 10: *Caudal diario de la Estación Ardilla periodo de 66 años (1951 a 2017)*



Nota. Obtenido de Seminario Gastelo, J. (2021)

Metodología

Las metodologías que se emplearon en el desarrollo de la investigación fueron suavizado exponencial y modelos autorregresivo integrado de media móvil (ARIMA). EL suavizado exponencial consiste en explicar las observaciones en la serie temporal con la media ponderada de los valores pasados de la serie con pesos geométricamente decrecientes. Por lo tanto, las últimas observaciones de la serie son más consistentes con el modelo que las observaciones anteriores. El estudio utilizó una función de optimización que encuentra los mejores parámetros que minimizan el error cuadrático medio de las predicciones. Por otro lado, los métodos ARIMA analizan las correlaciones entre las observaciones en diferentes momentos para definir un modelo óptimo. La función de autocorrelación simple y la función de autocorrelación parcial se utilizan para estimar la correlación entre la observación *t* en el tiempo *t-k.*

Técnica de Machine Learning

Para el desarrollo del modelo se consideró el método de suavizado exponencial en sus variantes de doble ecuación y de tres ecuaciones, así como el método ARIMA desarrollado por Box-Jenkins. Se utilizan varios modelos para realizar comparaciones entre ellos están: Modelo Holt, Modelo Holt Damped, Modelo Holt-Winters, Modelo Holt-Winters Damped, Modelo

ARIMA. Cada modelo se evaluó comparando el error cuadrático medio, el error absoluto medio y el coeficiente de determinación obtenido del modelo seleccionado en cada horizonte de pronóstico (h = 10).

Resultados

Se obtuvo el mejor rendimiento para la serie temporal diaria con el Modelo ARIMA (1,1,1) y para la serie temporal con frecuencia semanal el mejor rendimiento se obtuvo con el Modelo ARIMA $(1,1,1)(0,1,1)_{52}$.

Brenes Jimenez Anibal (2020). Predicción del caudal promedio horario de la estación hidrológica Palmar, utilizando modelos de Machine Learning basados en Árboles de decisión

Problema

Los registros de caudal medidos por estaciones hidrológicas suelen presentar grandes cantidades de datos faltantes debido a la ausencia temporal de observadores, fallas en el equipo y falta de recursos financieros. Además, no es raro que el equipo se descomponga debido a inundaciones extremas, se mueva a otro lugar o incluso se dañe debido al vandalismo. Esto genera la oportunidad de usar redes neuronales para modelar, completar y extender registros hidrológicos.

Objetivo

Comparar la capacidad predictiva de diferentes modelos estadísticos de datos de caudal promedio horario de la estación hidrológica Palmar del río Grande de Térraba.

Base de datos

Se utilizaron datos de información hidrometeorológica proporcionados por la División de Hidrología del Centro de Servicios de Investigación en Ingeniería del Instituto de Energía Eléctrica de Costa Rica. También, se utilizaron algunas capas de la edición 2014 del Atlas Digital de Costa Rica del Instituto Tecnológico de Costa Rica para la elaboración de mapas temáticos.

Metodología

El desarrollo del método comenzó con una revisión bibliográfica que identificó estudios previos que abordan el uso de modelos Machine Learning para la transmisión de datos basados en árboles de decisión. Después, se obtuvo datos hidrometeorológicos de Costa Rica y capas de atlas digitales, luego se realizó el análisis exploratorio de información hidrológica y meteorológica, se completó los datos faltantes y se interpoló usando ponderaciones. Siguiendo el método de tomar el cuadrado de la distancia bidireccional a las tres ubicaciones más cercanas, se modelaron los datos, se seleccionaron las variables, según su importancia, se ajustaron los parámetros del modelo y se comparó el rendimiento de los modelos. Además, se analizó la significancia de las variables con un modelo de bosque aleatorio. Para ello, se utilizó el criterio de ganancia de información, el estadístico independiente chi-cuadrado y la ganancia de información basada en permutaciones aleatorias del bosque. Por último, se ajustaron los parámetros de cada modelo y se afinaron para obtener una mejor predicción del caudal promedio.

Técnica de Machine Learning

La investigación utiliza tres modelos los cuales compara sus datos para identificar cual es el que mejor explica los datos, las técnicas usadas fueron el modelo de árboles de decisión, el modelo de bosques aleatorios, el modelo de ponderación y el modelo de aumento de gradiente extremo (XGBoosting).

Resultados

Al analizar el rendimiento de los modelos predictivos, se identificó que el aprendizaje automático basados en árboles de decisión tienen un rendimiento aceptable, puesto que todos los valores obtenidos para el Coeficiente de Nash-Sutcliffe (NSE) son superiores a 0,8. De acuerdo con el NSE, la técnica más adecuada para el modelado del caudal en la estación Palmar es el modelo de bosque aleatorio que utiliza un conjunto de variables imputadas para datos faltantes como predictores. La segunda técnica más adecuada es el modelo de ponderación y la tercera es el árbol de decisión.

Miranda Araya, F. (2019). Uso de redes neuronales artificiales calibradas en el periodo histórico para el pronóstico de caudales de deshielo proyectados en el periodo 2020-2050 en la Cuenca del río Maipo en el Manzano.

Problema

El cambio climático ha tenido diversas consecuencias a nivel mundial: aumento de la temperatura, aumento y/o disminución de la precipitación y derretimiento de glaciares. En Chile, los efectos del cambio climático se evidencian en la disminución de precipitación y en la oferta de agua potable; así como en la salinización y desertificación de áreas agrícolas.

Ante estos hechos, organismos del Gobierno de Chile estiman pronósticos para la temporada de deshielos; sin embargo, estos modelos no son realizados considerando un futuro no estacionario. En ese sentido, la investigación en mención buscó pronosticar los caudales de deshielo en base a información histórica mediante un modelo de redes neuronales.

Objetivo

El objetivo principal de esta investigación fue evaluar si el pronóstico de caudales de deshielo mediante modelos de redes neuronales puede ser aplicados en periodos futuros utilizando forzantes meteorológicas, como temperatura, precipitación y variables hidrológicas, como humedad del suelo, caudales mensuales, cobertura de nieve, entre otros.

Base de datos

Los datos utilizados fueron obtenidos mediante el modelo Capacidad de Infiltración Variable (VIC) y de ocho estaciones fluviométricas ubicadas en la cuenca Maipo. De estas bases de datos se adquirieron cinco variables medidas diariamente de 1980 a 2015: Caudal (m^3/s), equivalente en agua de la nieve (mm), precipitación (pp), temperatura media (°C), evapotranspiración (mm).

Metodología

Se obtuvieron los datos temporales diarios de las variables hidrometeorológicas de las estaciones fluviométricas de Maipo y luego fueron analizadas a través del modelo VIC. Se utilizaron dos modelos referentes a la disponibilidad de información (M1 y M2). Posteriormente, se separó la data en tres grupos: entrenamiento (50%), validación (25%) y prueba (25%) y se aplicó cada red con distintas combinaciones de variables. Adicionalmente,

se realizó cuatro entrenamientos con los modelos de circulación global CCSM4, CSIRO, MIROCESM e IPSL y esta fue aplicada en la red óptima utilizando información futura (2020-2050) obtenida del modelo VIC. Finalmente, se compararon los resultados de los volúmenes totales de deshielo y caudales obtenidos en el periodo de 2020 a 2050 con los simulados por el modelo VIC.

Figura 11: *Diagrama de flujo de la Metodología*

Nota. Obtenido de Miranda Araya, F. (2019)

Técnica de Machine Learning

Empleando Redes Neuronales Artificiales se pronosticó los caudales para la temporada de deshielo, este modelo fue calibrado utilizando información histórica de estaciones fluviométricas y el pronóstico de caudales fue realizado mediante el modelo hidrológico VIC usando forzantes climáticas de cuatro modelos climáticos desarrollados por el IPCC: MIROC-ESM, CSSM4, CSIRO e IPSL.

Resultados

Se calculó la correlación de los lags que tienen influencia sobre el caudal para cada agregación temporal y se obtuvo cinco lags a nivel semanal, cinco lags a nivel mensual y 1 lag a nivel semestral. Para obtener un análisis robusto, se generaron conjuntos aleatorios de

predictores (60 semanales, 300 mensuales y 60 mensuales), luego cada uno de estos fue calibrado calculando los indicadores NSE, Eficiencia de Kling-Gupta (KGE) y Error Cuadrático Medio (MSE). Todas las redes fueron entrenadas mediante el método Levenberg-Marquardt.

Los resultados mostraron una tendencia al descenso de los caudales de deshielo debido principalmente a la disminución de precipitaciones y equivalente de agua en nieve en el periodo estudiado. Por otro lado, los componentes utilizados señalaron un patrón repetitivo tanto a nivel semanal, como mensual y semestral.

Schumacher et al. (2021). From Random Forests to Flood Forecasts: A Research to Operations Success Story.

Problema

Las excesivas lluvias y las reiteradas inundaciones amenazan a la comunidad a nivel mundial, especialmente a las regiones que se ubican en los polos del planeta. En los Estados Unidos, entre el periodo que comprende desde 2010 y 2019, la inundaciones y precipitaciones extremas causaron 212 muertes y varios tipos de daños materiales importantes, lo que resultó en más de $60 mil millones en daños (NCIE, 2020). Para combatir ello, es importante contar con pronósticos oportunos que adviertan al público sobre la amenaza de lluvias excesivas. El Centro de Predicción Meteorológica (WPC) produce pronósticos de exceso de precipitación basados en la guía del modelo de Predicción Numérica del Tiempo (NWP), pero estos modelos aún no abordan eventos importantes de precipitación a largo y corto plazo que causan inundaciones (Schumacher et al., 2021).

Schumacher et al. (2021), de Universidad Estatal de Colorado y representante del Departamento de Ciencias Atmosféricas, utilizó técnicas de aprendizaje automático (bosques aleatorios y redes neuronales) como posprocesamiento de la salida del modelo y obtuvo resultados prometedores en el pronóstico de precipitaciones extremas. Además, desarrolló la herramienta US State Method, que ahora se utiliza por el Centro de Predicción del Tiempo de la Oficina de Meteorología, para predecir lluvias excesivas.

Objetivo

Desarrollar una herramienta que permita pronosticar las precipitaciones extremas haciendo uso de técnicas de Machine Learning.

Base de datos

Este documento utiliza eventos de precipitación extrema de una combinación de hasta tres conjuntos de datos:

(i) informes de inundaciones,

(ii) análisis de precipitación de fase IV del Centro de Predicción Ambiental de EE. UU. (NCEP) durante un período de acumulación de 24 horas,

(iii) 24 horas (*h)* acumulación de precipitaciones. *h* calibración climatológica.

El período de entrenamiento es de aproximadamente siete años, del 9 de junio de 2009 al 31 de agosto de 2016, y el período de prueba previsto es del 1 de enero de 2017 al 31 de diciembre de 2018.

Técnica de Machine Learning

La técnica utilizada es bosques aleatorios, que consisten en árboles de decisión que individualmente hacen predicciones de rango únicas (por ejemplo, 0 o 1) para eventos específicos basados en las propiedades del árbol (es decir, entradas). Considere un conjunto de etiquetas históricas (p. ej. eventos de lluvia) y las características correspondientes, cada árbol considera una submuestra aleatoria de la muestra de entrenamiento para generar un árbol único. A partir del nodo raíz del árbol, las ramas se recorren según el resultado de los criterios especificados en cada nodo. Para cada nodo, se evalúa un subconjunto aleatorio de características para seleccionar un criterio que minimice las impurezas del nodo para todos los ejemplos de entrenamiento restantes. En otras palabras, idealmente, los nodos subsiguientes en el árbol se vuelven más limpios y coinciden con una determinada etiqueta de clasificación (por ejemplo, inundado o no inundado) (Schumacher et al., 2021).

Este documento utiliza eventos de precipitación extrema de una combinación de hasta tres conjuntos de datos: (i) informes de inundaciones, (ii) análisis de precipitación de fase IV del Centro de Predicción Ambiental de EE. UU. (NCEP) durante un período de acumulación

de 24 horas, y (iii) 24 *h* acumulación de precipitaciones, calibración climatológica. El período de entrenamiento es de aproximadamente 7 años, del 9 de junio de 2009 al 31 de agosto de 2016, y el período de prueba previsto es del 1 de enero de 2017 al 31 de diciembre de 2018.

Cuando un subconjunto de eventos en un nodo está limpio o ya no es demasiado grande para dividirlo, se crea un nodo "hoja" que declara una clasificación de eventos. Para hacer una predicción, el árbol recibe nuevas entradas, como la salida del modelo NWP en tiempo real, y el árbol se recorre en los nodos de hoja. Suma todas las predicciones del árbol de decisión para producir un pronóstico de probabilidad de exceso de precipitación basado en la entrada dada (Schumacher et al., 2021).

Resultados

Los modelos entrenados con excedencias de intervalos de recurrencia promedio de 1 año definidas por el análisis de precipitación Stage-IV (ST4) funcionan mal en el norte de las Grandes Llanuras y el suroeste de los Estados Unidos, en parte debido a un alto sesgo en la cantidad de eventos de entrenamiento en estas regiones. Aumentar el umbral ST4 a dos años o eliminar los datos ST4 del entrenamiento, optimizar la habilidad de pronóstico geográficamente y promediar espacialmente las entradas meteorológicas para el entrenamiento generalmente da como resultado una habilidad de pronóstico de bosques aleatorios mejorada. Tanto los pronósticos de perspectivas de precipitaciones excesivas (ERO) como los de bosques aleatorios tienen habilidad estacional: pronósticos deficientes a fines del otoño e invierno y pronósticos hábiles en el verano y principios del otoño. Sin embargo, los ERO son consistentes y significativamente mejores que sus contrapartes de bosques aleatorios.

2.2 Bases Teóricas

2.2.1 Machine Learning

John McCarthy (1956) fue quien definió por primera vez el término Machine Learning en un taller llevado a cabo sobre Razonamiento Teórico Lógico en el año 1956, McCarthy sostiene que Machine Learning es la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cómputo inteligentes.

En ese sentido, de acuerdo con Bobadilla, J. (2010), Machine Learning o Aprendizaje Automático es la ciencia mediante la cual las computadoras "aprenden" de los datos. En lugar

de desarrollar gradualmente cada solución específica para cada necesidad, como se hace en los métodos de programación tradicionales, Machine Learning se enfoca en desarrollar algoritmos generales que puedan derivar patrones a partir de diferentes tipos de datos.

Aunque Machine Learning convierte los procesos más automatizados, eso no significa que todos los procesos que se apoyan en Machine Learning son fáciles de automatizar. Bobadilla, J. (2010), sostiene que los ingenieros de datos (científicos de datos) tienen que realizar muchas tareas específicas, como identificar fuentes de datos, limpiar datos, eliminar información repetitiva, realizando la normalización necesaria, para la aplicación determinando el tipo apropiado de solución de Machine Learning, seleccionando el algoritmo más adecuado, analizando los resultados, identificando comportamientos incorrectos, volviendo al proceso anterior para realizar los cambios necesarios para mejorar los resultados, etc.

Además, la aplicación de Machine Learning dependerá de acuerdo al tipo de clasificación de los problemas, tipo y clasificación de variables. Los tipos de aplicación de Machine Learning se pueden clasificar en: Aprendizaje supervisado que a su vez se divide en regresión y clasificación y Aprendizaje no supervisado que usa modelos llamados Clustering (agrupamiento) reducción de dimensiones.

2.2.2 Aprendizaje Supervisado

Es una modalidad de aprendizaje de Machine Learning, el cual se basa en el conocimiento previo de los datos. “El aprendizaje supervisado se concentra en patrones de aprendizaje mediante la conexión de la relación entre variables y resultados conocidos y el trabajo con conjuntos de datos etiquetados” (Theobald, 2018).

El aprendizaje supervisado “se refiere a un tipo de modelo de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida (outcome, eventos o labels) son conocidos” (Beunza, Puertas & Condés, 2019, p. 35).

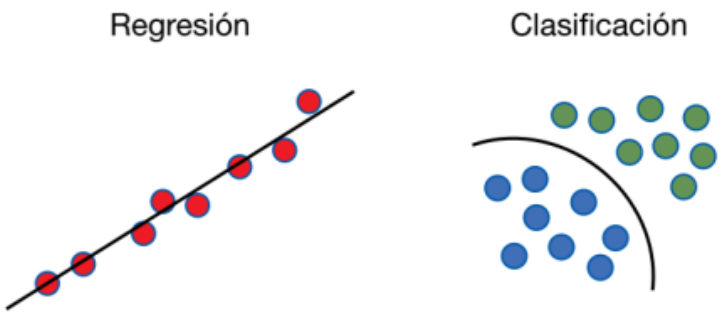
De acuerdo con Bobadilla (2020) el aprendizaje supervisado en Machine Learning se utiliza cuando un conjunto de datos o datos de entrada (muestra) tiene una etiqueta asociada. A modo de ejemplo, un conjunto de imágenes cada una de las cuales incluye algún tipo de metadato (generalmente una etiqueta): ((pict001.bmp, "perro"), (pict002.bmp, "búho"), (pict003.bmp, "ratón"). Con base en este conjunto de datos, el modelo se puede “entrenar”

utilizando varios algoritmos de clasificación de aprendizaje automático y, al final del entrenamiento, ser capaz de predecir un conjunto de características correspondientes a nuevas imágenes (no incluidas en los datos originales); es un problema de clasificación.

De manera similar, podemos usar conjuntos de datos que comprenden pruebas con datos numéricos asociados, pongamos el caso de un grupo de muestras de terremotos de los cuales sus datos comprenden la magnitud de vibraciones pasadas obtenidas de sensores para establecer la magnitud oficial de un terremoto ([8.1, 6.4, …], 5.8), ([3.8, 8.9,…], 7.5). Este es un problema de regresión y su utilidad puede generar información sobre la fuerza prevista del modelo de regresión cuando se proporcionan nuevas muestras (valores de terremotos recopilados en tiempo real).

De acuerdo a la naturaleza de la variable de respuesta, el aprendizaje supervisado puede ser de dos tipos: Clasificación y regresión (Ver Figura 12). Los algoritmos supervisados de clasificación se utilizan cuando la variable dependiente es etiquetada, discreta o sus valores no son consecutivos. Por otro lado, los algoritmos supervisados de regresión son usados cuando la variable dependiente o de respuesta es continua (Nuin et al., 2020).

Figura 12: *Aprendizaje supervisado para un algoritmo de regresión y un algoritmo de clasificación*



Nota. Obtenido de Nuin et al. (2020)

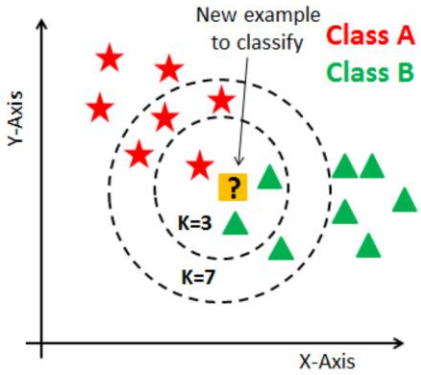
2.2.2.1 Algoritmos de aprendizaje supervisado

- ***Algoritmo K Nearest Neighbors (KNN)***

El algoritmo KNN se caracteriza por clasificar datos de forma efectiva y simple basado en instancias de datos para realizar el algoritmo de aprendizaje automático; en ese sentido, KNN calcula la medida de la distancia para cada dato en la base de datos (Harrington, 2012).

Esta técnica agrupa un conjunto de datos con valores cercanos a lo que se busca predecir y lo clasifica de acuerdo al tipo de dato que detecta.

Figura 13: *Algoritmo KNN*



Nota. Obtenido de Harrington (2012)

- ***Regresión***

La regresión es un proceso de predicción del valor objetivo similar a la clasificación. La diferencia entre regresión y clasificación es que las variables predictoras en regresión son continuas, mientras que en clasificación son discretas. La regresión es una de las herramientas más útiles en estadística. La minimización de la suma de errores cuadráticos se emplea para hallar los mejores pesos para las características de entrada de la ecuación de regresión. La regresión se puede efectuar en cualquier conjunto de datos siempre que se le proporcione una matriz de entrada X y se pueda calcular el inverso de $X^T.X$. El hecho de que pueda calcular una ecuación de regresión para un conjunto de datos no significa que los resultados sean buenos. Una prueba de cuán "bueno" o importante es un resultado es la correlación entre el valor predicho y los datos originales (Harrington, 2012).

La fórmula matemática para la regresión lineal se puede expresar como:

$$\hat{w} = (X^T X)^{-1} X^T y$$

Donde los datos de entrada están en la matriz X y los pesos de regresión en el vector $\hat{w}$.

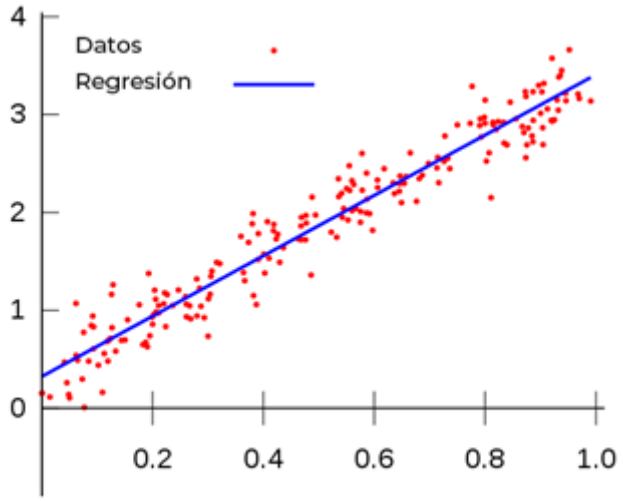
Los modelos de regresión se utilizan a menudo para realizar predicciones y pronósticos y para determinar la relación de causa y efecto entre variables independientes y dependientes. El análisis de regresión estima el valor de la variable dependiente "Y" dentro del rango del valor

de la variable independiente "X". Hay dos tipos de modelos de regresión: regresión lineal simple y regresión múltiple (Maulud & Abdulazeez, 2020). La regresión lineal simple es un modelo matemático que tiene como objetivo correlacionar la variable dependiente "Y" con valores independientes "Xi", como se muestra en la siguiente ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \varepsilon$$

Donde βi es el coeficiente y ε es el posible error. Un ejemplo de este tipo de regresión se muestra en la siguiente figura, donde los valores de Y se ajustan con una línea azul denominada línea de tendencia. El sistema se ocupa de asumir que la relación entre las variables es de forma lineal.

Figura 14: *Algoritmo de regresión*



Nota. Obtenido de Universidad Complutense de Madrid (2022)

Como se puede observar, este modelo se utiliza cuando sólo hay una variable. La regresión lineal múltiple es una extensión de esta regresión lineal simple; sin embargo, ésta se utiliza para correlacionar múltiples variables (Artacho, 2021). Este modelo de regresión se ve así en forma de matriz como se muestra a continuación:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p1} & \cdots & x_{pn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

De esta expresión se deduce:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_n X_{in} + \varepsilon \quad para\ i = 1,2,3,\ldots n$$

La aplicación de este modelo produce un error aleatorio ε, y n coeficientes βi, cada uno relacionado con una de las variables e indica cuánto influye esta variable en el resultado final del ajuste.

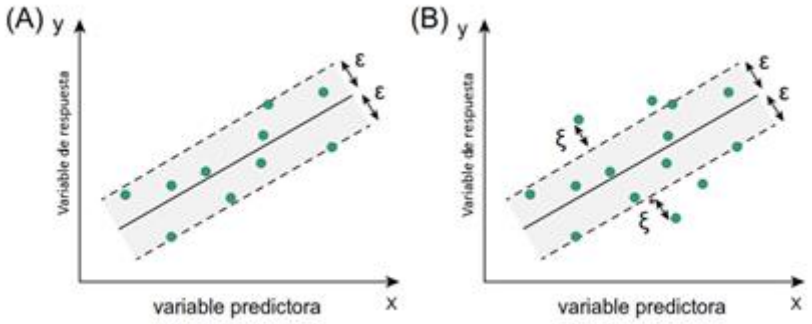
- ***Support Vector Regression (SVR)***

La regresión de vectores de soporte (SVR por sus siglas en inglés) es una técnica de aprendizaje supervisado de tipo regresión. Es eficaz para el análisis de relación entre una variable dependiente continua y una o más variables predictoras, porque equipara la complejidad del modelo y el error de predicción; además, cuenta con un buen rendimiento para manejar datos de alta dimensión. SVR es una extensión de la técnica Support Vector Machine (SVM), la cual es un algoritmo de clasificación que realiza una salida binaria, mientras que SVR se utiliza en problemas de regresión no lineales para producir estimaciones basadas en valores reales (Zhang & O'Donnell, 2020).

La optimización en SVR se representa mediante vectores de soporte, donde la solución de optimización no depende de la dimensión de los datos de entrada sino sólo del número de vectores de soporte. SVR proporciona una herramienta eficaz para datos de alta dimensión, asimismo, este es un método de aprendizaje automático que aprende un modelo para describir la importancia de una variable al caracterizar la relación entre la entrada y la salida, a diferencia de un método de regresión tradicional que depende de la suposición de un modelo (por ejemplo: distribución de datos lineal) que podría no ser preciso (Zhang & O'Donnell, 2020).

