



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Sistemas

**Modelo K vecinos más cercanos para optimizar la
clasificación de datos según el índice de la calidad del
agua de la cuenca superior de la ciudad de Huarmey**

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Jean Paul Fabian PAJUELO LEON

ASESOR

Dr. Hugo Froilan VEGA HUERTA

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Pajuelo, J. (2025). *Modelo K vecinos más cercanos para optimizar la clasificación de datos según el índice de la calidad del agua de la cuenca superior de la ciudad de Huarney*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios autor/ asesor

Datos de autor	
Nombres y apellidos	Jean Paul Fabian Pajuelo Leon
Tipo de documento de identidad	DNI
Número de documento de identidad	71660909
URL de ORCID	https://orcid.org/0009-0006-2003-1973
Datos de asesor	
Nombres y apellidos	Hugo Froilan Vega Huerta
Tipo de documento de identidad	DNI
Número de documento de identidad	06147737
URL de ORCID	https://orcid.org/0000-0002-4268-5808
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	David Mauricio Sánchez
Tipo de documento	DNI
Número de documento de identidad	06445495
Miembro del jurado 1	
Nombres y apellidos	Hugo David Calderón Vilca
Tipo de documento	DNI
Número de documento de identidad	01317748
Datos de investigación	
Línea de investigación	Inteligencia Artificial
Grupo de investigación	Innovando Sistemas Inteligentes

Agencia de financiamiento	Ninguno
Ubicación geográfica de la investigación	Cuenca del río Huarmey, provincia de Huarmey, región Áncash, Perú
Año o rango de años en que se realizó la investigación	2022 – 2025
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones http://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación de Tesis

Siendo las 15:00 horas del día 30 de abril del año 2025, se reunieron virtualmente los docentes designados como miembros de Jurado de Tesis, presidido por el Dr. David Mauricio Sánchez, Dr. Hugo Calderón Vilca (Miembro) y el Dr. Hugo F. Vega Huerta (Miembro Asesor), usando la plataforma Meet (<https://meet.google.com/cof-vbj-e-nbp>), para la sustentación Virtual de la tesis Intitulada: **“MODELO K VECINOS MÁS CERCANOS PARA OPTIMIZAR LA CLASIFICACIÓN DE DATOS SEGÚN EL ÍNDICE DE LA CALIDAD DEL AGUA DE LA CUENCA SUPERIOR DE LA CIUDAD DE HUARMEY”**, del Bachiller: **JEAN PAUL FABIAN PAJUELO LEON**; para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a responder las preguntas formuladas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez las preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de 14 (catorce)

A continuación, el Presidente del Jurado Dr. David Mauricio Sánchez, declara al Bachiller **Ingeniero de Sistemas**.

Siendo 16:00 horas, se levantó la sesión.

Dr. David Mauricio Sánchez
Presidente

Dr. Hugo Calderón Vilca
Miembro

Dr. Hugo F. Vega Huerta
Miembro Asesor



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Vicerrectorado de Investigación y Posgrado



Yo **Hugo Froilán Vega Huerta** en mi condición de **asesor** acreditado con la Resolución Directoral N° 00002-2024-EPISI-FISI/UNMSM de la **tesis**, cuyo título es: **MODELO K VECINOS MÁS CERCANOS PARA OPTIMIZAR LA CLASIFICACIÓN DE DATOS SEGÚN EL ÍNDICE DE LA CALIDAD DEL AGUA DE LA CUENCA SUPERIOR DE LA CIUDAD DE HUARMEY**, presentado por el Bachiller **JEAN PAUL FABIAN PAJUELO LEON** para optar el título de **INGENIERO DE SISTEMAS**, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de 14 % de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.

Firma del Asesor

DNI: 06147737

Nombres y apellidos del asesor: Dr. Hugo Froilán Vega Huerta

DEDICATORIA

A mi madre, por su apoyo incondicional, y a mis abuelos por estar conmigo en todo momento.

Jean Paul Pajuelo León.

AGRADECIMIENTOS

A Dios por concedernos la sabiduría, constancia, fortaleza, dedicación y compromiso necesarios para llevar a cabo esta tesis.

A mi familia, por ser el pilar fundamental, por comprender cada desvelo y por estar presentes en cada lágrima derramada mientras redactábamos cada párrafo.

Al Dr. Hugo Vega Huerta, por su tiempo, su apoyo, su disposición y su valiosa mentoría.

RESUMEN

La calidad del agua en la cuenca del río Huarmey enfrenta desafíos significativos debido a la contaminación por metales pesados y parámetros fisicoquímicos que superan los estándares de calidad ambiental del agua en Perú. Tradicionalmente, la evaluación del Índice de Calidad de Agua (ICA) se realiza mediante fórmulas y macros de Excel, lo que implica largos tiempos de procesamiento, alta propensión a errores y limitaciones frente a grandes volúmenes de datos.

La presente investigación propone la implementación de un modelo de clasificación basado en el algoritmo K-vecinos más cercanos (KNN) para optimizar la categorización de muestras de agua en las categorías de Consumo Humano (1-A2) y Bebida Animal (3-D2), conforme al ICA Perú. Además, se comparó su desempeño con los algoritmos Support Vector Machine (SVM) y Random Forest (RF), aplicando normalización Min-Max, validación cruzada 5-fold y ajuste de hiperparámetros.

Los resultados demostraron que KNN alcanzó una precisión (accuracy) del 95.2%, un F1-Score de 0.94 y un coeficiente de determinación (R^2) de 0.91, superando a SVM (accuracy 91.4%, F1-Score 0.89, R^2 0.85) y Random Forest (accuracy 93.7%, F1-Score 0.92, R^2 0.88). Estas métricas validan la efectividad y robustez de KNN en la clasificación multivariable de parámetros fisicoquímicos.

La automatización del proceso permitió disminuir los tiempos de clasificación y reducir significativamente los errores en el cálculo del ICA, lo que beneficia la toma de decisiones de la gestión de los recursos hídricos de la cuenca de Huarmey. Este algoritmo de clasificación contribuye al fortalecimiento de las herramientas de monitoreo ambiental y representa un aporte relevante para optimizar la gestión de los recursos hídricos en contextos similares.

PALABRAS CLAVE: CALIDAD DE AGUA, ÍNDICE DE CALIDAD DE AGUA (ICA), ALGORITMO KNN, MACHINE LEARNING, PARÁMETROS FISICOQUÍMICOS.

ABSTRACT

The water quality in the Huarmey River basin faces significant challenges due to contamination by heavy metals and physicochemical parameters that exceed the environmental water quality standards in Peru. Traditionally, the assessment of the Water Quality Index (WQI) is conducted using formulas and Excel macros, which involve lengthy processing times, a high propensity for errors, and limitations when handling large data volumes.

The present research proposes the implementation of a classification model based on the K-Nearest Neighbors (KNN) algorithm to optimize the categorization of water samples into the categories of Human Consumption (1-A2) and Animal Drinking (3-D2), in accordance with the ICA Peru. Additionally, its performance was compared with the Support Vector Machine (SVM) and Random Forest (RF) algorithms, applying Min-Max normalization, 5-fold cross-validation, and hyperparameter tuning.

The results demonstrated that KNN achieved an accuracy of 95.2%, an F1-Score of 0.94, and a coefficient of determination (R^2) of 0.91, outperforming SVM (accuracy 91.4%, F1-Score 0.89, R^2 0.85) and Random Forest (accuracy 93.7%, F1-Score 0.92, R^2 0.88). These metrics validate the effectiveness and robustness of KNN in the multivariable classification of physicochemical parameters.

The automation of the process reduced classification times and significantly decreased errors in the calculation of the WQI, benefiting decision-making in the management of water resources in the Huarmey basin. This classification algorithm contributes to the enhancement of environmental monitoring tools and represents a valuable contribution to optimizing water resource management in similar contexts.

KEYWORDS: WATER QUALITY, WATER QUALITY INDEX (WQI), KNN ALGORITHM, MACHINE LEARNING, PHYSICOCHEMICAL PARAMETERS.

ÍNDICE

DEDICATORIA	1
AGRADECIMIENTOS	2
ÍNDICE	5
ÍNDICE DE TABLAS	9
ÍNDICE DE FIGURAS	10
Introducción	13
CAPÍTULO I: Definición del problema	14
1.1 Antecedentes	14
1.1.1 Proceso actual AS-IS.....	20
1.2 Análisis y definición del problema	21
1.2.1 Problema general.....	21
1.2.2 Problemas específicos	22
1.3 Objetivos	24
1.3.1 Objetivo principal.....	24
1.3.2 Objetivos específicos	24
1.4 Justificación	26
1.4.1 Justificación práctica.....	26
1.4.2 Justificación teórica	27
1.4.3 Justificación metodológica.....	28
1.5 Propuesta	28
1.6 Alcances y limitaciones	29
1.6.1. Alcances.....	29
1.6.2. Limitaciones.....	30
1.7 Organización de la tesis	31
CAPÍTULO II: MARCO TEÓRICO	32
2.1 Conceptos principales	32
2.1.1 K vecinos más cercanos (KNN).....	32
2.1.2 Machine learning (ML)	32
2.1.3 Calidad de agua (CA).....	32
2.1.4 Cuenca hidrográfica (CH)	33
2.1.5 Parámetros de calidad del agua	33
2.1.6 Índice de calidad de agua (ICA).....	34
2.1.7 Autoridad Nacional de agua (ANA)	35
2.1.8 Random Forest (RF).....	35
2.1.9 Support vector machine (SVM)	35

CAPÍTULO III: ESTADO DE ARTE.....	36
3.1 Taxonomía.....	36
• ACM:.....	36
• IEEE:	36
• RR línea de investigación UNMSM:	36
3.2 Metodología de la revisión de la literatura.....	36
3.2.1 Criterios de inclusión	37
3.2.2 Criterios de exclusión	37
3.3 Revisión de la literatura.....	37
3.3.1 Revisión de artículos.....	37
3.3.1.1 Artículo 1: Prediction of groundwater quality using efficient machine learning technique (Singha et al., 2021).	37
3.3.1.2 Artículo 2: Advancing Water Quality Research: K-Nearest Neighbor Coupled with the Improved Grey Wolf Optimizer Algorithm Model Unveils New Possibilities for Dry Residue Prediction (Tahraoui et al., 2023).	39
3.3.1.3 Artículo 3: Prediction of hydrological and water quality data based on granular-ball rough set and k-nearest neighbor analysis (Dong et al., 2024a).	40
3.3.1.4 Artículo 4: Machine learning applications for anomaly detection in Smart Water Metering Networks: A systematic review (Kanyama et al., 2024).	41
3.3.1.5 Artículo 5: Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms (Hussein et al., 2024).	42
3.3.1.6 Artículo 6: Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data (Umar & Gray, 2023a). 43	
3.3.1.7 Artículo 7: Accuracy Assessment of RFerns, NB, SVM, and kNN Machine Learning Classifiers in Aquaculture (Çakir et al., 2023).	45
3.3.1.8 Artículo 8: Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques (Aish et al., 2023a).	45
3.3.1.9 Artículo 9: Drinking water potability prediction using machine learning approaches: a case study of Indian rivers (Ainapure et al., 2023).	46
3.3.1.10 Artículo 10: Ensemble machine learning paradigms in hydrology: A review (Zounemat-Kermani et al. 2021b).	47
3.3.1.11 Artículo 11: Water quality classification using machine learning algorithms (Nasir et al., 2022b).	48
3.3.1.12 Artículo 12: Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors (Naghibi et al., 2020).	50
3.3.1.13 Artículo 13: Performance analysis of the water quality index model for predicting water state using machine learning techniques (Uddin et al., 2023b).	51

3.3.1.14 Artículo 14: A combined water quality classification model based on kernel principal component analysis and machine learning techniques (Dilmi, 2022).	52
3.3.1.15 Artículo 15: Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water (Saddiqi et al., 2024).	53
3.3.2 Resumen de la revisión de los artículos	54
3.4.1 Aplicativos para el cálculo del índice de calidad de agua (ICA)	65
3.4.1.1. WQI Calculator.	65
3.4.1.2. Aquacheck.	66
3.4.1.3. Water Quality Index (WQI) Tool by USGS.	66
3.4.1.4. HACH WIMS.	66
3.4.1.5. Water Quality Index Calculator (WQIC).	66
3.4.1.6. AquaSpy.	67
3.4.1.7. Open Water Analytics (OWA) Tools.	67
CAPÍTULO IV: RESOLUCIÓN DEL PROBLEMA	68
4.1. Características de las cuencas de Huarmey	68
4.2. Fuentes contaminantes de la cuenca de Huarmey	73
4.3 Criterios de evaluación para el cálculo del índice de calidad de agua (ICA)	75
4.3.1. Estándares de Calidad Ambiental del Agua.	75
4.3.2. Clasificación de cuerpos de agua superficiales.	78
4.3.3. Índice de calidad de agua en Perú.	79
4.3.4. Cálculo del índice de calidad de agua en el Perú.	80
4.4. Uso del algoritmo K vecinos más cercanos (KNN) ante otros algoritmos de clasificación	87
4.4.1 El uso de K vecinos más cercanos (KNN).	87
4.4.2 Comparación metodológica de desarrollo	89
4.5 Implementación de la propuesta del algoritmo para la clasificación de los parámetros fisicoquímicos de las muestras de agua de la cuenca de Huarmey	91
4.5.1 Software y lenguaje utilizado.	91
4.5.1.1. Colaboratory como plataforma para codificar.	91
4.5.1.2. Python como lenguaje a implementar.	92
4.5.2. Desarrollo del algoritmo de entrenamiento para la función principal	94
4.5.2.1. Estructurar el dataset de entrenamiento para el cálculo del índice de calidad de agua.	95
4.5.2.2. Recopilación de datos para el entrenamiento.	97

4.5.2.3. Generación de datos faltantes para el Dataset del algoritmo de entrenamiento	99
4.5.2.4. Generación de modelos en base al algoritmo entrenado	101
4.5.3 Estructura de contenido del código a implementar.....	112
4.5.4 Desarrollo del código principal a implementar.....	114
4.5.5 Proceso de la Función generar_KNN_multiclass	128
4.5.6 Evitar el Overfitting con Validación Cruzada	129
4.5.7 Módulo del Algoritmo para R2 y su Uso.....	130
4.5.8 Impacto del Enfoque en el Cálculo del ICA	130
4.5.9 Formato de la data de entrada y ejecución del algoritmo	131
4.6 Comparación de precisión entre KNN, SVM y Random Forest	135
CAPÍTULO V: RESULTADOS	137
5.1. Proceso de resultados de los datos de la Cuenca Huarney del año 2020	139
5.2. Proceso de resultados de los datos de la Cuenca Huarney del año 2021	144
5.3. Proceso de resultados de los datos de la Cuenca Huarney del año 2023	150
5.4. Comparación de resultados de los monitoreos en base a las funciones del cálculo del ICA	155
5.5. Evaluación comparativa de desempeño del modelo KNN frente a métodos tradicionales	159
5.5.1 Comparativa de precisión y eficiencia: KNN frente a Macros de Excel	159
5.5.2 Evaluación del Índice de Calidad del Agua generado por el modelo	160
5.5.3 Evaluación del Índice de Calidad del Agua generado por el modelo	160
CAPÍTULO VI: DISCUSIONES.....	162
CAPÍTULO VII: CONCLUSIONES	163
CAPÍTULO VIII: CONCLUSIONES.....	165
REFERENCIAS BIBLIOGRÁFICAS	166

ÍNDICE DE TABLAS

Tabla 1 <i>Resumen de artículos revisados relacionados a calidad del agua y machine learning.</i>	54
Tabla 2 <i>Aspectos generales de la cuenca del río Huarmey.</i>	68
Tabla 3 <i>Fuentes contaminantes de las cuencas del río Huarmey.</i>	73
Tabla 4 <i>Parámetros de la muestra de las cuencas de Huarmey según su categoría. ...</i>	131
Tabla 5 <i>Métricas de evaluación.</i>	135
Tabla 6 <i>Cuadro comparativo con los resultados de cada monitorio con respecto al resultado del algoritmo principal.</i>	156
Tabla 7 <i>Cuadro comparativo con respecto al tiempo en clasificar los datos del monitoreo de las cuencas de Huarmey entre los resultados de cada monitorio con respecto al resultado del algoritmo principal.</i>	157
Tabla 8. <i>Comparación de métricas de desempeño entre KNN, Excel, SVM y Random Forest.</i>	159
Tabla 9. <i>Resultados de la validación cruzada 5-fold del modelo KNN.</i>	160

ÍNDICE DE FIGURAS

Figura 1 Informe sobre la calidad del agua en la cuenca del río Huarmey.	15
Figura 2 Contraloría entregó el informe a la Municipalidad Provincial de Huarmey sobre los ocho centros poblados de Huarmey que consumen agua sin cloración.	17
Figura 3 Errores comunes en el manejo de macros de Excel para cálculos técnicos.	18
Figura 4. Arquitectura del modelo ANFIS con 2 entradas y 5 capas.	19
Figura 5 Modelo AS-IS del tratamiento del agua en la ciudad de Huarmey.	20
Figura 6 Árbol del problema en la gestión de la calidad del agua en la ciudad de Huarmey.	24
Figura 7 Árbol de objetos para mejorar la clasificación de categorías de parámetros de agua según su ICA en la ciudad de Huarmey.	25
Figura 8 Modelo TO-BE propuesto para aumentar la precisión en hallar los contaminantes de agua en el tratamiento de la ciudad de Huarmey.	29
Figura 9 Localización sobre la cuenca del río Huarmey.	69
Figura 10 Descripción de agentes/fuentes contaminantes en el agua.	80
Figura 11 Parámetros considerados en la Categoría 1-A2 Poblacional y Recreacional: Aguas que pueden ser potabilizadas con tratamiento convencional.	82
Figura 12 Parámetros considerados en la Categoría 3 - Riego de vegetales y bebida de animales.	82
Figura 13 Interpretación de la Calificación ICA- PE.	84
Figura 14 Información base necesaria para la determinación del ICA.	85
Figura 15 Diagrama de flujo ECA para el algoritmo de entrenamiento basado en el cálculo del ICA – Perú.	95
Figura 16 Imagen de búsqueda de dataset relacionado a los parámetros del índice de calidad de agua para las categorías 1-A2 y 3-D2.	97
Figura 17 Imagen de recopilación del Dataset relacionada a los parámetros del cálculo de índice de calidad de agua extraídos desde la página web de Kaggle en tablas de Excel.	98
Figura 18 Diagrama de flujo de las funciones para generar datos basados en los parámetros del ICA para el algoritmo de entrenamiento del cálculo del ICA – Perú.	99
Figura 19 Código de entrenamiento para el algoritmo de K vecinos más cercanos (KNN) desarrollado en Colab.	101
Figura 20 Código de entrenamiento para el algoritmo de Support vector machine (SVM) desarrollado en Colab.	103
Figura 21 Código de entrenamiento para el algoritmo de Random Forest desarrollado en Colab.	105
Figura 22 Función principal del algoritmo de entrenamiento para SVM, Random Forest y KNN desarrollado en Colab.	106
Figura 23 Resultados del entrenamiento del algoritmo K vecinos más cercanos (KNN) en los datos para la categoría 1-A2 de consumo humano.	109

Figura 24 Resultados del entrenamiento del algoritmo K vecinos más cercanos (KNN) en los datos para la categoría de bebida animal 3-D2.	110
Figura 25 Estructura del código para la implementación del algoritmo principal para la clasificación de agua mediante el ICA-Perú en la plataforma Colab.	113
Figura 26 Importación de librerías en la elaboración del algoritmo de clasificación en Colab.	114
Figura 27 Estados iniciales para las clases ECA en la elaboración del algoritmo de clasificación en Colab.	116
Figura 28 Estados iniciales de la clase ECA para la categoría de Consumo Humano 1-A2 del algoritmo de clasificación en Colab.	117
Figura 29 Estados iniciales de la clase ECA para la categoría de Bebida de Animales 3-D2 del algoritmo de clasificación en Colab.	117
Figura 30 Clases y funciones principales del módulo de Clase KNN para el algoritmo de clasificación en Colab.	118
Figura 31 Clases y funciones principales del módulo de Clase ECA para el algoritmo de clasificación en Colab.	119
Figura 32 Clases y funciones principales del módulo de Clase Category1A2 para el algoritmo de clasificación en Colab.	120
Figura 33 Clases y funciones principales del módulo de Clase Category3 para el algoritmo de clasificación en Colab.	121
Figura 34 Clases y funciones principales del módulo de Clase AnimalDrinkingWater para el algoritmo de clasificación en Colab.	122
Figura 35 Módulo de la categoría de Consumo Humano para su clasificación en Colab.	124
Figura 36 Módulo de la categoría de Bebida de Animales para su clasificación en Colab.	124
Figura 37 Diagrama de flujo del algoritmo para la Categoría de Consumo Humano 1-A2.	125
Figura 38 Diagrama de flujo del algoritmo para la Categoría de Bebida de Animales 3-D2.	126
Figura 39 Función principal del algoritmo para mostrar los resultados de la clasificación de datos de muestras de agua de las cuencas de Huarmey.	127
Figura 40 Tabla de datos de entrada del algoritmo principal en formato de cuadrillas de Excel.	132
Figura 41 Cuadro elaborado con los datos del monitoreo de las muestras de agua del mar de Huarmey del año 2023.	133
Figura 42 Tabla de datos de entrada con la columna de Aluminio (al) conteniendo los datos del monitoreo de muestras de agua de mar de Huarmey del 2023.	134
Figura 43 Clasificación de los resultados en base a la puntuación del resultado de los parámetros del monitoreo establecidos por el Índice de Calidad de Agua (ICA) – Perú.	

.....	137
Figura 44 <i>Tabla de datos de las muestras de agua del monitoreo de las cuencas de Huarmey en Alto-Rio-Huarmey del año 2020.</i>	139
Figura 45 <i>Datos de entrada en el Template ICA del monitoreo del 2020 para el algoritmo del cálculo de ICA y clasificación según sus resultados.</i>	140
Figura 46 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2020 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.</i>	141
Figura 47 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2020 para la categoría 3 de bebida animal de las cuencas de Huarmey.</i>	143
Figura 48 <i>Tabla de datos de las muestras de agua inorgánicas del monitoreo de las cuencas de Huarmey Rio-Huarmey del año 2021.</i>	144
Figura 49 <i>Tabla de datos de las muestras de agua fisicoquímicos del monitoreo de las cuencas de Huarmey Rio-Huarmey del año 2021.</i>	145
Figura 50 <i>Datos de entrada en el Template ICA del monitoreo del 2021 para el algoritmo del cálculo de ICA y clasificación según sus resultados.</i>	146
Figura 51 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2021 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.</i>	146
Figura 52 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2021 para la categoría 3 de bebida animal de las cuencas de Huarmey.</i>	148
Figura 53 <i>Tabla de datos de las muestras de agua del monitoreo de las cuencas de Huarmey Rio-Huarmey del año 2023.</i>	150
Figura 54 <i>Datos de entrada en el Template ICA del monitoreo del 2023 para el algoritmo del cálculo de ICA y clasificación según sus resultados.</i>	151
Figura 55 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2023 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.</i>	152
Figura 56 <i>Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2023 para la categoría 3 de bebida animal de las cuencas de Huarmey.</i>	153

Introducción

La calidad del agua es un factor crítico para el bienestar humano, el desarrollo sostenible y la protección del medio ambiente. En el Perú, múltiples cuencas hidrográficas presentan altos niveles de contaminación debido a actividades industriales, mineras y agrícolas. Particularmente en la ciudad de Huarney, ubicada en la región de Áncash, las fuentes de agua enfrentan serias amenazas por la presencia de metales pesados y parámetros fisicoquímicos fuera de norma, afectando el suministro de agua para consumo humano y bebida animal.