Para funciones lineales, SVR introduce una función de pérdida insensible a ε para calcular un hiperplano tal que los valores de respuesta pronosticados de las muestras de entrenamiento tengan como máximo una desviación ε de su valor observado (real). El objetivo de ε--SVR es estimar una función con la restricción de que la estimación de cada punto de datos de entrada tenga una desviación máxima de ε de su valor de respuesta real, formando un tubo insensible a ε que abarque simétricamente la función estimada (Zhang & O'Donnell, 2020).

Figura 15: *Representación gráfica de modelos lineales de ε--SVR*



Nota. Obtenido de Zhang & O'Donnell (2020)

La formulación matemática de una ε--SVR lineal se expresa de la siguiente manera:

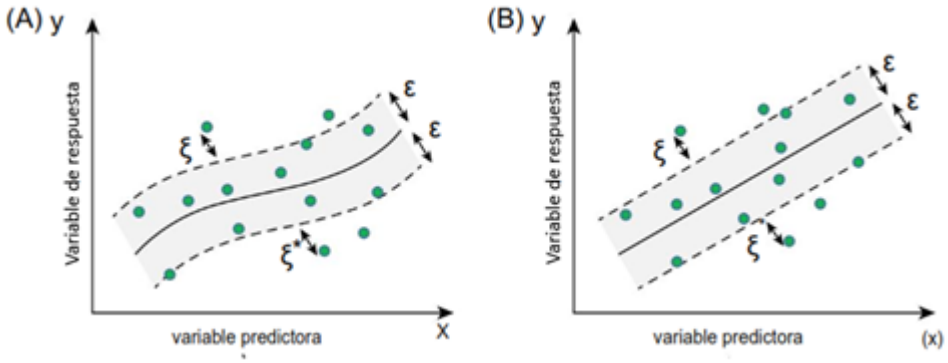
$$y = \sum_{i=1}^{N} (a_i - a_i^*).[x_i, x] + b$$

Por otro lado, para funciones no lineales, la función Kernel transforma los datos en una característica de espacio dimensional más alto para hacer posible ejecutar la separación lineal, mediante la siguiente fórmula matemática:

$$y = \sum_{i=1}^{N} (a_i - a_i^*).K[x_i, x] + b$$

En SVR, los datos se pueden discriminar usando una función lineal en el espacio del Kernel (*K*) y la optimización se puede resolver siguiendo el mismo cálculo del modelo no lineal. El uso de Kernel es uno de los enfoques más comunes en SVR, tanto para regresión y clasificación; en ese sentido, existen varias funciones habituales como linear kernels (modelo ε-SVR), polynomial kernels, radial basis function (RBF) kernels y ANOVA RB kernels. La elección de la función dependerá de la distribución de datos de entrada. Cuando las entradas son vectores de datos grandes y dispersos se usa linear kernels. Polynomial kernels es muy utilizado para el procesamiento de imágenes. RBF se aplica principalmente en ausencia de conocimientos previos. ANOVA RB suele utilizarse para tareas de regresión (Awad & Khanna, 2015).

Figura 16: *Representación gráfica de modelos no lineales SVR*



Nota. Obtenido de Zhang & O'Donnell (2020)

- ***Series temporales***

Una serie temporal es una colección de observaciones (datos) de una variable ordenadas y equidistantes sobre una (univariante o escalar) o varias (multivariante o vectorial) características en diferentes momentos (Peña, 2005).

La representación matemática de una serie temporal univariante puede ser:

$$y_1, y_2, \ldots, y_N \; ; (y_t)_{t=1}^{N} \; ; \; (y_t : \; t \; = 1, \ldots, N)$$

Por otro lado, la formulación matemática de series temporales multivariantes puede ser representada en una matriz *Y* de orden *N x M*

$$\mathbf{Y} \equiv \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_N' \end{bmatrix} \equiv \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix},$$

Según Hyndman & Athanasopoulos (2021), las series temporales tienen tres patrones: tendencia, estacional y cíclica. Una tendencia se refiere a un incremento o disminución, no necesariamente lineal, a largo plazo en los datos. Por otro lado, un patrón estacional sucede cuando una serie de tiempo es afectada por factores estacionales como una época del año. La estacionalidad es siempre de un período fijo y conocido. En cambio, el patrón cíclico ocurre cuando los datos exhiben subidas y bajadas que no tienen una frecuencia fija.

Si se puede predecir los valores exactos de una serie es determinística; en cambio, si solo se pueden determinar de manera parcial, es una serie estocástica (Peña, 2005). Existen dos tipos de procesos estocásticos:

- Procesos estocásticos estacionarios, su distribución de probabilidad varía a lo largo del tiempo de forma más o menos constante. Se cumple cuando las propiedades estadísticas son semejantes entre dos secuencias finitas de componentes *Yt* separadas por un número entero *h* cualquiera (Peris, F., 2022).
- Procesos estocásticos no estacionarios, su distribución de probabilidad varía a lo largo del tiempo de forma no constante. Se cumple cuando las propiedades estadísticas son diferentes entre dos secuencias finitas de componentes *Yt* para al menos un número entero $h > 0$ (Peris, F., 2022).

Para realizar un adecuado análisis de series de tiempo es necesario determinar si los datos tienen un comportamiento estacionario o no estacionario. En ese sentido, los estadísticos David Dickey y Wayne Fuller desarrollaron un método para verificar si un conjunto de datos sigue un proceso estacionario, el cual es conocido como la prueba Dickey Fuller o DF.

La prueba DF se basa en un contraste de hipótesis. La hipótesis nula es que la serie es No Estacionaria, cuando se contrastan las hipótesis se debe calcular un valor estadístico de prueba el cual se compara con un valor crítico en la tabla Dickey Fuller. Cuando el estadístico de prueba es inferior al valor crítico se rechaza la hipótesis nula, es decir los datos provienen de una serie estacionaria. Otra forma de obtener estas conclusiones es mediante el *p*-valor, el cual se interpreta como el valor que representa cuán probable es la hipótesis nula. Si el valor es muy cercano a cero significa que la hipótesis es poco probable y debe ser rechazada (Alonso, 2010).

Para predecir el futuro de una serie temporal se utilizan distintos modelos. Los modelos ARIMA proporcionan otro enfoque para el pronóstico de series de tiempo (Hyndman & Athanasopoulos, 2021). Hay distintos modelos ARIMA como el modelo autorregresivo, modelo de medias móviles, modelo autorregresivo de medias móviles, modelo autorregresivo integrado de medias móviles, modelos con componentes estacionales, entre otros.

En un modelo de autorregresión (AR) se pronostica la variable de interés mediante una combinación lineal de valores pasados de la variable. Autorregresión señala que se trata de una

regresión de una variable contra sí misma (Hyndman & Athanasopoulos, 2021). Un modelo autorregresivo de orden *p* se escribe de la siguiente manera:

$$xt = c + \varphi xt{-}1 + \epsilon t$$

Donde xt es el valor de interés, del periodo actual, c es la constante, φ es el coeficiente que se debe estimar, ϵt es el ruido en el periodo actual, y $xt{-}1$ es el valor de la serie en un periodo anterior.

El modelo de medias móviles (MA) usa errores de pronóstico pasados en un modelo similar a la regresión (Hyndman & Athanasopoulos, 2021). Un modelo MA de orden simple solo consideraría el valor del residuo en el periodo anterior y se expresaría de la siguiente manera:

$$xt = c + \theta\epsilon t{-}1 + \epsilon t$$

Donde xt es el valor de interés, del período actual, c es la constante, θ es el coeficiente a estimar, $\epsilon t{-}1$ es el valor del residuo en el período anterior y ϵt es el residuo en el período actual.

En ese sentido, el modelo autorregresivo de medias móviles (ARMA), combinación de los modelos AR y MA, tiene dos órdenes (*p, q*) donde *p* es el orden de la parte autorregresiva y *q* es el orden de la parte de medias móviles. Cabe resaltar que los modelos ARMA tienen mejor rendimiento si la serie es estacionaria (Hyndman & Athanasopoulos, 2021). Un modelo ARMA de orden simple, ARMA (1,1), se expresa de la siguiente forma:

$$xt = c + \varphi xt{-}1 + \theta\epsilon t{-}1 + \epsilon t$$

Por otro lado, el modelo autorregresivo integrado de medias móviles (ARIMA) resulta de integrar la serie temporal un número de veces determinado. Un modelo ARIMA de orden (*p,d,q*) consiste en integrar la serie original de veces, y luego ajustar el modelo ARMA*(p,q)* a la serie integrada. El objetivo de la integración es obtener una serie estacionaria (Hyndman & Athanasopoulos, 2021). A continuación, en la Tabla 2 se presentan casos especiales del modelo ARIMA.

Tabla 2: *Casos especiales de los modelos ARIMA*

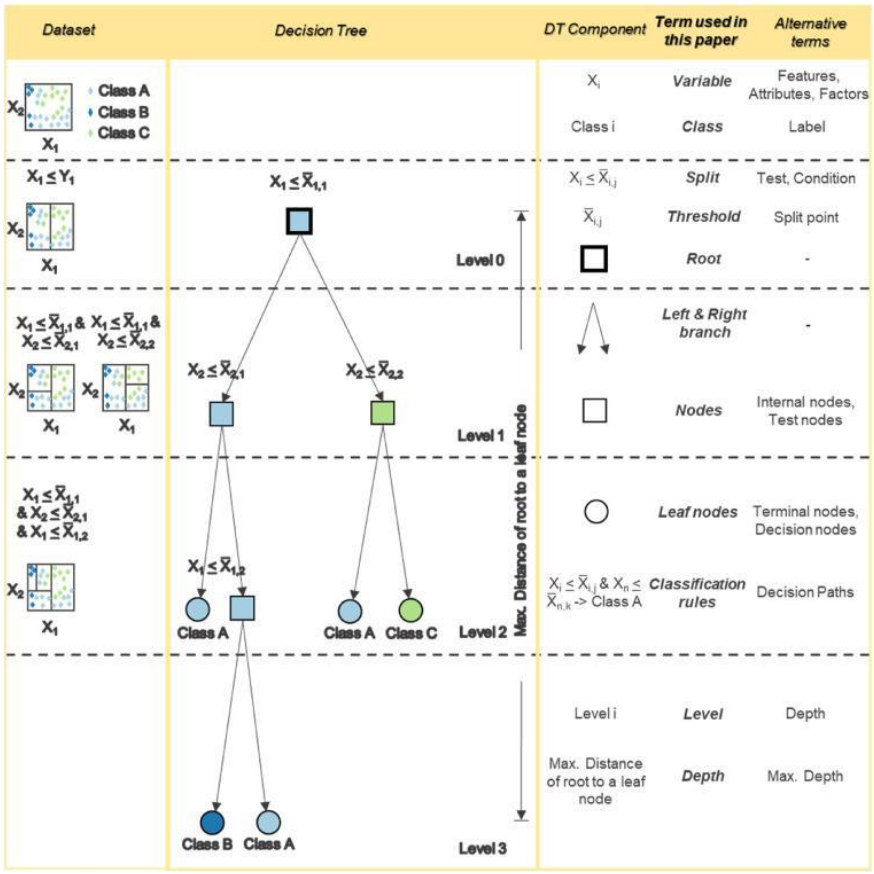
Casos especiales de los modelos ARIMA	
Ruido blanco	ARIMA(0,0,0) sin constante
Caminata aleatoria	ARIMA(0,1,0) sin constante
Caminata aleatoria con deriva	ARIMA(0,1,0) con una constante
Autorregresión	ARIMA(p,0,0)
Media móvil	ARIMA(0,0,q)

Nota. Obtenido de Hyndman & Athanasopoulos (2021)

- ***Árbol de decisión***

La clasificación mediante un árbol de decisión se basa en la idea de un diagrama de flujo con bloques terminales que representan a las decisiones de clasificación. La Figura 17 muestra cómo un algoritmo de árbol de decisión divide el espacio de las variables independientes en subespacios utilizando decisiones jerárquicas. Este modelo está compuesto por nodos y ramas. Cada nodo está ligado a un "split"; es decir a una expresión lógica; además cada nodo conducirá a dos ramas que representan los posibles resultados de la división (Sarailidis et al., 2022). Esta técnica puede ser utilizada para predecir clases para datos no vistos y para predecir valores continuos.

Figura 17: *Desarrollo de un algoritmo de árbol de decisión*

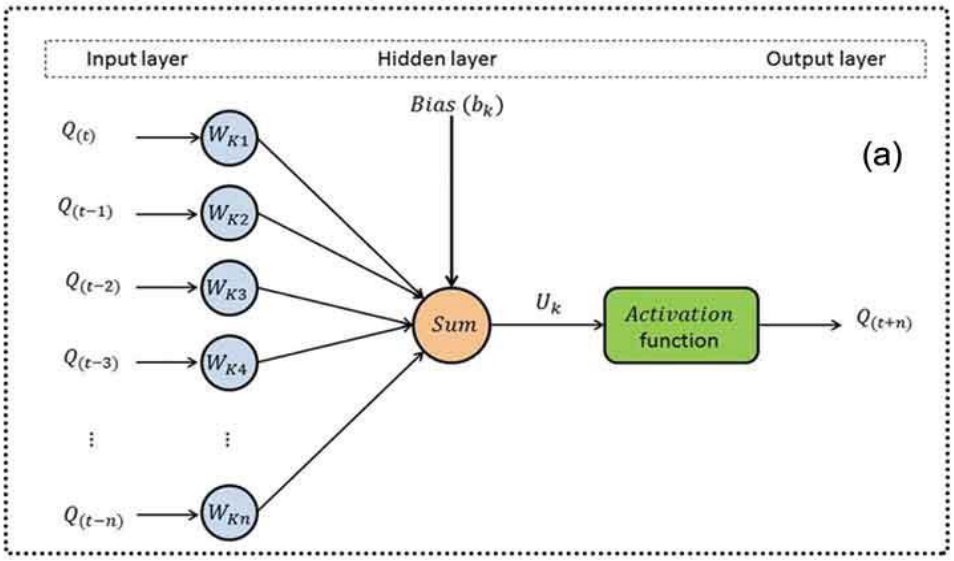


Nota. Obtenido de Sarailidis et al. (2022)

- ***Redes neuronales artificiales***

Esta técnica simula las características de un cerebro para lograr alcanzar una determinada capacidad de procesamiento de información. Una red neuronal es un conjunto de neuronas conectadas que interaccionan para obtener un estímulo de salida (Alba & Gonzáles, 2017). La red neuronal artificial está compuesta por un conjunto de nodos (neuronas) colocados en varias capas. Cada nodo recibe y procesa la entrada del promedio de la capa anterior; asimismo, conduce los nodos de salida a la siguiente capa mediante enlaces (Ver Figura 18). Cada enlace recibe un peso que dependerá del nivel de conexión (Rezaie-Balf et al., 2019).

Figura 18: *Estructura del modelo de red neuronal*



Nota. Obtenido de Rezaie-Balf et al. (2019)

2.2.4 Represas

Una represa se define como un depósito de agua natural o artificial en un punto determinado del flujo de agua de una corriente. De acuerdo con Paranjpye (1994), en los suburbios del este de Mesopotamia, los agricultores de las estribaciones de las montañas Zagros pueden haber construido las represas originarias del mundo hace unos 8000 años. Los sumerios 6500 a.C. edificaron una civilización basada en el riego entre los ríos Tigris y Éufrates.

De acuerdo con la Organización Mundial de la Salud (2001), abastecer de agua a una ciudad cuesta alrededor de $105 por persona; mientras que, en zonas rurales en promedio $50. El uso mundial del agua se ha triplicado desde 1950. Como consecuencia se ha generado una creciente necesidad por la construcción de suministros de agua cada vez mayores, especialmente represas y acueductos. Además, desde el punto de vista ambiental, las represas se utilizan para gestionar los recursos hídricos, mejorar la navegación fluvial y generar electricidad mediante un proceso de transformación mediante el uso de la fuerza de la caída del agua.

2.2.5 Parámetros hidrológicos y meteorológicos

Caudal efluente

En hidrología, se conoce como un efluente o "distributario", a un curso de agua que desde un punto denominado confluencia se separa de un lago o río como una derivación a menor escala y puede ser natural o artificial (Monkhouse, 1978).

Según Aguamarket (s.f.) un caudal efluente puede definirse como el flujo que abandona un curso de agua, embalse, lago, cuenca, formación acuífera, etc. En otras palabras, es el agua que sale de un cuerpo hídrico.

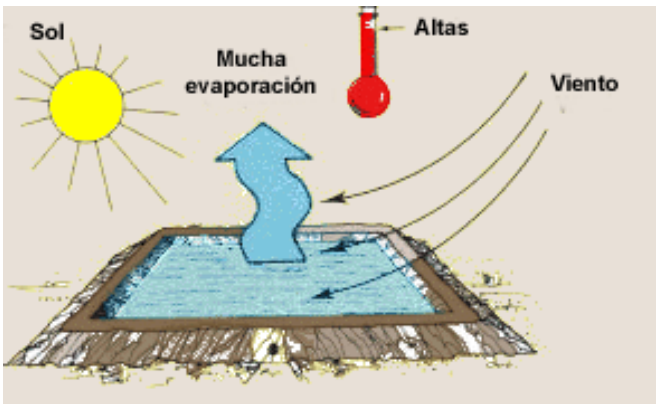
Caudal afluente

En hidrología, un afluente o "tributario" se refiere a una corriente de agua que no llega al mar sino que desemboca en otro río principal y se une a ese río en un punto llamado estuario (Monkhouse, 1978). Del mismo modo, un afluente puede entenderse como un río o corriente de agua que no desemboca en el mar, sino que desemboca en otro río mayor o en otro afluente con el que abastece su caudal como una represa.

Pérdidas por evaporación

Según la Organización de las Naciones Unidas para la Alimentación y la Agricultura (s.f.), la pérdida de agua desde la superficie del reservorio hacia la atmósfera se llama evaporación. La porción de agua que se pierde por evaporación depende mucho del clima local. Las altas temperaturas, la baja humedad, el viento fuerte y la luz solar aumentan la evaporación.

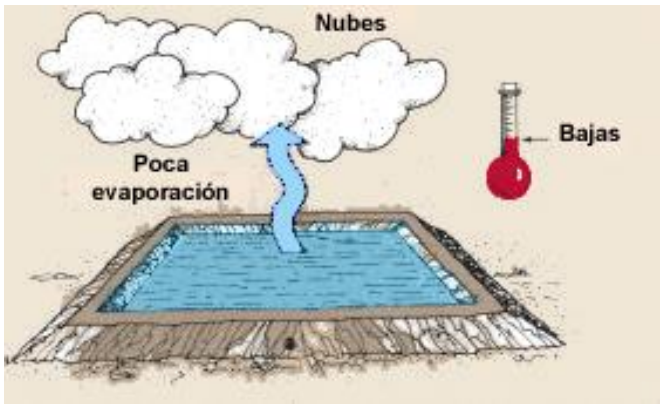
Figura 19*: Modelo de pérdida de evaporación - alta temperatura*



Nota. Obtenido de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (s.f.)

Por el contrario, las bajas temperaturas, la alta humedad, la precipitación y la nubosidad reducen la evaporación.

Figura 20: *Modelo de pérdida de evaporación - baja temperatura*



Nota. Obtenido de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (s.f.)

De acuerdo con Stambouli & Zapata (s.f.), la "pérdida por evaporación" se ve afectada por la humedad relativa, del agua de riego, la velocidad del viento, la altura de las gotas, la temperatura del aire, la presión de funcionamiento y el diámetro de las gotas.

Evaporación

La evaporación es un proceso por el cual la humedad pasa del estado líquido o sólido al gaseoso llamado volatilización. Este proceso es bastante delicado puesto que resulta afectado por un alto número de factores, como lo son: tensión de vapor, temperatura, viento, presión atmosférica, salinidad del agua, naturaleza del suelo (Manzur & Cardoso, 2015).

La evaporación depende del contenido de humedad del suelo, variable según el relieve orográfico y el microrrelieve que puede favorecer la escorrentía o el estancamiento del agua. Por otro lado, cada cultivo filtra el agua de manera diferente. El color del suelo es otro factor importante, así las tierras negras se calientan más y por tanto se evaporan más. A esto hay que añadir la posición de las laderas orográficas ya que las orientadas al sur reciben más radiación solar que las que están al norte. Otra influencia de gran importancia es la época del año (Manzur & Cardoso, 2015).

Precipitación

De acuerdo con Andrades, R., Muñez, L. (2012), la precipitación ocurre cuando una masa de aire se enfría provocando un proceso de condensación o congelación, generando como resultado la formación de pequeñas gotas de agua; estas pequeñas partículas crecen hasta alcanzar un tamaño suficiente y se precipitan o caen.

Según el tamaño de las partículas, su posición y cómo llegan al suelo, se experimentan diferentes tipos de precipitación líquida: llovizna (gotas pequeñas que caen uniformemente), chubasco (gotas grandes y que caen duras y fuertes), entre otras.

Además, como lo señala Rodriguez, B., Portela, L. (2004), la precipitación también puede ocurrir en forma sólida. Su origen es la formación de cristales de hielo en las nubes, cuyas cimas se encuentran a gran altura y a temperaturas extremadamente bajas (-40ºC). Estos cristales crecen a expensas de gotas de agua extremadamente frías que se congelan encima de ellos, o agregando otros cristales para formar copos de nieve.

Al alcanzar cierto tamaño y con la ayuda de la acción de la gravedad, se producen precipitaciones sólidas en la superficie. En ese sentido, Rodriguez, B., Portela, L. (2004) señalan que a veces, el granizo o copos de nieve que se desprenden de las nubes localizan una capa de aire caliente; mientras caen, se derriten antes de llegar al suelo y eventualmente forman una precipitación líquida.

Temperatura

Para Rodriguez, B., Portela, L. (2004), la temperatura es una de las magnitudes más usadas para explicar el estado de la atmósfera. Este parámetro cuantifica el contenido de energía cinética dentro de las partículas que conforman un cuerpo. Además, se sabe que la temperatura del aire oscila por rango diferente tanto el día como en la noche, y también entre una ubicación geográfica y otra. Así mismo, la temperatura es una magnitud asociada con la velocidad del movimiento de las partículas que constituyen la materia. Cuanta mayor agitación presente éstas, mayor será la temperatura.

Por otro lado, Andrades Rodriguez & Muñez León (2012), señalan que: “El calor no es más que una forma de energía susceptible de transformarse en trabajo y la temperatura puede

considerarse como un indicador del nivel de calor de un cuerpo, calor que se transmite desde los cuerpos de más temperatura a los de menos".

Temperatura Mínima

En meteorología, se acostumbra hablar de temperaturas mínimas y máximas, que representan los valores más bajos y más altos registrados durante un período de tiempo, por ejemplo, durante un mes (Rodriguez, B., Portela, L., 2004).

La temperatura más baja registrada en un día determinado se conoce como temperatura mínima. A menudo se reporta por la noche cuando las temperaturas tienden a bajar. Se puede expresar en grados Celsius (ºC), Kelvin (K) o Fahrenheit (°F).

Un termómetro de mínima es un instrumento para medir la temperatura más baja, está hecho de alcohol y tiene una aguja de esmalte en su interior que se sumerge en el líquido. Al aumentar la temperatura, el alcohol se moverá entre la pared del tubo y la flecha, y la flecha no se moverá. Por otro lado, cuando baja la temperatura, el alcohol tira del índice especificado en la dirección opuesta, ya que encuentra una gran resistencia para salir del líquido. Por lo tanto, la posición del índice señala la temperatura más baja alcanzada (Rodriguez, B., Portela, L., 2004).

Temperatura Máxima

La temperatura más alta registrada en un día determinado se llama temperatura máxima. Por lo general, durante el día la temperatura suele ser más alta que durante la noche. Se puede expresar en grados Celsius (ºC), Kelvin (K) o grados Fahrenheit (ºF).

Para medir la temperatura máxima, se utiliza un termómetro de máxima, que incluye un termómetro ordinario, cuyo tubo tiene un inductor en su interior junto al tanque; es así que, a medida que aumenta la temperatura, el mercurio se expande en el tanque. Por el contrario, cuando baja la temperatura y se comprime la masa de mercurio, la columna se rompe, de modo que su extremo libre permanece en la posición más alta que ocupa durante todo el período (Rodriguez, B., Portela, L., 2004).

Capítulo III: Entorno Empresarial

3.1 Descripción de la empresa

Autodema es una institución pública descentralizada del Gobierno Regional de Arequipa encargada de garantizar agua para todos los usos de población arequipeña, principalmente en las Cuencas Quilca – Chili y Colca – Camaná (Autoridad Autónoma de Majes, 2022). Con ello, se garantiza agua de calidad para consumo humano, generación de energía eléctrica, el consumo industrial, agrícola y ganadero fomentando una cultura de uso razonable del agua, una transición productiva a la agroexportación, inversiones privadas y alianzas empresariales para el desarrollo de la Región.

3.1.1 Reseña histórica y actividad económica

El 03 de octubre de 1971 inició el "Proyecto Regional Integral de Desarrollo Agrícola y Energético", conceptualizado como Autodema, cuyo objetivo principal es fortalecer la economía de la región sur del país (Autoridad Autónoma de Majes, 2022). Asimismo, se constituye en la única alternativa factible para disminuir los niveles de pobreza, acrecentar significativamente la producción de alimentos, generar divisas y alcanzar un desarrollo agroindustrial sostenible.

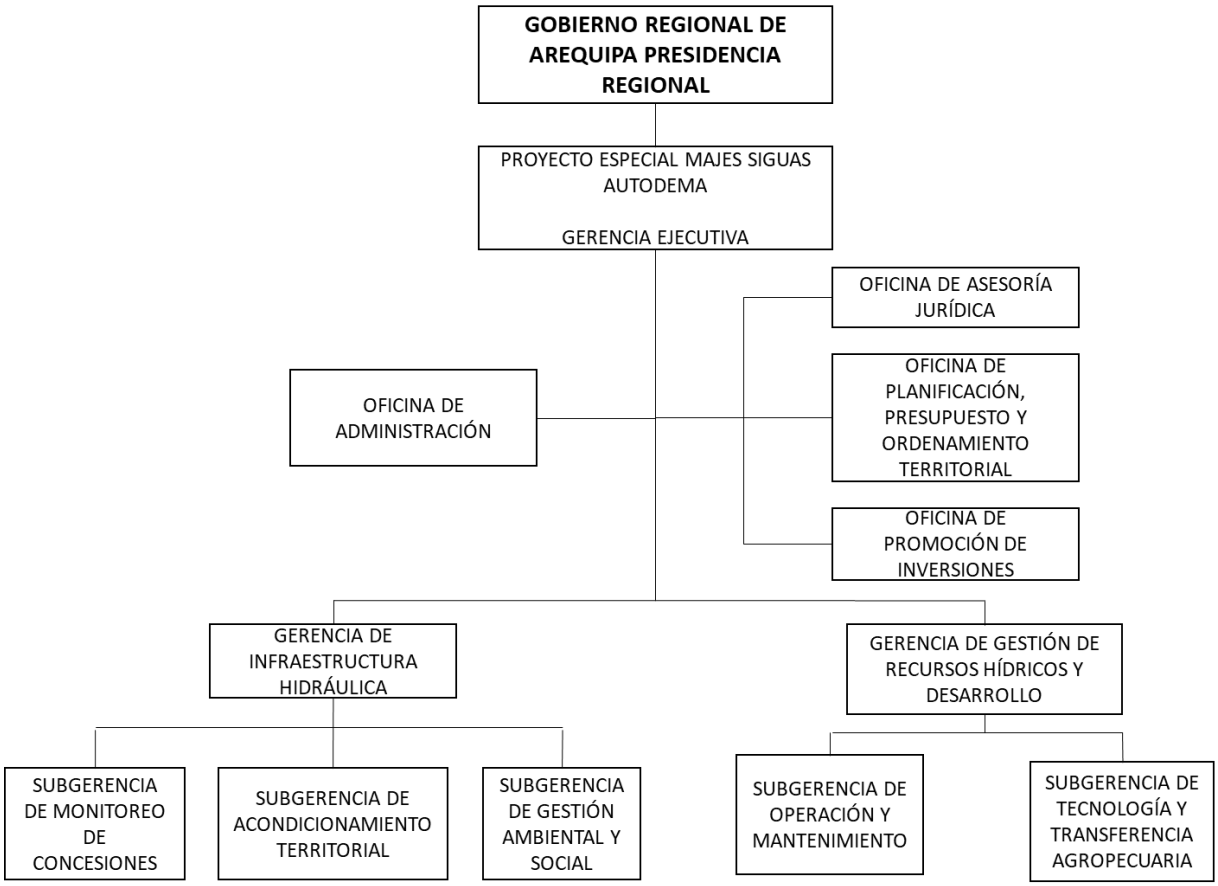
La actividad económica de Autodema es de comercio y servicios puesto que busca brindar y garantizar agua de calidad para las distintas actividades económicas de la región.

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

Autodema cuenta con tres gerencias: Gerencia Ejecutiva, Gerencia de Infraestructura Hidráulica y Gerencia de Gestión de Recursos Hídricos y Desarrollo, éstas a su vez están compuestas por oficinas o subgerencias las cuales se muestran a continuación.

Figura 21: *Organigrama Autoridad Autónoma de Majes (Autodema)*

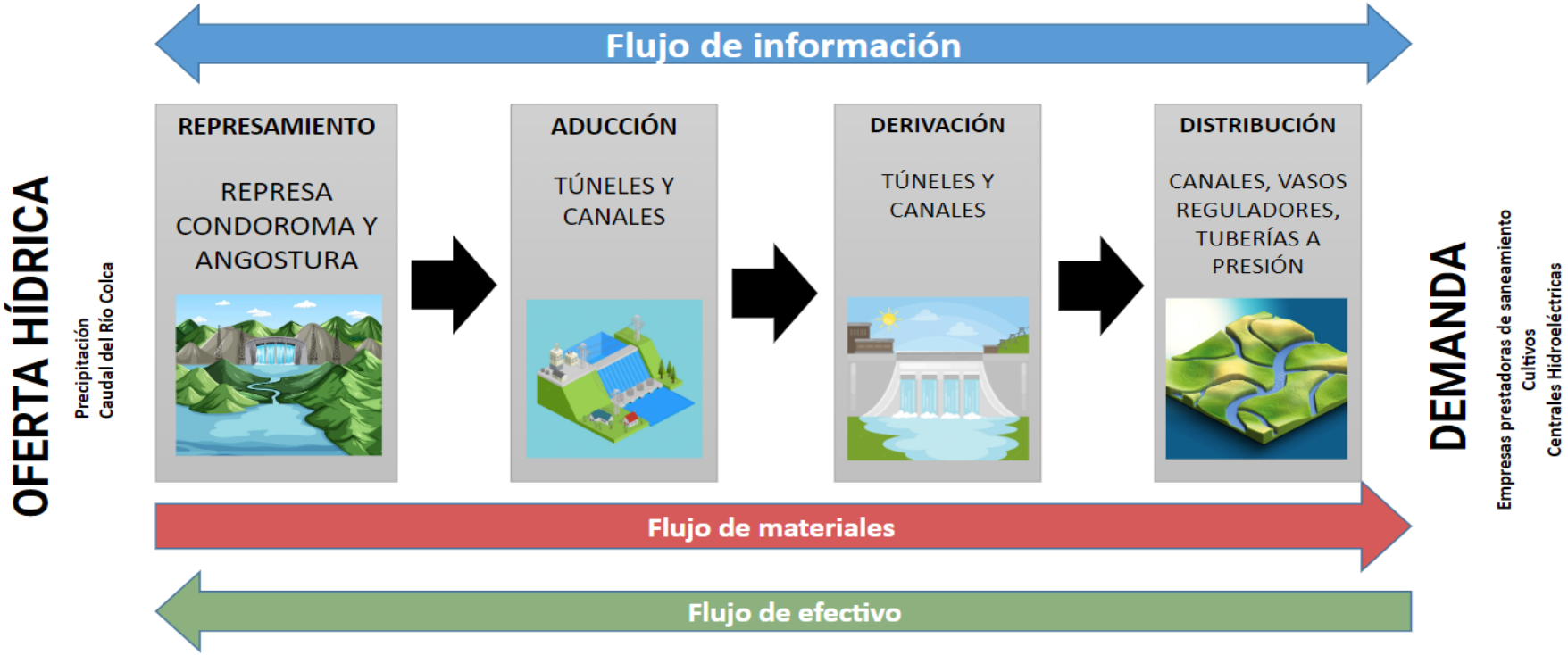


Nota. Obtenido de Autoridad Autónoma de Majes, 2015, p. 9

3.1.2.1 Cadena de suministros

A continuación, se presenta la cadena de suministro de Autodema.

Figura 22*: Cadena de Suministro de Autodema*



Nota. Elaboración propia.

3.1.3 Datos generales estratégicos de la empresa

3.1.3.1 Visión, misión y valores o principios

Visión

Seguridad hídrica para el desarrollo humano sostenible de Arequipa.

Misión

Somos una dependencia del Gobierno Regional de Arequipa que administra el Proyecto Especial Majes-Siguas, asegurando la disponibilidad de agua para los habitantes y las actividades económicas, fomentando una cultura de uso racional del agua, transformando la producción para la exportación de productos agrícolas, inversión privada y cooperación empresarial para el desarrollo de la Región.

Valores

Los valores que Autodema busca compartir con la población de Arequipa es el respeto al medio ambiente, solidaridad de los individuos para usar el agua de manera racional, disciplina para usar solo lo necesario, sabiduría para usar tecnología de vanguardia para reutilizar y disminuir el uso de agua.

3.1.3.2 Objetivos estratégicos

Los objetivos generales y específicos planteados en el Manual de Operaciones de la Autoridad Autónoma de Majes (2015) son los siguientes:

Objetivos Generales

a) Fortalecer el desarrollo socioeconómico de la primera fase del proyecto especial Majes - Siguas y promover y gestionar la ejecución de la segunda fase del proyecto como un proyecto integrado, contribuyendo a la promoción y desarrollo de la zona de Arequipa para promover el desarrollo sostenible en el territorio del Gobierno Regional de Arequipa.

b) Administrar y promover la construcción, operación y mantenimiento de las instalaciones clave que conectan la infraestructura de abastecimiento de agua y riego de los esteros de Majes y Siguas.
c) Facilitar la participación de la Inversión Privada en la construcción y operación de la infraestructura necesaria para el desarrollo de la agricultura, los complejos agroindustriales, la energía y otras industrias de la región.

Objetivos Específicos.

a) Garantizar la seguridad hídrica para el desarrollo sostenible de la región actuando como una institución eficaz y líder en la implementación y gestión de proyectos hidroeléctricos y energéticos.
b) Formalizar, vigilar y controlar de manera efectiva y oportuna las licencias de obras hidroeléctricas, energéticas y otras; además de incentivar la inversión privada en el desarrollo del Proyecto Especial Majes Siguas.
c) Gestión eficiente y sostenible de los sistemas hidráulicos de los ríos Colca y Chile.
d) Gestionar el área del Proyecto Especial Majes Siguas de acuerdo con el Plan de Ordenación del Territorio Armonizado y el Plan de Ordenación del Territorio Económico y Ambiental.
e) Gestionar el desarrollo del riego agrícola y diseñar áreas de influencia basadas en la transformación agrícola, las mejores prácticas agrícolas y las relaciones sociales o comunitarias de manera competitiva e impulsada por el mercado.

3.1.3.3 Evaluación interna y externa. FODA cuantitativo

Tabla 3: *Matriz de Efectos Internos (EFI)*

Factores internos determinantes de éxito	Peso	Calificación	Peso ponderado
Fortalezas			
Cumplimiento de las leyes y regulaciones aplicables.	0.1	4	0.40
Adquisición de nueva tecnología y maquinaria especializada.	0.12	3	0.36
Apoyo a la población de áreas cercanas a los proyectos.	0.05	4	0.20
Promoción de una cultura de uso racional de recursos hídricos.	0.1	4	0.40
Generación de recursos propios por concesiones y adjudicaciones.	0.1	4	0.40
Debilidades			
Daños en infraestructura debido a falta de mantenimiento.	0.12	1	0.12
Poco liderazgo e incidencia para hacer cumplir el plan de ordenamiento territorial en zonas de asentamiento.	0.1	1	0.10
Alta probabilidad de rebose de represas en temporada de lluvia.	0.15	1	0.15
Desarticulación con instituciones estatales.	0.08	2	0.16
Poca eficiencia y eficacia del sistema administrativo.	0.08	2	0.16
Total	**1.00**		**2.45**

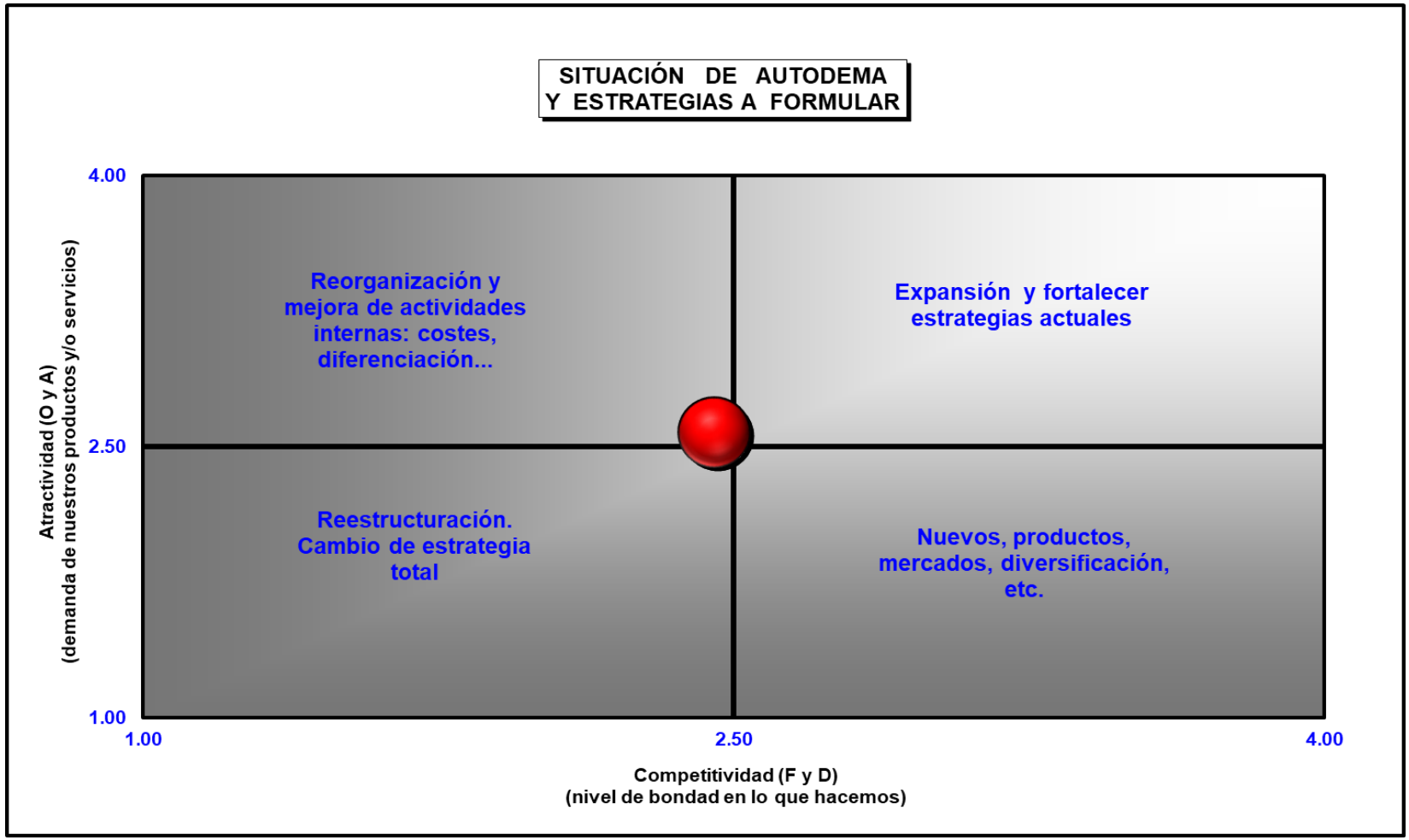
Nota. Elaboración propia.

Tabla 4: *Matriz de Efectos Externos (EFE)*

Factores externos determinantes de éxito	Peso	Calificación	Peso ponderado
Oportunidades			
Apoyo financiero por parte del Estado y concesionarios.	0.09	2	0.18
Potencial agrícola en el departamento de Arequipa debido a condiciones edafo climáticas.	0.08	3	0.24
Iniciativas y movimientos mundiales por la preservación de recursos hídricos.	0.09	2	0.18
Mercado globales para la agro exportación.	0.13	4	0.52
Priorizacion de proyectos de seguridad hídrica y alimentaria.	0.12	4	0.48
Amenazas			
Desconocimiento de los proyectos por parte de la población.	0.10	3	0.30
Alta y mediana probabilidad de ocurrencia de sismos.	0.09	2	0.18
Recurrentes conflictos sociales debido al propuesta de Majes Sihuas II	0.10	1	0.10
Impacto del cambio climático en la disponibilidad de agua.	0.10	2	0.20
Contaminación de fuentes de agua.	0.10	2	0.20
TOTAL	**1.00**		**2.58**

Nota. Elaboración propia.

Figura 23: *Situación interna y externa de Autodema*



Nota. Elaboración propia.

En base al análisis realizado, al desarrollar una matriz cuantitativa de impacto interno, identificamos las fortalezas y debilidades posterior a ellos se calificó cada factor en una escala de 1 a 4; donde 1 es el más débil, 2 es el más débil, 3 es el más débil y 4 es el más fuerte. Por ello, entre las fortalezas más puntuadas, destaca la oportunidad de obtener ingresos propios a través de concesiones y sentencias. Por otro lado, la debilidad destacada es la posibilidad de desbordamiento en las represas gestionadas por Autodema. Todos los años, durante la temporada de lluvias, la capacidad de almacenamiento de la represa alcanza un nivel alto, por lo que el agua que contiene debe drenarse, lo que aumenta el caudal de salida.

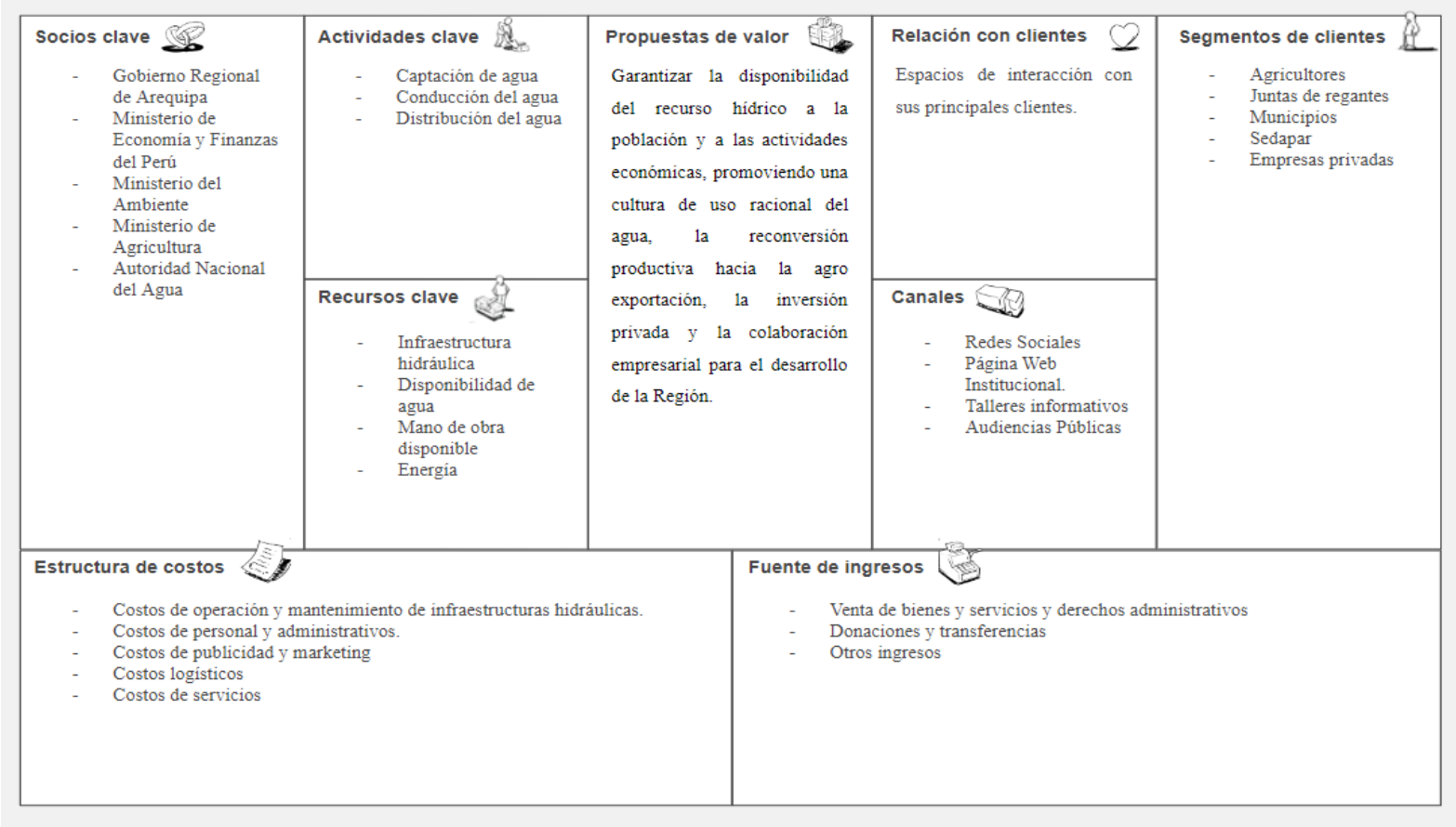
Por otra parte, mediante la matriz cuantitativa de influencias externas, se ha encontrado las oportunidades y amenazas externas que enfrenta la Autodema. Para el análisis, cada factor se puntúa en una escala de 1 a 4; donde 1 - sensibilidad baja al factor, 2 - sensibilidad promedio al factor, 3 - sensibilidad al factor superior al promedio, 4 - sensibilidad alta al factor. Entre las oportunidades identificadas se destaca la expansión del mercado global de exportación agrícola, ya que los principales clientes de Autodema son agricultores y el hecho de que existen oportunidades disponibles para ellos en el mercado que los benefician directamente. Por otro lado, la mayor amenaza es la falta de conocimiento público de los proyectos de Autodema, lo que genera conflicto social y desinformación.

Por último, como se observa en la Figura 23, el estado actual de Autodema de acuerdo al análisis interno y externo, se encuentra en el primer cuadrante, es decir que, la empresa debe encaminarse hacia la mejora de sus actividades internas y hacia una reorganización.

3.2 Modelo de negocio actual (CANVAS)

A continuación, se presenta el CANVAS de Autodema que permitirá conocer mejor su modelo de negocio.

Figura 24: *CANVAS Autodema*

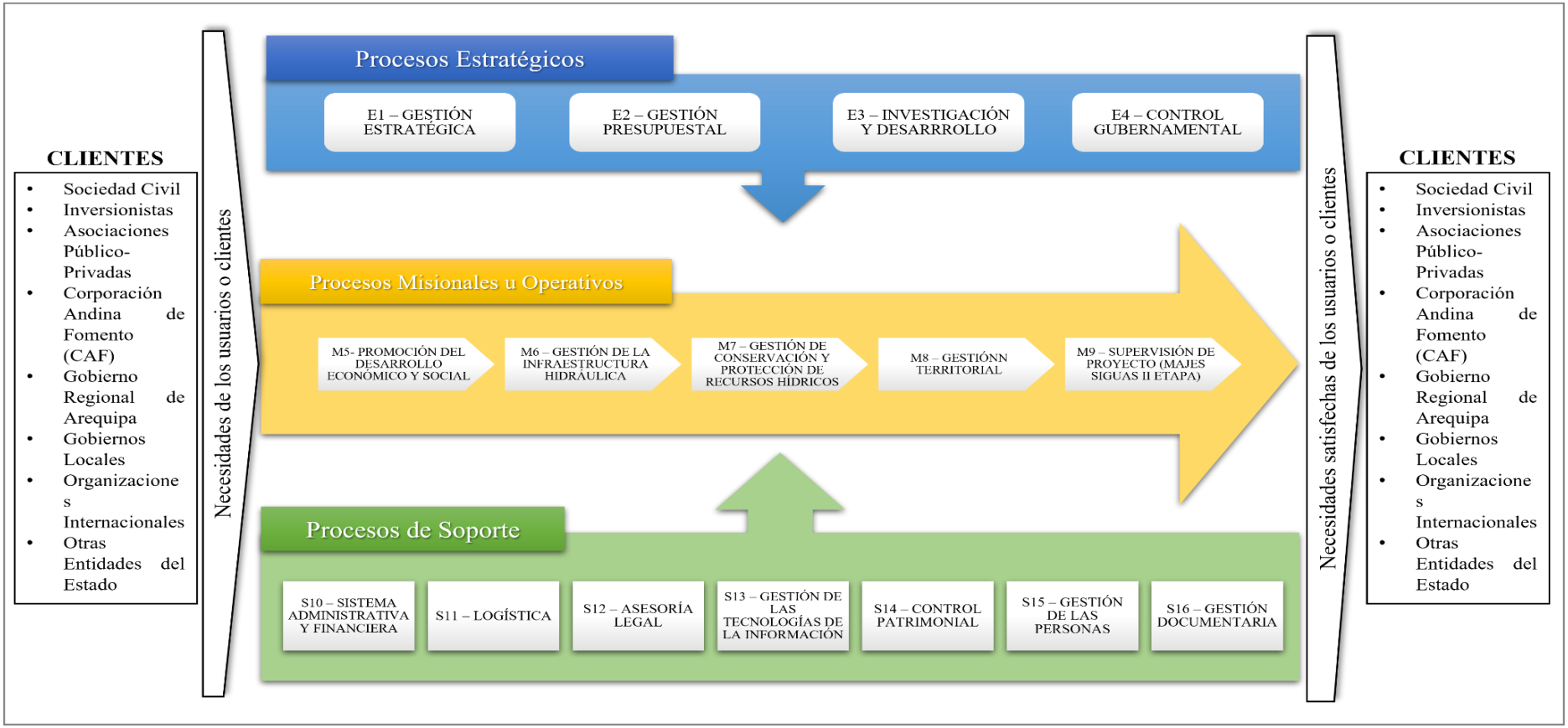


Nota. Elaboración propia.