Tradicionalmente, la evaluación del Índice de Calidad del Agua (ICA) en esta región ha dependido del uso de macros en hojas de cálculo de Excel. Si bien este enfoque ha sido funcional, presenta serias limitaciones en cuanto a precisión, escalabilidad y tiempo de procesamiento. La manipulación manual de datos y fórmulas incrementa el riesgo de errores, especialmente en contextos donde se manejan grandes volúmenes de información ambiental.

En respuesta a esta problemática, la presente investigación propone la implementación de un modelo de aprendizaje automático basado en el algoritmo K vecinos más cercanos (KNN), orientado a clasificar muestras de agua según las categorías establecidas por el ICA en Perú: Consumo Humano (1-A2) y Bebida Animal (3-D2). KNN ha demostrado ser eficaz en tareas de clasificación multivariada en estudios previos, gracias a su simplicidad, robustez y capacidad de adaptación a datos reales sin necesidad de una función de predicción explícita.

Esta tesis tiene como objetivo principal optimizar el proceso de clasificación de muestras de agua mediante el modelo KNN, reduciendo significativamente los errores manuales y mejorando la rapidez y confiabilidad del análisis. La solución se valida con datos reales de monitoreo de los años 2020, 2021 y 2023 en la cuenca superior de Huarney. El enfoque propuesto busca no solo acelerar el cálculo del ICA, sino también ofrecer una herramienta replicable para la gestión de recursos hídricos en otras zonas del país.

CAPÍTULO I: Definición del problema

1.1 Antecedentes

El agua es un valioso recurso utilizado en una gran cantidad de actividades en el mundo. La calidad del agua potable es fundamental para el bienestar social y ambiental, ya que influye directamente en la mitigación de riesgos para la salud pública, protege las cuencas de posibles contaminantes y asegura la preservación de la calidad de las aguas tanto superficiales como subterráneas. La calidad del agua potable llevada a cabo por el agua residual tratada en el Perú es uno de los grandes problemas a nivel nacional que tiene el país. La contaminación del agua en Perú proviene de varias fuentes, incluyendo la minería, la agricultura, y la falta de tratamiento de aguas residuales. Según la Organización Mundial de la Salud (OMS), "La minería en Perú, especialmente la extracción de oro y cobre ha causado un impacto significativo en los cuerpos de agua debido al uso de químicos tóxicos como el mercurio y el cianuro (Organización Mundial de la Salud, 2022)". La contaminación del agua tiene serias consecuencias para la salud de las personas. Según un informe del Ministerio de Medio Ambiente del Perú, "El agua contaminada con metales pesados y patógenos puede causar enfermedades gastrointestinales, problemas renales y efectos neurológicos" (Ministerio de medio ambiente, 2015). En las zonas rurales, la falta de infraestructura adecuada para el tratamiento de aguas residuales agrava el problema. Un estudio del Instituto Nacional de Estadística e Informática (INEI) señala que "el 73,1% de los peruanos enfrentó la falta de acceso a agua gestionada de manera segura, por lo que se evidenció un incremento del 1,6% (Instituto Nacional de Estadística e Informática, 2023)". En las áreas urbanas, la situación no es mucho mejor.

A pesar de estos esfuerzos, aún hay desafíos significativos. Un informe del Banco Mundial indica que "es necesario un enfoque más integrado que incluya la mejora de la infraestructura, la regulación estricta y la educación pública para enfrentar la crisis de contaminación del agua en Perú. Las sequías, inundaciones y deslizamientos de tierras asociados con lluvias torrenciales, la contaminación de las aguas y la falta de acceso a agua y saneamiento seguros le cuestan al Perú entre

8.400 millones y 13.400 millones de dólares al año” (Banco Mundial, 2024). Es muy importante gestionar de forma correcta el tratamiento de agua en el país.

La contaminación de las cuencas hidrológicas donde se extrae las fuentes de agua para ser tratadas en la ciudad de Huarney tiene ya casi un siglo por la mala gestión de la calidad de agua, la cual viene causando contaminación en el medio ambiente, los campos agrícolas, el consumo de agua potable y las enfermedades causadas por su consumo. Según el informe de Ancash Noticias en el año 2021, en donde se determinó que “la calidad del agua superficial en el río Huarney y la de sus afluentes está afectada por los efluentes de la zona alta del mismo río, principalmente por metales totales y parámetros físicoquímicos (DBO, DQO, sulfatos, fósforo total)” (Ancash Noticias, 2021). “La Autoridad Nacional del Agua efectuó la constatación de la calidad de agua en la provincia de Huarney durante dos semanas entre noviembre y diciembre de 2019, y dieron como resultados la contaminación por metales que superan los estándares de calidad del agua (ECA) permitidos en los diferentes puntos mencionados incluyen: cadmio, manganeso, plomo, zinc, aluminio, arsénico, cobre, hierro, mercurio y zinc” (Ancash Noticias, 2021). En la Figura 1 se presenta la portada del informe técnico correspondiente a este estudio.

Figura 1

Informe sobre la calidad del agua en la cuenca del río Huarney.



Nota. La imagen muestra el informe de la Autoridad Nacional del Agua (ANA)

sobre la calidad del agua en la cuenca del río Huarmey, donde se identificaron altos niveles de metales y parámetros fisicoquímicos. Recuperado de <https://ancashnoticias.com/2021/01/03/ancash-ana-constata-la-mala-calidad-de-la-cuenca-de-agua-del-rio-de-huarmey/>

La mala calidad del agua en la ciudad de Huarmey tiene diversos factores por los cuales se tornan en un problema mayor. La mala gestión en el tratamiento del agua, la política y las leyes en la minería ilegal a los bordes de las cuencas donde se extrae la fuente del agua a tratar y la gestión de la calidad del agua con métodos subjetivos que no hacen precisos ver los contaminantes fisicoquímicos y metales pesados del agua, llevan a cabo una cantidad de problemas al consumo de los pobladores de la ciudad de Huarmey. Contraloría General de la República comunicó en el año 2021 a la municipalidad provincial de Huarmey, sobre riesgos en la salud de la población, sobre el “Agua para consumo humano sin cloración previa, que podría afectar la salud de cientos de personas, se distribuye en ocho centros poblados donde existen Juntas Administradoras de Servicios de Saneamiento (JASS)” (La República, 2021). Según la Organización Mundial de la Salud (OMS), "la contaminación del agua puede derivarse de fuentes agrícolas, industriales y residenciales, y puede tener efectos devastadores en la salud humana y ecosistemas" (Organización Mundial de la Salud, 2022). La oficina de la Defensoría del Pueblo en Huancavelica en el año 2021 recomendó “a las municipalidades distritales, y en particular a los responsables de sus Áreas Técnicas Municipales (ATM), adoptar las medidas necesarias para proteger el derecho de la población a acceder a agua apta para el consumo humano. Asimismo, resaltó que los gobiernos locales no han emitido ordenanzas para la mejora de la prestación de servicios; y las ATM no han elaborado planes de supervisión e instrumentos para el seguimiento y evaluación de la aplicación de cloro del agua que se distribuye a la población. Por otra parte, dichas áreas carecen de un fondo rotatorio para la compra de cloro, y no se les asigna presupuesto para la adquisición de equipos, kits y mantenimiento de sistemas de agua colapsados” (Defensoría del Pueblo, 2021).

Desde las enfermedades hasta la contaminación ambiental, una mala gestión del

agua por el mal manejo de la calidad del agua, traen consigo en retroceso y detenimiento del avance en el tratamiento correcto del agua, por lo cual, se crean diferentes problemas en el consumo y derivación a los sectores de la población, por lo que, a través de los años, se ha tenido que soportar la mala gestión y prácticas en el tratamiento del agua que consumen los huarmeyanos. La Figura 2 muestra una imagen de una inspección previa a los servicios de saneamiento a cargo de la JASS Tayca en el año 2021, la Superintendencia Nacional de Servicios de Saneamiento (Sunass) constató que “no hay un sistema de precloración, así como tampoco la información sobre la calidad de las fuentes de agua, limpieza y desinfección del reservorio, análisis físico químico y del nivel de turbiedad del agua, controles de presión y continuidad del servicio” (La República, 2021).

Figura 2

Contraloría entregó el informe a la Municipalidad Provincial de Huarmey sobre los ocho centros poblados de Huarmey que consumen agua sin cloración.



Nota. La imagen muestra la entrega del informe por parte de la Contraloría a la Municipalidad Provincial de Huarmey sobre la calidad del agua en los centros poblados de la región. Recuperado de (La República, 2021).

Ante la situación descrita, “la comisión de control solicitó información al área técnica de la Municipalidad Provincial de Huarney e identificó que 8 de los 10 poblados que cuentan con JASS carecen de sistema de precloración; además, advirtió demora por parte de la comuna provincial en la implementación de las recomendaciones de Sunass sobre las deficiencias detectadas en Tayca. El informe de orientación de oficio fue elaborado por el Órgano de Control Institucional (OCI) de la Municipalidad Provincial de Huarney, y fue comunicado al titular de la entidad para las acciones que correspondan en salvaguarda de la salud de los habitantes de los ocho centros poblados afectados” (La República, 2021).

A nivel nacional, el problema se extiende a otras cuencas hidrográficas. Un estudio del Instituto Nacional de Estadística e Informática (INEI) indicó que, en 2023, el 73.1% de los peruanos no tenía acceso a agua gestionada de manera segura, lo que representó un incremento del 1.6% respecto al año anterior (Instituto Nacional de Estadística e Informática, 2023).

Por otro lado, los métodos convencionales para evaluar la calidad del agua mediante macros de Excel presentan limitaciones, como tiempos prolongados de procesamiento y alta propensión a errores. En la Figura 3 se muestra un ejemplo de errores comunes en el manejo de macros de Excel para cálculos técnicos.

Figura 3

Errores comunes en el manejo de macros de Excel para cálculos técnicos.

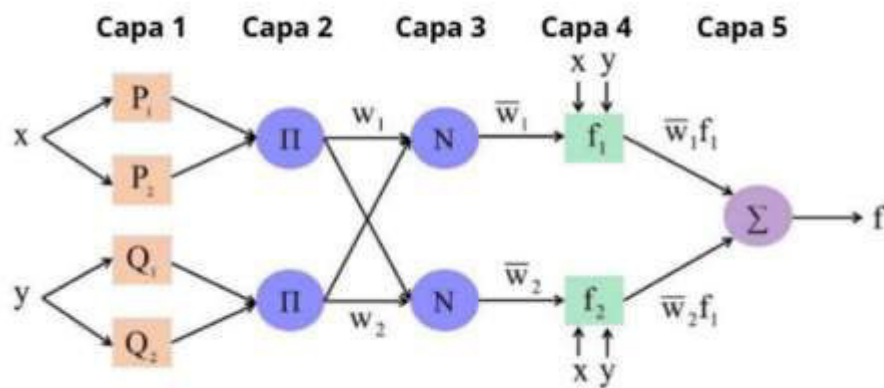
	A	B	C	D	E	F	G	H
1								
2								
3	Negocio de Acciones							
4								
5	Valor neto total	#DIV/0!						
6	Valor actual total	37.475						
7	Costo total	28.838						
8	Total acciones	1.000						
9								
10								
11	Empresas	Precio de compra	Núm de acciones	Costo total	Precio actual	Valor actual	Cambio neto	Valor neto
12	LAN	34,50	200	6.900	38,25	7.650	11%	750
13	Nestlé	22,13	500	11.063	29,5	14.750	33%	3.688
14	Quebrados S.A	0,00	300	10.875	50,25	15.075	#DIV/0!	#DIV/0!

Nota. *Guía de macros en Excel para análisis de datos: Ahorre tiempo y reduzca errores.* Recuperado de: <https://sigmacomputing.com/excel-macros-errors/>

En un contexto de avance tecnológico, varios estudios han propuesto el uso de técnicas de Machine Learning para optimizar la evaluación de la calidad del agua. Condorcuya R. (2025) propuso el uso de Adaptive Neuro-Fuzzy Inference System (ANFIS) para la clasificación de calidad de agua potable en la Universidad Nacional Micaela Bastidas de Apurímac (Condorcuya et al., 2025). En la Figura 4 se presenta la arquitectura del modelo ANFIS con dos entradas y cinco capas.

Figura 4.

Arquitectura del modelo ANFIS con 2 entradas y 5 capas.



Nota. *Modelo de Aprendizaje Automático para la clasificación de calidad de agua potable* (Condorcuya et al., 2025).

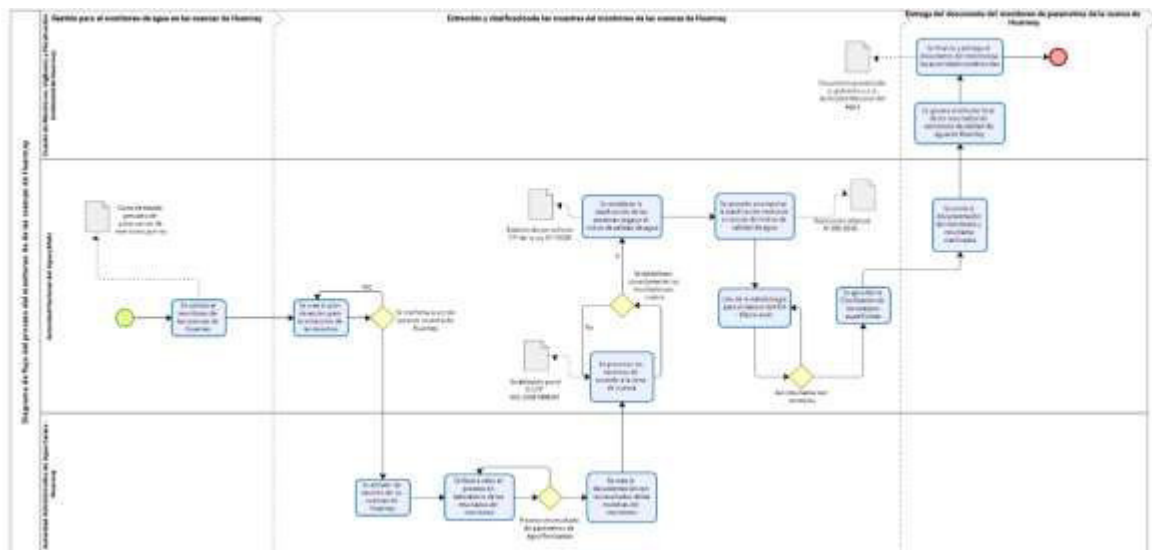
La implementación de sistemas automáticos para el análisis de la calidad del agua, como los modelos basados en Machine Learning, puede contribuir significativamente a reducir tiempos de procesamiento y minimizar errores, facilitando la toma de decisiones y mejorando la gestión de los recursos hídricos en Huarmey y otras regiones afectadas por la contaminación del agua.

1.1.1 Proceso actual AS-IS

El proceso AS-IS permite documentar y entender el estado actual de un sistema o proceso antes de proponer mejoras o cambios. Es una fotografía de cómo operan las cosas en el momento presente. Los flujos AS-IS son fundamentales para asegurar que las propuestas de mejora en una tesis sean pertinentes, bien fundamentadas y claramente articuladas, haciendo la investigación más robusta y sus conclusiones más válidas. Se muestra la Figura 5 para mostrar el flujo en el cual se calcula el índice de calidad de agua de un monitoreo en las cuencas de Huarmey.

Figura 5

Modelo AS-IS del tratamiento del agua en la ciudad de Huarmey.



Nota. Elaboración propia.

La imagen muestra un diagrama de flujo detallado que representa el proceso de gestión y monitoreo de la calidad del agua en las cuencas de Huarmey. El diagrama está estructurado en tres columnas principales, describiendo las etapas desde la recolección inicial de datos hasta la generación de informes finales. El proceso detalla como las autoridades encargadas de realizar el informe del monitoreo de la cuenca de Huarmey establecen mediante reglas y normas establecidas el cálculo del índice de calidad de agua (ICA) utilizando un macro Excel para poder categorizar y clasificar las muestras obtenidas en dicho monitoreo. Esto se presenta con factores

de tiempo en la aceptación y validez del cálculo del ICA y la clasificación de estas para el Consumo Humano y Bebida de Animales de la cuenca del río Huarmey.

1.2 Análisis y definición del problema

1.2.1 Problema general

Las cuencas fluviales en las áreas urbanas del Perú sufren la confluencia de múltiples fuentes de contaminación, incluyendo descargas industriales, aguas residuales no tratadas, escurrimientos agrícolas y residuos sólidos. Esto complica la evaluación precisa del índice de calidad del agua debido a la variedad y la variabilidad de los contaminantes presentes, lo que requiere métodos sofisticados de muestreo y análisis que pueden ser costosos y técnicamente desafiantes. El manejo de fórmulas tradicionales para el cálculo del índice de calidad de agua y un manejo de macro Excel como la validación manual de las fórmulas empleadas conlleva a un mayor tiempo de lo planificado y/o la extensión de la documentación de la clasificación de los parámetros encontrados en el muestreo del agua de las cuencas de Huarmey de los monitoreos establecidos por las entidades encargadas como la Autoridad Nacional del Agua (ANA) en Perú.

La gestión de la calidad del agua y el tratamiento del agua se desarrolla en base a los resultados establecidos por el índice de calidad de agua. Los resultados que conllevan al tratamiento de agua para las categorías de Bebida de Animales y para Consumo Humano, dependen de la clasificación según el resultado del análisis de los parámetros fisicoquímicos y biológicos encontrados en las cuencas del río de la ciudad de Huarmey. El Fondo de las Naciones Unidas para la Infancia (UNICEF) indicó que "Conocer el índice de calidad del agua empodera a las comunidades para tomar decisiones informadas sobre el consumo y la conservación del agua, promoviendo una mayor conciencia ambiental" (UNICEF, 2021). Los pobladores que utilizan el agua para el consumo humano y la distribución del consumo animal dependen de un correcto tratamiento de agua, pero los resultados por el tiempo de ejecución con las variables matemáticas manejadas por la Autoridad Nacional del Agua (ANA), la demora de los resultados y/o un tratamiento incorrecto de cloración del agua para su uso conlleve al malestar de la población. Además, en

una inspección previa a los servicios de saneamiento a cargo de la JASS Tayca, la Superintendencia Nacional de Servicios de Saneamiento (Sunass) constató que “no hay un sistema de precloración, así como tampoco la información sobre la calidad de las fuentes de agua, limpieza y desinfección del reservorio, análisis físico químico y del nivel de turbiedad del agua, controles de presión y continuidad del servicio” (La república, 2021). Defensoría del Pueblo reporta que "Las ciudades peruanas enfrentan problemas significativos con el tratamiento de aguas residuales, lo que resulta en la liberación de aguas residuales sin tratar en cuerpos de agua locales" (Defensoría del Pueblo, 2023). El gobierno peruano ha implementado varias medidas para abordar la contaminación del agua. Según el Ministerio del Ambiente (MINAM), "Se están llevando a cabo proyectos para mejorar el tratamiento de aguas residuales y fortalecer la regulación en la industria minera para reducir el impacto ambiental" (MINAM, 2024). Debido a que la categorización de estos parámetros físicoquímicos son usados para el manejo de la irrigación de agua en los campos agrícolas de la ciudad y manejo medio ambiental de los centros poblados de la provincia de Huarney, es indispensable tener los resultados correctos con un menor tiempo posible en la clasificación de los datos y el tratamiento respectivo continuo. "Los índices de calidad del agua permiten a los gestores de recursos tomar decisiones informadas sobre el tratamiento y la gestión del agua, ayudando a priorizar intervenciones y recursos de manera efectiva" (Instituto Internacional del Agua de Estocolmo, 2023). Reducir el tiempo en el manejo del cálculo del índice de calidad de agua en Perú es crucial por varias razones estratégicas y prácticas que impactan directamente en la salud pública, la gestión ambiental, y la planificación y respuesta a emergencias.

1.2.2 Problemas específicos

Manejar las fórmulas para calcular el índice de calidad de agua en un macro de Excel y validar manualmente los resultados de cada monitoreo puede presentar varios problemas significativos. Dentro de los problemas específicos para la clasificación de la calidad del agua, se tiene:

- Retrasos en la detección de contaminantes debido al procesamiento manual en

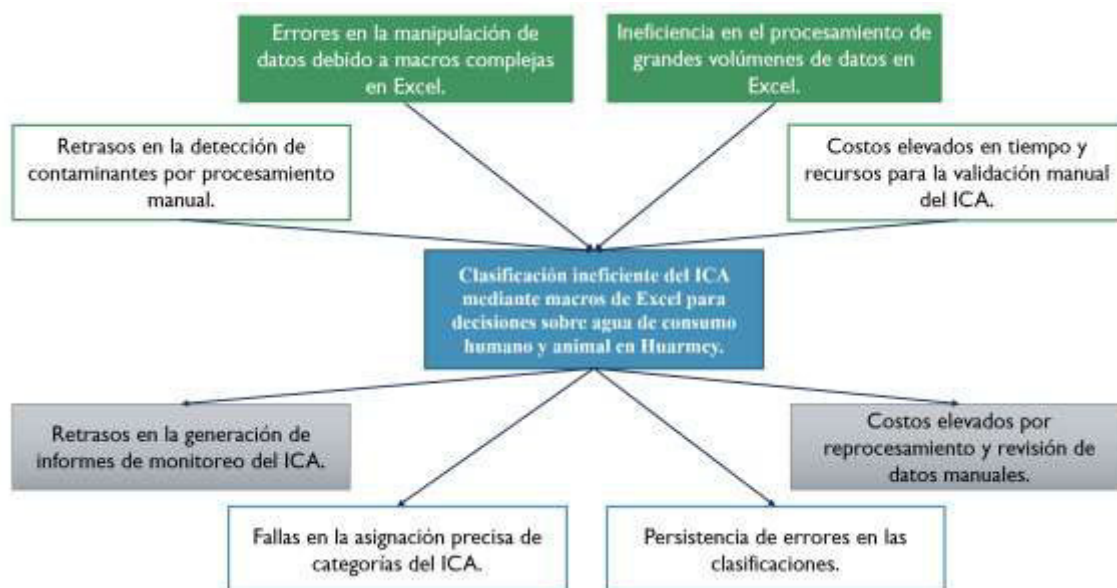
Excel: El uso de macros en Excel para el cálculo del ICA resulta en un proceso lento y poco eficiente, lo que retrasa la detección oportuna de parámetros fisicoquímicos críticos. Estos retrasos pueden permitir que contaminantes nocivos para el consumo humano y animal se distribuyan antes de ser identificados y tratados.

- Errores en la asignación de categorías del ICA por manejo manual de fórmulas: La gestión manual de datos y fórmulas en Excel incrementa significativamente el riesgo de errores, afectando la precisión en la clasificación de los parámetros del ICA. Esto puede llevar a decisiones erróneas sobre el tratamiento del agua y poner en riesgo a la población que depende de este recurso para consumo y bebida animal.
- Ineficiencia en el procesamiento de grandes volúmenes de datos: Los monitoreos del ICA en Huarmey abarcan múltiples puntos de muestreo y parámetros fisicoquímicos. Sin embargo, el procesamiento manual de grandes volúmenes de datos en Excel ralentiza el análisis y dificulta la gestión de alertas tempranas sobre calidad del agua, comprometiendo la capacidad de respuesta ante potenciales contaminaciones.
- Costos elevados en tiempo y recursos para la validación manual del ICA: La validación de los resultados del ICA requiere una revisión exhaustiva de los cálculos realizados en Excel, lo que incrementa el tiempo y los costos asociados a cada monitoreo. Además, la revisión manual prolonga la generación de informes, retrasando la implementación de medidas correctivas en caso de detectar contaminantes críticos.

La Figura 6 muestra un resumen estructural de los problemas específicos identificados, evidenciando cómo el proceso manual basado en macros de Excel genera ineficiencias que impactan negativamente en la gestión oportuna y precisa del ICA.

Figura 6

Árbol del problema en la gestión de la calidad del agua en la ciudad de Huarney.



Nota. Elaboración propia.

1.3 Objetivos

1.3.1 Objetivo principal

Implementar un modelo de K vecinos más cercanos (KNN) para optimizar la clasificación de las categorías de Bebida de Animales (3-D2) y Consumo Humano (1-A2) de acuerdo con el Índice de Calidad del Agua (ICA) de las muestras de las cuencas de Huarney. Utilizar un algoritmo de clasificación para el cálculo del ICA en lugar de fórmulas manuales y macros de Excel no solo mejora la precisión y eficiencia del proceso, sino que también ofrece una plataforma más robusta, escalable y flexible para la gestión de datos de calidad de agua. Esto es esencial para una gestión efectiva y proactiva de los recursos hídricos, especialmente en contextos donde las decisiones rápidas y basadas en datos precisos son críticas.

1.3.2 Objetivos específicos

Utilizar un algoritmo de clasificación para calcular el índice de calidad de agua presenta varias ventajas significativas en comparación con el manejo de fórmulas manuales y el uso de macros en Excel. Dentro de los objetivos específicos para la

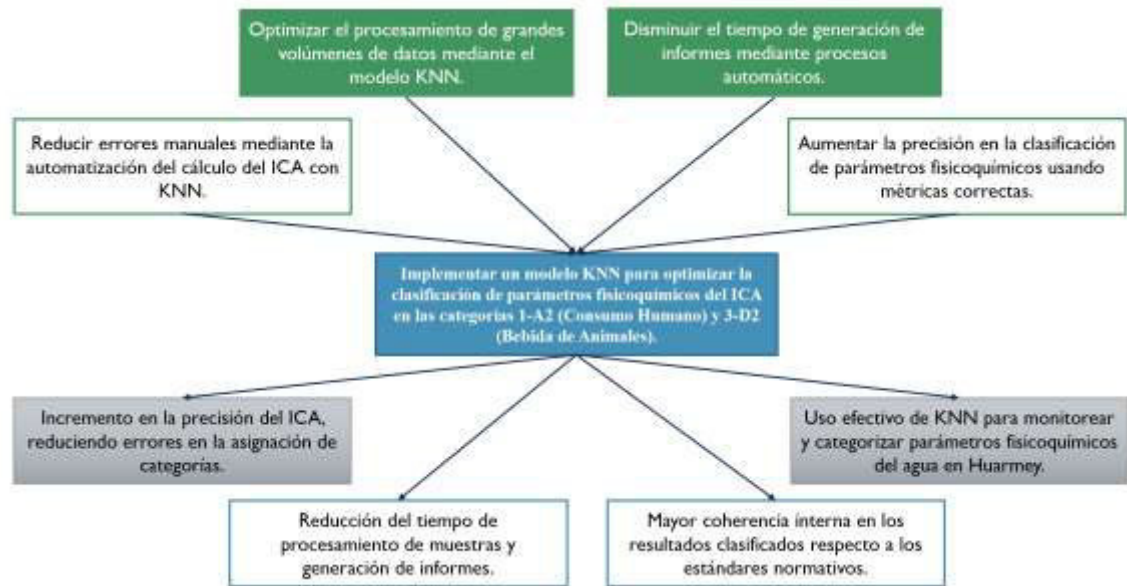
solución de la clasificación de la calidad del agua, se tiene:

- Reducir errores manuales en el cálculo del ICA: Implementar un algoritmo de clasificación que automatice completamente el proceso de cálculo del ICA, reduciendo significativamente los errores humanos asociados con la entrada manual de datos y la manipulación de fórmulas en Excel. Esto asegura que los cálculos sean más precisos y confiables.
- Incrementar la eficiencia en el procesamiento de datos extensos: Desarrollar un sistema que pueda manejar grandes volúmenes de datos de manera más eficiente que Excel. Mientras que Excel puede volverse lento o inestable con grandes conjuntos de datos, un algoritmo bien diseñado puede procesar rápidamente grandes cantidades de información sin comprometer el rendimiento del sistema.
- Disminuir el tiempo de generación de informes de monitoreo: Integrar el algoritmo de clasificación con herramientas modernas de visualización de datos y generación de informes, proporcionando los resultados de forma casi inmediata y reduciendo el tiempo y el coste para la elaboración del informe final del monitoreo de los parámetros analizados.
- Aumentar la precisión en la clasificación de parámetros fisicoquímicos según el ICA: Optimizar la clasificación del cálculo del ICA para las categorías de Consumo Humano y Bebida de Animales mediante el algoritmo de clasificación, ayudando a que sea más eficiente y preciso en sus resultados, así como reducir el tiempo en la elaboración de la entrega de la documentación del monitoreo de agua.

La Figura 7 muestra un resumen sobre los objetivos específicos del uso de un algoritmo de clasificación para calcular el ICA, el cual mejora la precisión, agiliza el procesamiento de datos, reduce costos en tiempo y optimiza la gestión del monitoreo del agua en comparación con los métodos manuales en Excel.

Figura 7

Árbol de objetos para mejorar la clasificación de categorías de parámetros de agua según su ICA en la ciudad de Huarney.



Nota. Elaboración propia.

El algoritmo de clasificación incorpora técnicas avanzadas de análisis y modelado estadístico que no son factibles de implementar en Excel. Por ejemplo, pueden aprender de los datos históricos para mejorar la precisión de las predicciones y adaptarse a nuevas tendencias o patrones en la calidad del agua. Los algoritmos de clasificación son fácilmente actualizables y escalables. Si cambian los estándares del índice de calidad de agua o se introducen nuevos parámetros, el algoritmo puede ser ajustado y extendido sin necesidad de modificar manualmente cada fórmula o macro en Excel, lo cual es un proceso propenso a errores y laborioso.

1.4 Justificación

1.4.1 Justificación práctica

Esta investigación se realiza porque existe la necesidad de mejorar el nivel de categorización, rendimiento y clasificación de los datos de parámetros fisicoquímicos en la calidad de agua en las cuencas de Huarmey, con el uso del algoritmo K vecinos más cercanos (KNN) en los puntos de Huarmey. El algoritmo ofrece una solución eficiente al problema práctico del tiempo y recursos que requieren los cálculos manuales del ICA. Esto es particularmente útil en situaciones donde se necesita monitorear una gran cantidad de datos de

diferentes cuencas hidrográficas. Monitorear cientos de muestras de agua provenientes de diversas cuencas sería extremadamente costoso y lento si se hiciera manualmente. Con el algoritmo, este proceso se realiza automáticamente, generando resultados rápidos y precisos. El modelo es capaz de manejar grandes volúmenes de datos, algo que es vital en un país como Perú, con diversas cuencas y diferentes condiciones ambientales. El uso de machine learning permite clasificar múltiples muestras simultáneamente, algo que sería poco práctico mediante métodos manuales. Al tener un sistema más rápido y preciso, las autoridades pueden tomar decisiones más informadas sobre la gestión del agua, responder rápidamente ante posibles problemas de contaminación, y planificar intervenciones adecuadas.

1.4.2 Justificación teórica

El índice de calidad de agua (ICA) se calcula a partir de una combinación de parámetros fisicoquímicos y biológicos, como el pH, oxígeno disuelto, nitratos, entre otros. Un algoritmo de clasificación es capaz de procesar estos datos multivariados de manera eficiente, identificando patrones en los datos que pueden ser difíciles de detectar con técnicas tradicionales. Esto permite clasificar cuerpos de agua en categorías claras de calidad (excelente, buena, regular, mala), facilitando la toma de decisiones para los gestores del recurso. Los algoritmos como KNN, SVM y Random Forest están bien fundamentados en la teoría del aprendizaje supervisado. Estos algoritmos han sido ampliamente utilizados en problemas de clasificación que involucran datos multivariados, como es el caso de los parámetros de calidad del agua (pH, oxígeno disuelto, turbidez, etc.). Estos modelos no solo clasifican los datos actuales, sino que también son capaces de predecir futuras clasificaciones basadas en datos históricos y patrones identificados. Esto aporta una ventaja teórica significativa sobre métodos más simples o tradicionales. Estos algoritmos aprenden a partir de un conjunto etiquetado de datos de entrenamiento (muestras de agua ya clasificadas) y luego aplican ese aprendizaje para predecir la clasificación de nuevas muestras. Esta aproximación es teóricamente sólida y ha demostrado ser eficaz en diversas

aplicaciones científicas y de ingeniería.

1.4.3 Justificación metodológica

La investigación se llevó a cabo debido a la necesidad de mejorar el índice de calidad de agua en los procesos de gestión del tratamiento de agua en la ciudad de Huarney, utilizando el modelo de K vecinos más cercanos (KNN) en la clasificación de los parámetros de agua de los monitoreos de las cuencas del río Huarney y su posterior tratamiento para consumo humano y bebida animal, y así los pobladores puedan gozar de un buen consumo de agua en los riegos y agricultura y aporten con el avance y prosperidad de la comunidad. La metodología incluye la selección de los parámetros de agua (pH, oxígeno disuelto, turbidez, etc.) basados en la documentación de la metodología del cálculo del Ica implementado por la Autoridad Nacional del Agua (ANA) para la clasificación. Esto no solo simplifica el modelo, sino que también mejora su rendimiento al reducir el ruido en los datos. La selección de características es clave para evitar el sobreajuste y mejorar la capacidad del modelo para generalizar a nuevos conjuntos de datos. Permite que el algoritmo se enfoque en las variables que más impactan la clasificación de la calidad del agua, de acuerdo con el reglamento del ICA.

1.5 Propuesta

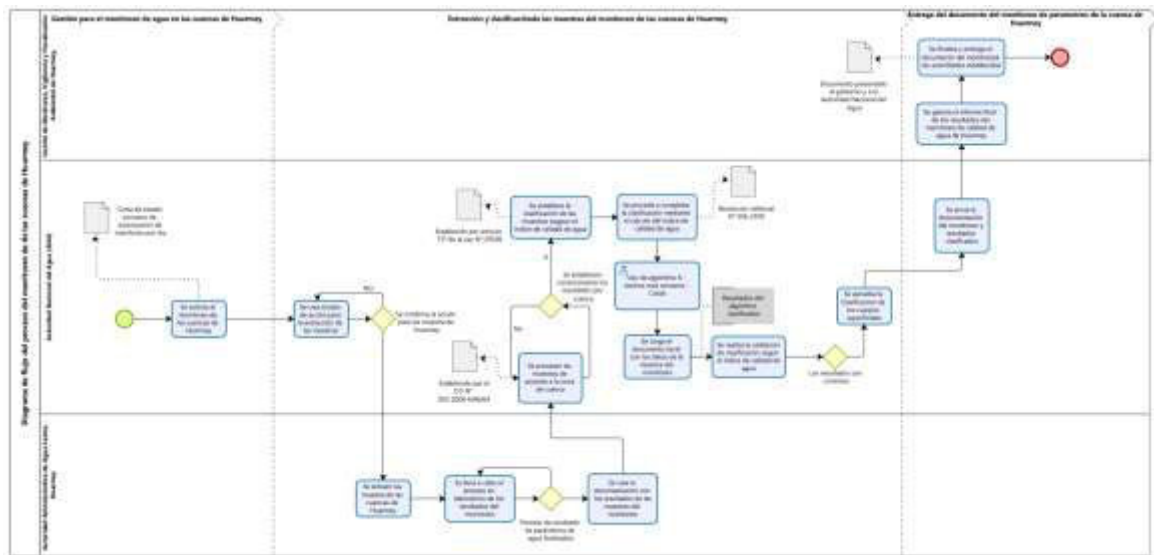
Siendo la calidad de agua el principal motor de la ciudad de Huarney se requiere de una correcta gestión en el tratamiento del agua. Por ello, se emplea los modelos de machine learning, en dicho caso, la técnica de K vecinos más cercanos (KNN), la cual es un modelo de clasificación para los contaminantes por parámetros fisicoquímicos en el agua, con lo cual, puede ayudar a gestionar, predecir y mejorar el tratamiento y la calidad del agua potable suministrada en la ciudad de Huarney. KNN es un algoritmo de clasificación intuitivo y fácil de implementar. No requiere muchos parámetros o ajustes complejos, lo que lo hace accesible para clasificar datos de calidad de agua a partir de múltiples parámetros como el pH, la turbidez y los niveles de oxígeno disuelto. Su enfoque basado en la distancia lo hace ideal para

trabajar con datos multivariados del ICA.

La Figura 8 muestra el diagrama de flujo que detalla el proceso de monitoreo y gestión de la calidad del agua en las cuencas de Huarmey, con la implementación de la propuesta, que incluye el algoritmo principal para el cálculo del índice de calidad de agua (ICA) y la clasificación según su categoría:

Figura 8

Modelo TO-BE propuesto para aumentar la precisión en hallar los contaminantes de agua en el tratamiento de la ciudad de Huarmey.



Nota. Elaboración propia.

Este diagrama es esencial para visualizar el procedimiento completo desde la planificación hasta la ejecución y reporte del monitoreo de la calidad del agua, asegurando que todas las etapas estén bien documentadas y que el proceso sea transparente y conforme a los estándares regulatorios.

1.6 Alcances y limitaciones

1.6.1. Alcances

El alcance principal de la investigación es resolver el problema del tiempo y precisión en la clasificación de categorías de Consumo Humano y Bebida de Animales

según el índice de calidad de agua (ICA) de los parámetros de agua de los monitoreos de la cuenca del río Huarmey mediante la implementación del algoritmo de clasificación K vecinos más cercanos (KNN) y los algoritmos de Support Vector Machine (SVM) y Random Forest de la clasificación de los datos de agua y el cálculo del ICA.

- La propuesta será usada para mejorar la calidad de agua en la ciudad de Huarmey, con la posible implementación en otros sectores de la población provincial.
- El uso de las técnicas de machine learning con el algoritmo de K vecinos más cercanos (KNN) mejorará significativamente el alcance de las tecnologías para operaciones futuras.
- Las medidas y métricas de las variables en el algoritmo de K vecinos más cercanos (KNN) se usarán conforme al reglamento establecido por el Ministerio del agua sobre el cálculo del índice de calidad de agua (ICA).
- El procedimiento y uso de las técnicas del algoritmo como su operacionalizado estarán en manos de los encargados del manejo de la calidad de agua, así como los supervisores del tratamiento de agua y los contaminantes más propensos a causar un problema en el consumo de agua potable.

1.6.2. Limitaciones

Los límites de estos procesos de solución los campos a los que se lleven a cabo dentro de las políticas e implementación general de este método de K vecinos más cercanos (KNN), así mismo también los límites de adoptar este modelo en otros campos.

- Limitación espacial: El algoritmo, las métricas y los procedimientos solo cubren los datos dados por la municipalidad de Huarmey sobre los contaminantes y parámetros fisicoquímicos del año 2020, 2021 y 2023.
- Limitación Social: La aplicación del algoritmo y los métodos de esta tesis son en beneficio de la ciudad de Huarmey, en cuyo caso se aplique en otros sectores del departamento o país, los resultados diferirán de la cantidad y variables de los datos en la calidad de agua de su sector correspondiente.

- Limitación técnica: Debido al conocimiento actual, el algoritmo utilizado para la clasificación y predicción de datos será por medio de un software de código abierto, por lo que, para una mayor precisión en otros campos, se deberá extender y agregar otros procedimientos en dicho algoritmo.

1.7 Organización de la tesis

La organización en base a los conceptos se resume de la siguiente manera:

- En el capítulo uno se presentar el planteamiento del problema, la técnica del problema, el problema principal, la justificación para resolver el problema. los objetivos principales y específicos y los alcances y limitaciones de este proceso
- En el capítulo dos, se presentan los conceptos fundamentales que se llevarán a cabo a través de la tesis para entender mejor los conceptos y los temas que se tomarán en cuenta en todo el desarrollo de la solución del problema
- En el capítulo tres, se establece la taxonomía de la tesis, las referencias conceptuales, los criterios y la literatura utilizados en el proceso de los artículos.
- En el capítulo cuatro, se establecen los puntos sobre la resolución del problema, estableciendo las bases de desarrollo y resolución del tema establecido.
- En el capítulo cinco, se toman los datos de las muestras de las cuencas de Huarmey para analizar, proceder, y mostrar los resultados del algoritmo diseñado para resolver el problema principal y consecuentes.
- En el capítulo seis se toma prioridad a las discusiones acerca del alcance de los resultados sobre las muestras de parámetros de agua evaluados en esta investigación, así como también mostrar una mejora en el cálculo del índice de calidad de agua frente a otras aplicaciones actuales que utilizan métricas para el cálculo del ICA, como documentos de autores que reflejan la utilidad de algoritmos de clasificación implementados en esta tesis.
- En el capítulo siete, se establecen las conclusiones en base a los resultados, la resolución del problema y se toman los puntos iniciales del objetivo principal y consecuentes para finalizar con el resultado optimo obtenido para futuras aplicaciones en el campo de agua establecidos en este documento.

CAPÍTULO II: MARCO TEÓRICO

En este capítulo se darán los conceptos básicos de los términos principales usados en esta tesis. Para esta tesis centrada en la clasificación de datos de agua, proporciona el sustento necesario para el diseño del algoritmo, asegurando que el estudio esté bien fundamentado y alineado con las necesidades de la investigación científica.

2.1 Conceptos principales

2.1.1 K vecinos más cercanos (KNN)

La implementación de metodologías de aprendizaje por conjuntos en diversas áreas de la ingeniería, como la hidrología, con fines de simulación y predicción mejora la diversidad de técnicas de conjuntos disponibles para su implementación en ciencias hidrológicas ha llevado al desarrollo y utilización de diferentes estrategias en la implementación (Chen et al., 2021). Los clasificadores de vecinos más cercanos se utilizan como clasificadores base. Este clasificador tiene una baja complejidad computacional, por lo tanto, es apropiado para grandes datos dispersos, pero no se construye un modelo general. Se genera un vector sobre clases de decisión basado en cada tabla local para el objeto de prueba actualmente considerado (Przybyła-Kasperek, 2022).

2.1.2 Machine learning (ML)

El aprendizaje automático (ML) se refiere a una parte de la inteligencia artificial que pertenece a un conjunto de técnicas que hacen posible el manejo de datos través del entrenamiento con grandes volúmenes de datos. Una predicción precisa de los parámetros relacionados con la calidad del agua (WQ) se considera fundamental herramienta decisiva en la gestión sostenible de los recursos hídricos (Al-Sulttani et al., 2021). El monitoreo de la calidad del agua es esencial para proteger la salud humana y el medio ambiente y controlar la calidad del agua (Nasir et al., 2022).

2.1.3 Calidad de agua (CA)

La calidad del agua hace referencia a sus características químicas, físicas, biológicas

y radiológicas. Es una forma de evaluar el estado del agua según los requisitos de una o más especies vivas o con base en un propósito u objetivo humano. Normalmente, se utiliza para describir un conjunto de estándares que permiten medir su condición. Los estándares más comunes para evaluar la calidad del agua son la salud del ecosistema, la seguridad contra el contacto humano y el agua potable (Uddin et al., 2023a).

2.1.4 Cuenca hidrográfica (CH)

Una cuenca hidrográfica es un área de la superficie terrestre sobre la cual (cuando es anormal) cae agua de lluvia, esta es desviada por un sistema de drenaje a una salida similar. La zona tiene una especie de "embudo" por el que fluye el agua desde la parte superior hasta un punto común desde el que fluye toda el agua hacia el otro lado (Omer and Omer, 2019).

2.1.5 Parámetros de calidad del agua

Los parámetros son propiedades tanto físicas como químicas utilizadas para evaluar la calidad de agua u otras sustancias (Autoridad Nacional del Agua, 2018). Los parámetros utilizados en esta investigación incluyen:

- **Aluminio (Al):** Un elemento con número atómico 13, que indica que cada átomo de aluminio tiene 13 protones en su núcleo (Carbotecnia, 2023).
- **Cromo (Cr):** Metal blanco plateado, brillante, duro y frágil, resistente a la corrosión, con número atómico 24 (National Institutes of Health (NIH), 2024).
- **Arsénico (As):** Material sólido de color gris acero en su forma elemental, pero en el ambiente se encuentra combinado con otros elementos como oxígeno, cloro y azufre (Carbotecnia, 2023).
- **Boro (B):** Elemento químico no metálico con número atómico 5 (National Institutes of Health (NIH), 2024).
- **Cadmio (Cd):** Es un elemento con número atómico 48, ubicado en el grupo 12 de la tabla periódica (Instituto Nacional del Cancer, 2015).
- **Cobre (Cu):** Es un elemento con número atómico 29, cuyo símbolo es Cu,

derivado del latín "cuprum" (Carbotecnia, 2023).

- **Hierro (Fe):** Es un metal duro y maleable de color gris, con características magnéticas (Carbotecnia, 2023).
- **Mercurio (Hg):** Elemento químico natural conocido como mercurio metálico o elemental, con símbolo Hg (Carbotecnia, 2023).
- **Manganeso (Mn):** Metal presente naturalmente en rocas, aunque en su forma pura es de color plateado (López Lozano et al., 2015).
- **Oxígeno disuelto (O):** Representa la cantidad de oxígeno gaseoso (O_2) presente en el agua, expresado en miligramos por litro o como porcentaje de saturación (Omer and Omer, 2019).
- **Plomo (Pb):** Metal pesado de color azulado que adquiere un tono gris mate al empañarse (Carbotecnia, 2023).
- **Zinc (Zn):** Metal similar al magnesio y al berilio, clasificado como metal de transición (López Lozano et al., 2015).
- **Demanda bioquímica de oxígeno (DBO):** Mide el oxígeno que los microorganismos consumen durante la descomposición de sustancias orgánicas en el agua (Omer and Omer, 2019).
- **Potencial de hidrógeno (pH):** Mide la concentración de iones de hidrógeno para determinar si una sustancia es ácida, alcalina o neutra (Omer and Omer, 2019).
- **Conductividad eléctrica:** Mide la capacidad de un material para conducir electricidad, siendo los mejores conductores la plata, cobre, oro, aluminio, entre otros (Omer and Omer, 2019).
- **Coliformes termotolerantes:** Subgrupo de coliformes fecales que fermentan lactosa a altas temperaturas, conocidos también como coliformes termotolerantes (Carbotecnia, 2023).

2.1.6 Índice de calidad de agua (ICA)

El Índice de Calidad del Agua (ICA) es un método que integra diferentes parámetros fisicoquímicos, biológicos y microbiológicos para evaluar y medir la calidad del agua de manera simplificada. "El Índice de Calidad del Agua (ICA) indica el grado de contaminación del agua a la fecha del muestreo y está expresado

como porcentaje del agua pura; así, agua altamente contaminada tendrá un ICA cercano o igual a cero por ciento, en tanto que en el agua en excelentes condiciones el valor del índice será cercano a 100% (Autoridad Nacional del Agua, 2018)”.

2.1.7 Autoridad Nacional de agua (ANA)

La Autoridad Nacional del Agua (ANA) en Perú es la entidad que gestiona y regula los recursos hídricos del país. Su función principal es priorizar la sostenibilidad del agua, supervisando su distribución y calidad, además de coordinar el uso de las fuentes hídricas entre diversos sectores como agricultura, industria y consumo humano. “La Autoridad Nacional del Agua (ANA) es el ente rector y la máxima autoridad técnico-normativa del Sistema Nacional de Gestión de los Recursos Hídricos (Autoridad Nacional del Agua, 2018)”.

2.1.8 Random Forest (RF)

El algoritmo Random Forest es un método de aprendizaje automático que se utiliza ampliamente para tareas de clasificación y regresión. “Una de las principales características de Random Forest es su capacidad para manejar grandes conjuntos de datos y una gran cantidad de variables, lo que lo hace ideal para aplicaciones en campos como la predicción financiera, la clasificación de riesgos de crédito, y la predicción de la calidad del agua. Además, es particularmente útil para seleccionar características importantes, ya que puede medir la importancia de cada variable en el proceso de clasificación (Naghbi et al., 2020)”.