3.3 Mapa de procesos actual

En la Figura 25, se presenta el mapa actual de procesos de Autodema.

Figura 25: *Mapa de Procesos de Autodema*



Nota. Obtenido de Autodema (2022)

Capítulo IV: Metodología de la Investigación

En este capítulo se establece el enfoque, alcance y diseño de investigación. De igual modo, se indica la metodología a seguir y el cronograma de actividades.

4.1 Diseño de la Investigación

Para alcanzar los objetivos de investigación propuestos, se deben presentar opciones prácticas y específicas que den respuesta a las preguntas de investigación. Un diseño de investigación es un plan o estrategia para alcanzar la información deseada. (Hernández & Mendoza, 2018).

4.1.1 Enfoque de la investigación

El enfoque de la presente investigación es cuantitativo, debido a que está orientado a la recolección y análisis de datos numéricos para predecir escenarios con el uso de la técnica de Machine Learning. Así mismo, se mide y analiza las variables para extraer conclusiones.

4.1.2 Alcance de la investigación

El alcance de un estudio cuantitativo puede ser: descriptivo, exploratorio, correlacional y explicativo; cabe señalar que, estos no son excluyentes entre sí, es decir, un mismo estudio puede abarcar uno o más tipos de alcances (Hernández & Mendoza, 2018).

El alcance de esta investigación es correlacional porque tiene como propósito determinar la relación entre las variables independientes: nivel de embalse, volumen útil, caudal afluente, pérdidas por evapotranspiración, evaporación, precipitación, temperatura mínima y temperatura máxima para predecir el caudal efluente (variable dependiente).

4.1.3 Tipo de investigación

Según Hernandez & Mendoza (2018), la investigación que emplea métodos cuantitativos puede utilizar dos categorías de diseño: experimental y no experimental.

El diseño experimental se divide en pre experimental, cuasi experimental y experimental puro. En esa misma línea, el diseño no experimental puede ser horizontal o vertical.

Esta investigación es de tipo experimental, puesto que se mide las variables independientes, mencionadas anteriormente, para ver su efecto en la variable dependiente.

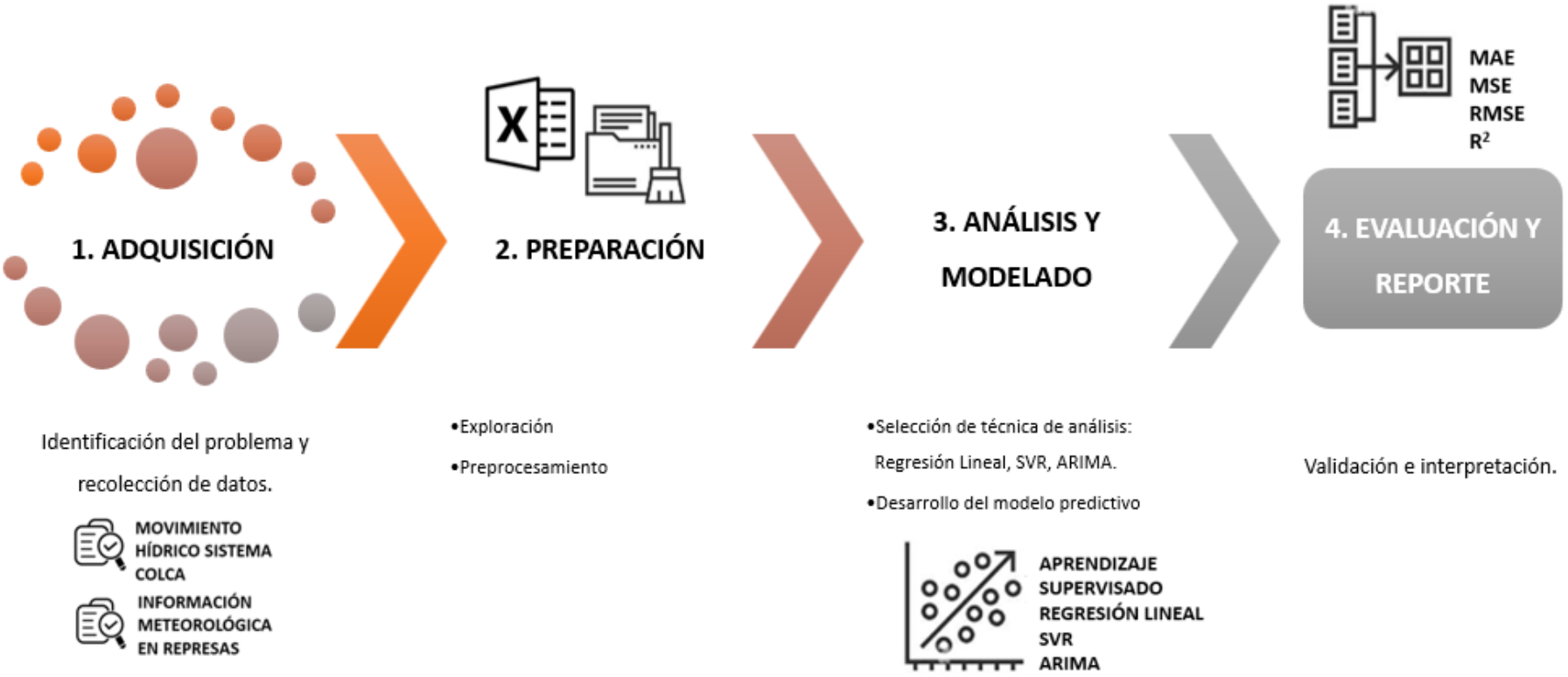
4.1.4 Población y muestra

Población: Base de datos Autodema del Movimiento Hídrico del Sistema Colca, el cual comprende 04 represas: Condoroma, Bocatoma Tuti, Bocatoma Pitay y Bocatoma Santa Rita.

Muestra: Base de datos de la represa Condoroma desde diciembre de 2009 hasta febrero de 2023.

4.2 Metodología de implementación de la solución

Figura 26: *Diseño de la implementación*



Nota. Elaboración propia.

4.2.1 Adquisición de datos

En esta primera etapa se reconoce la problemática que existe respecto a la operación y gestión de riesgos que hay en la represa Condoroma. Posteriormente, se recolectan datos e información de dos plataformas de Autodema: Movimiento Hídrico Sistema Colca e Información Meteorológica en Represas.

4.2.2 Preparación

En esta etapa, se analiza la calidad de los datos identificando si es necesario realizar una limpieza de datos, o si se detectan anomalías como datos duplicados, datos ruidosos que estén fuera del rango de valores esperados y/o datos desconocidos que no corresponden a un valor real. Si se detecta alguno de estos casos, se reemplaza por el valor más cercano usando métricas como media, moda, mínimo y máximo.

4.2.3 Análisis

Se aplican técnicas de Machine Learning para un modelo supervisado, ya que se cuenta con variables predictoras y de respuesta. Se desarrollan modelos mediante las técnicas de Regresión Lineal, SVR y ARIMA, con el fin de predecir los valores futuros de la variable objetivo caudal efluente usando variables independientes.

4.2.4 Reporte

Se interpreta los modelos de acuerdo al conocimiento del dominio y se validan mediante distintas métricas estadísticas para comprobar su funcionamiento y de esta manera determinar si los modelos son confiables para predecir el caudal efluente en la represa Condoroma.

4.3 Metodología para la medición de resultados de la implementación

Para la evaluación del resultado de la aplicación de los modelos desarrollados, se emplean cuatro métricas estadísticas: Mean Squared Error (MSE) , Mean Absolute Error (MAE), Root Mean Square Error (RMSE) y Varianza (R^2).

- Mean Squared Error (MSE): En español error cuadrático medio, es una medida de la diferencia cuadrática media entre los valores original y esperado en un conjunto de datos. Es decir, mide la varianza de los residuos. El valor resultante deriva de la fórmula presentada a continuación:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

- Mean Absolute Error (MAE): En español error absoluto medio, es la diferencia absoluta media entre los valores reales y predichos en el conjunto de datos. Mide la media de los residuos en el conjunto de datos. Esta métrica se calcula mediante la siguiente fórmula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

- Error de raíz cuadrada media (RMSE) es la desviación estándar de los valores residuales (errores de predicción). Los valores residuales son una medida de la distancia de los puntos de datos de la línea de regresión; RMSE me da como resultado la medida de cuál es el nivel de dispersión de estos valores residuales. Se expresa mediante la siguiente ecuación:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y})^2}{N}}$$

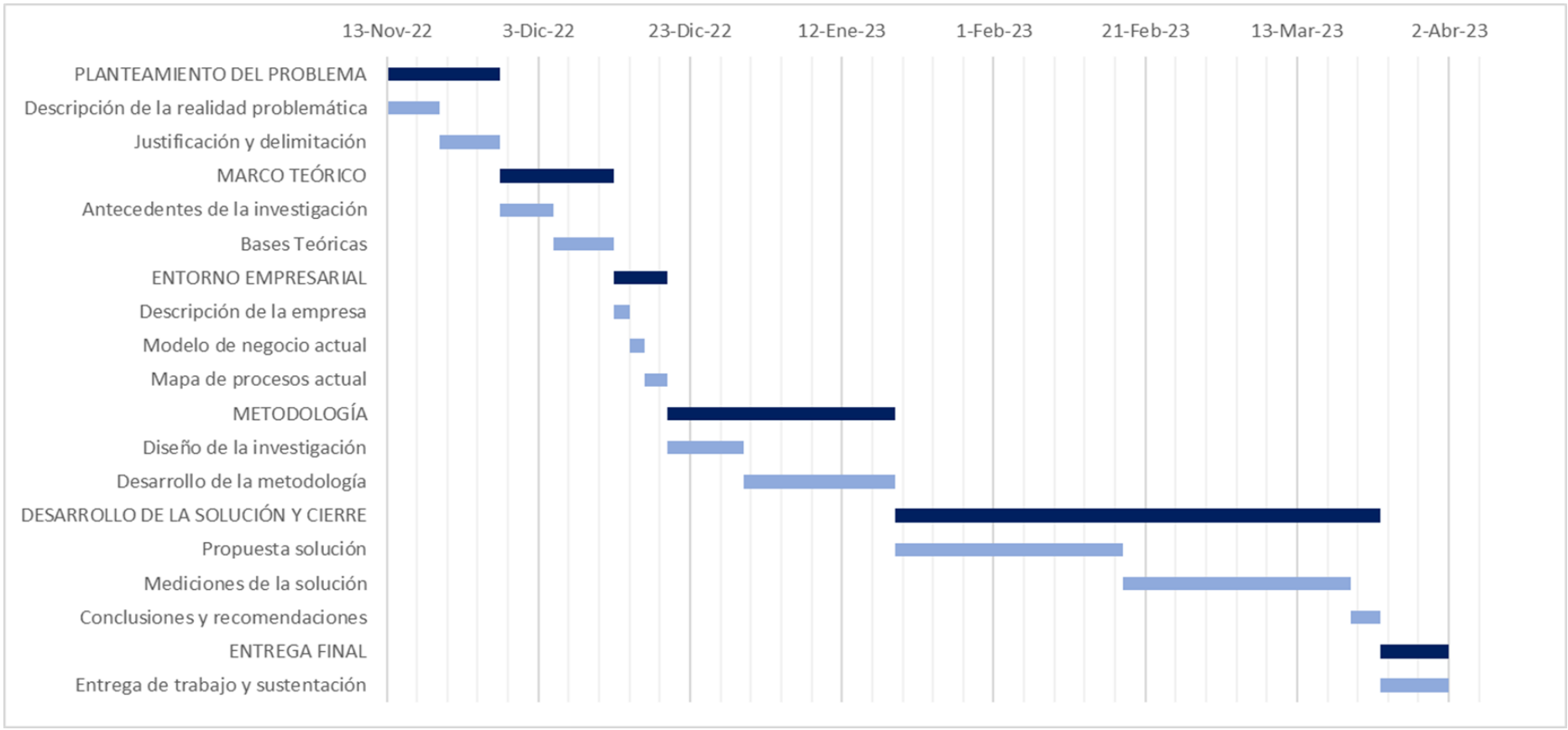
- Varianza (R^2): Representa la proporción de varianza de la variable dependiente explicada por el modelo de regresión lineal. Esta es una estimación sin escala, lo que significa que no importa cuán pequeños o grandes sean los valores, el valor de R cuadrado será menor que uno.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \underline{y})^2}$$

4.4 Cronograma de actividades y presupuesto

4.4.1 Cronograma

Figura 27: *Cronograma de actividades*



Nota. Elaboración propia.

4.4.2 Presupuesto

Tabla 5: *Presupuesto de Actividades*

Recursos	Unidad de medida	Cantidad	Costos Unitario	Costo Total
Herramienta tecnológica (Laptop)	Unidad	3	S/.3.000,00	S/.9.000,00
Software (Google Meet, Microsoft Office, Jupyter Notebook)	Unidad	3	S/.0,00	S/.0,00
Servicios de Internet	Mes	3	S/.120,00	S/.360,00
Energía eléctrica	Mes	3	S/.60,00	S/.180,00
			TOTAL DE RECURSOS	**S/.9.540,00**

Nota. Elaboración propia.

Capítulo V: Desarrollo de la Solución

5.1 Propuesta solución.

5.1.1 Planteamiento y descripción de Actividades

Mediante técnicas de Machine Learning de aprendizaje supervisado se buscó predecir el caudal efluente de la represa Condoroma respecto a nueve parámetros medidos por Autodema: Nivel de embalse (m.s.n.m), volumen útil (m^3), caudal afluente (m^3/s), pérdidas por evaporación (m^3), evaporación (m.m.), precipitación (m.m.), temperatura máxima (°C), temperatura mínima (°C) y fecha diaria. Para ello, se trabajó con datos diarios desde diciembre de 2009 hasta febrero de 2023 y se ejecutaron las técnicas Regresión Lineal, SVR y Series de Tiempo - ARIMA, haciendo uso del lenguaje de programación Python.

5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución

5.1.2.1 Adquisición de datos

Se inició esta etapa descargando los datos de dos Plataformas de Autodema: i) Movimiento Hídrico Sistema Colca (Ver Figura 28), para los parámetros nivel de embalse, volumen útil, caudal afluente, caudal efluente, pérdidas por evaporación, evaporación y precipitación y ii) Meteorología Represas (Ver Figura 29) para obtener los datos de los parámetros temperatura máxima y temperatura mínima.

Figura 28: *Plataforma Movimiento Hídrico Cuenca Regulado*

Seleccionar Fecha de Información : Diciembre 2022

1 / 1 Main Report 100% BusinessObjects

Reporte al: 09/12/2022

GOBIERNO REGIONAL DE AREQUIPA
PROYECTO ESPECIAL MAJES - SIGUAS
GERENCIA DE GESTIÓN DE RECURSOS HÍDRICOS
SUB GERENCIA DE OPERACIÓN Y MANTENIMIENTO

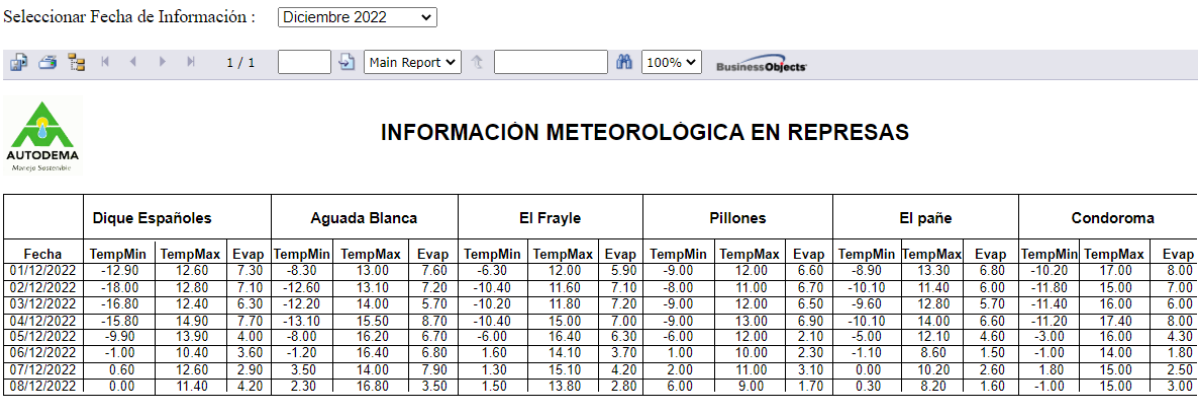
MOVIMIENTO HÍDRICO SISTEMA COLCA
diciembre 2022.

AUTODEMA Manejo Sostenible

Fecha	REPRESA CONDOROMA Vol. Util Max 259 hm³							BOCATOMA TUTI Max. Caudal Circulante 1300 m³/s					BOCATOMA PITAY			BOCATOMA SANTA RITA
	Nivel de Embalse (m.s.n.m)	Volumen Util (m³)	Caudal Afluente (m³/s)	Caudal Efluente (m3/s)	Perdidas evapora. (m³)	Evapora-ción (m.m.)	Precipita-ción (m.m.)	Aporte C. Inter. (m³/s)	Caudal Recibido (m³/s)	Caudal canal 2 (m³/s)	Caudal A.Abajo (m³/s)	Precipi-tación (m.m.)	Caudal Recib. (m³/s)	Caudal Irrigac. (m³/s)	Aguas Abajo (m³/s)	Caudal Recib. (m³/s)
01/12/2022	4,129.25	83,758,667	0.84	13.69	41,252.00	8.00	0.00	1.27	14.99	14.94	0.05	0.00	13.80	11.91	1.90	1.87
02/12/2022	4,129.04	82,607,059	0.86	13.67	35,817.00	7.00	0.00	1.30	14.99	14.94	0.05	0.00	13.83	11.93	1.90	1.90
03/12/2022	4,128.84	81,464,313	0.61	13.58	30,481.00	6.00	0.00	1.32	14.99	14.94	0.05	0.00	13.88	11.98	1.90	1.89
04/12/2022	4,128.64	80,290,306	0.85	13.62	40,348.00	8.00	0.00	1.41	14.99	14.94	0.05	0.00	13.87	11.96	1.91	1.80
05/12/2022	4,128.44	79,169,612	0.65	13.60	21,520.00	4.30	4.30	1.37	14.99	14.94	0.05	0.10	13.87	11.97	1.90	1.83
06/12/2022	4,128.23	78,029,170	0.24	13.57	23,854.00	1.80	1.80	1.39	14.99	14.94	0.05	0.10	13.80	11.90	1.90	1.86
07/12/2022	4,128.02	76,853,058	1.01	13.55	12,332.00	2.50	8.00	1.48	15.05	14.94	0.12	22.10	13.94	12.03	1.91	1.86
08/12/2022	4,127.82	75,756,894	1.73	13.53	14,695.00	3.00	19.50	1.83	15.38	14.94	0.44	22.30	14.18	12.25	1.94	1.88
09/12/2022	4,127.63	74,722,835														
Promedio	———	———	0.85	13.60	27,537.38	5.08	4.20	1.42	15.04	14.94	0.11	5.58	13.90	11.99	1.91	1.86
Total	———	74,722,835	585,619	9,401,443	220,299.00	40.60	33.60	981,124	10,398,637	10,324,593	74,045	44.60	9,605,487	8,288,247	1,317,240	14.88

Nota. Obtenido de Plataforma Movimiento Hídrico Cuenca Regulado (2023).

Figura 29: *Plataforma Meteorología Represas*



Seleccionar Fecha de Información : Diciembre 2022

1 / 1 Main Report 100%

AUTODEMA

INFORMACIÓN METEOROLÓGICA EN REPRESAS

	Dique Españoles			Aguada Blanca			El Frayle			Pillones			El pañe			Condoroma		
Fecha	TempMin	TempMax	Evap	TempMin	TempMax	Evap	TempMin	TempMax	Evap	TempMin	TempMax	Evap	TempMin	TempMax	Evap	TempMin	TempMax	Evap
01/12/2022	-12.90	12.60	7.30	-8.30	13.00	7.60	-6.30	12.00	5.90	-9.00	12.00	6.60	-8.90	13.30	6.80	-10.20	17.00	8.00
02/12/2022	-18.00	12.80	7.10	-12.60	13.10	7.20	-10.40	11.60	7.10	-8.00	11.00	6.70	-10.10	11.40	6.00	-11.80	15.00	7.00
03/12/2022	-16.80	12.40	6.30	-12.20	14.00	5.70	-10.20	11.80	7.20	-9.00	12.00	6.50	-9.60	12.80	5.70	-11.40	16.00	6.00
04/12/2022	-15.80	14.90	7.70	-13.10	15.50	8.70	-10.40	15.00	7.00	-9.00	13.00	6.90	-10.10	14.00	6.60	-11.20	17.40	8.00
05/12/2022	-9.90	13.90	4.00	-8.00	16.20	6.70	-6.00	16.40	6.30	-6.00	12.00	2.10	-5.00	12.10	4.60	-3.00	16.00	4.30
06/12/2022	-1.00	10.40	3.60	-1.20	16.40	6.80	1.60	14.10	3.70	1.00	10.00	2.30	-1.10	8.60	1.50	-1.00	14.00	1.80
07/12/2022	0.60	12.60	2.90	3.50	14.00	7.90	1.30	15.10	4.20	2.00	11.00	3.10	0.00	10.20	2.60	1.80	15.00	2.50
08/12/2022	0.00	11.40	4.20	2.30	16.80	3.50	1.50	13.80	2.80	6.00	9.00	1.70	0.30	8.20	1.60	-1.00	15.00	3.00

Nota. Obtenido de Plataforma Meteorología Represas (2023).

Cabe destacar que cada plataforma generaba reportes a nivel mensual y los datos eran descargados en archivos excel. En ese sentido, se obtuvieron 318 hojas de cálculo correspondientes a cada mes desde diciembre de 2009 hasta febrero de 2023.

5.1.2.2 Preparación de datos

Las hojas de cálculo obtenidas de las bases de datos de Autodema contenían información de distintas represas, por lo que primero se eliminó la información de cualquier represa diferente a Condoroma y luego se unió los datos correspondientes a esta última en una sola hoja de cálculo como se puede ver en la Figura 30.

Figura 30: *Recopilación de datos descargados en una hoja de cálculo*

REPRESA CONDOROMA

Fecha	Caudal_Efluente	Caudal_Afluente	Nivel_de_Embalse	Volumen_util	Perdidas_evapora	Evaporacion	Precipitacion	TempMin	TempMax
01/12/2009	9.87	8.51	4,123.75	54,990,662	25,875.00	6.00	0.00	3.20	17.80
02/12/2009	9.97	8.49	4,123.72	54,846,913	19,361.00	4.50	0.00	-1.80	16.00
03/12/2009	8.90	3.98	4,123.63	54,440,565	12,862.00	3.00	0.00	-1.00	0.00
04/12/2009	8.89	3.23	4,123.54	54,002,320	27,754.00	6.50	0.00	-3.00	18.00
05/12/2009	9.44	2.00	4,123.43	53,485,198	31,861.00	7.50	0.00	-2.80	20.00
06/12/2009	9.89	1.24	4,123.29	52,810,225	21,869.00	7.00	0.00	-3.00	20.00
07/12/2009	9.86	0.86	4,123.13	52,040,939	31,445.00	7.50	0.00	-3.20	20.50
...									
21/02/2023	0.00	10.27	4,135.25	121,289,174	21,635.00	3.50	0.00	1.80	14.20
22/02/2023	0.00	8.84	4,135.38	122,154,576	26,056.00	4.20	0.20	2.00	15.00
23/02/2023	0.00	9.05	4,135.48	122,892,133	16,286.00	2.60	3.60	0.80	16.60
24/02/2023	0.00	8.00	4,235.59	123,657,736	20,009.00	3.20	0.20	0.50	14.80
25/02/2023	0.00	16.70	4,235.70	124,387,223	18,848.00	3.00	1.50	-0.60	16.00
26/02/2023	0.00	8.40	4,135.90	125,811,273	23,986.00	3.80	0.80	-0.40	16.80

Nota. Obtenido de Autodema (2023)

Una vez elaborada la base de datos se procedió a identificar la naturaleza de las variables, especialmente la variable dependiente utilizando el lenguaje de programación Python. En este caso, la variable dependiente caudal efluente, es numérica, por lo que correspondía aplicar el tipo de aprendizaje supervisado - regresión. Por otro lado, se realizó un análisis estadístico por cada variable mediante la función *describe* para determinar si había datos faltantes. En ese sentido, como se aprecia en la Figura 31, todas las variables tienen 4836 datos por lo que no fue necesario realizar procedimientos adicionales para completar la base de datos.