2.1.9 Support vector machine (SVM)

El Support Vector Machine (SVM) es un algoritmo de aprendizaje automático utilizado para la clasificación y regresión de datos. Support vector machine (SVM) es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz (Sayedmohammadi et al., 2023).

CAPÍTULO III: ESTADO DE ARTE

El estado del arte permite situar el trabajo dentro del panorama académico actual. Es un puente entre el conocimiento existente y la investigación que se va a desarrollar. Para ello se tienen los puntos de la taxonomía del tema de investigación, la metodología de la revisión literaria y la revisión de los artículos científicos para sustentar el trabajo de investigación en base a evidencias científicas de otros autores relacionados con el tema principal.

3.1 Taxonomía

- ACM:

Computing methodologies

Machine learning

Machine learning algorithms

- IEEE:

Computational and artificial Intelligence

Learning systems

- RR línea de investigación UNMSM:

C.0.5 Diseño y aplicación de nuevas tecnologías

C.0.5.8 Inteligencia artificial

3.2 Metodología de la revisión de la literatura

En el presente trabajo de investigación utilizó los siguientes motores de búsqueda:

Tesis:

- Google académico

Papers, artículos, libros y revistas indexadas:

- ElSevier
- Google académico
- IEEE
- Scopus

- ScienceDirect

3.2.1 Criterios de inclusión

- Revistas indexadas.
- Artículos de los años 2020-2025.
- Palabras clave:
 - Machine learning
 - Contaminación de agua
 - Agua potable
 - Calidad de agua
 - Inteligencia artificial
 - Cuenca de agua

3.2.2 Criterios de exclusión

- Revistas no indexadas.
- Artículos publicados antes del año 2020.

3.3 Revisión de la literatura

3.3.1 Revisión de artículos

A continuación, se presenta un breve análisis de los artículos revisados tomando el resumen correspondiente y el aporte que brindan a esta tesis.

3.3.1.1 Artículo 1: Prediction of groundwater quality using efficient machine learning technique (Singha et al., 2021).

Resumen: El artículo menciona que, para garantizar agua potable segura, se requiere comprender la calidad y contaminación de las aguas subterráneas. Este estudio propone un modelo basado en aprendizaje profundo (DL) para predecir la calidad del agua subterránea y lo compara con tres modelos de aprendizaje automático: Bosque Aleatorio (RF), XGBoost y Redes Neuronales Artificiales (ANN). Se recolectaron 226 muestras de agua subterránea en un área agrícola de India, evaluando parámetros fisicoquímicos para calcular el índice de calidad del

agua subterránea (EWQI).

Los resultados muestran que el modelo DL es el más preciso, con un coeficiente de determinación con la mayor precisión en términos de R^2 , es decir $R^2 = 0.996$ contra RF ($R^2 = 0.886$), XGBoost ($R^2 = 0.927$) y ANN ($R^2 = 0.917$). La incertidumbre del modelo DL se verificó mediante validación cruzada con datos aleatorizados, mostrando desviaciones menores. Además, el análisis de importancia de variables confirma que el modelo DL es el enfoque más realista y preciso para predecir la calidad del agua subterránea.

Para garantizar fuentes seguras de agua potable en el futuro, es imperativo comprender la calidad y el nivel de contaminación de las aguas subterráneas existentes. La predicción de la calidad del agua con alta precisión es la clave para controlar la contaminación del agua y mejorar la gestión del agua. En este estudio, se propone un modelo basado en aprendizaje profundo (DL) para predecir la calidad del agua subterránea y se compara con otros tres modelos de aprendizaje automático (ML), a saber, bosque aleatorio (RF), aumento de gradiente extremo (XGBoost) y red neuronal artificial (ANA). Se recolecta un total de 226 muestras de agua subterránea de un área de agricultura intensiva Arang del distrito de Raipur, Chhattisgarh, India, y se miden varios parámetros físicoquímicos para calcular el índice de calidad del agua subterránea (EWQI) basado en el peso de la entropía. Los rendimientos de predicción de los modelos se determinan mediante la introducción de cinco métricas de error. Los resultados mostraron que el modelo DL es el mejor modelo de predicción. La incertidumbre de la salida del modelo DL se verifica de forma cruzada ejecutando el algoritmo propuesto con un nuevo conjunto de datos aleatorios diez veces, donde se observan desviaciones menores en el valor medio de las métricas de rendimiento. Además, la importancia de la variable de entrada calculada por los modelos de predicción destaca que el modelo DL es el enfoque más realista y preciso en la predicción de la calidad del agua subterránea.

Aporte: El aporte principal de este artículo es el manejo de la calidad de las

aguas subterráneas con los algoritmos de aprendizaje automático (ML) los cuales se utilizan para predecir la calidad de las aguas subterráneas. En este artículo se pretende demostrar que la evaluación mediante el método EWQI no es tan preciso y por ello se compara el rendimiento de predicción de los modelos RF, XGBoost, ANN y DL. Estos algoritmos sirven como base de esta tesis para demostrar que el modelo de predicción de la calidad basado en Deep learning funciona mucho mejor que otros modelos ML.

3.3.1.2 Artículo 2: Advancing Water Quality Research: K-Nearest Neighbor Coupled with the Improved Grey Wolf Optimizer Algorithm Model Unveils New Possibilities for Dry Residue Prediction (Tahraoui et al., 2023).

Resumen: Se han establecido estaciones de monitoreo para combatir la contaminación del agua, mejorar el ecosistema, promover la salud humana y facilitar la producción de agua potable. Sin embargo, el monitoreo continuo y extenso del agua es costoso y requiere mucho tiempo, lo que genera conjuntos de datos limitados y obstaculiza la investigación sobre la gestión del agua. Este estudio se centra en el desarrollo de un modelo de vecino más cercano K (KNN) optimizado utilizando el algoritmo de optimización de Grey Wolf mejorado (I-GWO) para predecir las cantidades de residuos secos. El modelo incorpora 20 parámetros físicos y químicos derivados de un conjunto de datos de 400 muestras. Se emplea una validación cruzada para evaluar el rendimiento del modelo, optimizar los parámetros y mitigar el riesgo de sobreajuste. Se crean cuatro pliegues y cada pliegue se optimiza utilizando 11 métricas de distancia y sus funciones de ponderación correspondientes para determinar la mejor configuración del modelo. Entre los modelos evaluados, la métrica de distancia de Jaccard con función de ponderación del cuadrado inverso demuestra consistentemente el mejor desempeño en términos de errores estadísticos y coeficientes para cada pliegue. Al promediar las predicciones de los modelos en los cuatro pliegues, se obtiene una estimación del desempeño general del modelo. El modelo resultante exhibe alta eficiencia, con errores notablemente

bajos reflejados en los valores de R, R², R²ADJ, RMSE y EPM, que se informan como 0.9979, 0.9958, 0.9956, 41.2639 y 3.1061, respectivamente. Este estudio revela una correlación no lineal convincente entre los atributos fisicoquímicos del agua y el contenido de relaves secos, lo que indica la capacidad de predecir con precisión las cantidades de relaves secos. Al emplear la metodología propuesta para mejorar los modelos de calidad del agua, se hace posible superar las limitaciones en la gestión de la calidad del agua y mejorar significativamente la precisión de las predicciones con respecto a los parámetros críticos del agua.

Aporte: El aporte principal de este artículo es el desarrollo de un modelo de vecino más cercano K (KNN) optimizado para predecir las cantidades de residuos secos, así como también buscar evaluar el rendimiento del modelo, optimizar los parámetros y mitigar el riesgo de sobreajuste dando como resultado la alta eficiencia, con errores notablemente bajos reflejados en los valores y hace posible superar las limitaciones en la gestión de la calidad del agua.

3.3.1.3 Artículo 3: Prediction of hydrological and water quality data based on granular-ball rough set and k-nearest neighbor analysis (Dong et al., 2024a).

Resumen: Los conjuntos de datos hidrológicos y de calidad del agua suelen abarcar una gran cantidad de variables características, pero no todas ellas influyen significativamente en los resultados analíticos. Por lo tanto, al seleccionar sabiamente variables características con un rico contenido de información y eliminar características redundantes, no solo se puede mejorar la eficiencia del análisis, sino que también se puede simplificar la complejidad del modelo. Este documento considera la introducción del algoritmo de conjuntos aproximados de bolas granulares para la selección de variables características y su combinación con el método de k vecinos más cercanos y la red de retropropagación para analizar datos hidrológicos y de calidad del agua, promoviendo así la inspección general y fusionada. Los resultados del análisis de datos hidrológicos de calidad del agua muestran que el método propuesto

produce mejores resultados en comparación con el uso de un regresor de k vecinos más cercanos independiente.

Aporte: El artículo muestra el uso del método de k vecinos más cercanos y la red de retro propagación. Los resultados del análisis de datos hidrológicos de calidad del agua muestran que el método propuesto produce mejores resultados en comparación con el uso de un regresor de k vecinos más cercanos independiente. Esto es un aporte importante ya que se enfoca en una combinación de ambos métodos el cual puede implementarse dentro de futuras investigaciones y puede aportar a esta tesis el uso preciso del método independiente para la clasificación de parámetros de agua.

3.3.1.4 Artículo 4: Machine learning applications for anomaly detection in Smart Water Metering Networks: A systematic review (Kanyama et al., 2024).

Resumen: La digitalización en el sector del agua ha impulsado el desarrollo de redes inteligentes de medición de agua (SWMN, por sus siglas en inglés), que permiten el monitoreo automatizado y continuo del consumo de agua. No obstante, persisten desafíos en la gestión y transmisión eficiente de la gran cantidad de datos que generan estas redes. Para enfrentar este problema, los investigadores han propuesto técnicas de detección de anomalías para identificar irregularidades en las SWMN. A pesar de los avances en este campo, existe una notable falta de estudios que integren de manera exhaustiva las aplicaciones del aprendizaje automático (ML) en la detección de anomalías en SWMN. Para llenar este vacío, se evaluaron 32 artículos de investigación publicados entre 2016 y 2023, centrados en el uso de ML para detectar anomalías en redes de agua inteligentes, sistemas de distribución de agua y otras redes relacionadas. De una selección inicial de 725 artículos, se escogieron aquellos más relevantes sobre técnicas de detección de anomalías basadas en ML mediante un análisis cuantitativo y cualitativo.

El estudio encontró que técnicas de ML como k-Nearest Neighbors (kNN),

autocodificadores, el método de agrupamiento espacial DBSCAN, bosques aleatorios, árboles de decisión y máquinas de vectores de soporte (SVM) han contribuido significativamente a la detección de anomalías en estas redes. Con frecuencia, los investigadores utilizaron varias técnicas de ML de manera combinada para mejorar la precisión. El análisis de métricas de rendimiento mostró que la puntuación F1, la precisión, la exactitud y la recuperación eran las más utilizadas para evaluar la efectividad de estas técnicas de detección de anomalías. Esta revisión tiene como objetivo ofrecer recomendaciones a los investigadores sobre la selección adecuada de técnicas de ML para la detección de anomalías en SWMN. Los resultados destacan el potencial prometedor del uso de ML en este campo y subrayan la necesidad de continuar investigando.

Aporte: El principal aporte de este artículo es la comparación y análisis de aplicaciones de aprendizaje automático (ML) para la detección de anomalías en SWMN. El estudio evaluó treinta y dos artículos de investigación escritos entre 2016 y 2023, centrándose en aplicaciones de ML para la detección de anomalías en sistemas de agua inteligentes, redes de agua inteligentes, redes de distribución de agua y redes de agua.

3.3.1.5 Artículo 5: Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms (Hussein et al., 2024).

Resumen: La evaluación de la calidad de las aguas subterráneas es esencial para su uso en riego; sin embargo, en los países en desarrollo, las limitaciones financieras hacen que estas evaluaciones se realicen con poca frecuencia, lo que dificulta obtener una visión completa. Por ello, este estudio combina enfoques de aprendizaje automático con el índice de calidad del agua de riego (IWQI) para evaluar la calidad de las aguas subterráneas en Naama, una región del suroeste de Argelia. Se utilizaron parámetros hidroquímicos (cationes, aniones, pH y CE), índices cualitativos y representaciones geoespaciales para determinar la adecuación del agua subterránea para riego en el área. Además, se aplicaron

modelos de aprendizaje automático como XGBoost, Regresión de Vectores de Soporte (SVR) y K vecinos más cercanos (KNN) para predecir el IWQI.

En este estudio, se analizaron 166 muestras de agua subterránea para calcular el índice de riego. Los resultados indicaron que el 42,18% de las muestras eran de excelente calidad, el 34,34% de muy buena calidad, el 6,63% de buena calidad, el 9,64% satisfactorias, y el 4,21% no aptas para riego. Asimismo, los modelos de aprendizaje automático mostraron un desempeño destacado, con XGBoost sobresaliendo en precisión y estabilidad, obteniendo un RMSE de 2,8272 y un R elevado de 0,9834. Por su parte, SVR, utilizando solo cuatro entradas (Ca^{2+} , Mg^{2+} , Na^{+} y K), también mostró una alta capacidad predictiva, con un RMSE de 2,6925 y un R de 0,98738, mientras que KNN demostró un rendimiento sólido. Las diferencias entre estos modelos tienen implicaciones importantes para la toma de decisiones informadas en la gestión del agua agrícola y la distribución de recursos en la región.

Aporte: El aporte de este artículo es la implementación de varios modelos de machine learning con el objetivo de evaluar la calidad de las aguas subterráneas en Naama, implementando enfoques eficientes de aprendizaje automático para pronosticar el IWQI utilizando modelos de Impulso de Gradiente Extremo (XGBoost), Regresión de Vectores de Soporte (SVR) y K vecinos más cercanos (KNN).

3.3.1.6 Artículo 6: Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data (Umar & Gray, 2023a).

Resumen: Los valores faltantes en los datos de nivel de agua son un problema persistente en el modelado de datos y especialmente común en los países en desarrollo. La imputación de datos ha recibido considerable atención en la investigación, con el fin de aumentar la calidad de los datos en el estudio de eventos extremos como inundaciones y sequías. Este artículo evalúa los métodos

de imputación simple y múltiple utilizados en datos mensuales univariados y multivariados del nivel de agua de cuatro estaciones de agua en los ríos Benue y Níger en Nigeria. Se consideraron los mecanismos de datos faltantes completamente aleatorios, faltantes aleatorios y faltantes no aleatorios. El mejor método de imputación se identifica utilizando dos métricas de error: error cuadrático medio y error porcentual absoluto medio. Para el caso univariado, el método de descomposición estacional es mejor para imputar valores faltantes en varios niveles de faltantes para los tres mecanismos faltantes, seguido del suavizado de Kalman, mientras que la imputación aleatoria es mucho peor. Por ejemplo, para el 5% de datos faltantes para la estación de agua Kainji, faltantes completamente aleatorios, los métodos de suavizado de Kalman, descomposición aleatoria y estacional tuvieron errores cuadráticos medios promedio de 13,61, 102,60 y 10,46, respectivamente. Para el caso multivariado, missForest es el mejor, seguido de cerca por k vecino más cercano para los mecanismos de falta completamente al azar y falta al azar, y k vecino más cercano es el mejor, seguido por missForest, para el mecanismo de falta no al azar. Los métodos de bosque aleatorio y de coincidencia de media predictiva tienen un rendimiento deficiente en términos de las dos métricas consideradas. Por ejemplo, para un 10% de datos faltantes que faltan completamente al azar para la estación de agua de Ibi, los errores cuadráticos medios promedio para bosque aleatorio, k vecino más cercano, missForest y coincidencia de media predictiva fueron 22,51, 17,17, 14,60 y 25,98, respectivamente. Los resultados indican que el método de descomposición estacional y los métodos missForest o k vecino más cercano pueden imputar datos faltantes de nivel de agua univariados y multivariados, respectivamente, con mayor precisión que los otros métodos considerados.

Aporte: El artículo evalúa los métodos de imputación simple y múltiple utilizados en datos mensuales univariados y multivariados del nivel de agua de cuatro estaciones de agua en los ríos Benue y Níger en Nigeria. Si bien missForest es el mejor, seguido de cerca por k vecino más cercano, demuestra también el uso de métodos de bosque aleatorio y de coincidencia de media

predictiva brindando un mayor enfoque en la evaluación de niveles de agua como sus calcificaciones.

3.3.1.7 Artículo 7: Accuracy Assessment of RFerns, NB, SVM, and kNN Machine Learning Classifiers in Aquaculture (Çakir et al., 2023).

Resumen: El artículo aborda el problema de las enfermedades bacterianas en piscifactorías, que representan una causa significativa de pérdidas económicas. Tradicionalmente, la detección de estas enfermedades es un proceso manual lento, pero el avance tecnológico permite recopilar datos a velocidades y niveles de detalle sin precedentes. Se emplearon cuatro algoritmos de aprendizaje automático (SVM, Naïve Bayes, Random Ferns y k-NN) para identificar parámetros fisicoquímicos del agua asociados con brotes de enfermedades. Los datos se recopilaron durante un año en tres piscifactorías, aplicando validación cruzada de 10 pliegues. Los algoritmos SVM y Random Ferns se destacan por su capacidad para identificar de manera precisa condiciones de agua que pueden propiciar enfermedades, ayudando a mejorar la gestión ambiental en piscifactorías. Esto resalta el potencial de los modelos de aprendizaje automático en aplicaciones relacionadas con la calidad del agua.

Aporte: Este artículo aporta estrategias y enfoques relevantes para desarrollar algoritmos que clasifiquen datos de calidad de agua, mejorando la precisión y eficiencia en la evaluación ambiental. Las técnicas de validación cruzada y las métricas utilizadas en el estudio pueden implementarse para mejorar el entrenamiento y evaluación de un algoritmo del ICA. Los resultados pueden integrarse en sistemas de monitoreo automatizados, optimizando la clasificación de la calidad del agua y la toma de decisiones.

3.3.1.8 Artículo 8: Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques (Aish et al., 2023a).

Resumen: Este artículo examina el desempeño de cinco algoritmos de

aprendizaje automático supervisado, que incluyen máquinas de vectores de soporte (SVM), regresión logística (LogR), árboles de decisión (DT), redes neuronales de perceptrón multicapa (MLP-NN) y K vecinos más cercanos (KNN) para predecir el índice de calidad del agua (WQI) y la clasificación de la calidad del agua (WQC) en el acuífero costero de la Franja de Gaza. Se recolectaron 2448 muestras de agua subterránea de dicho acuífero, midiendo diversos parámetros físicos y químicos para calcular el WQI basado en su ponderación. La precisión de las predicciones se evaluó utilizando cinco métricas de error. Los resultados indicaron que MLP-NN fue el modelo más preciso, con un valor R entre 0,9945 y 0,9948, superando a SVM (0,9897-0,9880), LogR (0,9784-0,9800), KNN (0,9464-0,9247) y DT (0,9301-0,9064).

La clasificación con SVM reveló que el 78,32% del área de estudio presentaba agua de calidad deficiente a inadecuada, mientras que en la parte norte la calidad del agua variaba de buena a excelente. Los sólidos disueltos totales (TDS) resultaron ser el parámetro más relevante en las predicciones del WQI, mientras que HCO_3^- y SO_4^{2-} fueron los menos significativos. En conclusión, MLP-NN y SVM destacaron como los modelos más precisos para predecir y clasificar el WQI en el acuífero costero de Gaza.

Aporte: El aporte principal de este artículo es usar la metodología y la estructura que se utilizó, ya que muestra la comparación del rendimiento de cinco algoritmos de aprendizaje automático supervisado, que incluyen máquinas de vectores de soporte (SVM), regresión logística (LogR), árboles de decisión (DT), redes neuronales de perceptrones múltiples (MLP-NN) y K vecinos más cercanos (KNN) para predecir el índice de calidad del agua (WQI) y la clase de calidad del agua (WQC) en el acuífero costero de la Franja de Gaza.

3.3.1.9 Artículo 9: Drinking water potability prediction using machine learning approaches: a case study of Indian rivers (Ainapure et al., 2023).

Resumen: El agua potable es el recurso más valioso en el planeta. En las

últimas décadas, su calidad ha disminuido considerablemente debido a la contaminación. Evaluar la calidad del agua es crucial para la salud de las personas, ya que la presencia de contaminantes puede generar serios problemas de salud. En países en desarrollo como la India, la calidad del agua no se evalúa de manera adecuada. Este estudio utiliza técnicas de aprendizaje automático para predecir la calidad del agua de los ríos en India, con el objetivo de determinar su potabilidad al analizar los factores clave empleados en el cálculo del índice de calidad del agua para cada muestra.

Entre los parámetros importantes que se utilizan para calcular el índice de calidad del agua están la temperatura del agua, el pH, la conductividad eléctrica, el oxígeno disuelto, los coliformes fecales, el recuento total de coliformes y la demanda bioquímica de oxígeno. El estudio exploró el uso de K-nearest neighbor (KNN), Random Forest y XGBoost, con y sin ajuste de hiperparámetros, así como una red neuronal artificial secuencial, para identificar cuál de los modelos ofrece las predicciones más precisas sobre la potabilidad del agua. XGBoost resultó ser el modelo más eficiente, con una precisión del 98,93%.

Aporte: Este artículo aportó la utilización de técnicas de aprendizaje automático para predecir la calidad del agua de los ríos de la India. Los enfoques que se exploran incluyen el uso de K-nearest neighbor (KNN), Random Forest y XGBoost, con y sin ajuste de hiperparámetros, y el uso de una red neuronal artificial secuencial aportan en el proceso de la predicción de potabilidad de agua.

3.3.1.10 Artículo 10: Ensemble machine learning paradigms in hydrology: A review (Zounemat-Kermani et al. 2021b).

Resumen: Recientemente, ha habido una tendencia notable hacia el empleo de metodologías de aprendizaje por conjuntos en diversas áreas de la ingeniería, como la hidrología, con fines de simulación y predicción. La diversidad de

técnicas de conjuntos disponibles para su implementación en ciencias hidrológicas ha llevado al desarrollo y utilización de diferentes estrategias en la implementación. Este artículo de revisión explora y se refiere al avance de los métodos de conjunto, incluidos los métodos de conjunto de remuestreo (por ejemplo, ensacado, refuerzo y dagging), promediado de modelos y apilamiento a saber. apilado generalizado, en diferentes campos de aplicación de la hidrología. Los principales temas hidrológicos de este estudio de revisión cubren temas como hidrología superficial, calidad del agua de los ríos, escorrentía de lluvia, flujo de escombros, engelamiento de ríos, transporte de sedimentos, aguas subterráneas, inundaciones y modelado y pronóstico de sequías. Los hallazgos generales de esta encuesta demuestran la superioridad absoluta del uso de estrategias de conjunto sobre el aprendizaje del modelo regular (individual) en hidrología. Además, las técnicas de refuerzo (por ejemplo, refuerzo, AdaBoost y refuerzo de gradiente extremo) se han implementado con más frecuencia y con éxito en problemas hidrológicos que los enfoques de ensacado, apilamiento y dagging.

Aporte: El gran aporte de este artículo a esta tesis, es poder tener en cuenta los diferentes algoritmos usados en el campo de la hidrología en la última década, así también para tomar en cuenta los modelos más usados y poder compararlos con el algoritmo principal de esta tesis, el cual es el de Deep learning por ser el más eficaz en cuanto a algoritmo predictivo de contaminantes de agua.

3.3.1.11 Artículo 11: Water quality classification using machine learning algorithms (Nasir et al., 2022b).

Resumen: El control de la calidad del agua es crucial para proteger la salud humana y el medio ambiente. La inteligencia artificial (IA) ofrece grandes oportunidades para mejorar la clasificación y predicción de la calidad del agua (CA). En este estudio, se evaluaron varios algoritmos de IA para gestionar datos de CA recopilados durante un largo período, con el fin de desarrollar un enfoque

confiable que permita predecir la calidad del agua con la mayor precisión posible. Se utilizaron diversos clasificadores de aprendizaje automático y modelos de conjunto apilados para clasificar los datos de CA utilizando el índice de calidad del agua (ICA). Los clasificadores analizados incluyeron máquinas de vectores de soporte (SVM), bosques aleatorios (RF), regresión logística (LR), árboles de decisión (DT), CATBoost, XGBoost y perceptrones multicapa (MLP).

El conjunto de datos empleado en el estudio constaba de 1679 muestras y metadatos recopilados a lo largo de nueve años. Para evaluar el rendimiento de los clasificadores, se utilizaron curvas de precisión-recuperación y curvas ROC. Los resultados mostraron que CATBoost fue el clasificador más preciso, alcanzando una precisión del 94,51 %. Además, al aplicar modelos de conjunto apilados con todos los clasificadores, la precisión llegó al 100 % en varios metaclasificadores. CATBoost destacó tanto como algoritmo principal de refuerzo de gradiente como metaclasificador. Por lo tanto, se sugiere el uso de este algoritmo de refuerzo como una solución confiable para la clasificación de la calidad del agua. El análisis presentado en este estudio ofrece un marco que puede apoyar a los investigadores en la mejora de la calidad del agua mediante el uso de inteligencia artificial.

Aporte: En este estudio, se evalúan varios algoritmos de IA para manejar datos WQ recopilados durante un período prolongado y desarrollar un enfoque confiable para pronosticar la calidad del agua con la mayor precisión posible. Específicamente, se usaron varios clasificadores de aprendizaje automático y sus modelos de conjuntos de apilamiento para clasificar los datos de WQ a través del Índice de calidad del agua (WQI). Los clasificadores estudiados incluyeron Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost y Multilayer Perceptron (MLP). El conjunto de datos utilizado en el estudio incluyó 1679 muestras y sus metadatos recopilados durante nueve años. Además, se utilizaron curvas de recuperación de precisión y curvas características operativas del receptor (ROC)

para evaluar el rendimiento de los diversos clasificadores.

3.3.1.12 Artículo 12: Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors (Naghibi et al., 2020).

Resumen: Los recursos de agua subterránea (GW) satisfacen una parte significativa de la demanda mundial de agua para sectores como la agricultura, la industria y el consumo humano. Esto es especialmente importante en regiones áridas y semiáridas, donde la escasez de agua superficial y la alta evaporación hacen que el agua subterránea sea un recurso valioso. Sin embargo, la información sobre los recursos de agua subterránea a menudo es incompleta o inexistente, lo que dificulta su evaluación. En este estudio, se desarrollaron métodos para crear mapas de potencial de agua subterránea (GWPM) confiables utilizando únicamente datos derivados de un modelo de elevación digital (DEM) como insumo.

Se seleccionó una región de estudio en Irán, y se extrajeron 13 factores del DEM. Se recopilaron datos de ubicación de manantiales proporcionados por organismos del sector del agua, y junto con puntos sin manantiales, se utilizaron para entrenar y validar algoritmos de aprendizaje automático. Se emplearon algoritmos como bosque aleatorio (RF), su versión mejorada RF paralelo (PRF) y refuerzo de gradiente extremo (XGB) con distintos tipos de refuerzo para identificar el potencial de agua subterránea de manera confiable.

Los resultados, medidos mediante el área bajo la curva de características operativas del receptor, mostraron que tanto PRF como XGB con refuerzo lineal proporcionaron una precisión similar y alta (alrededor del 86%) en la creación de los mapas de potencial de agua subterránea. Los factores más relevantes para un GWPM preciso fueron la convergencia, el índice de humedad topográfica, la densidad de los ríos y la altitud. En resumen, se concluyó que es posible generar mapas de potencial de agua subterránea de alta precisión solo con factores derivados del DEM, con una precisión aceptable. Esta metodología puede ser

utilizada para proporcionar información inicial sobre la explotación de agua subterránea en áreas con escasez de datos.

Aporte: El artículo contribuye al campo de la clasificación de datos de agua al implementar algoritmos de aprendizaje automático, como Extreme Gradient Boosting y Parallel Random Forest, para predecir el potencial de manantiales subterráneos. Utilizando factores derivados de modelos digitales de elevación (DEM), los algoritmos logran alta precisión en la creación de mapas de potencial hídrico, destacando la relevancia de variables geográficas clave. Esta metodología permite una evaluación eficiente en áreas con datos limitados sobre recursos hídricos, mejorando la gestión del agua.

3.3.1.13 Artículo 13: Performance analysis of the water quality index model for predicting water state using machine learning techniques (Uddin et al., 2023b).

Resumen: Los modelos de índice de calidad del agua (WQI) existentes evalúan la calidad del agua con diferentes esquemas de clasificación, lo que genera interpretaciones diversas y cierta incertidumbre en la clasificación precisa de la calidad del agua. Este estudio tuvo como objetivo evaluar el rendimiento de un modelo WQI para clasificar correctamente la calidad del agua costera utilizando un nuevo esquema de clasificación. Se utilizaron datos del puerto de Cork, recopilados por la Agencia de Protección Ambiental de Irlanda (EPA), y se probaron cinco algoritmos de aprendizaje automático: máquinas de vectores de soporte (SVM), Naïve Bayes (NB), random forest (RF), k-nearest neighbor (KNN) y gradient boosting (XGBoost), para predecir la calidad del agua utilizando siete modelos WQI, tres de ellos propuestos recientemente.

Los mejores resultados se obtuvieron con los algoritmos KNN (100 % de precisión) y XGBoost (99,9 % de precisión). XGBoost fue el mejor clasificador con exactitud, precisión, sensibilidad, especificidad y puntuación F1 cercanas al 1.0. Además, los modelos de media cuadrática ponderada (WQM) y de raíz

cuadrada media no ponderada (RMS) resultaron ser más precisos y confiables para la clasificación de la calidad del agua en comparación con otros modelos.

Aporte: Nos muestra como se utilizaron cuatro algoritmos de clasificación de aprendizaje automático, incluidas las máquinas de vectores de soporte (SVM), Naïve Bayes (NB), random forest (RF), k-nearest neighbor (KNN) y gradient boosting (XGBoost), para identificar el mejor clasificador para predecir las clases de calidad del agua utilizando siete modelos WQI ampliamente utilizados, si bien mostraron que los modelos WQM y RMS podrían ser efectivos y confiables para evaluar la calidad del agua costera, sirve para determinar el uso de los algoritmos para la clasificación del índice de la calidad de agua.

3.3.1.14 Artículo 14: A combined water quality classification model based on kernel principal component analysis and machine learning techniques (Dilmi, 2022).

Resumen: El monitoreo de la calidad del agua es fundamental para la gestión ambiental y la protección de los recursos hídricos. Sin embargo, los crecientes riesgos de contaminación han complicado y encarecido el uso de métodos convencionales de monitoreo. En la actualidad, la implementación de procesos automatizados mediante inteligencia artificial y técnicas de aprendizaje automático se ha vuelto indispensable para mejorar el control de la calidad del agua y reducir los costos operativos. Este artículo compara tres técnicas de aprendizaje automático: K-nearest neighbors (KNN), árbol de decisión (DT) y máquina de vectores de soporte (SVM), para clasificar la calidad del agua de la presa de Tilesdit en Argelia. Además, se aplicó la técnica de análisis de componentes principales de kernel (KPCA) para seleccionar las variables más relevantes en la clasificación. Los modelos se entrenaron y validaron con datos históricos de la estación de monitoreo de la presa, recopilados durante 3 años (2016-2018). Los resultados mostraron que la combinación de KPCA y DT ofreció el mejor rendimiento, alcanzando una precisión del 99,68% en la clasificación.

Aporte: El artículo aporta al campo de la clasificación de datos de agua al combinar el análisis de componentes principales con kernel (KPCA) y técnicas de aprendizaje automático, como las máquinas de soporte vectorial (SVM) y las redes neuronales, para mejorar la precisión en la clasificación de la calidad del agua. La reducción de dimensionalidad mediante KPCA optimiza el procesamiento de datos, permitiendo una identificación más eficiente de los parámetros clave que afectan la calidad del agua, lo que resulta en predicciones más robustas y precisas.

3.3.1.15 Artículo 15: Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water (Saddiqi et al., 2024).

Resumen:

Este estudio se centra en mejorar la eficiencia y confiabilidad de los sistemas de filtración de membranas utilizados para tratar el agua producida, mediante la predicción precisa de parámetros clave como la fuerza de elevación, la presión transmembrana (TMP) y la resistencia total (TR) utilizando modelos de aprendizaje automático, incluyendo modelos lineales, k-Nearest Neighbors (KNN) y redes neuronales artificiales (ANN). Los resultados indican que KNN es particularmente eficiente en la predicción precisa de estos parámetros de membrana y podría ser una herramienta robusta para optimizar los procesos de filtración de membranas. Además, se emplea el modelado de superficie de respuesta (RSM) para aclarar la relación entre los parámetros de entrada y salida, identificando condiciones óptimas para mantener una TMP baja y una filtración efectiva. Los hallazgos sugieren que el uso de modelos predictivos puede mejorar significativamente la eficiencia y confiabilidad del tratamiento de agua producida, mejorando la eliminación de contaminantes, reduciendo el consumo de energía y costos, y potencialmente disminuyendo los impactos ambientales, lo que contribuye a la sostenibilidad en la industria del petróleo y gas.

Aporte: El artículo evalúa la eficiencia de varios modelos de aprendizaje automático, incluyendo K-Nearest Neighbors (KNN), para predecir parámetros clave en sistemas de filtración de membranas utilizados en el tratamiento de agua producida. Los resultados indican que KNN es particularmente eficiente en la predicción precisa de parámetros de membrana y podría ser una herramienta robusta para optimizar los procesos de filtración de membranas. Por lo tanto, este artículo aporta información valiosa sobre la aplicación de KNN en el contexto de la clasificación de datos de agua basados en el ICA, destacando su potencial para mejorar la precisión y eficiencia en la evaluación de la calidad del agua.

3.3.2 Resumen de la revisión de los artículos

En la Tabla 1 se muestra el resumen de la revisión de los quince artículos de diferentes fuentes indicando el aporte para esta investigación de los últimos cuatro años:

Tabla 1

Resumen de artículos revisados relacionados a calidad del agua y machine learning.

No	Título	Autor(es)	Fuente	Año	Palabras claves	Aporte
1	Prediction of groundwater quality using efficient machine learning technique	(Singha et al., 2021)	ScienceDirect	2021	Aprendizaje automático. Contaminación del agua. Agua potable	El aporte principal de este artículo es el manejo de la calidad de las aguas subterráneas con los algoritmos de aprendizaje automático (ML) los cuales se utilizan para predecir la calidad de las aguas subterráneas. En este artículo se

						pretende demostrar que la evaluación mediante el método EWQI no es tan preciso y por ello se compara el rendimiento de predicción de los modelos RF, XGBoost, ANN y DL. Estos algoritmos sirven como base de esta tesis para demostrar el modelo de predicción de la calidad basado en Deep learning funciona mucho mejor que otros modelos ML
2	Advancing Water Quality Research: K-Nearest Neighbor Coupled with the Improved Grey Wolf Optimizer Algorithm Model Unveils New Possibilities for Dry Residue Prediction	(Tahraoui et al., 2023)	Multidisciplinary Digital Publishing Institute (MPDI)	2023	Agua; parámetros físicoquímicos; residuo seco; K-vecino más cercano; optimizador lobo gris	El aporte principal de este artículo es el desarrollo de un modelo de vecino más cercano K (KNN) optimizado para predecir las cantidades de residuos secos, así como también buscar evaluar el rendimiento del modelo, optimizar los parámetros y mitigar el riesgo de sobreajuste dando como resultado la alta eficiencia, con errores notablemente bajos reflejados en los valores y hace posible superar las

						limitaciones en la gestión de la calidad del agua.
3	Prediction of hydrological and water quality data based on granular-ball rough set and k-nearest neighbor analysis	(Dong et al., 2024b; Nasir et al., 2022b)	PLOS ONE	2024	Calidad de agua. Agua potable. Monitorio de agua. Agua; parámetros físicoquímicos; K-vecino más cercano.	El artículo muestra el uso del método de k vecinos más cercanos y la red de retro propagación. Los resultados del análisis de datos hidrológicos de calidad del agua muestran que el método propuesto produce mejores resultados en comparación con el uso de un regresor de k vecinos más cercanos independiente. Esto es un aporte importante ya que se enfoca en una combinación de ambos métodos el cual puede implementarse dentro de futuras investigaciones y puede aportar a esta tesis el uso preciso del método independiente para la clasificación de parámetros de agua.
4	Machine	(Kanyama	ScienceDirect	2024	Redes de	El principal aporte

	learning applications for anomaly detection in Smart Water Metering Networks: A systematic review	et al., 2024)			agua. Aprendizaje automático (ML). Tratamiento de agua. Agua potable. Calidad de agua.	de este artículo es la comparación y análisis de aplicaciones de aprendizaje automático (ML) para la detección de anomalías en SWMN. El estudio evaluó treinta y dos artículos de investigación escritos entre 2016 y 2023, centrándose en aplicaciones de ML para la detección de anomalías en sistemas de agua inteligentes, redes de agua inteligentes, redes de distribución de agua y redes de agua.
5	Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms	(Hussein et al., 2024)	Multidisciplinary Digital Publishing Institute (MPDI)	2024	Aguas subterráneas ; parámetros hidroquímicos; índice de calidad del agua de riego (IWQI); aprendizaje automático; gestión del agua agrícola.	El aporte de este artículo es la implementación de varios modelos de machine learning con el objetivo de evaluar la calidad de las aguas subterráneas en Naama, implementando enfoques eficientes de aprendizaje automático para pronosticar el IWQI utilizando modelos de Impulso de Gradiente Extremo

						(XGBoost), Regresión de Vectores de Soporte (SVR) y K vecinos más cercanos (KNN).
6	Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data	(Umar & Gray, 2023b)	Multidisciplinary Digital Publishing Institute (MPDI)	2023	lagunas de datos; datos de nivel de agua; series temporales; univariado; multivariado; imputación.	El artículo evalúa los métodos de imputación simple y múltiple utilizados en datos mensuales univariados y multivariados del nivel de agua de cuatro estaciones de agua en los ríos Benue y Níger en Nigeria. Si bien missForest es el mejor, seguido de cerca por k vecino más cercano, demuestra también el uso de métodos de bosque aleatorio y de coincidencia de media predictiva brindando un mayor enfoque en la evaluación de niveles de agua como sus clasificaciones.
7	Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture	(Çakir et al., 2023)	ScienceDirect	2023	Aquaculture Machine Learning RFerns NB SVM kNN	Este artículo aporta estrategias y enfoques relevantes para desarrollar algoritmos que clasifiquen datos de calidad de agua, mejorando la

						precisión y eficiencia en la evaluación ambiental. Las técnicas de validación cruzada y las métricas utilizadas en el estudio pueden implementarse para mejorar el entrenamiento y evaluación de un algoritmo del ICA. Los resultados pueden integrarse en sistemas de monitoreo automatizados, optimizando la clasificación de la calidad del agua y la toma de decisiones.
8	Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques	(Aish et al., 2023b)	ResearchGate	2023	Recursos hídricos · Aprendizaje automático · SVR · MLP · DTR.	El aporte principal de este artículo es usar la metodología y la estructura que se utilizó, ya que muestra la comparación del rendimiento de cinco algoritmos de aprendizaje automático supervisado, que incluyen máquinas de vectores de soporte (SVM), regresión logística (LogR), árboles de decisión (DT), redes neuronales de perceptrones múltiples (MLP-

						NN) y K vecinos más cercanos (KNN) para predecir el índice de calidad del agua (WQI) y la clase de calidad del agua (WQC) en el acuífero costero de la Franja de Gaza.
9	Drinking water potability prediction using machine learning approaches: a case study of Indian rivers	(Ainapure et al., 2023)	ScienceDirect	2023	Calidad del agua. Modelos de aprendizaje automático. Búsqueda en cuadrícula. Índice de calidad del agua. Clasificación de la calidad del agua	Este artículo aportó la utilización de técnicas de aprendizaje automático para predecir la calidad del agua de los ríos de la India. Los enfoques que se exploran incluyen el uso de K-nearest neighbor (KNN), Random Forest y XGBoost, con y sin ajuste de hiperparámetros, y el uso de una red neuronal artificial secuencial aportan en el proceso de la predicción de potabilidad de agua.
10	Ensemble machine learning paradigms in hydrology: A review	(Zounemat-Kermani et al., 2021b)	ScienceDirect	2021	Aprendizaje automático. Calidad de agua. Contaminación del agua. Agua potable.	El gran aporte de este artículo es validar diferentes algoritmos usados en el campo de la hidrología en la última década, así también para tomar en cuenta los

						modelos más usados y poder compararlos con el algoritmo principal de esta tesis, el cual es el de Deep learning por ser el más eficaz en cuanto a algoritmo predictivo de contaminantes de agua
11	Water quality classification using machine learning algorithms	(Nasir et al., 2022b)	ScienceDirect	2022	Calidad del agua Índice de calidad del agua Aprendizaje automático Modelado de pilas Meta clasificador Modelos de conjunto	En este estudio, se evalúan varios algoritmos de IA para manejar datos WQ recopilados durante un período prolongado y desarrollar un enfoque confiable para pronosticar la calidad del agua con la mayor precisión posible. Específicamente, se usaron varios clasificadores de aprendizaje automático y sus modelos de conjuntos de apilamiento para clasificar los datos de WQ a través del Índice de calidad del agua (WQI). Los clasificadores estudiados incluyeron Support Vector Machine (SVM),

						Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost y Multilayer Perceptron (MLP). El conjunto de datos utilizado en el estudio incluyó 1679 muestras y sus metadatos recopilados durante nueve años. Además, se utilizaron curvas de recuperación de precisión y curvas características operativas del receptor (ROC) para evaluar el rendimiento de los diversos clasificadores.
12	Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors	(Naghibi et al., 2020)	ScienceDirect	2020	Recursos hídricos. Sistema de agua. Contaminación de agua. Extreme Gradient Boosting, Parallel Random Forest.	El artículo contribuye al campo de la clasificación de datos de agua al implementar algoritmos de aprendizaje automático, como Extreme Gradient Boosting y Parallel Random Forest, para predecir el potencial de manantiales subterráneos.

						Utilizando factores derivados de modelos digitales de elevación (DEM), los algoritmos logran alta precisión en la creación de mapas de potencial hídrico, destacando la relevancia de variables geográficas clave. Esta metodología permite una evaluación eficiente en áreas con datos limitados sobre recursos hídricos, mejorando la gestión del agua.
13	Performance analysis of the water quality index model for predicting water state using machine learning techniques	(Uddin et al., 2023a)	ScienceDirect	2023	Índice de calidad del agua Clasificación de la calidad del agua costera Incertidumbre del modelo Algoritmo de clasificación Puerto de Cork	Nos muestra como se utilizaron cuatro algoritmos de clasificación de aprendizaje automático, incluidas las máquinas de vectores de soporte (SVM), Naïve Bayes (NB), random forest (RF), k-nearest neighbor (KNN) y gradient boosting (XGBoost), para identificar el mejor clasificador para predecir las clases de calidad del agua utilizando siete modelos WQI

						ampliamente utilizados, si bien mostraron que los modelos WQM y RMS podrían ser efectivos y confiables para evaluar la calidad del agua costera, sirve para determinar el uso de los algoritmos para la clasificación del índice de la calidad de agua.
14	A combined water quality classification model based on kernel principal component analysis and machine learning techniques	(Dilmi, 2022)	ScienceDirect	2022	Clasificación de la calidad del agua; Técnicas de aprendizaje automático; Análisis de componentes principales del kernel; Selección de variables	El artículo aporta al campo de la clasificación de datos de agua al combinar el análisis de componentes principales con kernel (KPCA) y técnicas de aprendizaje automático, como las máquinas de soporte vectorial (SVM) y las redes neuronales, para mejorar la precisión en la clasificación de la calidad del agua. La reducción de dimensionalidad mediante KPCA optimiza el procesamiento de datos, permitiendo una identificación más eficiente de los parámetros clave

						que afectan la calidad del agua, lo que resulta en predicciones más robustas y precisas.
15	Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water	(Saddiqi et al., 2024)	ScienceDirect	2024	Índice de calidad del agua Clasificación de la calidad del agua costera Incertidumbre del modelo Algoritmo de clasificación Puerto de Cork	Este artículo aporta información valiosa sobre la aplicación de KNN en el contexto de la clasificación de datos de agua basados en el ICA, destacando su potencial para mejorar la precisión y eficiencia en la evaluación de la calidad del agua.

Nota. Elaboración propia con base en los artículos revisados.

3.4.1 Aplicativos para el cálculo del índice de calidad de agua (ICA)

El cálculo del Índice de Calidad del Agua (ICA) es crucial para evaluar la salud y la calidad de las fuentes de agua. Existen varias aplicaciones y herramientas diseñadas para calcular y analizar el ICA con sus diversos parámetros.

3.6.1.1. WQI Calculator.

WQI Calculator es una herramienta en línea que permite calcular el Índice de Calidad del Agua basándose en parámetros como pH, turbidez, oxígeno disuelto y otros indicadores clave. Esta aplicación es útil para profesionales y entidades que necesitan evaluar rápidamente la calidad del agua en diferentes ubicaciones.” La WQI Calculator proporciona una manera sencilla y eficiente de calcular el Índice

de Calidad del Agua, ayudando a los usuarios a interpretar los datos de calidad del agua de manera rápida y precisa" (Gestión sostenible del agua, 2024).

3.4.1.2. Aquacheck.

Aquacheck es una aplicación móvil que permite a los usuarios realizar análisis de calidad del agua en el campo y calcular índices de calidad del agua. La aplicación está diseñada para ser usada por técnicos y científicos en la recopilación de datos de agua y en el campo. "Aquacheck facilita la evaluación de la calidad del agua en el campo al permitir el cálculo del índice de calidad del agua de manera rápida y accesible a través de dispositivos móviles" (AquaChek, 2024).

3.4.1.3. Water Quality Index (WQI) Tool by USGS.

La Water Quality Index (WQI) Tool proporcionada por el United States Geological Survey (USGS) permite calcular el índice de calidad del agua basado en una serie de parámetros de calidad. Esta herramienta está diseñada para ayudar a los gestores de recursos hídricos a evaluar la calidad del agua en ríos y lagos. "La herramienta de Water Quality Index (WQI) de USGS ofrece un método estandarizado para calcular el índice de calidad del agua, facilitando la evaluación y comparación de la calidad del agua en diferentes cuerpos de agua" (USGS Water Resources, 2024).

3.4.1.4. HACH WIMS.

HACH WIMS (Water Information Management Solution) es una plataforma avanzada para la gestión y análisis de datos para la calidad del agua. Permite el cálculo del ICA a partir de datos recolectados de múltiples fuentes, y proporciona informes detallados sobre la calidad del agua. "HACH WIMS proporciona una solución integral para el cálculo y análisis del Índice de Calidad del Agua, con capacidades avanzadas de gestión de datos y generación de informes" (HACH, 2023).

3.4.1.5. Water Quality Index Calculator (WQIC).

Water Quality Index Calculator (WQIC) es una aplicación en línea que permite a los usuarios calcular el Índice de Calidad del Agua usando diferentes métodos y parámetros. Está diseñada para ser accesible a profesionales del agua y al público

en general. "Water Quality Index Calculator (WQIC) ofrece una plataforma accesible para calcular el ICA, adaptándose a distintos métodos de cálculo y parámetros de calidad del agua" (WQIC, 2024).