Figura 31: *Análisis de datos faltantes*

```
In [4]:
1 ###Análisis estadísticos de los datos
2 datos.describe(include='all')
```

Out[4]:

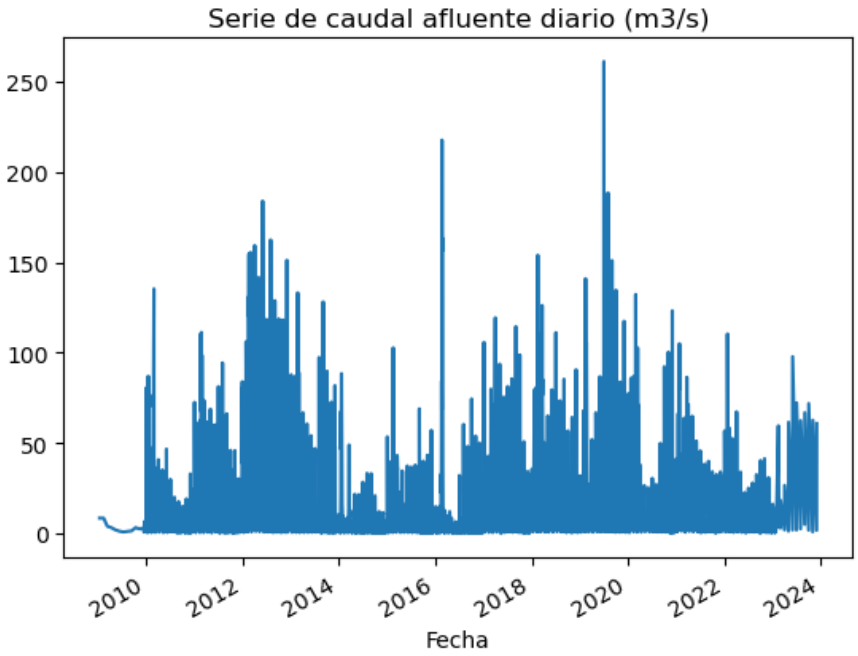
	Fecha	Caudal_Efluente	Caudal_Afluente	Nivel_de_Embalse	Volumen_util	Perdidas_evapora	Evaporacion	Precipitacion	TempMin	TempMax
count	4836	4836.000000	4836.000000	4836.000000	4.836000e+03	4836.000000	4836.000000	4836.000000	4836.000000	4836.000000
unique	4836	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	01/12/2009	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	10.416579	10.963287	4141.219916	1.748907e+08	30794.043011	4.192928	2.021588	-2.412864	15.300116
std	NaN	13.252188	22.330695	8.633883	6.662916e+07	10421.726916	1.537985	4.828726	3.837808	2.329144
min	NaN	0.000000	0.040000	4116.860000	2.636843e+07	4285.000000	0.000000	0.000000	-12.600000	5.500000
25%	NaN	5.190500	1.123000	4135.082250	1.202275e+08	23438.500000	3.000000	0.000000	-5.500000	13.800000
50%	NaN	8.774000	1.958500	4143.266000	1.835335e+08	30946.000000	4.000000	0.000000	-2.000000	15.200000
75%	NaN	11.914250	8.555750	4149.064750	2.386209e+08	38276.500000	5.000000	1.200000	0.925000	16.800000
max	NaN	123.494000	261.168000	4151.763000	2.673119e+08	62231.000000	16.600000	53.300000	8.000000	21.800000

Nota. Elaboración propia.

5.1.2.3 Análisis de datos

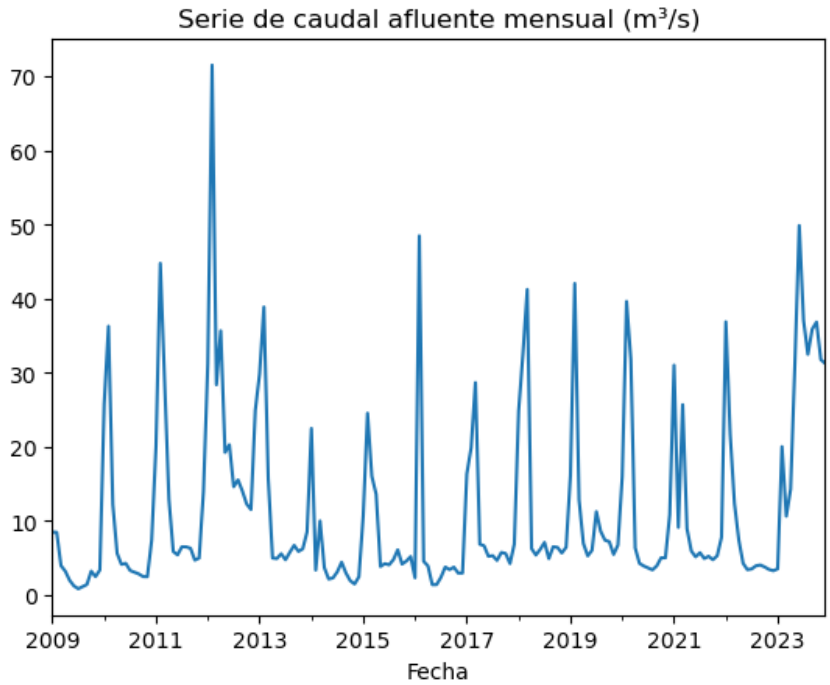
Se inició el análisis de datos graficando las variables independientes, con el fin de conocer su comportamiento a lo largo del tiempo.

Figura 32: *Caudal afluente diario (m^3/s)*



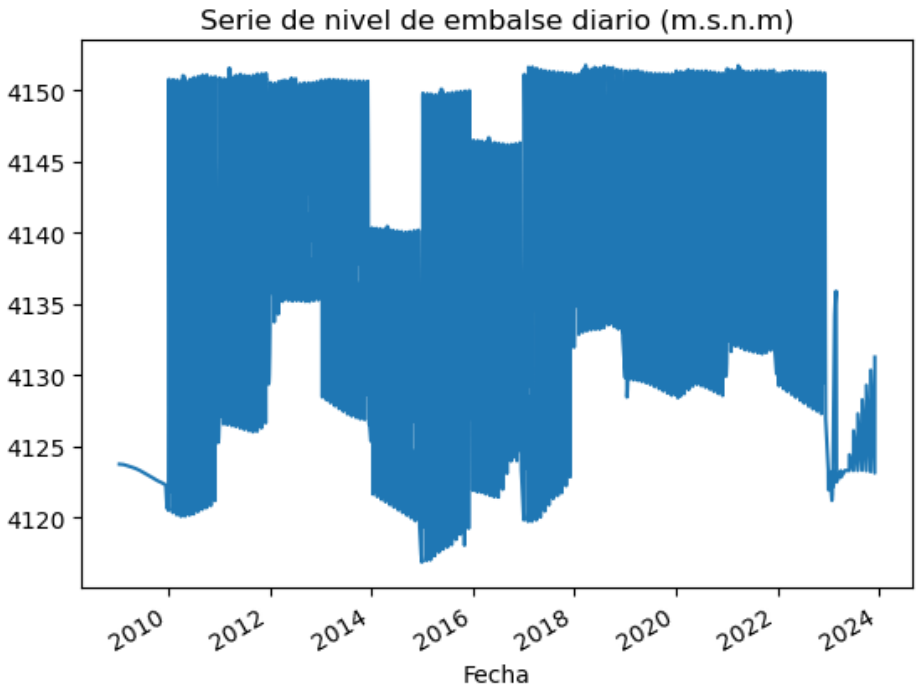
Nota. Elaboración propia.

Figura 33: *Caudal afluente promedio mensual (m^3/s)*



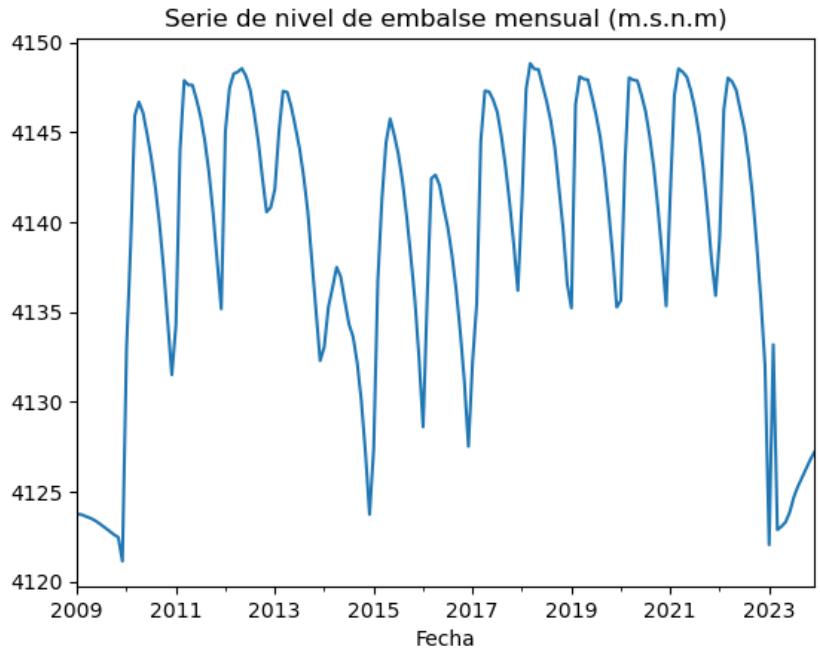
Nota. Elaboración propia.

Figura 34: *Nivel de embalse diario (m.s.n.m)*



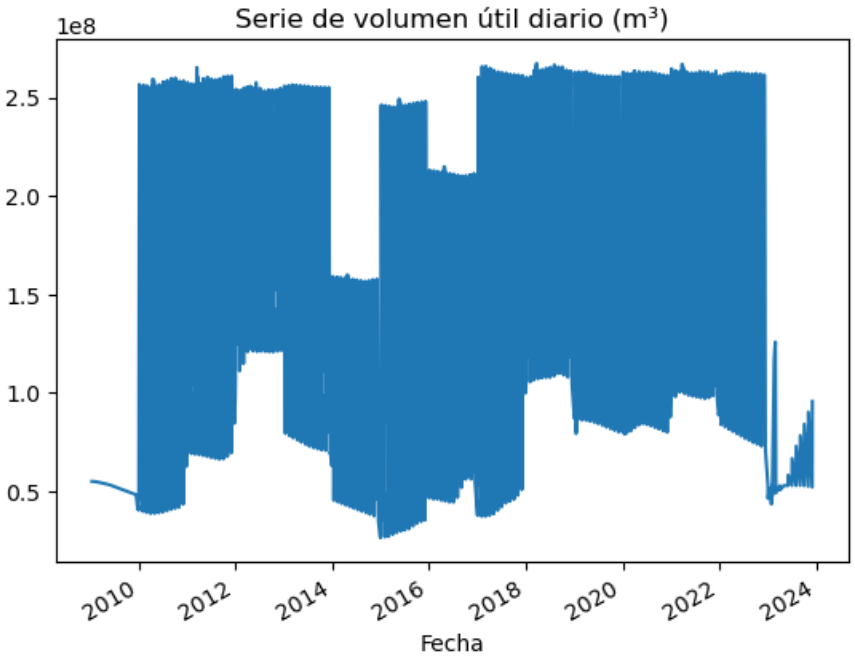
Nota. Elaboración propia.

Figura 35: *Nivel de embalse promedio mensual (m.s.n.m)*



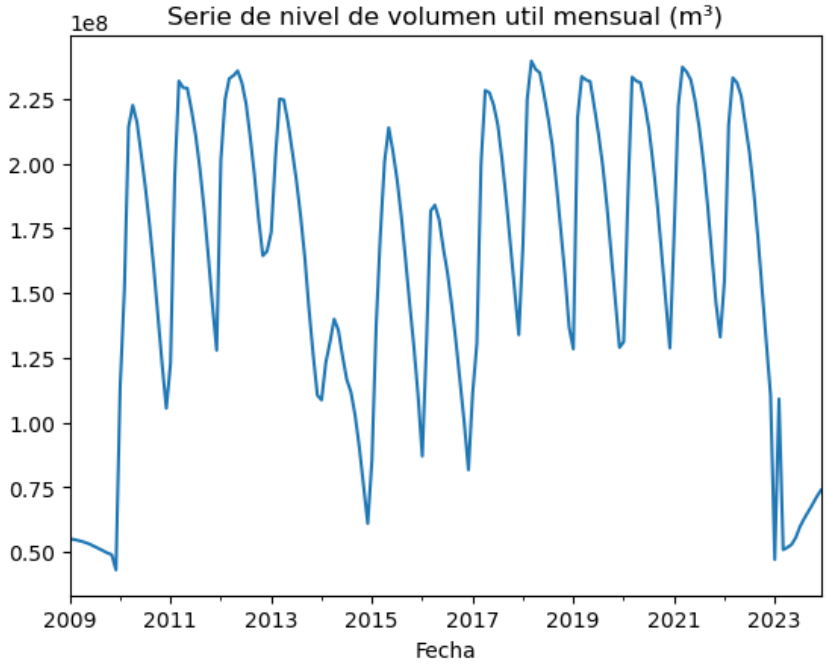
Nota. Elaboración propia.

Figura 36: *Volumen útil diario (m^3)*



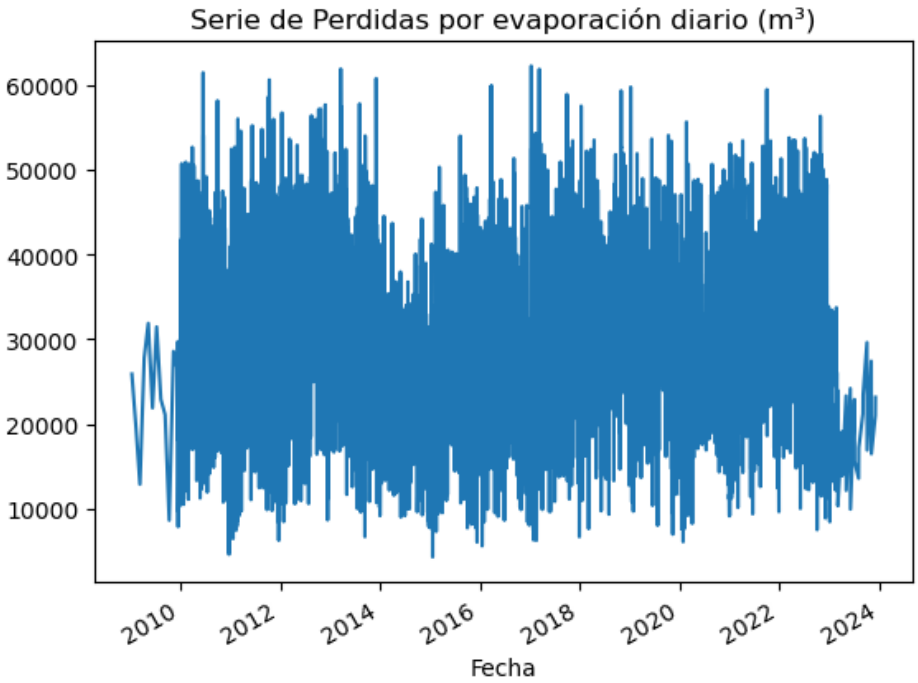
Nota. Elaboración propia.

Figura 37: *Volumen útil promedio mensual (m^3)*



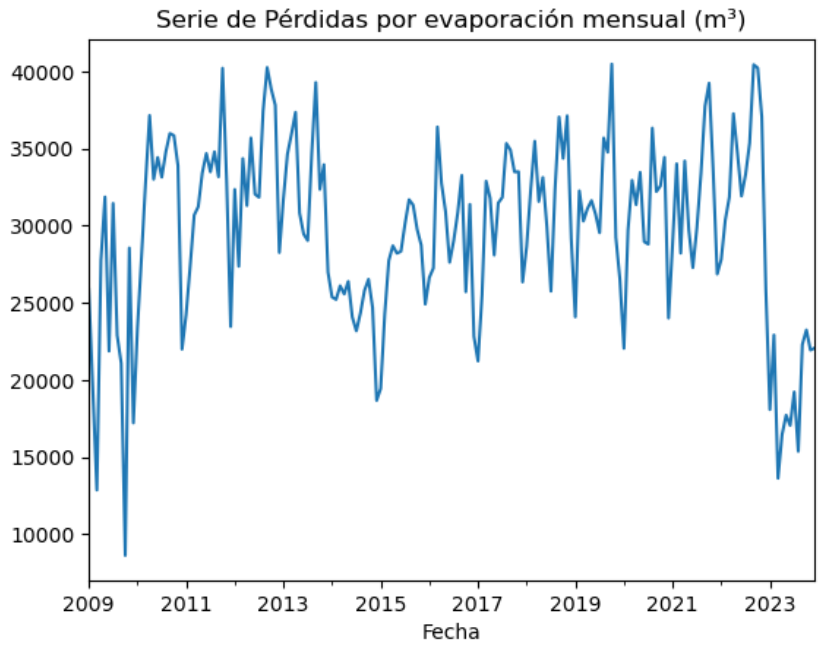
Nota. Elaboración propia.

Figura 38: *Pérdidas por evaporación diario (m^3)*



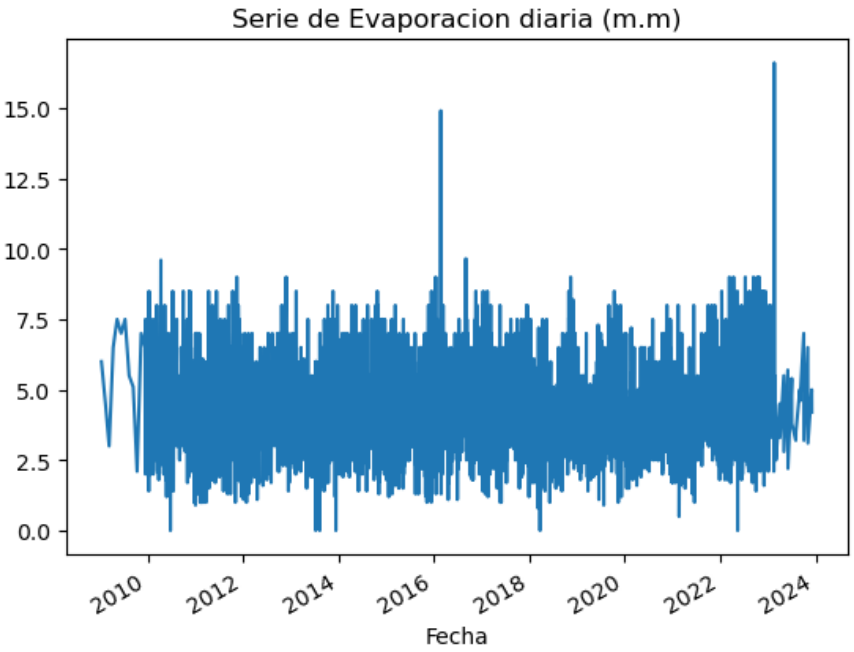
Nota. Elaboración propia.

Figura 39: *Pérdidas por evaporación promedio mensual (m^3)*



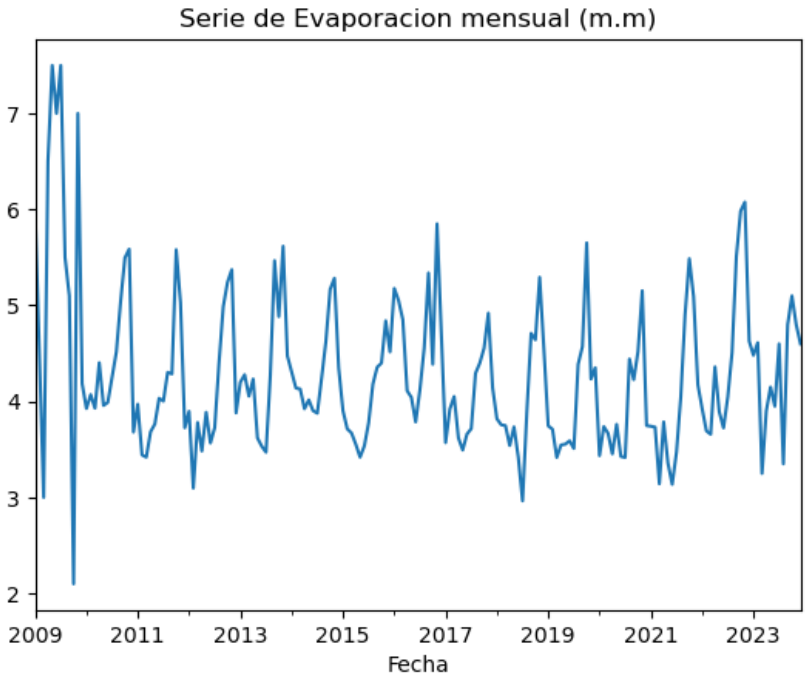
Nota. Elaboración propia.

Figura 40: *Evaporación diaria (m.m)*



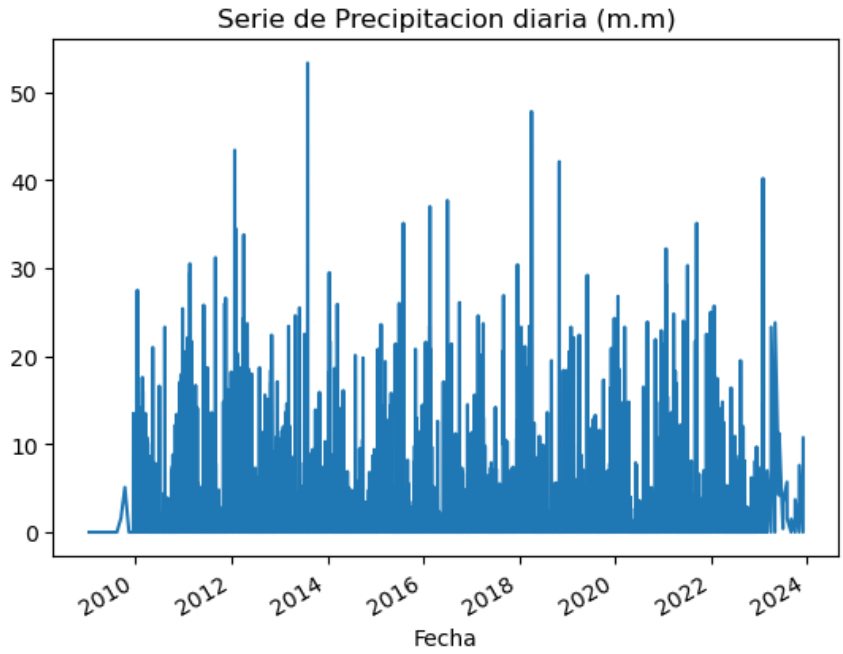
Nota. Elaboración propia.

Figura 41: *Evaporación promedio mensual (m.m)*



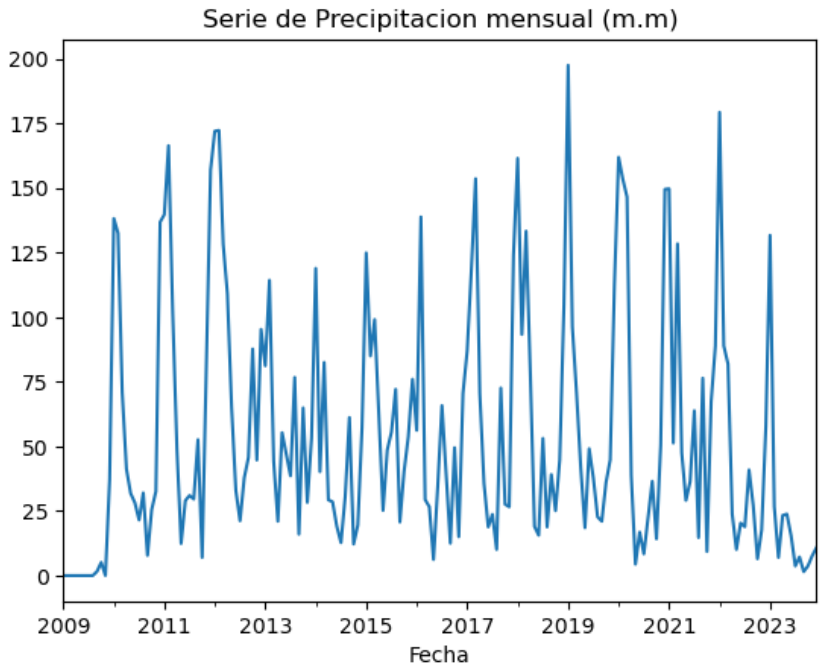
Nota. Elaboración propia.

Figura 42: *Precipitación diaria (m.m)*



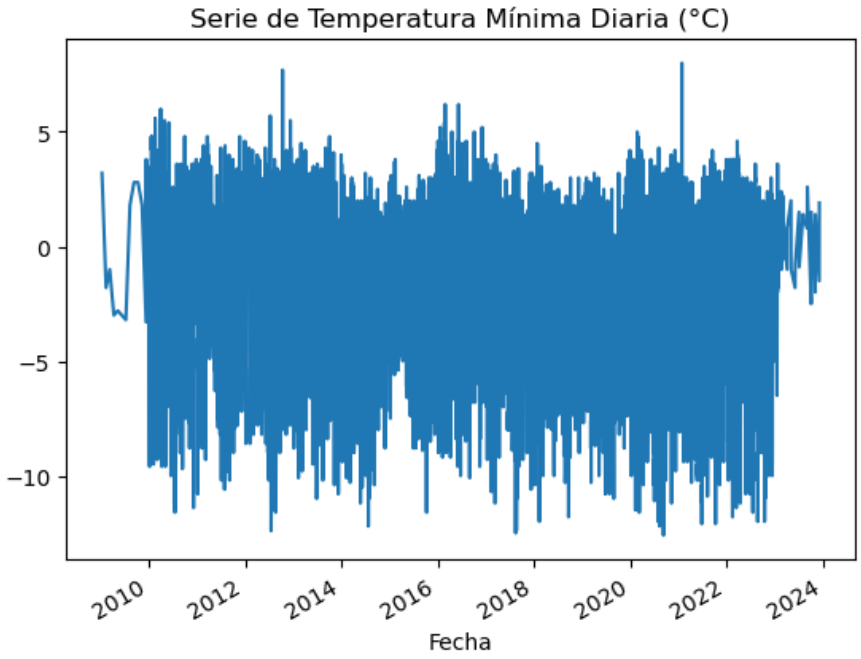
Nota. Elaboración propia.