3.4.1.6. AquaSpy.

AquaSpy es una herramienta basada en la web para la gestión de la calidad del agua que también incluye funcionalidades para calcular el índice de calidad del agua. Está diseñada para ser utilizada en el contexto de la gestión de recursos hídricos y en la evaluación ambiental. "AquaSpy integra el cálculo del Índice de Calidad del Agua con otras funcionalidades de gestión, facilitando una evaluación exhaustiva de la calidad del agua en proyectos ambientales y de recursos" (AquaSpy, 2024).

3.4.1.7. Open Water Analytics (OWA) Tools.

Las herramientas de Open Water Analytics (OWA), como EPANET y SWMM, aunque no están diseñadas específicamente para calcular el ICA, pueden integrarse con módulos que permiten evaluar la calidad del agua y calcular índices relacionados. "Open Water Analytics (OWA) proporciona herramientas flexibles que, aunque no están centradas exclusivamente en el cálculo del ICA, permiten la integración y análisis de datos de calidad del agua en modelos de simulación" (OWA, 2023).

CAPÍTULO IV: RESOLUCIÓN DEL PROBLEMA

Huarmey, ubicada en la región costera del departamento de Áncash, tiene características ambientales propias de un área árida y desértica, marcada por la presencia del río Huarmey. Aunque el entorno es mayormente seco, el río y sus pequeños afluentes ofrecen recursos hídricos limitados y estacionales.

Reducir el tiempo y maximizar la precisión en el cálculo del Índice de Calidad del Agua en las cuencas de Huarmey es muy importante ya que, al mejorar la precisión, se pueden enfocar esfuerzos y recursos en las áreas más críticas, gestionando mejor las limitadas fuentes hídricas disponibles. Huarmey enfrenta problemas como la contaminación agrícola y minera. Una evaluación rápida y precisa ayuda a mitigar impactos negativos y a prevenir daños irreversibles. El agua segura es esencial para la salud pública. Un monitoreo eficaz garantiza que las fuentes de agua potable cumplan con los estándares de calidad necesarios para evitar enfermedades. Mejorar el proceso del cálculo del ICA contribuye a la gestión sostenible del agua en una región vulnerable como Huarmey.

4.1. Características de las cuencas de Huarmey

La cuenca del río Huarmey está ubicada en la región costera del Perú, en el departamento de Áncash. Este río es uno de los sistemas hidrográficos más importantes de la zona, y la cuenca en sí presenta una serie de características notables que afectan la calidad y cantidad de agua disponible para los diferentes usos, como la agricultura, el abastecimiento de agua potable y la industria. Los aspectos generales de la cuenca se muestran en la Tabla 2:

Tabla 2

Aspectos generales de la cuenca del río Huarmey.

Aspectos	Descripción
Ubicación geográfica	Departamento de Áncash, costa central del Perú, desemboca en el océano Pacífico.

Clima	Clima árido, con precipitaciones bajas concentradas en los meses de verano.
Extensión	Aproximadamente 2,751 km ² de extensión.
Hidrología	Río intermitente, con caudal estacional, más abundante en los meses lluviosos.
Uso del agua	Agricultura, consumo humano, pesca y minería.
Problemas ambientales	Erosión, sedimentación, contaminación agrícola y minera, escasez de agua en estiaje.
Entidades encargadas	ANA, ALA, Juntas de Usuarios, EPS, municipalidades, sector minero.
Actividades económicas principales	Agricultura, minería, pesca artesanal.

Nota. Adoptado de (Autoridad Nacional del Agua, 2018).

El manejo del agua en la cuenca del río Huarmey es un esfuerzo colaborativo entre entidades gubernamentales como la ANA y ALA, las Juntas de Usuarios, las empresas de saneamiento y otros actores como el sector minero. La coordinación entre estas instituciones es esencial para asegurar que el uso del agua sea eficiente, equitativo y sostenible, considerando los desafíos ambientales y la escasez hídrica que afecta a la región. Como se muestra en la Figura 9, la cuenca del río Huarmey está localizada en la región de Áncash y es clave para la gestión hídrica en la zona.

Figura 9

Localización sobre la cuenca del río Huarmey.



Nota. Localización sobre la cuenca del río Huarney. Adoptado de (Autoridad Nacional del Agua, 2018).

El manejo de los recursos hídricos en las cuencas del río Huarney involucra tanto entidades gubernamentales como organizaciones locales y empresas privadas. Estas instituciones trabajan en la regulación, supervisión y distribución del agua, así como en la implementación de políticas y estrategias de conservación.

1. Ubicación geográfica y extensión:

- La cuenca del río Huarney se extiende en una región predominantemente desértica y costera, con escasa vegetación.
- Se encuentra en el departamento de Áncash y desemboca en el océano Pacífico, cerca de la ciudad de Huarney.
- La cuenca tiene una extensión de aproximadamente 2,751 km².

2. Clima y precipitaciones:

- El clima de la cuenca es árido, con precipitaciones muy bajas en la zona costera (menos de 100 mm anuales), lo que contribuye a la escasez hídrica en la mayor parte del año.

- Las lluvias suelen concentrarse en los meses de verano (diciembre a marzo), lo que genera picos de caudal durante esos meses.
- Durante la temporada de estiaje (seca), el caudal del río se reduce significativamente, y en algunos tramos el río puede quedar completamente seco.

3. Hidrología:

- El río Huarmey es un río intermitente en su recorrido, lo que significa que, en ciertas épocas del año, sobre todo durante el estiaje, sus aguas no llegan hasta el mar.
- Existen varios afluentes pequeños en la cuenca, pero su aporte hídrico es limitado y estacional.
- La cuenca presenta una alta variabilidad en el caudal, lo que depende en gran medida de las precipitaciones que se producen en la sierra durante el verano.

4. Geomorfología:

- La cuenca está ubicada en una zona de la cordillera de los Andes, con un relieve accidentado en su parte alta, y un descenso brusco hacia la llanura costera.
- La cuenca presenta pendientes pronunciadas y una topografía que favorece la erosión y el arrastre de sedimentos.

5. Usos del agua:

- El principal uso del agua en la cuenca de Huarmey es la agricultura. Los sistemas de riego en la cuenca son mayormente tradicionales, lo que conlleva a un uso poco eficiente del recurso hídrico.
- También hay un uso para consumo humano, aunque de manera limitada, y en algunas áreas es necesario recurrir a fuentes subterráneas o aguas importadas desde otras cuencas.
- La minería y la industria en la región también demandan agua, aunque en menor proporción comparada con la agricultura.

6. Calidad del agua:

- La calidad del agua en la cuenca es afectada principalmente por el arrastre

de sedimentos, la contaminación agrícola (uso de fertilizantes y pesticidas) y, en algunos casos, por actividades mineras.

- La escasez de infraestructura de tratamiento de aguas residuales en algunas áreas rurales también puede ser un factor que comprometa la calidad del agua.

7. Problemas ambientales:

- La deforestación en la parte alta de la cuenca y la falta de cobertura vegetal agravan los problemas de erosión y sedimentación.
- La sobreexplotación de las aguas subterráneas en algunos sectores, para compensar la falta de agua superficial, puede generar problemas de agotamiento del recurso hídrico a largo plazo.
- Durante los eventos de El Niño, la cuenca es vulnerable a fuertes precipitaciones, lo que genera inundaciones y desbordes que afectan a la población y la infraestructura.

8. Gestión de recursos hídricos:

- Existe una necesidad de un manejo más eficiente del recurso hídrico en la cuenca, con la implementación de tecnologías de riego más eficientes y la promoción de la reforestación y conservación de suelos en la parte alta de la cuenca.
- La planificación hídrica a nivel local debe considerar la variabilidad del caudal y la demanda creciente de agua, tanto para uso humano como para la agricultura.

El río Huarmey es intermitente, con caudales que dependen de las lluvias en la sierra. Existen también algunas fuentes subterráneas. Dada su naturaleza desértica, la vegetación y fauna están adaptadas a condiciones secas. A pesar de su limitación, el agua del río y de fuentes subterráneas es vital para la agricultura local. La costa cercana a Huarmey proporciona sustento a una parte de la población a través de la pesca. Las playas y paisajes desérticos de la costa ofrecen atractivo para el turismo local y recreación. Sin embargo, el manejo adecuado de los recursos hídricos y la protección ambiental son esenciales para el bienestar a largo plazo de los pobladores de Huarmey.

4.2. Fuentes contaminantes de la cuenca de Huarmey

Las principales fuentes contaminantes en las cuencas del río Huarmey son las actividades agrícolas, mineras y urbanas, siendo las más preocupantes el uso de fertilizantes y pesticidas, los residuos mineros con metales pesados, etc. La combinación de estas fuentes afecta tanto la calidad y los ecosistemas del agua, así como también la salud de la población local. La Tabla 3 muestra las fuentes contaminantes de las cuencas del río Huarmey, así como también el impacto que tienen en la calidad de agua:

Tabla 3

Fuentes contaminantes de las cuencas del río Huarmey.

Cuenca de Huarmey	Fuentes contaminantes	Impacto en la calidad de agua
Cuenca alta	Actividad minera (metales pesados, aguas ácidas), erosión natural.	Contaminación por metales pesados y ácido de mina, aumento de sedimentos.
Cuenca media	Agricultura (fertilizantes, pesticidas), ganadería (excrementos animales), sedimentación por deforestación.	Aumento de nutrientes (nitratos, fosfatos), presencia de patógenos, incremento de sólidos suspendidos.
Cuenca baja	Descargas domésticas (aguas residuales no tratadas), residuos sólidos urbanos, industria local.	Presencia de coliformes fecales, residuos químicos, aumento de basura y microplásticos.

Nota. Las principales fuentes contaminantes en las cuencas del río Huarmey. Adaptado de (Autoridad Nacional del Agua, 2018).

Las principales fuentes contaminantes de las cuencas del río Huarmey son variadas y están vinculadas tanto a actividades humanas como a condiciones naturales. A continuación, se detallan las principales fuentes contaminantes:

1. Agricultura intensiva:

- **Uso de pesticidas y fertilizantes:** Los agroquímicos utilizados en la agricultura se filtran en el suelo y alcanzan los cuerpos de agua, aumentando la concentración de nitratos y fosfatos, lo que puede causar eutrofización en cuerpos de agua cercanos.

- **Sedimentación:** La deforestación y las prácticas agrícolas inadecuadas contribuyen a la erosión del suelo, lo que aumenta la carga de sedimentos en el agua, afectando la calidad del río.

2. Actividades mineras:

- **Metales pesados:** Las operaciones mineras en la región de Áncash pueden liberar residuos tóxicos como arsénico, mercurio, plomo y otros metales pesados, que contaminan las fuentes de agua superficiales y subterráneas.

- **Aguas ácidas:** En algunos casos, la minería puede generar drenaje ácido de minas (DAM), un proceso en el que los minerales sulfurosos se oxidan y forman ácido, contaminando los cursos de agua.

3. Descargas de aguas residuales domésticas:

- **Falta de tratamiento de aguas residuales:** En áreas rurales y urbanas, la falta de infraestructura para el tratamiento de aguas residuales provoca que las descargas domésticas (con coliformes fecales y otros contaminantes) terminen en el río, comprometiendo la calidad del agua para el consumo humano y agrícola.

- **Contaminación por residuos sólidos:** En algunas zonas, los residuos sólidos, incluyendo plásticos y materiales orgánicos, se desechan en los cuerpos de agua o cerca de ellos, afectando el ecosistema.

4. Industria:

- Aunque en menor medida que la agricultura y la minería, algunas industrias locales vierten residuos no tratados en los ríos, contribuyendo a la presencia de contaminantes químicos, aceites y otros materiales tóxicos.

5. Contaminación natural:

- Las actividades humanas agravan el problema, pero la topografía montañosa y las intensas lluvias también provocan erosión y arrastre de sedimentos, afectando la calidad del agua.

6. Ganadería:

- El ganado cercano a los ríos puede contaminar el agua al arrastrar sus excrementos, que contienen coliformes y patógenos que impactan la calidad del agua.

7. Urbanización descontrolada:

- La expansión desordenada de asentamientos humanos cerca de los ríos aumenta la presión sobre los recursos hídricos, generando vertimientos de aguas residuales sin tratar y residuos sólidos.

4.3 Criterios de evaluación para el cálculo del índice de calidad de agua (ICA)

La evaluación de la calidad del agua disponible es fundamental para asegurar su uso en el consumo humano, el abrevadero de animales, el riego, la industria, el ocio o la vida acuática. El estado de conservación de los recursos hídricos refleja el cuidado y la atención tanto hacia el entorno natural como hacia las actividades humanas a largo plazo. Por lo tanto, la gestión sostenible de los recursos hídricos está estrechamente relacionada con la capacidad de garantizar tanto su cantidad como su calidad física, química y biológica. Este apartado incluye el reglamento para la determinación del Índice de Calidad del Agua (ICA) en el Perú, así como la clasificación de las categorías empleadas en los estudios de los parámetros de las muestras de las cuencas de agua. Por último, se describe el proceso actual para el cálculo del ICA, su categorización y clasificación.

4.3.1. Estándares de Calidad Ambiental del Agua

El Perú, en lo que respecta al uso y aprovechamiento de los recursos hídricos, se rige desde hace más de setenta años por un código de aguas cuyos orígenes son desde siglo XIX. Este código fue la base que ayudó a consolidar el control de la agricultura por parte de los grandes terratenientes. “En julio de 1969, se estableció la aplicación de la Ley General de Aguas, mediante D.L. N°17752, sufrió algunas modificaciones en cuanto a la regulación del uso y aprovechamiento de los recursos hídricos hasta la creación del Ministerio del Ambiente – MINAM y la posterior aprobación de las Normas Ambientales de Calidad de las Aguas” (Autoridad

Nacional del Agua, 2018). La referida ley y su reglamento decretan siete (07) clasificaciones de aguas, sean terrestres o marítimas, vinculadas al uso, las cuales son:

- I. Aguas destinadas al abastecimiento doméstico con desinfección básica.
- II. Aguas para abastecimiento doméstico tratadas mediante procesos combinados de mezcla, coagulación, sedimentación, filtración y cloración, aprobados por el Ministerio de Salud.
- III. Aguas utilizadas para el riego de vegetales que se consumen crudos y para el abrevadero de animales.
- IV. Aguas en áreas recreativas de contacto primario, como baños y actividades similares.
- V. Aguas en zonas dedicadas a la pesca de mariscos bivalvos.
- VI. Aguas en áreas de preservación de fauna acuática y pesca, ya sea recreativa o comercial.

También define veintitrés (23) parámetros, agrupados en sustancias bacteriológicas potencialmente peligrosas y parámetros potencialmente nocivos. “Esta ley de aguas estuvo vigente durante 39 años. Con la creación del Ministerio del Ambiente en el año 2008, posteriormente se aprobaron las Normas Nacionales de Calidad Ambiental (ECA) de aguas, con el objetivo de establecer el nivel o grado de concentración de elementos, sustancias o parámetros físicos, químicos y biológicos presentes. ...en el agua, como cuerpo receptor y constituyente fundamental de los ecosistemas acuáticos, que no represente un riesgo significativo para la salud humana o el medio ambiente. Esta norma es aplicable a los cuerpos de agua del territorio nacional en su estado natural. Estos estándares han sido agrupados en cuatro (04) categorías: Categoría 1: Población y ocio; Categoría 2: Actividades marítimas costeras; Categoría 3: Riego de puntos de agua vegetales y animales y Categoría 4: Conservación del medio acuático. Asimismo, estas categorías estuvieron compuestas por quince (15) subcategorías e incluyeron 104 parámetros. Francés En diciembre de 2015, el MINAM estableció la modificación de los parámetros y valores de las Normas Nacionales de Calidad Ambiental (CEA) para el agua, aprobadas mediante Decreto Supremo N°002-2008-MINAM5. La

modificación de las Normas de Calidad Ambiental (ECA) del agua se clasifica en 4 categorías y 16 subcategorías. En 2017, el MINAM, mediante Decreto Supremo N°004-2017-MINAM, aprobó las normas de calidad ambiental del agua y estableció disposiciones adicionales. El presente decreto tiene por objeto recopilar las disposiciones aprobadas por el Decreto Supremo N°002-2008-MINAM, Decreto Supremo N°023-2009-MINAM y Decreto Supremo N°015-2015-MINAM, que aprueba el ECA-Agua, sujeto a lo dispuesto en el presente decreto supremo. Esta recopilación normativa modifica y elimina ciertos valores, parámetros, categorías y subcategorías del ECA-Agua y mantiene otros, los cuales han sido aprobados por los decretos supremos antes mencionados. Este ECA-Agua, aprobado en 2017, tiene 4 categorías y 17 subcategorías (Autoridad Nacional del Agua, 2018)”. En esta investigación, la evaluación de los parámetros fisicoquímicos permite clasificar el índice de calidad del agua en categorías de agua potable (1A2) y agua utilizable como bebida para animales (3D2):

Categoría 1: Poblacional y Recreacional

- a) Subcategoría A: Aguas superficiales destinadas a la producción de agua potable:
 - A1: Aguas que pueden ser potabilizadas con desinfección.
 - A2: Aguas que pueden ser potabilizadas con tratamiento convencional.
 - A3: Aguas que pueden ser potabilizadas con tratamiento avanzado.
- b) Subcategoría B: Agua superficiales destinadas para recreación:
 - B1: Contacto primario.
 - B2: Contacto secundario.

Categoría 3: Riego de vegetales y Bebida Animal

Subcategoría D1: Riego de vegetales:

- Agua para riego restringido.
- Agua para riego no restringido.

Subcategoría D2: Bebida de animales.

El ECA - Aguas es de cumplimiento obligatorio para definir los usos de los cuerpos de agua, considerando las condiciones naturales o niveles de referencia. Además, es esencial para diseñar la normativa legal, conforme a lo estipulado en la

Ley N°28611, Ley General del Ambiente. Por lo tanto, es una herramienta valiosa para evaluar la calidad de los cuerpos de agua naturales en las cuencas hidrográficas del país.

4.3.2. Clasificación de cuerpos de agua superficiales

La clasificación de los cuerpos de agua superficiales es un paso esencial para calcular el Índice de Calidad del Agua (ICA), ya que permite definir los parámetros y criterios específicos que deben aplicarse según el tipo de cuerpo de agua y su uso. Esta clasificación se realiza considerando el tipo de cuerpo de agua, su ubicación geográfica y los usos previstos del recurso.

Clasificar los cuerpos de agua conforme a estándares estandarizados para su evaluación ambiental es crucial, ya que permite asignar a cada espacio marino, estuario, río, arroyo, lago o laguna una categoría definida. “Al respecto, la Dirección de Calidad y Evaluación de los Recursos Hídricos de la ANA, dentro de sus competencias, aprobó la “Clasificación de los Cuerpos de Agua Marinos y Costeros” del Perú mediante la resolución principal N° 030-2016-ANA, que delimita el perímetro del Mar de Grau (dominio marítimo, aguas interiores y zona estuarina) clasificando 133 unidades marinas en relación a los 10 departamentos costeros, en 45 secciones a lo largo del litoral y proyectadas en 3 franjas; y de acuerdo al ECA – Estándares de Aguas, se consideraron tres (3) categorías y siete (7) subcategorías: Categoría 1: Población y ocio con dos subcategorías, 23 unidades clasificadas (18% de las unidades marinas), Categoría 2: Actividades Costeras y de Extracción y Cultivo Marino Continental con tres subcategorías, existen 59 unidades clasificadas (44%), y Categoría 4: Conservación del medio acuático con dos subcategorías, 51 unidades clasificadas (38%)” (Autoridad Nacional del Agua, 2018).

“Se estableció una Resolución Principal N°056-2018-ANA, en la cual se aprobó la Clasificación de Cuerpos de Agua Superficiales Continentales, referida a los principales cuerpos de agua naturales (lagos, lagunas, ríos y quebradas) e incluso represas en el marco de las 159 unidades hidrográficas del Perú, tomando en cuenta

las actividades de explotación de la población y las diferentes actividades productivas. Corresponden a 1.154 cursos de agua clasificados que incluyen los principales ríos del país, los cuales albergan 1.224 secciones o registros clasificados según el ECA-Agua vigente” (Autoridad Nacional del Agua, 2018).

4.3.3. Índice de calidad de agua en Perú

La evaluación de los parámetros de calidad del agua debe realizarse siguiendo normas físicas, químicas y biológicas, que consideren tanto los procesos dinámicos como los factores que los afectan, además de la capacidad del ecosistema para resistir la presión y su capacidad de autodepuración. En Perú, esta evaluación se efectúa comparando los resultados de los parámetros con los valores establecidos en los Estándares de Calidad Ambiental (ECA), de acuerdo con la categoría del cuerpo de agua superficial. Este proceso permite determinar si los estándares se cumplen o no, tomando en cuenta las características del entorno y su uso previsto. No obstante, existe cierta ambigüedad al clasificar la calidad del agua como excelente, buena, regular, mala o muy mala.

Los índices de calidad del agua (ICA) son herramientas matemáticas que integran diversos parámetros, permitiendo resumir grandes volúmenes de datos en una escala única que mide la calidad del agua. Según la Organización para la Cooperación y el Desarrollo Económico (OCDE), los índices ambientales cumplen funciones clave:

1. Reducen el número de mediciones y parámetros necesarios para representar con precisión una situación.
2. Facilitan la comunicación de los resultados obtenidos.

Los índices de calidad del agua son herramientas esenciales para transmitir de manera clara y accesible el estado del agua a las autoridades y al público en general. Además, permiten identificar, comparar y analizar las condiciones del agua, así como sus tendencias a lo largo del tiempo y en distintos lugares. La evaluación de la calidad del agua se presenta en una escala de 0 a 100, donde 0 indica una calidad muy baja y 100 una calidad excelente. Por estas razones, el índice ha sido adoptado

y utilizado por muchos países desde su creación.

4.3.4. Cálculo del índice de calidad de agua en el Perú

El ICA permite monitorear y determinar si el agua es adecuada para el consumo humano, ayudando a detectar parámetros que exceden los límites permitidos, como la presencia de coliformes o metales pesados. Este índice transforma diversos parámetros fisicoquímicos, microbiológicos y de metales pesados en un solo valor numérico, lo que facilita la interpretación y uso de la información tanto para las autoridades como para la población. Esto permite tomar decisiones rápidas sobre la calidad del agua en zonas rurales y urbanas. La figura 10 muestra los agentes contaminantes y su descripción relacionados a la calidad de agua:

Figura 10

Descripción de agentes/fuentes contaminantes en el agua.

Agentes / Fuentes Contaminantes	Descripción
Sólidos en suspensión	Los sólidos en suspensión pueden dar lugar al desarrollo de depósitos de fango y de condiciones anaerobias cuando se vierte agua residual sin tratar al entorno acuático.
Materia orgánica biodegradable	Compuesta principalmente por proteínas, carbohidratos, grasas animales, la materia orgánica biodegradable. Esta última se mide en la mayoría de las ocasiones en función de la DBO (Demanda Bioquímica de Oxígeno) y la DQO (Demanda Química de Oxígeno). Si se descargan al entorno sin tratar su estabilización biológica puede llevar al agotamiento de los recursos naturales de oxígeno y al desarrollo de condiciones sépticas.
Patógenos (bacterias, virus, protozoarios, gusanos)	Transmiten enfermedades contagiosas tales como el cólera, tífus, disentería, gastroenteritis, hepatitis, poliomielitis, esquistomiasis. Estos agentes pueden causar altas tasas de morbilidad y mortalidad si no se toman las medidas adecuadas de higiene y desinfección, o de búsqueda de otras fuentes de agua de mejor calidad.
Nutrientes	Tanto el nitrógeno como el fósforo, junto con el carbono son nutrientes esenciales para el crecimiento de algas. Cuando se vierten al entorno acuático, estos nutrientes pueden favorecer el crecimiento de algas produciendo las floraciones algales. Cuando se vierten al terreno en cantidades excesivas, pueden provocar la contaminación de agua subterránea.

Materia orgánica refractaria	Esta materia orgánica tiende a resistir los métodos convencionales de tratamiento. Ejemplos típicos son los agentes tensoactivos, los fenoles y los pesticidas agrícolas.
Metales pesados	Los metales pesados son frecuentemente añadidos al agua residual en el curso de ciertas actividades comerciales e industriales, y puede ser necesario eliminarlos si se pretende reutilizar el agua residual.
Sólidos inorgánicos disueltos	Los constituyentes inorgánicos tales como el calcio, sodio, y los sulfatos se añaden de suministro como consecuencia del uso del agua y es posible que se deban eliminar si se va a reutilizar el agua residual.
Sólidos sedimentables	Entre los contaminantes sólidos se encuentran arena, arcillas, tierra, cenizas, materia vegetal agrícola, grasas, brea, papel, hule, plásticos, madera y metales. Las partículas del suelo o sólidos de basura se acumulan en el cauce de los ríos, perjudicando a la biota existente. Si los sedimentos acarrean sustancias tóxicas, estos pueden ser transmitidos a otros organismos a través de la cadena alimentaria ocasionando la muerte de los organismos acuáticos. Otras partículas flotan cerca de la superficie enturbando el agua y obstaculizando la penetración de la luz y por ende el proceso de fotosíntesis.
Energía radiactiva	Producen muerte de especies de flora y fauna, problemas en la salud humana, alteraciones genéticas y cáncer.
Energía térmica	Los procesos industriales producen en numerosos casos aguas a elevadas temperaturas. Cuando éstas llegan a canales, ríos, lagos o mares causan varios efectos químicos, físicos y biológicos. Uno de los más graves es la descomposición del agua, agotando el oxígeno que ésta contiene. El aumento notable de la temperatura del agua afecta, además, los ciclos reproductivos, la digestión y la respiración de los organismos que habitan las aguas y cuando la temperatura es demasiado elevada, se presenta incluso muertes de peces.

Nota. Adaptado de (Autoridad Nacional del Agua, 2018).

Según el análisis de los datos de monitoreo de la calidad de los cuerpos de agua superficiales realizado por la ANA, se llevó a cabo una evaluación continua de los parámetros establecidos por el Protocolo Nacional de Monitoreo de la Calidad de los Recursos Hídricos Superficiales. “Además, se consideró la posibilidad de cambios en los recursos hídricos y el riesgo para la salud y el medio ambiente. Teniendo en cuenta las categorías asignadas al cuerpo de agua, se optó por utilizar la "Clasificación de cuerpos de agua superficiales continentales" y los "Estándares de Calidad del Agua (ECA - Agua)" como parámetros para aplicar el método ICA – PE (Autoridad Nacional del Agua, 2018)“. La figura 11 muestra los parámetros

considerados en la clasificación de la categoría de Consumo Humano según la ANA:

Figura 11

Parámetros considerados en la Categoría 1-A2 Poblacional y Recreacional: Aguas que pueden ser potabilizadas con tratamiento convencional.

N°	Parámetro	Unidades
01	Oxígeno disuelto (valor mínimo)	mg/L
02	Demanda Bioquímica de Oxígeno (DBO ₅)	mg/L
03	Arsénico	mg/L
04	Cadmio	mg/L
05	Cobre	mg/L
06	Cromo Total	mg/L
07	Hierro	mg/L
08	Manganeso	mg/L
09	Plomo	mg/L
10	Mercurio	mg/L
11	Zinc	mg/L
12	Potencial de Hidrógeno (pH)	Unid. de pH
13	Coliformes Termotolerantes (44,5 °C)	NMP/100 ml

Nota. Adaptado de (Autoridad Nacional del Agua, 2018).

La figura 12 muestra los parámetros considerados en la clasificación de la categoría de Riego de vegetales y bebida de animales según la ANA:

Figura 12

Parámetros considerados en la Categoría 3 - Riego de vegetales y bebida de animales.

N°	Parámetro	Unidades
01	Cloruros	mg/L
02	Conductividad	mg/L
03	Demanda Bioquímica de Oxígeno (DBO ₅)	mg/L
04	Oxígeno Disuelto (valor mínimo)	mg/L
05	Potencial de Hidrógeno (pH)	Unid. de pH
06	Aluminio	mg/L
07	Arsénico	mg/L
08	Boro	mg/L
09	Cadmio	mg/L
10	Cobre	mg/L
11	Hierro	mg/L
12	Manganeso	mg/L
13	Mercurio	mg/L
14	Plomo	mg/L
15	Zinc	mg/L
16	Coliformes Termotolerantes	NMP/100 ml
17	Huevos y larvas helmintos	Huevos/L

Nota. Adaptado de (Autoridad Nacional del Agua, 2018).

Para que el índice de calidad del agua pueda medir adecuadamente un punto de monitoreo como un acueducto, río o cuenca, es esencial disponer de una cantidad suficiente de datos. Esta información proviene de las actividades de vigilancia y conservación del recurso hídrico realizadas por la Autoridad Nacional del Agua (ANA). “En este sentido, el monitoreo de la calidad del agua en la cuenca es una de las primeras medidas que implementa la ANA en la gestión de los recursos hídricos. Los resultados se reflejan en informes técnicos que evalúan la calidad del agua en la cuenca o en tramos del río, comparándolos con los valores establecidos en los Estándares de Calidad Ambiental (ECA) (Autoridad Nacional del Agua, 2018)“.

Por lo tanto, es fundamental establecer una red de puntos de monitoreo a lo largo del cuerpo de agua para obtener una visión clara de la distribución espacial en el área de estudio. Esta red es diseñada, evaluada y aprobada por la ANA para estudiar la calidad y el comportamiento del agua en el lago, río u otra masa de agua. Se sugiere analizar al menos cuatro parámetros en un mínimo de cuatro periodos de monitoreo. No obstante, no hay un número máximo de parámetros, y con un solo

monitoreo se pueden evaluar múltiples puntos en la región, permitiendo abarcar toda la cuenca. La figura 13 muestra cómo se interpreta el índice de calidad de agua con respecto al resultado obtenido en el cálculo de los parámetros de las muestras del monitoreo respectivo:

Figura 13

Interpretación de la Calificación ICA- PE.

ICA - PE	Calificación	Interpretación
90-100	Excelente	La calidad del agua está protegida con ausencia de amenazas o daños. Las condiciones son muy cercanas a niveles naturales o deseados.
75-89	Bueno	La calidad del agua se aleja un poco de la calidad natural del agua. Sin embargo las condiciones deseables pueden estar con algunas amenazas o daños de poca magnitud.
45-74	Regular	La calidad del agua natural ocasionalmente es amenazada o dañada. La calidad del agua a menudo se aleja de los valores deseables. Muchos de los usos necesitan tratamiento.
30-44	Malo	La calidad del agua no cumple con los objetivos de calidad, frecuentemente las condiciones deseables están amenazadas o dañadas. Muchos de los usos necesitan tratamiento.
0-29	Pésimo	La calidad de agua no cumple con los objetivos de calidad, casi siempre está amenazada o dañada. Todos los usos necesitan previo tratamiento.

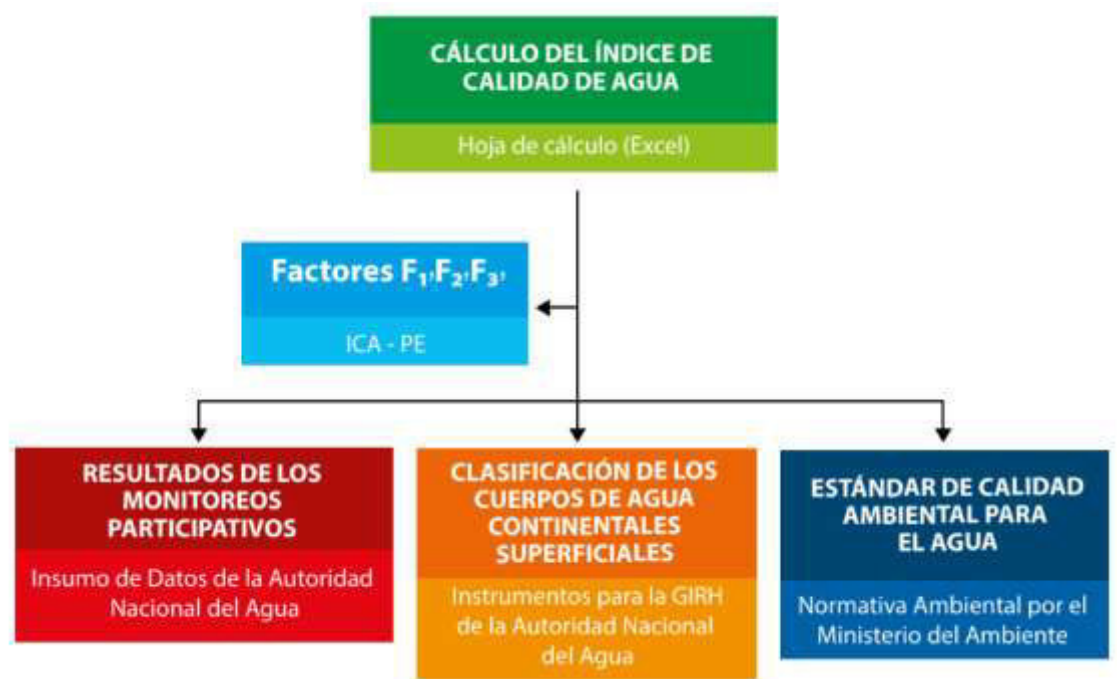
Nota. Adaptado de (Autoridad Nacional del Agua, 2018).

“Este sistema de evaluación de la calidad del agua está vinculado a una escala de colores, donde cada nivel de calidad se representa con un color específico, lo que facilita la interpretación del estado del agua. El indicador, aplicado en un periodo determinado, analiza los resultados de parámetros físicos, químicos y microbiológicos. Mediante una fórmula matemática, convierte una gran cantidad de datos (concentraciones de los parámetros analizados) en una escala única, expresada en porcentaje. Un valor de ICA cercano al 0 por ciento indica un alto impacto negativo en la calidad del agua, mientras que un valor próximo al 100 por ciento refleja condiciones ideales. Este enfoque es útil para gestionar grandes volúmenes de datos y presentar el ICA de un área o cuerpo de agua bajo diversas condiciones, como: (1) monitoreo general, (2) estación seca, y (3) estación lluviosa

(Autoridad Nacional del Agua, 2018)“. La figura 14 muestra la información resumida para la determinación del índice de calidad de agua:

Figura 14

Información base necesaria para la determinación del ICA.



Nota. Estructura de cuadros sobre la categorización del ICA. Adaptado de (Autoridad Nacional del Agua, 2018).

“Para calcular el índice de calidad del agua, se utiliza la fórmula canadiense, la cual se basa en tres factores: alcance, frecuencia y amplitud. Esta fórmula genera un valor único, que oscila entre 0 y 100, y que representa y describe el estado de la calidad del agua en un punto de monitoreo, un cuerpo de agua, un río o una cuenca (Autoridad Nacional del Agua, 2018)”. **Estas fórmulas son extraídas directamente de la documentación de la Autoridad Nacional del Agua** (Autoridad Nacional del Agua, 2018) sobre la metodología para la definición y determinación de los factores para el cálculo del índice de calidad de agua. Estos factores son tres (F1, F2 y F3) que se describen a continuación:

F1 - Alcance: representa la cantidad de parámetros de calidad que no cumplen los

valores establecidos en la normativa de Estándares de Calidad Ambiental para Agua (ECA- Agua) vigente, respecto al total de parámetros a evaluar:

$$F_1 = \frac{\text{Número de parámetros que no cumplen los ECA – Agua}}{\text{Número Total de parámetros a evaluar}} \quad (1)$$

F2 - Frecuencia: indica el porcentaje de datos que no cumplen con la normativa ambiental (ECA-Agua) en relación con el total de datos de los parámetros evaluados, basados en los resultados de al menos 4 monitoreos:

$$F_2 = \frac{\text{Número de parámetros que no cumplen los ECA – Agua de los datos evaluados}}{\text{Número Total de Datos Evaluados}} \quad (2)$$

Datos = Resultados de los monitoreos

F3 - Amplitud: es una medida que refleja la magnitud de la desviación en los datos, calculada a través de la suma normalizada de los excedentes, es decir, los valores que superan los límites establecidos en relación con el total de datos:

$$F_3 = \frac{\text{Suma Normalizada de Excedentes}}{\text{Suma Normalizada de Excedentes} + 1} * 100 \quad (3)$$

La Suma Normalizada de Excedentes (nse) es igual a:

$$nse = \text{Suma Normalizada de Excedentes} = \frac{\sum_{i=1} \text{Excedente}_i}{\text{Total de datos}} \quad (4)$$

El "Excedente" se refiere a cada parámetro y representa la diferencia entre el valor del ECA y el valor del dato en relación con el ECA - Agua. Caso 1: Cuando la concentración del parámetro excede el valor establecido en el ECA - Agua, el excedente se calcula de la siguiente manera:

$$\text{Excedente}_i = \frac{\text{Valor del parámetro que no cumpla el ECA – Agua}}{\text{Valor establecido del parámetro en ECA – Agua}} - 1 \quad (5)$$

Caso 2: Cuando la concentración de un parámetro es inferior al valor establecido en el ECA - Agua, no cumpliendo con la condición especificada, como en el caso del Oxígeno Disuelto (> 4) o el pH (>6.5, <8.5), el cálculo del excedente se lleva a cabo de la siguiente forma:

$$\text{Excedente}_i = \frac{\text{Valor establecido del parámetro en el ECA – Agua}}{\text{Valor del parámetro que no cumpla el ECA – Agua}} - 1 \quad (6)$$

Una vez obtenidos los valores de los factores F1, F2 y F3, se procede al cálculo del

Índice de Calidad de Agua (ICA). Este se determina restando de 100 la raíz cuadrada del promedio de los cuadrados de los tres factores (F1, F2 y F3). El valor resultante se presenta en un rango de 100, indicando una excelente calidad de agua, hasta 0, que representa una calidad de agua muy deficiente. La fórmula se expresa de la siguiente manera:

$$ICA - PE = 100 - \sqrt{\frac{F_1^2 + F_2^2 + F_3^2}{3}} \quad (7)$$

La documentación del ANA menciona que se utiliza un macro de Microsoft Excel basado en las fórmulas del ICA, en la que se ingresan los datos y las fórmulas matemáticas necesarias para obtener los factores (F1, F2 y F3). De esta forma, se calcula el valor del Índice de Calidad del Agua (ICA - PE). El resultado es un número decimal que se sitúa dentro de un rango de 0 a 100, permitiendo clasificar la calidad del agua en cinco categorías: Pésimo, Malo, Regular, Bueno y Excelente, lo que facilita evaluar su estado.

4.4. Uso del algoritmo K vecinos más cercanos (KNN) ante otros algoritmos de clasificación

El algoritmo ofrece varias ventajas para la clasificación de datos, aunque su idoneidad puede variar según el contexto. En este apartado se detallan las razones por las cuales KNN se ha considerado como óptimo para la clasificación de la categorización de los datos de parámetros según el ICA correspondiente.

4.4.1 El uso de K vecinos más cercanos (KNN).

El algoritmo K vecinos más cercanos (KNN) ofrece varias ventajas para la clasificación de datos, aunque su idoneidad puede variar según el contexto. Las razones por las cuales KNN se ha considerado como óptimo para la clasificación de los datos de parámetros fisicoquímicos en el cálculo del índice de calidad de agua (ICA).

4.4.1.1. Simplicidad y Facilidad de Implementación

El algoritmo K vecinos más cercanos (KNN) es conocido por su simplicidad

en la implementación. "El algoritmo de los K vecinos más cercanos es uno de los algoritmos de clasificación más simples. Su principal ventaja es su simplicidad y facilidad de comprensión" (Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2019). Según H. Zhang y D. Liu (2010), "El algoritmo KNN es uno de los métodos más simples y fáciles de entender en el aprendizaje automático. Su implementación básica no requiere un proceso de entrenamiento explícito, lo que facilita su adopción y prueba" (Zhang & Liu, 2010).

4.4.1.2. Sin Suposiciones sobre la Distribución de los Datos

Una de las principales ventajas de KNN es que no hace suposiciones sobre la distribución de los datos. T. M. Mitchell (1997) señala que "KNN no asume ninguna forma particular para la distribución de los datos, lo que le permite ser flexible y aplicable a una amplia variedad de problemas, especialmente cuando las distribuciones de datos no son conocidas" (Mitchell, 1997).

4.4.1.3. Adaptabilidad a Fronteras de Decisión Complejas

KNN es capaz de manejar fronteras de decisión no lineales. C. M. Bishop (2006) afirma que "KNN puede aprender y ajustar fronteras de decisión complejas basadas en la densidad local de los datos, lo que lo hace adecuado para problemas donde las relaciones entre las clases no son lineales" (Bishop, 2006).

4.4.1.4. Versatilidad en Diferentes Tipos de Datos

El algoritmo KNN puede manejar tanto datos numéricos como categóricos, dependiendo de la métrica de distancia utilizada. J. Han, M. Kamber, y J. Pei (2012) mencionan que "KNN puede aplicarse a diferentes tipos de datos utilizando varias medidas de distancia, lo que lo hace versátil para diferentes dominios de datos" (Han, Kamber, & Pei, 2012).

4.4.1.5. Menor Requerimiento de Ajuste de Parámetros

En comparación con otros algoritmos, KNN requiere relativamente poco ajuste de parámetros. I. H. Witten y E. Frank (2005) destacan que "El principal parámetro en KNN es el número de vecinos (k), y su ajuste suele ser el principal desafío. Sin embargo, a diferencia de muchos modelos complejos, KNN no requiere un ajuste exhaustivo de múltiples parámetros" (Witten & Frank, 2005).

4.4.1.6. Robustez Frente al Ruido con un Valor Adecuado de k

Cuando se elige un valor adecuado de k, KNN puede ser robusto frente a datos ruidosos. R. O. Duda, P. E. Hart, y D. G. Stork (2001) explican que "El uso de un valor mayor de k puede ayudar a mitigar el impacto del ruido y las variaciones en los datos, ya que la clasificación se basa en la mayoría de los vecinos cercanos" (Duda, Hart, & Stork, 2001).

4.4.2 Comparación metodológica de desarrollo

Comparar el algoritmo K vecinos más cercanos (KNN) con otros algoritmos de clasificación puede ayudar a entender en qué contextos KNN es más efectivo y cuándo otros métodos podrían ser más apropiados. Para determinar la metodología de desarrollo de este proceso, se opta por detallar el uso del algoritmo K vecinos más cercanos (KNN) frente a otros algoritmos de clasificación.

- **K vecinos más cercanos (KNN) vs. Máquinas de Vectores de Soporte (SVM):** El uso de la Máquinas de Vectores de Soporte (SVM) resulta eficaz en espacios de alta dimensión y puede encontrar la frontera óptima entre clases. Utiliza un margen máximo para mejorar la generalización, pero esto requiere un proceso de entrenamiento que puede ser computacionalmente costoso y necesita la selección de parámetros, como el tipo de núcleo y el valor de C. Cortes y Vapnik (1995) destacan que "SVM busca maximizar el margen entre las clases, lo que permite una mayor robustez a datos de alta dimensión" (Cortes & Vapnik, 1995). H. Zhang y D. Liu (2010) mencionan que "KNN no requiere un proceso de entrenamiento explícito, a diferencia de SVM, que requiere optimización de parámetros y entrenamiento del modelo" (Zhang & Liu, 2010). Debido a que KNN cuenta con una simplicidad y no requiere entrenamiento explícito, la clasificación se basa en la similitud local de los datos.
- **K vecinos más cercanos (KNN) vs. Árboles de Decisión:** Los Árboles de Decisión son intuitivos y fáciles de interpretar. Pueden manejar tanto datos

numéricos como categóricos y no requieren normalización de los datos, pero pueden ser propensos al sobreajuste si no se poda adecuadamente. R. O. Duda, P. E. Hart, y D. G. Stork (2001) sugieren que "KNN puede ser ineficaz en términos de tiempo de computación cuando se enfrenta a grandes conjuntos de datos, en comparación con los árboles de decisión que son más rápidos una vez construidos" (Duda, Hart, & Stork, 2001). Ya que el algoritmo de K vecinos más cercanos (KNN) es capaz de manejar datos no lineales y no necesita suposiciones sobre la distribución de los datos.

- **K vecinos más cercanos (KNN) vs. Redes Neuronales Artificiales (ANN):** El uso de K vecinos más cercanos (KNN) no requiere entrenamiento explícito y es simple de implementar a diferencia del ANN que requiere entrenamiento extensivo y es computacionalmente costoso. Necesita un ajuste detallado de los hiperparámetros. I. H. Witten y E. Frank (2005) observan que "Mientras que KNN ofrece una implementación sencilla y rápida en términos de desarrollo, las redes neuronales requieren más recursos computacionales y un proceso de entrenamiento extenso" (Witten & Frank, 2005).
- **K vecinos más cercanos (KNN) vs. Regresión Logística:** El uso de K vecinos más cercanos (KNN) es flexible y no requiere suposiciones sobre la relación entre características. C. M. Bishop (2006) indica que "La regresión logística es un modelo lineal que proporciona probabilidades de clase y es más eficiente que KNN en términos de tiempo de inferencia, pero puede no capturar relaciones no lineales" (Bishop, 2006). El proceso de la regresión logística supone una relación lineal entre las características y la variable objetivo.

Esta comparación resalta las fortalezas y debilidades de KNN en relación con otros algoritmos de clasificación, ayudando a decidir cuál método puede ser más adecuado según las características y requisitos del problema de clasificación en cuestión. Los modelos matemáticos tradicionales para poder visualizar las ventajas de este modelo frente al uso de esta matemática y modelo manual que se usa en el campo de la calidad de agua, además de su adecuada implementación en

la que se implican los datos del inicio que le de agua y se puede deducir el mejor rendimiento frente al campo de la calidad de agua.

4.5 Implementación de la propuesta del algoritmo para la clasificación de los parámetros fisicoquímicos de las muestras de agua de la cuenca de Huarmey

4.5.1 Software y lenguaje utilizado

4.5.1.1. Colaboratory como plataforma para codificar.

Colaboratory, o Colab, es una plataforma de Google para la ejecución de código Python en notebooks, ofreciendo muchas ventajas como el manejo de datos, el aprendizaje automático y la colaboración en proyectos. Entre sus ventajas, se tiene lo siguiente:

- **Acceso Gratuito a Recursos de Computación Potentes:** Colab proporciona acceso gratuito a GPUs y TPUs, lo cual es beneficioso para los cálculos y entrenar algoritmos de aprendizaje automático sin necesidad de invertir en hardware costoso. Según Google AI Blog, "Colab ofrece acceso a GPUs y TPUs sin costo, lo que democratiza el acceso a recursos computacionales avanzados y facilita la investigación y el desarrollo de modelos complejos" (Google AI Blog, 2023).
- **Entorno Basado en la Nube:** Como plataforma basada en la nube, Colab permite a los usuarios acceder a sus notebooks desde cualquier lugar y en cualquier momento, lo que mejora la flexibilidad y la colaboración. TechCrunch destaca que "la naturaleza basada en la nube de Colab permite a los usuarios trabajar desde diferentes ubicaciones sin necesidad de instalaciones locales" (TechCrunch, 2023).
- **Facilidad de Compartición y Colaboración:** Colab facilita la compartición de notebooks y la colaboración en tiempo real, permitiendo que varios usuarios trabajen simultáneamente en un mismo documento. Según Towards Data Science, "la capacidad de compartir notebooks y colaborar en tiempo real es una característica fundamental que mejora la productividad y facilita el trabajo en equipo" (Towards Data Science, 2023).