Figura 43: *Precipitación acumulada mensual (m.m)*



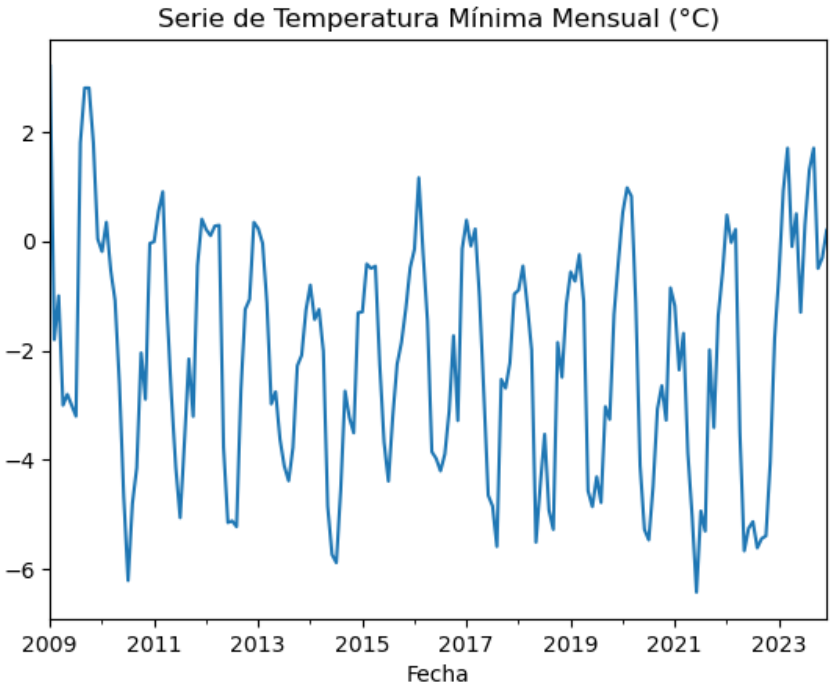
Nota. Elaboración propia.

Figura 44: *Temperatura mínima diaria (°C)*



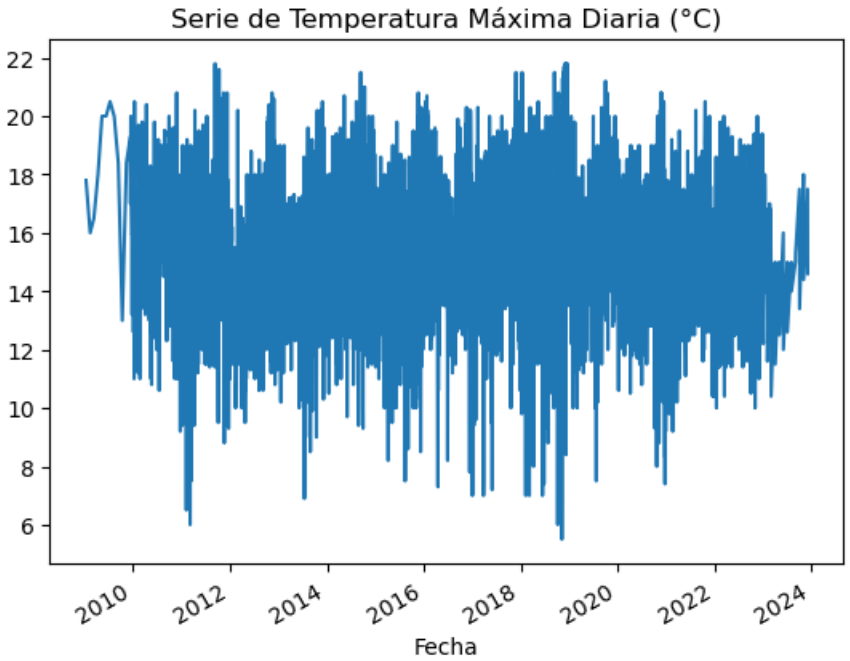
Nota. Elaboración propia.

Figura 45: *Temperatura mínima promedio mensual (°C)*



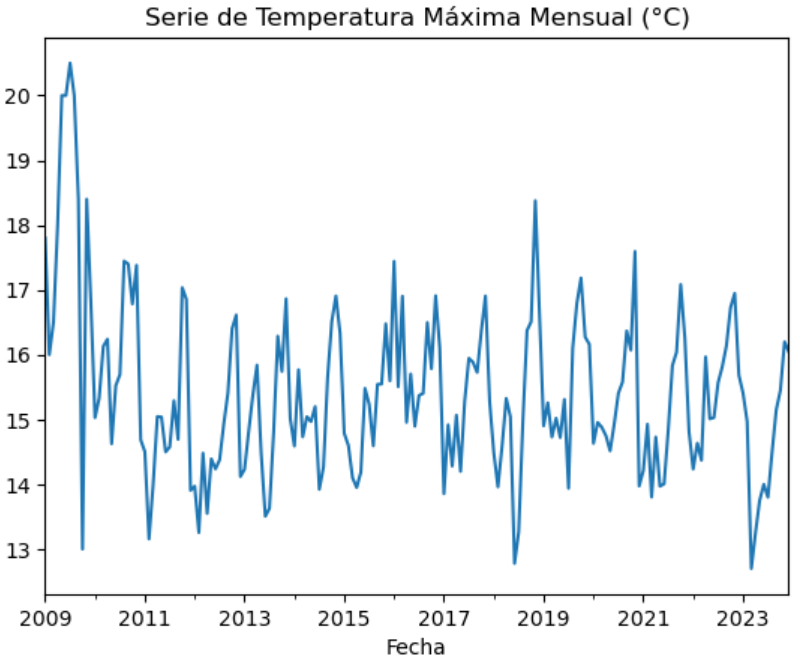
Nota. Elaboración propia.

Figura 46: *Temperatura máxima diaria (°C)*



Nota. Elaboración propia.

Figura 47: *Temperatura máxima promedio mensual (°C)*



Nota. Elaboración propia.

De acuerdo con las Figuras 33, 34, 38 y 39, se observa que las variables Nivel de Embalse (m.s.n.m) y Pérdidas de evaporación (m^3) presentaron un comportamiento constante a lo largo del tiempo, en comparación del resto de variables que tienden a variar de acuerdo a determinadas temporadas del año. Por ello, se optó por realizar pruebas, para las técnicas de Regresión Lineal y SVR, excluyendo dichas variables primero de forma individual y, luego de forma conjunta.

Asimismo, se consideró como antecedente la investigación realizada por Marín y Pineda (2019), en el que se realizó la predicción del caudal de descarga en una represa recurriendo a la utilización de datos históricos hidrometeorológicos, debido a la relevancia de estos parámetros para entender la variabilidad climática de una región en un determinado periodo de tiempo. Esta investigación realizó un modelo de predicción considerando únicamente variables hidrometeorológicas: caudal efluente, precipitación, evaporación, temperatura máxima y temperatura mínima para las técnicas de Regresión Lineal y SVR.

Por otro lado, teniendo en cuenta que el rendimiento del algoritmo SVR depende de la configuración de los parámetros C, Kernel y Epsilon, se seleccionó la función de base radial (rbf) para el parámetro Kernel y 2 para el parámetro C, de acuerdo con la investigación de Dongsheng et al. (2023), en el que se realizaron distintas pruebas para la predicción de series temporales de temperatura y caudales y obtuvieron los mejores resultados utilizando estos parámetros.

Respecto al parámetro Épsilon, se realizaron pruebas considerando cinco valores: 0.2, 0.4, 0.6, 0.8 y 1, ya que el presente parámetro puede tomar valores mayores a cero y menores o iguales a uno.

En la Tabla 6 se resumen las pruebas realizadas para las técnicas Regresión Lineal y SVR.

Tabla 6: *Pruebas aplicadas para Regresión Lineal y SVR*

N° de Prueba	Regresión Lineal	Regresión de Vectores de Soporte
1	Data completa	Data completa - Epsilon: 0.2
2	Excluyendo la variable "Nivel de embalse"	Excluyendo la variable "Nivel de embalse" - Epsilon: 0.2
3	Excluyendo la variable "Pérdidas de evaporación"	Excluyendo la variable "Pérdidas de evaporación" - Epsilon: 0.2
4	Excluyendo variables "Nivel de embalse" y "Pérdidas de evaporación"	Excluyendo variables "Nivel de embalse" y "Pérdidas de evaporación" - Epsilon: 0.2
5	Trabajando solo con variables hidrometeorológicas (caudal, temperatura, precipitación y evaporación)	Trabajando solo con variables hidrometeorológicas (caudal, temperatura, precipitación y evaporación) - Epsilon: 0.2
6	-	Considerando Epsilon: 0.4
7	-	Considerando Epsilon: 0.6
8	-	Considerando Epsilon: 0.8
9	-	Considerando Epsilon: 1

Nota. Elaboración propia.

El análisis de datos se realizó mediante el lenguaje de programación Python utilizando Jupyter Notebook de la plataforma Anaconda. Se inició asignando la variable dependiente (*Y*) a caudal efluente y las variables independientes (*X*) a nivel de embalse, volumen útil, caudal afluente, caudal efluente, pérdidas por evaporación, evaporación, precipitación, temperatura máxima y temperatura mínima y las distintas variaciones detalladas en la Tabla 6, para las técnicas de regresión lineal y regresión de vectores de soporte. Para ello se utilizó sklearn de la biblioteca scikit-learn. Las variables se dividieron en una proporción 80:20 para ser utilizadas como entrenamiento y prueba. Al respecto, se obtuvieron 3868 datos de entrenamiento y 968 datos para el test. Posteriormente, se utilizó la técnica de regresión lineal y regresión de vectores de soporte. Con el fin de obtener mejores resultados se normalizaron los parámetros correspondientes a las variables dependientes e independientes. Para ello, primero se estandarizó los datos con el objetivo de eliminar la media y escalar los datos para que la varianza sea igual a 1 y luego se continuó con la aplicación de las técnicas de regresión lineal y SVR.

Por otro lado, para la técnica ARIMA se asignó la fecha como variable independiente (*X)* y caudal efluente como variable dependiente (*Y*), las cuales fueron divididas en la misma proporción (80:20) para delimitar datos de entrenamiento y prueba.

5.1.2.4 Reporte

En esta etapa se evaluó los resultados obtenidos de los modelos. En ese sentido, de acuerdo al análisis de los resultados se pudo concluir que las variables predictoras utilizadas en el modelo de regresión de vectores de soporte normalizado obtuvo los mejores resultados de acuerdo con las métricas MAE, MSE y RMSE y Varianza.

5.2 Medición de la solución

La presente investigación ha utilizado tres técnicas de modelado para predecir el caudal efluente en la represa Condoroma: regresión lineal, regresión de vectores de soporte y ARIMA. Para medir el rendimiento de estas técnicas se utilizaron cuatro métricas: error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y Varianza.

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo.

5.2.1.1 Métricas obtenidas en Regresión Lineal

Tabla 7: *Métricas estadísticas para Regresión Lineal*

		Regresión Lineal							
		Datos No Normalizados				**Datos Normalizados**			
		MAE	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
1	Data completa	6.8486209	154.1023253	12.4137958	0.2692163	6.4259020	123.2465735	11.1016473	0.2915173
2	Excluyendo la variable "Nivel de embalse"	6.8337969	163.8920577	12.8020333	0.1553415	6.8751291	152.7019463	12.3572629	0.1224084
3	Excluyendo la variable "Pérdidas de evaporación"	6.7963826	154.5846569	12.4332078	0.2448982	6.6471042	135.1287641	11.6244898	0.2124984
4	Excluyendo variables "Nivel de embalse" y "Pérdidas de evaporación"	5.8612455	102.5938771	10.1288636	0.1217619	6.2506542	118.0363437	10.8644532	0.1607144
5	Trabajando solo con variables hidrometeorológicas (caudal, temperatura, precipitación y evaporación)	6.1097213	137.7152540	11.7352143	0.1645498	**5.5357040**	**109.9134931**	**10.4839636**	**0.2239825**

Nota. Elaboración propia.

De acuerdo con la Tabla 7 y luego de aplicar la técnica de Regresión Lineal, los mejores resultados se obtuvieron en la quinta prueba que trabajó únicamente con los datos normalizados de las variables hidrometeorológicas.

Error absoluto medio (mean absolute error - MAE)

Se puede afirmar que la prueba 5, modelo que considera datos normalizados de las variables hidrometeorológicas, presentó un menor MAE respecto al resto de pruebas realizadas. El MAE obtenido fue 5.5357040 lo cual significa que la diferencia entre los valores pronosticados y el valor real u observado fue +/- 5.54 unidades.

Error medio cuadrado (mean square error - MSE)

Para el MSE, al igual que los resultados del MAE, el menor error se obtuvo en la prueba 5 con datos normalizados y su valor fue 109.9134931 unidades cuadradas.

Raíz del error medio cuadrado (root mean square error - RMSE)

De las cinco pruebas realizadas y en línea con los resultados del MAE y MSE, el menor RMSE resultó de la prueba 5 con datos normalizados y su valor fue 10.4839636 unidades.

Varianza - R^2

De los cinco resultados obtenidos se deduce que los modelos propuestos son factibles, pero poco aceptables ya que ninguno tuvo un resultado cercano a 1. En ese sentido, la varianza más alta fue 0.2915173 y se obtuvo en la prueba N° 1 con datos normalizados; y la segunda varianza más alta resultó de la prueba N° 5 con datos normalizados cuyo valor fue 0.2239825.

5.2.1.2 Métricas obtenidas en Regresión de Vectores de Soporte

Tabla 8: *Métricas estadísticas para Regresión de Vectores de Soporte*

		Regresión de Vectores de Soporte							
		Datos No Normalizados				Datos Normalizados			
		MAE	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
1	Data completa	5.5353261	190.7389901	13.8108287	-0.0691864	3.7034306	95.8650454	12.7100543	0.3030571
2	Excluyendo la variable "Nivel de embalse"	5.7477428	199.9917076	14.1418424	-0.0745725	3.9375153	106.2321530	10.3068983	0.3169651
3	Excluyendo la variable "Pérdidas de evaporación"	5.3775784	189.2672316	149.4524907	-0.0687910	3.8321032	118.7833350	10.8987768	0.4092654
4	Excluyendo variables "Nivel de embalse" y "Pérdidas de evaporación"	5.2898103	161.6451261	12.7139737	-0.0679686	**3.6766613**	**83.7013535**	**9.1488444**	**0.4273347**
5	Trabajando solo con variables hidrometeorológicas (caudal, temperatura, precipitación y evaporación)	4.3068202	141.1447489	11.8804356	0.0091921	4.9315007	181.6333441	13.4771415	0.0031501

		Regresión de Vectores de Soporte							
		Datos No Normalizados				Datos Normalizados			
		MAE	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
6	Considerando Epsilon: 0.4	4.8129580	139.2148424	11.7989340	-0.0611729	3.8039162	94.5650710	9.1959269	0.3688137
7	Considerando Epsilon: 0.6	5.2227447	177.5292816	13.3240115	-0.0625607	3.9057901	100.1149986	11.9506982	0.2688250
8	Considerando Epsilon: 0.8	5.8390677	204.7443556	14.3088908	-0.0631058	3.7167921	94.7628833	10.7346229	0.3468454
9	Considerando Epsilon: 1	5.8631923	236.4809232	15.3779363	-0.0631193	3.9215864	109.4877789	10.4636408	0.2361236

Nota. Elaboración propia.

De acuerdo con la Tabla 8, que presenta las métricas estadísticas obtenidas mediante la aplicación de la técnica de Regresión de Vectores de Soporte, los mejores resultados se obtuvieron en la cuarta prueba que no se consideró las variables "Nivel de embalse" y "Pérdidas de evaporación".

Error absoluto medio (mean absolute error - MAE)

Se puede afirmar que la prueba 4 llevada a cabo con datos normalizados, presentó un menor MAE respecto al resto de pruebas realizadas. El MAE obtenido fue 3.6766613 lo cual significa que la diferencia entre los valores predichos y el valor real u observado fue +/- 3.67 unidades.

Error medio cuadrado (mean square error - MSE)

Para el MSE, al igual que los resultados del MAE, el menor error se obtuvo en la prueba 4 con datos normalizados y su valor fue 83.7013535 unidades cuadradas.

Raíz del error medio cuadrado (root mean square error - RMSE)

De las nueve pruebas ejecutadas y en línea con los resultados del MAE y MSE, el menor RMSE resultó de la prueba 4 normalizada y su valor fue 9.1488444 unidades.

Varianza - R^2

De los cinco resultados obtenidos se dedujo que los modelos propuestos son factibles (menores a 1). En ese sentido, la varianza más alta fue 0.4273347 y se obtuvo en la prueba N° 4.

5.2.1.3 Métricas obtenidas para Series de Tiempo - ARIMA

Tabla 9: *Métricas estadísticas - Técnica ARIMA*

Métrica estadística	ARIMA (1,1,1)	ARIMA (2,0,2)	Auto ARIMA (3,1,2)
MAE	4.4764379	3.9981646	4.5153956
MSE	37.0332460	34.2724830	37.4053388
RMSE	6.0854947	5.8542705	6.1159904
Varianza (R^2)	-0.1099166	-0.0271743	-0.1210685

De acuerdo con la Tabla 9, que muestra las métricas estadísticas obtenidas de la aplicación de la técnica ARIMA, se obtuvieron los siguientes resultados:

Error absoluto medio (mean absolute error - MAE)

Se puede afirmar que el modelo ARIMA (2,0,2) presentó un menor MAE respecto al resto de pruebas realizadas. El MAE obtenido fue 3.9981646 lo cual significa que la diferencia entre los valores predichos y el valor real u observado fue +/- 4 unidades.

Error medio cuadrado (mean square error - MSE)

Para el MSE, al igual que los resultados del MAE, el menor error se obtuvo del modelo ARIMA (2,0,2) y su valor fue 34.2724830 unidades cuadradas.

Raíz del error medio cuadrado (root mean square error - RMSE)

De las pruebas ejecutadas y en línea con los resultados del MAE y MSE, el menor RMSE resulta del modelo ARIMA (2,0,2) y su valor fue 5.8542705 unidades.

Varianza - R^2

Se obtuvieron resultados negativos para las tres pruebas realizadas, lo cual significa que los modelos son menos ajustados que el promedio.

A continuación, se presenta una tabla que resume los mejores resultados de las métricas obtenidas mediante la aplicación de tres técnicas de Machine Learning: Regresión lineal, SVR y ARIMA.

Tabla 10: *Resumen de métricas estadísticas para el pronóstico del caudal efluente*

Métrica estadística	Regresión Lineal	SVR	ARIMA
MAE	5.5357040	3.6766613	3.9981646
MSE	109.9134931	83.7013535	34.2724830
RMSE	10.4839636	9.1488444	5.8542705
Varianza (R^2)	0.2239825	0.4273347	-0.0271743

Nota. Elaboración propia.

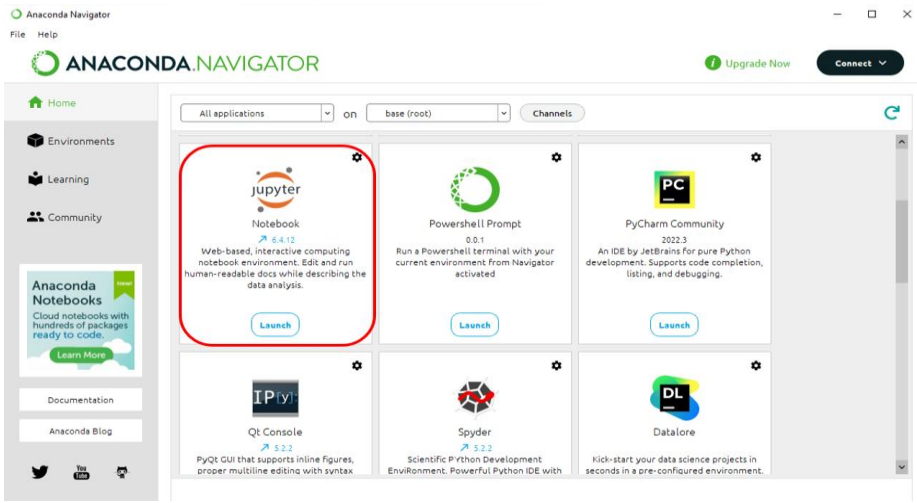
Como se puede observar en la Tabla 10, en el modelo SVR se obtuvo la varianza más alta; es decir, puede capturar de manera más efectiva las correlaciones espacio-tiempo, presentando un mejor rendimiento de predicción, en comparación con el resto de modelos.

5.2.2 Simulación de solución. Aplicación de Software

En este apartado se describen los pasos seguidos para realizar la simulación de la solución aplicando tres técnicas de aprendizaje supervisado.

Una vez obtenida la base de datos final, se procedió a ingresar a la interfaz de Jupyter Notebook 6.4.12 mediante la plataforma de Anaconda Navigator para utilizar el lenguaje de programación Python, en el que se realizaron distintos modelos aplicando las tres técnicas mencionadas: regresión lineal, regresión de vectores de soporte y ARIMA.

Figura 48: *Plataforma Anaconda Navigator*



Nota. Obtenido de Plataforma Anaconda Navigator (2023)

Figura 49: *Interfaz de Jupyter Notebook 6.4.12*

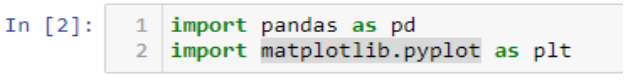
Nota. Obtenido de Interfaz de Jupyter Notebook 6.4.12

5.2.2.1 Aplicación de Software para Regresión lineal y Regresión de Vectores de Soporte

5.2.2.1.1 Importación de librerías

Se inició importando las librerías pandas y matplotlib.pyplot, para leer las bases de datos y para elaborar gráficos.

Figura 50: *Importación de Librerías pandas y matplotlib.pyplot*



Nota. Elaboración propia.

5.2.2.1.2 Lectura de datos

Posteriormente, se procedió a subir la base de datos en formato excel que recopila las variables utilizadas, esta fue nombrada como “datos”.

Figura 51: *Lectura del dataset - Regresión Lineal y SVR*

```
In [2]:
1 datos=pd.read_excel('BD_CONDOROMA_AUTODEMA_27_02.xlsx')
2 datos
```

Out[2]:

	Fecha	Caudal_Efluente	Caudal_Afluente	Nivel_de_Embalse	Volumen_util	Perdidas_evapora	Evaporacion	Precipitacion	TempMin	TempMax
0	01/12/2009	9.872	8.508	4123.749	54990662.0	25875	6.0	0.0	3.2	17.8
1	02/12/2009	9.965	8.486	4123.719	54846913.0	19361	4.5	0.0	-1.8	16.0
2	03/12/2009	8.902	3.977	4123.634	54440565.0	12862	3.0	0.0	-1.0	16.5
3	04/12/2009	8.892	3.228	4123.542	54002320.0	27754	6.5	0.0	-3.0	18.0
4	05/12/2009	9.438	1.995	4123.433	53485198.0	31861	7.5	0.0	-2.8	20.0
...	...	...	...	...	...	...	...	...	...	...
4831	22/02/2023	0.000	8.838	4135.375	122154576.0	26056	4.2	0.2	2.0	15.0
4832	23/02/2023	0.000	9.050	4135.482	122892133.0	16286	2.6	3.6	0.8	16.6
4833	24/02/2023	0.000	8.000	4135.592	123657736.0	20009	3.2	0.2	0.5	14.8
4834	25/02/2023	0.000	16.700	4135.697	124387223.0	18848	3.0	1.5	-0.6	16.0
4835	26/02/2023	0.000	8.395	4135.901	125811273.0	23986	3.8	0.8	-0.4	16.8

4836 rows × 10 columns

Nota. Elaboración propia.

5.2.2.1.3 Pre-procesamiento de información

El siguiente paso fue analizar los datos identificando el tipo (numérico o nominal) y realizando una descripción estadística.

Figura 52: *Identificación de variables numéricas*

```
In [3]:
1 ###Tipos de datos
2 datos.dtypes
```

```
Out[3]: Fecha              object
        Caudal_Efluente    float64
        Caudal_Afluente    float64
        Nivel_de_Embalse   float64
        Volumen_util       float64
        Perdidas_evapora     int64
        Evaporacion        float64
        Precipitacion      float64
        TempMin            float64
        TempMax            float64
        dtype: object
```

Nota. Elaboración propia.

Figura 53: *Descripción estadística y verificación de data*

```
In [4]:
1 ###Análisis estadísticos de los datos
2 datos.describe(include='all')
```

Out[4]:

	Fecha	Caudal_Efluente	Caudal_Afluente	Nivel_de_Embalse	Volumen_util	Perdidas_evapora	Evaporacion	Precipitacion	TempMin	TempMax
count	4836	4836.000000	4836.000000	4836.000000	4.836000e+03	4836.000000	4836.000000	4836.000000	4836.000000	4836.000000
unique	4836	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	01/12/2009	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	10.416579	10.963287	4141.219916	1.748907e+08	30794.043011	4.192928	2.021588	-2.412864	15.300116
std	NaN	13.252188	22.330695	8.633883	6.662916e+07	10421.726916	1.537985	4.828726	3.837808	2.329144
min	NaN	0.000000	0.040000	4116.860000	2.636843e+07	4285.000000	0.000000	0.000000	-12.600000	5.500000
25%	NaN	5.190500	1.123000	4135.082250	1.202275e+08	23438.500000	3.000000	0.000000	-5.500000	13.800000
50%	NaN	8.774000	1.958500	4143.266000	1.835335e+08	30946.000000	4.000000	0.000000	-2.000000	15.200000
75%	NaN	11.914250	8.555750	4149.064750	2.386209e+08	38276.500000	5.000000	1.200000	0.925000	16.800000
max	NaN	123.494000	261.168000	4151.763000	2.673119e+08	62231.000000	16.600000	53.300000	8.000000	21.800000

Nota. Elaboración propia.

5.2.2.1.4 Visualización de series de tiempo para variables independientes

Para designar a la columna ‘Fecha’ como el índice de las series de tiempo, se utilizó las funciones to_datetime e inplace. Posteriormente, con el fin de realizar un análisis diario y mensual de las variables, se transformó los datos diarios a datos mensuales con la función resample. Finalmente, se generaron gráficos de series de tiempo para cada variable.

Figura 54: *Designación de columna Fecha como índice*

```
In [5]:
1 ###designando la columna fecha como índice de la información
2 datos['Fecha'] = pd.to_datetime(datos.Fecha)
3 datos.set_index('Fecha',inplace=True)
```

Nota. Elaboración propia.

Figura 55: *Transformación de datos diarios a mensuales*

```
In [6]:
1 ###transformando datos diarios a datos mensuales
2 datos_mensual = datos.resample(rule='M').sum()
```

Nota. Elaboración propia.

Figura 56: *Función para graficar variables*

```
In [7]:
1 ###gráficos variables independientes
2 ##Caudal afluente
3 datos['Caudal_Afluente'].plot().set_title('Serie de caudal afluente diario (m3/s)')
```

Nota. Elaboración propia.

5.2.2.1.5 Procesamiento de datos

i) Pruebas con base de datos aplicando la Técnica de Regresión Lineal

El presente apartado muestra el desarrollo de la Técnica de Regresión Lineal para datos no normalizados y normalizados de la prueba 5 que trabaja solo con variables hidrometeorológicas: caudal efluente, evaporación, precipitación, temperatura máxima y temperatura mínima; ya que fue la en la que se obtuvo mejores resultados (Ver ítem 5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo).

Regresión lineal con datos no normalizados

Se procede a definir las variables dependientes e independientes que, para este caso, son caudal afluente, evaporación, precipitación, temperatura mínima y temperatura máxima, como variables independientes y, caudal efluente como variable dependiente.

Figura 57: *Separación de variables - Regresión Lineal No Normalizada*

Procesamiento

```
In [20]:
1 #Separando X Y
2 X = datos[['Caudal_Afluente','Evaporacion','Precipitacion','TempMin','TempMax']]
3 Y = datos[['Caudal_Efluente']]
```

Nota. Elaboración propia.

Figura 58: *Asignación de variables X - Regresión Lineal No Normalizada*

```
In [21]:
1 X
```

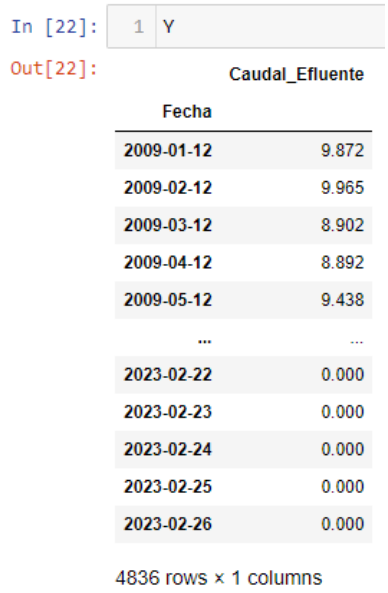
Out[21]:

Fecha	Caudal_Afluente	Evaporacion	Precipitacion	TempMin	TempMax
2009-01-12	8.508	6.0	0.0	3.2	17.8
2009-02-12	8.486	4.5	0.0	-1.8	16.0
2009-03-12	3.977	3.0	0.0	-1.0	16.5
2009-04-12	3.228	6.5	0.0	-3.0	18.0
2009-05-12	1.995	7.5	0.0	-2.8	20.0
...	...	...	...	...	...
2023-02-22	8.838	4.2	0.2	2.0	15.0
2023-02-23	9.050	2.6	3.6	0.8	16.6
2023-02-24	8.000	3.2	0.2	0.5	14.8
2023-02-25	16.700	3.0	1.5	-0.6	16.0
2023-02-26	8.395	3.8	0.8	-0.4	16.8

4836 rows × 5 columns

Nota. Elaboración propia.

Figura 59: *Asignación de variable Y - Regresión Lineal No Normalizada*



```
In [22]: 1 Y
```

Out[22]:

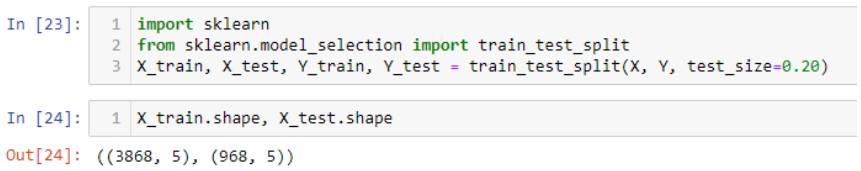
Fecha	Caudal_Efluente
2009-01-12	9.872
2009-02-12	9.965
2009-03-12	8.902
2009-04-12	8.892
2009-05-12	9.438
...	...
2023-02-22	0.000
2023-02-23	0.000
2023-02-24	0.000
2023-02-25	0.000
2023-02-26	0.000

4836 rows × 1 columns

Nota. Elaboración propia.

De sklearn se importó el modelo train_test_split para dividir los datos. Se tomó el 20% de datos para la prueba final y la diferencia como datos de entrenamiento.

Figura 60: *Datos train y test - Regresión Lineal No Normalizada*

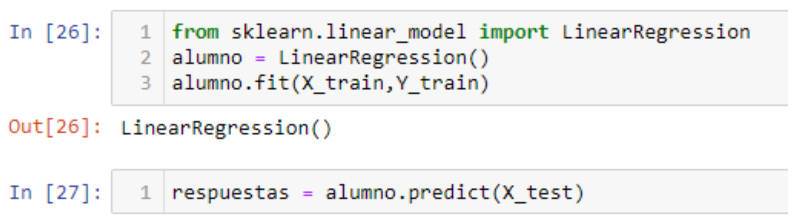


```
In [23]: 1 import sklearn
         2 from sklearn.model_selection import train_test_split
         3 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20)

In [24]: 1 X_train.shape, X_test.shape

Out[24]: ((3868, 5), (968, 5))
```

Nota. Elaboración propia.

Posteriormente, se aplicó el logaritmo de sklearn, se importó el modelo de regresión lineal y se indicó que lleve a cabo la predicción del caudal efluente.

Figura 61: *Modelamiento de la Regresión Lineal No Normalizada*



```
In [26]: 1 from sklearn.linear_model import LinearRegression
         2 alumno = LinearRegression()
         3 alumno.fit(X_train,Y_train)

Out[26]: LinearRegression()

In [27]: 1 respuestas = alumno.predict(X_test)
```

Nota. Elaboración propia.

Luego de aplicar el modelo, se procedió a evaluarlo mediante métricas de precisión, para lo cual se utilizó MAE, MSE, RMSE y R^2

Figura 62: *Métricas - Regresión Lineal No Normalizada*

```
In [28]:
#MAE
from sklearn.metrics import mean_absolute_error
mean_absolute_error(Y_test,respuestas)

Out[28]: 6.109721339183873

In [29]:
#MSE
from sklearn.metrics import mean_squared_error
mean_squared_error(Y_test,respuestas)

Out[29]: 137.71525402316183

In [30]:
#RMSE
mean_squared_error(Y_test,respuestas)**0.5

Out[30]: 11.73521427257133

In [31]:
#Varianza
from sklearn.metrics import r2_score
r2_score(Y_test,respuestas)

Out[31]: 0.16454980567895916
```

Nota. Elaboración propia.

Solicitamos los coeficientes de las variables.

Figura 63: *Coeficientes - Regresión Lineal No Normalizada*

```
In [32]:
print('Coefficients: \n', alumno.coef_)

Coefficients:
 [[ 0.32811663  0.29652829 -0.30429447 -0.14849894  0.421295  ]]
```

Nota. Elaboración propia.

Regresión lineal con datos normalizados

Para la predicción del caudal efluente mediante la técnica de regresión lineal con datos normalizados se siguieron los siguientes pasos:

De sklearn se importó la función StandarScarler para normalizar los datos

Figura 64: *Función StandarScarler - Regresión lineal*

```
In [33]:
from sklearn.preprocessing import StandardScaler
```

Nota. Elaboración propia.

Luego se definió las variables y se separó los datos en aprendizaje y testeo

Figura 65: *Asignación de variables X - Regresión Lineal Normalizada*

```
In [34]:
1 Xnor = datos[['Caudal_Afluente','Evaporacion','Precipitacion','TempMin','TempMax']]
2 Yn = datos[['Caudal_Efluente']]
```

Nota. Elaboración propia.

Posteriormente, se procedió a normalizar las variables para que puedan ser aplicadas en la técnica de regresión lineal.

Figura 66: *Normalización de datos - Regresión Lineal*

```
In [35]:
1 normalizador = StandardScaler()
2 Xn = normalizador.fit_transform(Xnor)

In [36]:
1 Xn

Out[36]: array([[-0.10996259,  1.1750825 , -0.41870201,  1.46266949,  1.0734172 ],
                [-0.11094789,  0.19967931, -0.41870201,  0.15970766,  0.30052102],
                [-0.31288814, -0.77572388, -0.41870201,  0.36818156,  0.5152144 ],
                ...,
                [-0.1327139 , -0.64567012, -0.37727893,  0.7590701 , -0.21474311],
                [ 0.25692465, -0.77572388, -0.10802892,  0.4724185 ,  0.30052102],
                [-0.11502342, -0.25550884, -0.25300969,  0.52453698,  0.64403043]])
```

Nota. Elaboración propia.

De sklearn se importó nuevamente train_test_plit para separar los datos normalizados en train y test.

Figura 67: *Datos train y test - Regresión Lineal Normalizada*

```
In [37]:
1 import sklearn
2 from sklearn.model_selection import train_test_split
3 Xn_train, Xn_test, Yn_train, Yn_test = train_test_split(Xn, Yn, test_size=0.20)
```

Nota. Elaboración propia.

Se importó la regresión lineal indicando que realice la predicción con la muestra normalizada.

Figura 68: *Modelamiento de la Regresión Lineal Normalizada*

```
In [38]:
1 from sklearn.linear_model import LinearRegression
2 alumno = LinearRegression()
3 alumno.fit(Xn_train,Yn_train)

Out[38]: LinearRegression()

In [39]:
1 respuestasN = alumno.predict(Xn_test)
```

Nota. Elaboración propia.

Luego, se evaluó el modelo de regresión lineal con datos normalizados, mediante cuatro métricas de precisión: MAE, MSE, RMSE y R^2.

Figura 69: *Métricas - Regresión Lineal Normalizada*

```
In [40]:
#MAE
from sklearn.metrics import mean_absolute_error
mean_absolute_error(Yn_test,respuestasN)

Out[40]: 5.535704022192262

In [41]:
#MSE
from sklearn.metrics import mean_squared_error
mean_squared_error(Yn_test,respuestasN)

Out[41]: 109.91349310054257

In [42]:
#RMSE
mean_squared_error(Yn_test,respuestasN)**0.5

Out[42]: 10.483963615949007

In [43]:
#Varianza
from sklearn.metrics import r2_score
r2_score(Yn_test,respuestasN)

Out[43]: 0.2239825002914505
```

Nota. Elaboración propia.

ii) Pruebas con base de datos aplicando la Técnica Regresión de Vectores de Soporte

A continuación, se presenta el desarrollo de la Técnica Regresión de Vectores de Soporte para datos no normalizados y normalizados de la prueba 4 en la que se excluye las variables Nivel de embalse y Pérdidas de evaporación, ya que fue la en la que se obtuvieron mejores resultados (Ver ítem 5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo).

Regresión de Vectores de Soporte con datos no normalizados

Se define las variables dependientes e independientes que son volumen útil, caudal afluente, evaporación, precipitación, temperatura mínima y temperatura máxima, como variables independientes y, caudal efluente como variable dependiente.

Figura 70: *Separación de variables - SVR No Normalizada*

Procesamiento

```
In [20]:
#Separando X Y
X = datos[['Volumen_util','Caudal_Afluente','Evaporacion','Precipitacion','TempMin','TempMax']]
Y = datos[['Caudal_Efluente']]
```

Nota. Elaboración propia.

Figura 71: *Asignación de variables X - SVR No Normalizada*

```
In [21]: 1 X
```

Out[21]:

Fecha	Volumen_util	Caudal_Afluente	Evaporacion	Precipitacion	TempMin	TempMax
2009-01-12	54990662.0	8.508	6.0	0.0	3.2	17.8
2009-02-12	54846913.0	8.486	4.5	0.0	-1.8	16.0
2009-03-12	54440565.0	3.977	3.0	0.0	-1.0	16.5
2009-04-12	54002320.0	3.228	6.5	0.0	-3.0	18.0
2009-05-12	53485198.0	1.995	7.5	0.0	-2.8	20.0
...	...	...	...	...	...	...
2023-02-22	122154576.0	8.838	4.2	0.2	2.0	15.0
2023-02-23	122892133.0	9.050	2.6	3.6	0.8	16.6
2023-02-24	123657736.0	8.000	3.2	0.2	0.5	14.8
2023-02-25	124387223.0	16.700	3.0	1.5	-0.6	16.0
2023-02-26	125811273.0	8.395	3.8	0.8	-0.4	16.8

4836 rows × 6 columns

Nota. Elaboración propia.

Figura 72: Asignación de variable Y - SVR No Normalizada

```
In [22]: 1 Y
```

Out[22]:

Fecha	Caudal_Efluente
2009-01-12	9.872
2009-02-12	9.965
2009-03-12	8.902
2009-04-12	8.892
2009-05-12	9.438
...	...
2023-02-22	0.000
2023-02-23	0.000
2023-02-24	0.000
2023-02-25	0.000
2023-02-26	0.000

4836 rows × 1 columns

Nota. Elaboración propia.

De sklearn se importó el modelo train_test_split para dividir los datos. Se tomó el 20% de datos para la prueba final y la diferencia como datos de entrenamiento.

Figura 73: *Datos train y test - SVR No Normalizada*

```
In [23]: 1 import sklearn
         2 from sklearn.model_selection import train_test_split
         3 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20)

In [24]: 1 X_train.shape, X_test.shape
Out[24]: ((3868, 6), (968, 6))
```

Nota. Elaboración propia.

Posteriormente, se aplicó el logaritmo de sklearn para importar el modelo SVR y se indica que lleve a cabo la predicción del caudal efluente.

Figura 74: *Modelamiento de SVR No Normalizada*

```
In [26]:
from sklearn.svm import SVR

In [27]:
#Definimos el algoritmo a utilizar
svr = SVR(kernel='rbf', C=2.0, epsilon=0.2)

In [28]:
#Entrenamos el modelo
svr.fit(X_train, Y_train)
Out[28]: SVR(C=2.0, epsilon=0.2)

In [29]:
#Realizamos una predicción
Y_pred = svr.predict(X_test)
```

Nota. Elaboración propia.

Luego de aplicar el modelo SVR, se procedió a evaluarlo mediante métricas, para lo cual se utilizó MAE, MSE, RMSE y R^2

Figura 75: *Métricas - SVR No Normalizada*

```
In [30]:
#MAE
from sklearn.metrics import mean_absolute_error
mean_absolute_error(Y_test,Y_pred)
Out[30]: 5.289810265603133

In [31]:
#MSE
from sklearn.metrics import mean_squared_error
mean_squared_error(Y_test,Y_pred)
Out[31]: 161.64512612077198

In [32]:
#RMSE
mean_squared_error(Y_test,Y_pred)**0.5
Out[32]: 12.713973655815556

In [33]:
#Varianza
from sklearn.metrics import r2_score
r2_score(Y_test,Y_pred)
Out[33]: -0.06796858786138738

In [32]:
print('Coefficients: \n', alumno.coef_)
Coefficients:
 [[ 0.32811663  0.29652829 -0.30429447 -0.14849894  0.421295  ]]
```

Nota. Elaboración propia.

Regresión de Vectores de Soporte con datos normalizados

Para la predicción del caudal efluente mediante SVR normalizada, se realizaron los siguientes pasos:

De sklearn se importó la función StandarScarler para normalizar los datos.

Figura 76: *Función StandarScarler - SVR*

```
In [35]:  1 from sklearn.preprocessing import StandardScaler
```

Nota. Elaboración propia.

Luego se definieron las variables y se separaron los datos en aprendizaje y testeo.

Figura 77: *Asignación de variables X - SVR Normalizada*

```
In [36]:  1 X = datos[['Volumen_util','Caudal_Afluente','Evaporacion','Precipitacion','TempMin','TempMax']]
          2 Yn = datos[['Caudal_Efluente']]
```

Nota. Elaboración propia.

Posteriormente, se procedió a normalizar las variables para que puedan ser aplicadas en la técnica de SVR Normalizada.

Figura 78: *Normalización de datos - SVR*

```
In [37]:  1 normalizador = StandardScaler()
          2 Xn = normalizador.fit_transform(X)

In [38]:  1 Xn

Out[38]: array([[-1.79969937, -0.10996259,  1.1750825 , -0.41870201,  1.462
         66949,
                  1.0734172 ],
                [-1.80185704, -0.11094789,  0.19967931, -0.41870201,  0.159
         70766,
                  0.30052102],
                [-1.80795632, -0.31288814, -0.77572388, -0.41870201,  0.368
         18156,
                  0.5152144 ],
                ...,
                [-0.76900689, -0.1327139 , -0.64567012, -0.37727893,  0.759
         0701 ,
                 -0.21474311],
                [-0.75805729,  0.25692465, -0.77572388, -0.10802892,  0.472
         4185 ,
                  0.30052102],
                [-0.73668231, -0.11502342, -0.25550884, -0.25300969,  0.524
         53698,
                  0.64403043]])
```

Nota. Elaboración propia.

De sklearn se utilizó nuevamente train_test_plit para separar los datos normalizados en train y test.

Figura 79: *Datos train y test - SVR Normalizada*

```
In [40]:
import sklearn
from sklearn.model_selection import train_test_split
Xn_train, Xn_test, Yn_train, Yn_test = train_test_split(Xn, Yn, test_size=0.20)
```

Nota. Elaboración propia.

Se importó el modelo SVR indicando que realice la predicción con la muestra normalizada.

Figura 80: *Modelamiento SVR Normalizada*

```
In [40]:
import sklearn
from sklearn.model_selection import train_test_split
Xn_train, Xn_test, Yn_train, Yn_test = train_test_split(Xn, Yn, test_size=0.20)

In [41]:
#Definimos el algoritmo a utilizar
svrN = SVR(kernel='rbf', C=2.0, epsilon=0.2)

In [42]:
#Entrenamos el modelo
svrN.fit(Xn_train, Yn_train)

Out[42]: SVR(C=2.0, epsilon=0.2)

In [43]:
#Realizamos una predicción
Yn_pred = svrN.predict(Xn_test)
```

Nota. Elaboración propia.

Luego se evaluó el modelo SVR con datos normalizados, mediante cuatro métricas de precisión: MAE, MSE, RMSE y R^2.

Figura 81: *Métricas - SVR Normalizada*

```
In [44]:
#MAE
from sklearn.metrics import mean_absolute_error
mean_absolute_error(Yn_test,Yn_pred)

Out[44]: 3.676661322532572

In [45]:
#MSE
from sklearn.metrics import mean_squared_error
mean_squared_error(Yn_test,Yn_pred)

Out[45]: 83.70135345902959

In [46]:
#RMSE
mean_squared_error(Yn_test,Yn_pred)**0.5

Out[46]: 9.148844378337058

In [47]:
#Varianza
from sklearn.metrics import r2_score
r2_score(Yn_test,Yn_pred)

Out[47]: 0.4273346636175488
```

Nota. Elaboración propia.

5.2.2.2 Aplicación de Software para la Técnica ARIMA

5.2.2.2.1 Instalación e importación de librerías

Se instaló la librerías pmdarima e importó las librerías pandas, matplotlib.pyplot, os y numpy.

Figura 82: *Instalación de librería pmdarima*

```
In [2]:  1  !pip install pmdarima
```

Nota. Elaboración propia.

Figura 83: *Librerías - Técnica ARIMA*

```
In [2]:  1  import os
         2  import pandas as pd
         3  import numpy as np
         4  import matplotlib.pyplot as plt
```

Nota. Elaboración propia.

5.2.2.2.2 Lectura de datos

Posteriormente, se procedió a subir la base de datos en formato excel que recopila las variables utilizadas, esta fue nombrada como “df”.

Figura 84: *Lectura del dataset - Técnica ARIMA*

```
In [27]:  1  df=pd.read_excel('BD_CONDOROMA_AUTODEMA_efluente.xlsx')
          2  df
```

Out[27]:

	Fecha	Caudal_Efluente
0	01/12/2009	9.872
1	02/12/2009	9.965
2	03/12/2009	8.902
3	04/12/2009	8.892
4	05/12/2009	9.438
...	...	...
4831	22/02/2023	0.000
4832	23/02/2023	0.000
4833	24/02/2023	0.000
4834	25/02/2023	0.000
4835	26/02/2023	0.000

4836 rows × 2 columns

Nota. Elaboración propia.

5.2.2.2.3 Fecha como índice

El siguiente paso se transformó la columna Fecha de la base de datos como una variable de tipo fecha, para ello se usó la función datetime. Posteriormente, se aplicó el método set_index del paquete pandas, se indicó que la variable Fecha es el atributo que debe convertirse

en índice. Luego, se señaló que la frecuencia del conjunto de datos es diaria mediante el método asfreq.

Figura 85: *Fecha como índice*

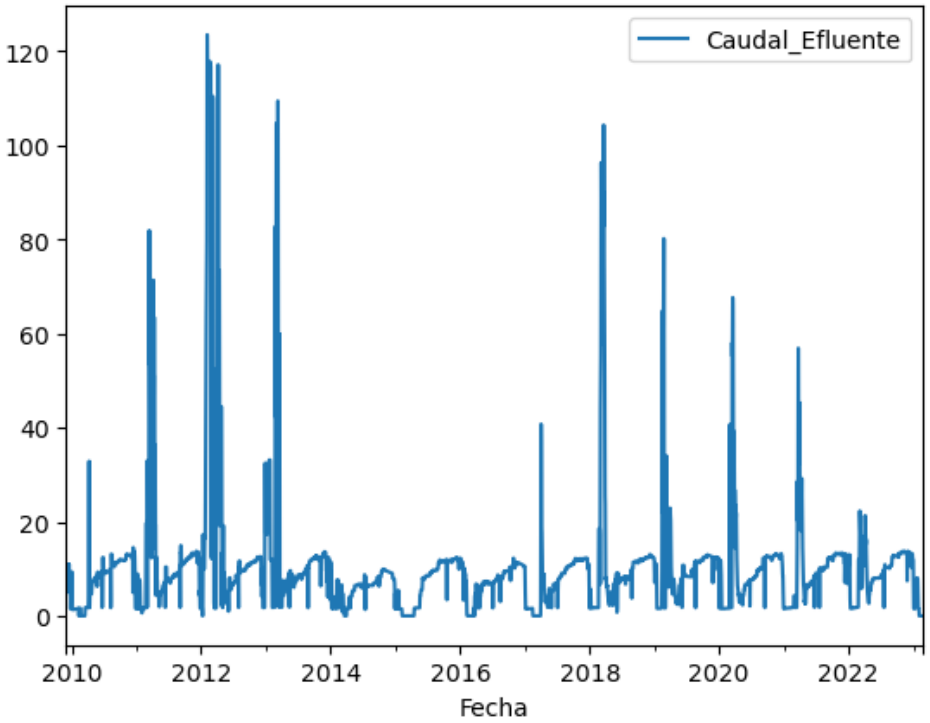
```
In [4]:
1 df.Fecha = pd.to_datetime(df.Fecha, dayfirst = True)
2 df.set_index("Fecha", inplace=True)
3 ts=df.asfreq('d')
4
```

Nota. Elaboración propia.

5.2.2.2.4 Estacionariedad

Para evaluar la estacionariedad de la serie primero se graficó la variable Caudal Efluente

Figura 86: *Gráfico de Caudal Efluente*

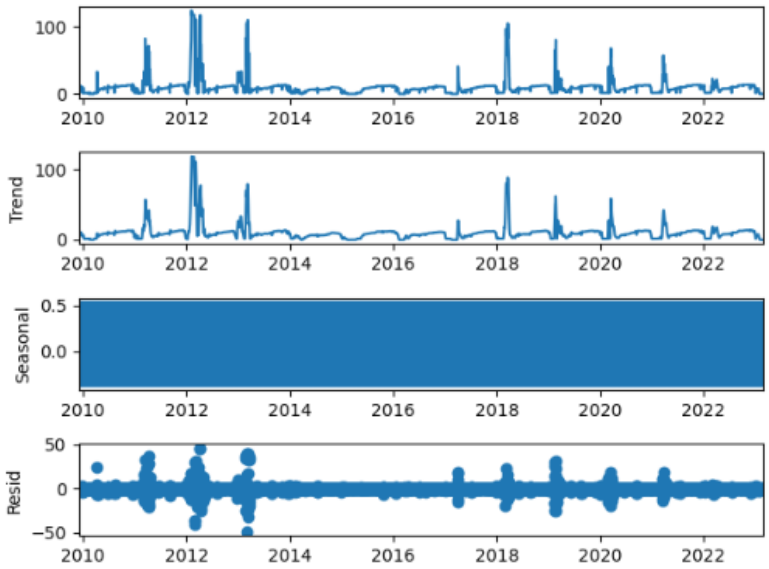


Nota. Elaboración propia.

Luego, se graficó la descomposición aditiva del Caudal Efluente

Figura 87: *Descomposición aditiva*

```
In [7]:
1 import statsmodels.api as sm

In [8]:
1 decomposition = sm.tsa.seasonal_decompose(ts, model='additive')
2 fig = decomposition.plot()
3 plt.show()
```



Nota. Elaboración propia.

Finalmente, se aplicó la prueba de Dickey-Fuller y se obtuvo un valor menor a cero lo cual indica que la serie es estacionaria.

Figura 88: *Prueba de Dickey-Fuller*

```
In [11]:
1 from statsmodels.tsa.stattools import adfuller

In [12]:
1 adftest=adfuller(ts)

In [13]:
1 print('pvalue of adfuller test is:', adftest[1])

pvalue of adfuller test is: 1.6616126275858408e-12
```

Nota. Elaboración propia.

5.2.2.2.5 División de datos en entrenamiento y prueba

Se dividieron los datos en dos conjuntos: entrenamiento y prueba, teniendo en cuenta el orden cronológico de los datos. Se destinó el 80% de los datos para entrenamiento y el 20% de los datos para prueba.

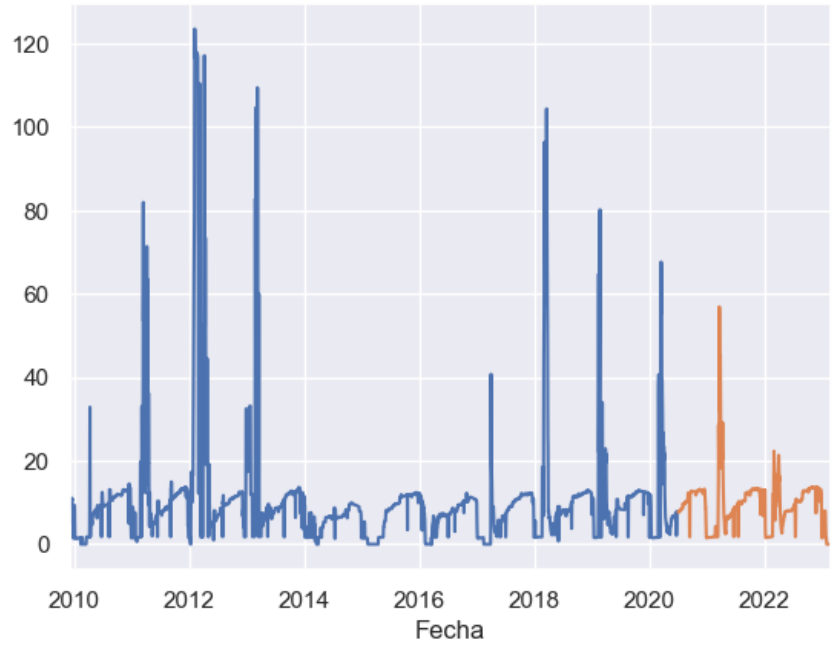
Figura 89: *División de datos en entrenamiento y prueba*

```
In [14]:  1 size = int(len(ts)*0.8)
          2 size

Out[14]: 3868

In [15]:  1 train=ts.iloc[:size]
          2 test = ts.iloc[size:]
```

Nota. Elaboración propia.

Figura 90: *Muestra gráfica de los datos de entrenamiento y prueba*

```
In [16]:  1 train['Caudal_Efluente'].plot()
          2 test['Caudal_Efluente'].plot()

Out[16]: <AxesSubplot:xlabel='Fecha'>
```



Nota. Elaboración propia.

5.2.2.2.6 Construcción del Modelo ARIMA

Se inició modelando ARIMA con el orden 1,1,1; luego con el orden 2,0,2

Figura 91: *ARIMA (1,1,1)*

```
In [16]: 1 from statsmodels.tsa.arima.model import ARIMA

In [17]: 1 model=ARIMA(train,order=(1,1,1)).fit()

In [28]: 1 model.summary()
```

Out[28]:

SARIMAX Results

Dep. Variable:	Caudal_Efluente	**No. Observations:**	3868
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-11660.522
Date:	Fri, 10 Mar 2023	**AIC**	23327.045
Time:	18:27:22	**BIC**	23345.826
Sample:	12-01-2009	**HQIC**	23333.714
	- 07-03-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1935	0.031	-6.312	0.000	-0.254	-0.133
ma.L1	0.3097	0.029	10.554	0.000	0.252	0.367
sigma2	24.3590	0.101	241.575	0.000	24.161	24.557

Ljung-Box (L1) (Q):	0.21	**Jarque-Bera (JB):**	619244.71
Prob(Q):	0.65	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.26	**Skew:**	0.36
Prob(H) (two-sided):	0.00	**Kurtosis:**	64.99

Nota. Elaboración propia.

Figura 92: *ARIMA (2,0,2)*

```
In [16]: 1 from statsmodels.tsa.arima.model import ARIMA

In [17]: 1 model=ARIMA(train,order=(2,0,2)).fit()

In [18]: 1 model.summary()
```

Out[18]:

SARIMAX Results

Dep. Variable:	Caudal_Efluente	**No. Observations:**	3868
Model:	ARIMA(2, 0, 2)	**Log Likelihood**	-11579.136
Date:	Fri, 10 Mar 2023	**AIC**	23170.272
Time:	18:31:57	**BIC**	23207.835
Sample:	12-01-2009	**HQIC**	23183.610
	- 07-03-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	10.6132	2.860	3.711	0.000	5.009	16.218
ar.L1	1.6436	0.024	67.130	0.000	1.596	1.692
ar.L2	-0.6562	0.023	-28.576	0.000	-0.701	-0.611
ma.L1	-0.5828	0.024	-23.892	0.000	-0.631	-0.535
ma.L2	-0.1685	0.004	-43.814	0.000	-0.176	-0.161
sigma2	23.3056	0.119	195.494	0.000	23.072	23.539

Ljung-Box (L1) (Q):	0.16	**Jarque-Bera (JB):**	551165.19
Prob(Q):	0.69	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.27	**Skew:**	1.79
Prob(H) (two-sided):	0.00	**Kurtosis:**	61.37

Nota. Elaboración propia.

Posteriormente, se utilizó la técnica auto ARIMA para calcular los parámetros que mejor se ajusten al modelo.

Figura 93: *Auto ARIMA*

```
In [17]: 1 from pmdarima.arima import auto_arima

In [18]: 1 model_auto = auto_arima(train.Caudal_Efluente[1:])

In [19]: 1 model_auto

Out[19]: ARIMA(order=(3, 1, 2), scoring_args={}, suppress_warnings=True,
         with_intercept=False)
```

Nota. Elaboración propia.

Figura 94: *ARIMA (3,1,2)*

```
In [20]: 1 model_auto.summary()
```

Out[20]:

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	3867
Model:	SARIMAX(3, 1, 2)	**Log Likelihood**	-11591.504
Date:	Fri, 10 Mar 2023	**AIC**	23195.007
Time:	18:36:36	**BIC**	23232.567
Sample:	12-02-2009	**HQIC**	23208.344
	- 07-03-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0226	0.028	0.803	0.422	-0.033	0.078
ar.L2	0.6134	0.025	24.208	0.000	0.564	0.663
ar.L3	-0.1978	0.005	-36.266	0.000	-0.208	-0.187
ma.L1	0.0600	0.029	2.041	0.041	0.002	0.118
ma.L2	-0.7002	0.024	-29.125	0.000	-0.747	-0.653
sigma2	23.5400	0.105	225.123	0.000	23.335	23.745

Ljung-Box (L1) (Q):	0.17	**Jarque-Bera (JB):**	548319.21
Prob(Q):	0.68	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.27	**Skew:**	0.88
Prob(H) (two-sided):	0.00	**Kurtosis:**	61.32

Nota. Elaboración propia.

5.2.2.2.7 Evaluación de los Modelos ARIMA

Luego de aplicar los modelos, se procedió a evaluarlos mediante métricas, para lo cual se utilizó MAE, MSE, RMSE y R^2.

Figura 95: *Métricas - ARIMA (1,1,1)*

```
In [32]:
#MAE
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(test,pred)
mae

Out[32]: 4.4764379078994905

In [41]:
#MSE
from sklearn.metrics import mean_squared_error
mse= mean_squared_error(test,pred)
mse

Out[41]: 37.03324597128662

In [42]:
#RMSE
mean_squared_error(test,pred)**0.5

Out[42]: 6.085494718696798

In [43]:
#Varianza
from sklearn.metrics import r2_score
r2_score(test,pred)

Out[43]: -0.10991656702803354
```

Nota. Elaboración propia.

Figura 96: *Métricas - ARIMA (2,0,2)*

```
In [32]:
#MAE
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(test,pred)
mae

Out[32]: 3.998164643759228

In [33]:
#MSE
from sklearn.metrics import mean_squared_error
mse=mean_squared_error(test,pred)
mse

Out[33]: 34.2724830186151

In [34]:
#RMSE
mean_squared_error(test,pred)**0.5

Out[34]: 5.85427049414486

In [35]:
#Varianza
from sklearn.metrics import r2_score
r2_score(test,pred)

Out[35]: -0.02717425107811189
```

Nota. Elaboración propia.

Figura 97: *Métricas - ARIMA (3,1,2)*

```
In [28]: #MAE
         from sklearn.metrics import mean_absolute_error
         mae = mean_absolute_error(test,pred)
         mae

Out[28]: 4.515395574881734

In [29]: #MSE
         from sklearn.metrics import mean_squared_error
         mse=mean_squared_error(test,pred)
         mse

Out[29]: 37.40533881157893

In [30]: #RMSE
         mean_squared_error(test,pred)**0.5

Out[30]: 6.115990419513337

In [31]: #Varianza
         from sklearn.metrics import r2_score
         r2_score(test,pred)

Out[31]: -0.12106849273914055
```

Nota. Elaboración propia.

Capítulo VI: Conclusiones y recomendaciones

6.1 Conclusiones

Durante el desarrollo del presente trabajo, aplicando técnicas de Machine Learning se compara diferentes modelos de predicción como SVR, ARIMA y Regresión Lineal, para identificar la técnica más adecuada que ejecute el modelo y sirva de guía para la predicción del caudal como información base para la gestión de riesgos en represas. En este caso, se ha comprobado que los métodos de inteligencia artificial, especialmente los relacionados con la previsión de series temporales, son de gran beneficio para la predicción de caudal en las represas operadas por Autodema.

Inicialmente, en la etapa de adquisición, se identificó el problema objeto de estudio y se recolectó las bases de datos que contenían la información de las distintas variables utilizadas en un periodo de tiempo de los años 2009 hasta 2022 y parte de 2023. Luego, para la etapa de preparación de datos, se evaluó la información de forma estadística y se definió las variables dependientes e independientes.

En la etapa de análisis, se utilizaron las técnicas de Regresión Lineal y SVR, las cuales fueron entrenadas con datos no normalizados y normalizados, luego, teniendo en cuenta la temporalidad de la variable a predecir, se aplicó la técnica ARIMA.

Respecto a la última etapa de reporte, se aplicaron cuatro métricas estadísticas. En ese sentido, de acuerdo con las métricas obtenidas presentadas en la Tabla N° 10, a nivel de la métrica MAE, el modelo de Regresión Lineal obtuvo un valor igual a 5.5357040, el modelo SVR obtuvo un valor igual a 3.6766613, el modelo ARIMA obtuvo un valor igual a 3.9981646. A nivel de la métrica MSE, el modelo de Regresión Lineal obtuvo un valor igual a 109.9134931, el modelo SVR obtuvo un valor igual a 83.7013535, el modelo ARIMA obtuvo un valor igual a 34.2724830. A nivel de la métrica RMSE, el modelo de Regresión Lineal obtuvo un valor igual a 10.4839636, el modelo SVR obtuvo un valor igual a 9.1488444, el modelo ARIMA obtuvo un valor igual a 5.8542705. A nivel de la métrica Varianza (R^2), el modelo de Regresión Lineal obtuvo un valor igual a 0.2239825, el modelo SVR obtuvo un valor igual a 0.4273347, el modelo ARIMA obtuvo un valor igual a -0.0271743. De acuerdo a estos valores se concluye que la técnica SVR trabajada con datos normalizados cuenta con las

mejores métricas para predecir la variable dependiente en comparación con la técnica de Regresión Lineal y ARIMA.

Para concluir, es importante recalcar la importancia de vigilar y evaluar las variables hidrológicas en las represas, apoyándose de herramientas que ayuden a predecir su comportamiento y permitan organizar calendarios de descarga gradual de agua de las represas antes que esta llegue a su nivel máximo de embalse, y la descarga descontrolada o estrepitosa provoque inundaciones u otro tipo de afectaciones en las zonas y comunidades cercanas.

6.2 Recomendaciones

Mediante la aplicación de técnicas de Machine Learning se generaron modelos de predicción del caudal efluente para la represa Condoroma; sin embargo, debido al periodo de tiempo en el que fue desarrollado este trabajo, no se realizaron pruebas con otras técnicas de aprendizaje supervisado. Por lo que futuras investigaciones pueden explorar más a fondo la relación entre las variables utilizadas y aplicar distintas técnicas que incorporen, analicen y predigan múltiples series temporales que aporten información unas a otras como el modelo del tipo Vector Autorregresivo (VAR) y redes neuronales como la red perceptrón multicapa (MLP), red de memoria a corto y largo plazo (LSTM) y red convolucional (CNN).

Además, se sugiere adicionar, a las pruebas realizadas, los datos disponibles de otras instituciones como el SENAMHI y la Autoridad Nacional del Agua que cuentan con estaciones meteorológicas e hidrológicas instaladas en zonas aledañas a la ubicación de la represa.

Finalmente, se recomienda replicar esta metodología, que utiliza técnicas de Machine Learning, en el resto de represas que opera Autodema con el objetivo de implementar un Sistema de Alerta Temprana integrado que pueda prevenir desbordes de represas y posibles inundaciones.

Referencias bibliográficas

Alba, F., & Gonzáles, A. (2017). Machine Learning en la industria: el caso de la siderurgia. https://dialnet.unirioja.es/servlet/articulo?codigo=6207513

Aguamarket. (s/f). *Caudal efluente*. Aguamarket.com. Recuperado el 9 de diciembre de 2022, de https://www.aguamarket.com/diccionario/terminos.asp?Id=4779&termino=Caudal+efluente

Aguilar, A. & Obando, F. (2020). Aprendizaje Automático para la Predicción de Calidad de Agua Potable. *Ingeniare*, *2*(28), 47-62. https://doi.org/10.18041/1909-2458/ingeniare.28.6215

Alonso, J. (2010). Tutorial para Pruebas de Raíces Unitarias: Dickey-Fuller Aumentado y PhillipsPerron en EasyReg . https://ideas.repec.org/p/col/000131/009100.html

Andrades Rodriguez, M., & Muñez León, C. (2012). FUNDAMENTOS DE CLIMATOLOGIA. Logroño: IBERUS.

Annandale, G., Morris, G., & Karki, P. (2016). Extending the Life of Reservoirs : Sustainable Sediment Management for Dams and Run-of-River Hydropower. World Bank. Recuperado de: https://openknowledge.worldbank.org/handle/10986/25085

Artacho, C. (2021). *Desarrollo y aplicación de técnicas de Machine Learning para la predicción de contagios por Covid-19* . https://idus.us.es/bitstream/handle/11441/114871/TFG-3386-ARTACHO%20GOMEZ.pdf?sequence=1&isAllowed=y

Autoridad Nacional del Agua. (2015). Inventario de presas en el Perú (p. 97). Recuperado de: https://www.ana.gob.pe/sites/default/files/publication/files/inventario_de_presas.pdf

Autoridad Autónoma de Majes . (noviembre de 2022). *Autodema*. Recuperado de https://www.autodema.gob.pe/

Autoridad Autónoma de Majes . (2015). *Manual de Operaciones* . Recuperado de https://autodema.gob.pe/wp-content/uploads/2017/08/Manual_de_Operaciones.pdf

Autoridad Autónoma de Majes. (2023). P*lataforma Movimiento Hídrico Sistema Colca de la Autoridad Autónoma de Majes* . Recuperado de https://autodema.gob.pe/reportesom/frmRptcolca.aspx

Awad, M., & Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. Apress.

Babaei, M., Moeini, R., & Ehsanzadeh, E. (2019). Artificial Neural Network and Support Vector Machine Models for Inflow Prediction of Dam Reservoir (Case Study: Zayandehroud Dam Reservoir). *Water Resources Management*, *33*(6), 2203–2218. https://doi.org/10.1007/s11269-019-02252-5

Banco Mundial. (2014). Gestión de los recursos hídricos: Resultados del sector. 2014. Recuperado de: https://www.bancomundial.org/es/results/2013/04/15/water-resources-management-results-profile

Beunza, J., Puertas, E., y Condés, E. (2019). Manual práctico de Inteligencia Artificial en entornos sanitarios. Elsevier España. https://books.google.com.pe/books?id=88nSDwAAQBAJ&pg=PA35&dq=aprendizaje+supervisado+machine+learning&hl=es&sa=X&ved=2ahUKEwj90bDen4D0AhWeGbkGHey7A0U4FBC7BXoECAkQBg#v=onepage&q=aprendizaje%20supervisado%20machine%20learning&f=false

Bobadilla, J. (2020) Machine Learning y Deep Learning. Bogotá. Ra-ma Editorial

Brenes Jimenez Anibal (2020). Predicción del caudal promedio horario de la estación hidrológica Palmar, utilizando modelos de Machine Learning basados en Árboles de decisión. Recuperado de: https://www.kerwa.ucr.ac.cr/handle/10669/81896

Confederación Hidrográfica del Ebro. (2021). Predicción de caudales y gestión de embalses en las crecidas del Ebro. Recuperado de: https://confederaciondelebro.wordpress.com/2021/06/14/prediccion-de-caudales-y-gestion-de-embalses-en-las-crecidas-del-ebro/

Dongsheng, L., Jinfeng, M., & Kaifeng, R. (2023). Prediction of rainfall time series using the hybrid DWT-SVR-Prophet model. https://doi.org/https://doi.org/10.21203/rs.3.rs-2578458/v1

Esayase, T. (2022). Predicting the Peak Flow and Assessing the Hydrologic Hazard of Kessem Dam, Ethiopia using Machine Learning and RMC-RFA Software. Recuperado de: https://assets.researchsquare.com/files/rs-1746769/v2/9f0a46ca-35ff-47b3-9eca-03bf8aa2a327.pdf?c=1663808548

Hancco, N. (2019, February 13). Represa de Condoroma se convierte en peligro para Majes y Camaná. Diario Correo.

Harrington, P. (2012). Machine Learning in Action.

Hernández, R., & Mendoza, C. (2018). Metodología de la Investigación: Las rutas cuantitativas, cualitativas y mixtas. (McGraw-Hill Interamericana Editores S.A de C.V, Ed.). Ciudad de México.

Hong, J., Lee, S., Bae, J. H., Lee, J., Park, W. J., Lee, D., Kim, J. & Lim, K. J. (2020). Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow. *Water*, *12*(10), 2927. https://doi.org/10.3390/w12102927

Hyndman, R.J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3

Lozano-Parra, J. (2018). Recursos hídricos. Disponibilidad, variabilidad y gestión . In Revista de geografía Norte Grande (pp. 5–8).

Manzur, A., & Cardoso, J. (2015). *Velocidad de evaporación del agua*. Revista Mexicana de Fisica E, 61(1), 31–34. Recuperado de: https://www.scielo.org.mx/pdf/rmfe/v61n1/v61n1a7.pdf

Marín, D., & Pineda, I. (2019). Modelo predictivo Machine Learning aplicado a análisis de datos Hidrometeorológicos para un SAT en Represas [Universidad Tecnológica del Perú]. https://repositorio.utp.edu.pe/handle/20.500.12867/3300?show=full

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1(4), 140–147. https://doi.org/10.38094/jastt1457

Miranda Araya, F. (2019). *Uso de redes neuronales artificiales calibradas en el periodo histórico para el pronóstico de caudales de deshielo proyectados en el periodo 2020-2050 en la Cuenca del río Maipo en el Manzano* [Universidad de Chile]. https://repositorio.uchile.cl/handle/2250/170680

Monkhouse, F. (1978). *Diccionario de términos geográficos*. Barcelona: Oikos-Tau Editores.

Nuin, J. J. B., Sanz, E. P., & Moreno, E. C. (2020). Manual práctico de inteligencia artificial en entornos sanitarios. Elsevier Health Sciences. https://books.google.com.py/books?id=88nSDwAAQBAJ

Organización de la Naciones Unidas para la Alimentación y la Agricultura. (s/f). *Estimaciones de las necesidades de agua*. Fao.org. Recuperado el 9 de

diciembre de 2022, de https://www.fao.org/fishery/static/FAO_Training/FAO_Training/General/x6705s/x6705s02.htm

Organización Mundial de la Salud. (2001). Informe sobre la Evaluación Mundial del Abastecimiento de Agua y el Saneamiento en 2000. Recuperado de: https://apps.who.int/iris/bitstream/handle/10665/42420/9243562029_spa.pdf;jsessionid=07B51E7906EFF808CE37CACE171724B4?sequence=1

Orihuela, R. (2020, March 3). Represas de Arequipa incrementan descargas. La República. Recuperado de: https://larepublica.pe/sociedad/2020/03/03/represas-de-arequipa-incrementan-descargas-lrsd/

Paranjpye, V. (1994) History of Large Dam Controversy. A Third World Perspective. En: http://www.riostropicales.com/periodico/struggle.htm

Peña, D. (2005). Análisis de series temporales. Madrid: Alianza Editorial.

Peris, F. (2022). Modelos de predicción en series temporales: Un estudio comparativo entre métodos estadísticos y machine learning [Universidad Jaume I]. https://repositori.uji.es/xmlui/handle/10234/201358

Rezaie-Balf, M., Naganna, S. R., Kisi, O., & El-Shafie, A. (2019). Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: case study of Aswan High Dam. Hydrological Sciences Journal, 64(13), 1629–1646. https://doi.org/10.1080/02626667.2019.1661417

Rodríguez Jiménez, R. M., Benito Capa, Á., & Portela Lozano, A. (2004). Meteorología y Climatología. Madrid: Fundación Española para la Ciencia y la Tecnología.

Sarailidis, G., Wagener, T., & Pianosi, F. (2022). Integrating scientific knowledge into machine learning using interactive decision trees. Computers & Geosciences, 170, 105248. https://doi.org/https://doi.org/10.1016/j.cageo.2022.105248

Schumacher, R. S., Hill, A. J., Klein, M., Nelson, J. A., Erickson, M. J., Trojniak, S. M., & Herman, G. R. (2021). From Random Forests to Flood Forecasts: A Research to Operations Success Story. Bulletin of the American Meteorological Society, 102(9), E1742–E1755. https://doi.org/10.1175/BAMS-D-20-0186.1

Seminario Gastelo, J. (2021) Modelos de predicción para el caudal del río Chira en la estación Ardilla. Recuperado de: https://pirhua.udep.edu.pe/handle/11042/4986

Stambouli, T., & Zapata, N. (s/f). *Evaluación de las pérdidas por evaporación y arrastre y de los cambios microclimáticos durante el riego por aspersión de*

alfalfa. Csic.es. Recuperado el 9 de diciembre de 2022, de http://digital.csic.es/bitstream/10261/43449/1/Martinez-CobA_ComCong2011.pdf

Theobald, O. (2017). *Machine Learning for Absolute Beginners: A Plain English Introduction [Aprendizaje Automático para los Novatos Absolutos: Una introducción plana en inglés]*. *Independently Published.*

WWAP UNESCO. (2021). Informe Mundial de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos 2021: el valor del agua (p. 207). UNESCO. Recuperado de: https://unesdoc.unesco.org/notice?id=p::usmarcdef_0000378890

Zhang, F., & O'Donnell, L. J. (2020). Chapter 7 - Support vector regression (A. Mechelli & S. B. T.-M. L. Vieira (eds.); pp. 123–140). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-815739-8.00007-9