- Integración con Google Drive: La integración con Google Drive permite una gestión sencilla de archivos y una sincronización automática de los notebooks. Forbes señala que "la integración con Google Drive hace que la gestión y el almacenamiento de archivos sean más simples y eficientes" (Forbes, 2023).
- Soporte para Bibliotecas y Herramientas Populares: Colab soporta una amplia gama de bibliotecas de Python y herramientas populares para el análisis de datos y el aprendizaje automático, como TensorFlow y PyTorch. Según Medium, "el soporte para una amplia variedad de bibliotecas permite a los usuarios trabajar con herramientas avanzadas sin necesidad de configuraciones complejas" (Medium, 2023).
- Facilidad de Uso y Configuración: La plataforma ofrece un entorno fácil de usar, con una configuración mínima, lo que la hace accesible incluso para aquellos que no son expertos en administración de sistemas. Data Science Central menciona que "la simplicidad de configuración y el entorno amigable para el usuario son aspectos que hacen de Colab una herramienta atractiva para investigadores y desarrolladores" (Data Science Central, 2023).

4.5.1.2. Python como lenguaje a implementar.

Python es un lenguaje de programación ampliamente utilizado para la clasificación de datos, gracias a su flexibilidad, simplicidad y la gran cantidad de bibliotecas disponibles. A continuación, se mencionan algunas ventajas de usar Python en la clasificación de datos. El machine learning en Python es el enfoque más común para desarrollar estos modelos. Python fue lanzado como lenguaje de programación en 1991 y, desde entonces, ha ido ganando popularidad de manera gradual, convirtiéndose en el quinto lenguaje más utilizado en 2021. Este éxito se debe, en parte, a su gran eficiencia en comparación con otros lenguajes, pero también a su activa comunidad y a la gran cantidad de recursos disponibles, como sus bibliotecas especializadas en machine learning.

- Amplia Gama de Bibliotecas y Herramientas: Python cuenta con una

extensa colección de bibliotecas y herramientas para el análisis y la clasificación de datos, como scikit-learn, TensorFlow y Keras. Según Towards Data Science, "la disponibilidad de bibliotecas especializadas en Python facilita la implementación de algoritmos de clasificación y el desarrollo de modelos complejos de manera eficiente" (Towards Data Science, 2023).

- **Simplicidad y Legibilidad del Código:** Python es conocido por su sintaxis clara y legible, lo que facilita la escritura y el mantenimiento del código. Forbes destaca que "la sintaxis simple de Python permite a los desarrolladores escribir código limpio y comprensible, lo cual es crucial para la implementación y depuración de modelos de clasificación" (Forbes, 2023).
- **Gran Comunidad y Soporte:** La comunidad de Python es amplia y activa, lo que proporciona un valioso soporte y una gran cantidad de recursos educativos. Según Data Science Central, "una comunidad activa asegura que los problemas sean resueltos rápidamente y que haya una gran cantidad de tutoriales y ejemplos disponibles para aprender y aplicar técnicas de clasificación" (Data Science Central, 2023).
- **Facilidad para la Integración de Datos:** Python facilita la integración y manipulación de datos mediante bibliotecas como Pandas y NumPy. Medium señala que "las herramientas de manipulación de datos en Python permiten una integración fluida de diversas fuentes de datos, lo cual es esencial para preparar conjuntos de datos para la clasificación" (Medium, 2023).
- **Soporte para Técnicas Avanzadas de Aprendizaje Automático:** Python es compatible con técnicas avanzadas de aprendizaje automático y aprendizaje profundo, lo que permite la construcción de modelos de clasificación más sofisticados. Google AI Blog menciona que "Python es la herramienta preferida para implementar modelos de aprendizaje automático avanzados gracias a su soporte para bibliotecas de vanguardia y su capacidad para manejar grandes volúmenes de datos" (Google AI Blog, 2023).

- Visualización de Datos: Las bibliotecas de visualización en Python, como Matplotlib y Seaborn, ayudan a interpretar los resultados de los modelos de clasificación de manera efectiva. Según KDnuggets, "las herramientas de visualización en Python son cruciales para entender y comunicar los resultados de los modelos de clasificación, lo que permite una mejor toma de decisiones basada en datos" (KDnuggets, 2023).
- Flexibilidad y Adaptabilidad: Python es flexible y adaptable a diferentes necesidades y requisitos de los proyectos de clasificación de datos, desde aplicaciones simples hasta sistemas complejos. Harvard Business Review destaca que "la versatilidad de Python permite a los analistas y científicos de datos adaptar sus enfoques de clasificación según las necesidades específicas del proyecto" (Harvard Business Review, 2023).

4.5.2. Desarrollo del algoritmo de entrenamiento para la función principal

El uso de programación orientada a objetos (POO) en algoritmos de clasificación es importante porque facilita la modularidad y la reutilización del código, lo que permite a los desarrolladores encapsular datos y funciones en clases reutilizables. Esto es particularmente útil para implementar y mantener algoritmos de clasificación complejos, como los que se utilizan para analizar datos de agua. La encapsulación también garantiza que los datos estén protegidos y que solo las funciones autorizadas puedan manipularlos, lo que mejora la seguridad y la capacidad de mantenimiento del código.

Entrenar el algoritmo garantiza que mejore su capacidad de predicción y precisión. A través del entrenamiento continuo con conjuntos de datos variados y bien estructurados, el modelo ajusta sus parámetros para minimizar errores, lo que es crucial en la clasificación de grandes volúmenes de literatura científica. "Los algoritmos de clasificación como las Máquinas de Soporte Vectorial (SVM) se benefician del entrenamiento porque permite al modelo aprender fronteras de decisión complejas para separar datos no lineales. Entrenar el algoritmo asegura que el modelo pueda distinguir correctamente entre clases que no se separan fácilmente en el espacio de

entrada original” (Seyedmohammadi et al., 2023). Entrenar el algoritmo en un conjunto de datos representativo permite minimizar los errores de clasificación. “En un estudio sobre el uso de Random Forests, se encontró que el tamaño y la calidad de los datos de entrenamiento influyen significativamente en la precisión del modelo, especialmente en escenarios complejos como la clasificación de ecosistemas y mapeo” (Hamzaoui et al., 2023). Entrenar un algoritmo de clasificación es crucial para asegurar que pueda generalizar bien a nuevos datos no vistos. “Esto se debe a que el algoritmo aprende patrones relevantes durante el entrenamiento, lo que le permite hacer predicciones precisas en datos futuros” (Dong et al., 2024a).

4.5.2.1. Estructurar el dataset de entrenamiento para el cálculo del índice de calidad de agua

La primera fase consistió en el análisis técnico de las fórmulas encontradas en el cálculo del índice de calidad de agua (ICA) de Perú (ver el apartado 4.4.4), y estructurar la forma de los Datasets y los Dataframes correspondientes. Estructurar adecuadamente un dataset para el entrenamiento del algoritmo es crucial porque influye directamente en la capacidad del modelo para aprender de manera efectiva y generalizar correctamente en datos nuevos. Una mala estructuración puede llevar a un modelo deficiente, sobreajustado o subajustado, lo que afectaría su rendimiento.

Estructurar correctamente el dataset no solo asegura que el algoritmo pueda aprender patrones útiles y generalizables, sino que también optimiza el rendimiento, mejora la eficiencia y facilita la interpretación de los resultados. Toda esta lógica para el entrenamiento de datos para la clasificación de las muestras de los parámetros de agua se encapsula en la clase Eca, siguiendo los datos de entrada de cada categoría. Esta clase contiene un conjunto de atributos y métodos que se utilizaron para modelar datos relacionados a los parámetros para el cálculo del ICA (ver Figura 15).

Figura 15

Diagrama de flujo ECA para el algoritmo de entrenamiento basado en el cálculo del ICA

– Perú.

Eca
+_o: float
+_dbo: float
+_as_: float
+_cd: float
+_cu: float
+_cr: float
+_fo: float
+_mn: float
+_pb: float
+_hg: float
+_zn: float
+_ph: float
+_al: float
+_b: float
+_coliformes: float
+_cloruros: float
+_conductividad: float
+_huevos_de_emintos: float
+propertynumber: int
+NOMINALS: Dict[str, Union[float, int]]
+COMPARES: Dict[str, Callable[[float, float], bool]]
+PROBES: List[Dict[str, Union[float, int]]]
+HEADERSA: List[Dict[str, str]]
+HEADERSB: List[Dict[str, str]]
+HEADERSC: List[Dict[str, str]]
+HEADERSD: List[Dict[str, str]]
+params_gnd_SVM: List[Dict[str, Union[str, float]]]

Nota. Lógica para el entrenamiento de datos para la clasificación de las muestras de los parámetros de agua. Elaboración propia.

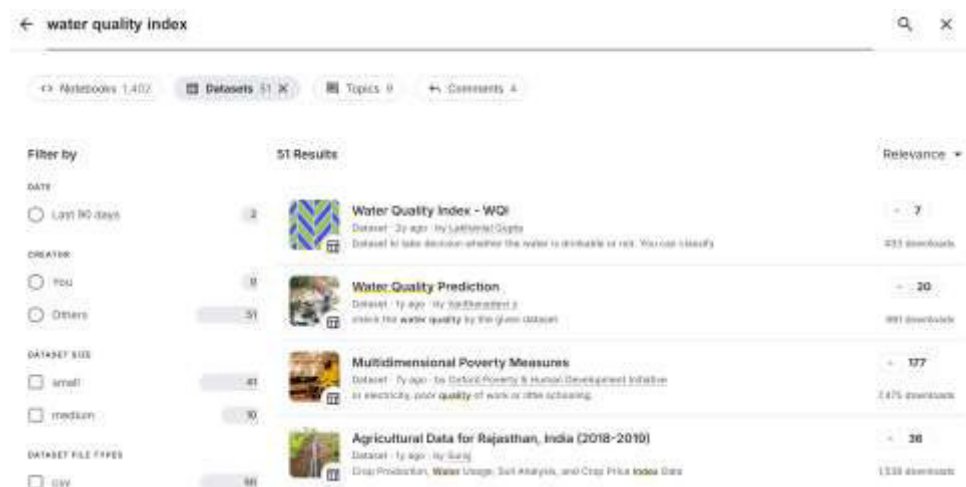
Esta estructura representa un modelo de datos orientado al cálculo del índice de calidad de agua, con diferentes atributos los cuales describen la concentración de elementos químicos, indicadores de contaminación (como coliformes), y otras medidas importantes (pH, conductividad, etc.). Además, incluye elementos para la comparación de datos, listas de encabezados para organizar la información, y parámetros utilizados en el proceso de entrenamiento. El resultado de esta etapa fue la definición de una tabla Excel para cargar los datos de entrada en el algoritmo principal y para el algoritmo de entrenamiento definidos como `templateIca` y `trainKNN` respectivamente.

4.5.2.2. Recopilación de datos para el entrenamiento

Recopilar una gran cantidad de datos para el entrenamiento de un algoritmo es fundamental para garantizar un buen desempeño y la capacidad de generalización del modelo. Cuando un modelo tiene pocos datos, corre el riesgo de sobreajustarse, es decir, de aprender patrones específicos de los datos de entrenamiento en lugar de aprender patrones generales. Con una mayor cantidad de datos, el modelo puede capturar una representación más precisa de los datos subyacentes y no simplemente memorizar casos particulares. La segunda fase consistió en la obtención de datos de concentraciones de los materiales que participen para el cálculo del ICA. Los datos fueron obtenidos de Kaggle como se muestra en la Figura 16.

Figura 16

Imagen de búsqueda de dataset relacionado a los parámetros del índice de calidad de agua para las categorías 1-A2 y 3-D2.



Nota. Extraída de Kaggle (<https://www.kaggle.com/>).

Kaggle es una plataforma en línea que reúne a la comunidad de ciencia de datos más grande del mundo, con más de 536.000 miembros activos en 194 países, que aloja más de 150.000 publicaciones por mes y proporciona

Nota. La imagen representa la variedad de tablas de datos relacionados con los parámetros de agua para el entrenamiento del algoritmo de clasificación del ICA. Tomado de diversas fuentes de la página web Kaggle con dirección de <https://www.kaggle.com/search?q=water+quality+dataset>.

Recopilar grandes cantidades de datos para el entrenamiento de un algoritmo es esencial para mejorar la capacidad de generalización, reducir el sesgo, aumentar la precisión, capturar patrones complejos y robustecer el modelo ante ruidos y outliers. Además, permite entrenar modelos más complejos y realizar mejores evaluaciones de desempeño. En resumen, cuantos más datos tengas, mejor podrá el modelo aprender y adaptarse a situaciones del mundo real.

4.5.2.3. Generación de datos faltantes para el Dataset del algoritmo de entrenamiento

El overfitting sucede cuando un modelo se adapta excesivamente a los datos de entrenamiento, lo que impide que generalice correctamente a datos nuevos. Para prevenir este problema, es crucial controlar la complejidad del modelo, recolectar más datos, aplicar regularización y emplear técnicas de validación efectivas. Un modelo balanceado es aquel que encuentra el punto adecuado entre aprender los patrones importantes sin sobreajustarse al ruido. Después de la recolección de datos se obtuvieron un 70% de datos para 8000 lecturas, se generaron dos funciones que toman un error permitido y rellenan los espacios faltantes en los datasets (`cutDataset()`, `randomDfList()`). Para ello, en la Figura 18 se muestra el apartado del código de entrenamiento para los tres algoritmos utilizados.

Figura 18

Diagrama de flujo de las funciones para generar datos basados en los parámetros del ICA para el algoritmo de entrenamiento del cálculo del ICA – Perú.

Estas funciones están diseñadas para realizar experimentos de clasificación con algoritmos KNN y Random Forest, incluyendo la generación de conjuntos de datos aleatorios, el corte de datasets, y la posterior clasificación de nuevos datos utilizando estos modelos. Además, cuenta con un mecanismo para esperar el entrenamiento de un modelo antes de utilizarlo, lo cual fue útil en procesos asíncronos y distribuidos. En este caso, se evitó el overfitting teniendo primero la mayor cantidad de datos posibles de cada categoría siempre validando con cross-validation, y posteriormente seleccionando un conjunto de pruebas de datos pertenecientes al área de cada categorización con la distribución más uniforme posible.

4.5.2.4. Generación de modelos en base al algoritmo entrenado

Una vez ejecutando la primera generación de datos teniendo errores permitidos $0 < \text{valorNominal} < 2 * \text{valorNominal}$, se generaron datos tendiendo a que estén todos dentro de los rangos o todos fuera de rango para obtener las distintas categorías, las cuales se muestran gráficamente más adelante. Durante la etapa de generación de datos hubo categorizaciones con menos frecuencia que otras eso conllevó la necesidad de seleccionar una cantidad homogénea de muestras por categorización.

Una vez teniendo los Dataset seleccionados se procedió a la generación de los modelos productivos. El underfitting (o subajuste) en el campo de la inteligencia artificial ocurre cuando un modelo es demasiado simple para captar los patrones o la estructura subyacente de los datos. Esto provoca que el modelo no se ajuste adecuadamente ni a los datos de entrenamiento ni a los de prueba, lo que resulta en un bajo rendimiento predictivo. En este caso, se evitó el underfitting durante el análisis y las pruebas del ICA, ya que fue en esta fase donde se evaluó el algoritmo, concluyendo que era necesario utilizar el cálculo de F1, F2 y F3 como entradas para los modelos (ver Figura 19).

Figura 19

Código de entrenamiento para el algoritmo de K vecinos más cercanos (KNN)

desarrollado en Colab.

```

@classmethod
def generar_KNN_multiclass(cls) -> None:
    trainDf = cls.outDataset(1000)

    # Extracción de las características y transposición de la matriz para KNN
    X = trainDf[['f1', 'f2', 'f3']].values.T

    # Extracción de las etiquetas
    C = trainDf['labels'].values

    # Convertir las etiquetas a números
    etiquetas_unicas = list(set(C))
    etiquetas_map = {etiqueta: i for i, etiqueta in enumerate(etiquetas_unicas)}
    C = np.array([etiquetas_map[etiqueta] for etiqueta in C])

    # Entrenar el clasificador KNN
    clasificador = KNN(k=3)
    clasificador.aprendizaje(X, C)

    # Crear directorio para la clase si no existe
    class_dir = f'./{cls.__name__}'
    if not os.path.exists(class_dir):
        os.makedirs(class_dir)
        print(f'Directorio "{class_dir}" creado.")

    # Guardar el modelo
    modelo_filename = "modelo_knn.pkl"
    modelo_filepath = os.path.join(class_dir, modelo_filename)
    joblib.dump(clasificador, modelo_filepath)
    print(f'Modelo exportado como "{modelo_filepath}"')

    # Guardar el mapeo de etiquetas
    etiquetas_filename = "label_encoder_knn.pkl"
    etiquetas_filepath = os.path.join(class_dir, etiquetas_filename)
    joblib.dump(etiquetas_map, etiquetas_filepath)
    print(f'Codificador de etiquetas exportado como "{etiquetas_filepath}"')

    # Verificar si los archivos fueron guardados correctamente
    exito_modelo = cls.esperar_modelo(clase_nombre=cls.__name__, archivo=modelo_filename)
    exito_etiquetas = cls.esperar_modelo(clase_nombre=cls.__name__, archivo=etiquetas_filename)

    if exito_modelo and exito_etiquetas:
        print("Ambos archivos fueron encontrados y están listos para su uso.")
    else:
        print("Hubo un problema al guardar los archivos dentro del tiempo de espera especificado.")

```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

Para determinar el valor óptimo de k en el algoritmo de K vecinos más cercanos (KNN) utilizado para clasificar los parámetros del agua en las categorías de consumo humano y bebida animal según el índice de calidad de agua, se puede seguir un enfoque sistemático y basado en datos. Primero, se divide el conjunto de datos en dos partes: entrenamiento y prueba. Esto permite evaluar la capacidad del modelo para generalizar a nuevos datos que no fueron vistos durante el entrenamiento. Dado que KNN es sensible a la magnitud de los datos, se normalizan las características para asegurar que cada una contribuya equitativamente al cálculo de la distancia. Se prueban varios valores de k para determinar cuál proporciona el mejor rendimiento. En un inicio se estableció como un valor de “ $k=4$ ” ya que la normativa para la clasificación de parámetros de agua especifica el uso mínimo de 4 monitoreos por parámetro. Esto se hace aplicando KNN a los datos de entrenamiento y utilizando una métrica de evaluación, como la precisión, para medir el rendimiento en el conjunto de datos de prueba. Para cada valor de k , se realiza una validación cruzada,

típicamente de 5 o 10 pliegues, para asegurar que los resultados son estables y no dependen de cómo se partitionaron los datos originalmente. La validación cruzada implica dividir el conjunto de entrenamiento en varias subpartes, entrenar el modelo en algunas de estas partes y validar en otras, rotando las subpartes utilizadas para la validación. El valor de k que muestra el mejor equilibrio entre precisión y generalización (evitando tanto el sobreajuste como el subajuste) es seleccionado como el valor óptimo. Un valor de k muy bajo puede hacer que el modelo sea demasiado sensible a los outliers, mientras que un valor muy alto puede hacer que sea demasiado general, ignorando las sutilezas de los datos. El valor óptimo que se ha seleccionado mediante las pruebas de entrenamiento de los datos de parámetros de agua fue “ $k=8$ ”, ya que se utilizó luego para entrenar el modelo final en todo el conjunto de entrenamiento y se evaluó en el conjunto de prueba para verificar su rendimiento. El uso de $k=8$ indica que el modelo considera los 8 vecinos más cercanos para la clasificación, lo cual es una elección para este caso específico y requirió de pruebas de rendimiento y validación, ya que al revisar la distancia euclidiana y realizar las pruebas de entrenamiento, se concluyó en la selección de k como 8. Este proceso ayudó a asegurar que el valor de k elegido sea el más adecuado para la clasificación de los parámetros de agua en las categorías específicas, basado en el índice de calidad de agua en el contexto de tus datos específicos (ver Figura 20).

Figura 20

Código de entrenamiento para el algoritmo de Support vector machine (SVM)

desarrollado en Colab.

```

#-----
@classmethod
def generar_SVM_multiclas(cls):
    # data_etiquetas=cls.leer_para_etiquetado() #Lista de objetos Category1A2
    # # print(data_etiquetas)
    # if data_etiquetas==None:
    #     return None
    X,labels=cls.ecaListFromRndDataset(1600)
    print("Lista de factores ",X)
    print("Lista de etiquetas ",labels)
    # Codificar etiquetas de clase
    label_encoder = LabelEncoder()
    Y = label_encoder.fit_transform(labels)

    # Crear un clasificador SVM con la estrategia "one-vs-one"
    clf = svm.SVC(decision_function_shape='ovo')
    grid_search= GridSearchCV(clf,cls.params_grid_SVM,cv=5,scoring="f1_weighted", return_train_score=True)

    # Entrenar el modelo
    # clf.fit(X, Y)
    grid_search.fit(X, Y)
    print(grid_search.best_params_)
    print(grid_search.best_estimator_)
    # Exportar el modelo
    if not os.path.exists(f'./{cls.__name__}'):
        os.makedirs(f'./{cls.__name__}')
        print(f'Directorio '{f'./{cls.__name__}'}' creado.")

    modelo="modelo_svm.pkl"
    joblib.dump(grid_search, f'./{cls.__name__}/{modelo}')
    print(f'Modelo exportado como '{f'./{cls.__name__}/{modelo}'}')
    exito = cls.esperar_modelo(clase_nombre=f'{cls.__name__}', archivo=f'{modelo}')
    if exito:
        print("Archivo encontrado y listo para su uso.")
    else:
        print("El archivo no fue encontrado dentro del tiempo de espera especificado.")

```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

El código de la Figura 18 describe el método de clase para entrenar al clasificador Support Vector Machine (SVM) usando la estrategia "one-vs-one" (uno contra uno).

La función leer_para_etiquetado() está destinada a cargar datos y devolver una lista de objetos, cada objeto contiene atributos de una muestra de agua. Si la función no devuelve datos, la función generar_SVM_multiclas() retorna None, finalizando su ejecución. Se extraen características (X) y etiquetas (labels) utilizando la función ecaListFromRndDataset(1600), que procesa los datos para formar un conjunto de entrenamiento adecuado para el SVM.

Luego las etiquetas se transforman de categorías textuales a numéricas usando LabelEncoder, una práctica común para adaptar datos categóricos a algoritmos que requieren entradas numéricas. Se crea una instancia de SVM configurada para usar la estrategia de clasificación "one-vs-one". Esto significa que el clasificador compara cada par de clases en el entrenamiento para construir un modelo. Se utiliza GridSearchCV para automatizar la búsqueda de los mejores parámetros para el modelo SVM. Los parámetros probados incluyen diferentes valores de 'C', 'gamma', y tipos de 'kernel'. Además, se configura para utilizar la métrica de puntuación

"fl_weighted", que considera el desbalance de clases y la precisión de la clasificación. Se entrena el modelo con `grid_search.fit(X, Y)` sobre los datos codificados y transformados. Se verifica si existe un directorio para almacenar el modelo entrenado; si no existe, se crea. El modelo y el codificador de etiquetas se guardan en el sistema de archivos usando `joblib.dump`, facilitando su uso posterior sin necesidad de reentrenamiento.

La función espera y verifica la existencia del archivo del modelo, asegurando que el modelo esté listo para su uso antes de proceder. Se maneja adecuadamente la posibilidad de errores y se asegura que todos los componentes necesarios estén en su lugar antes de finalizar el proceso. Este algoritmo es robusto y efectivo para aumentar la precisión y reproducibilidad en los resultados de clasificación de parámetros de agua. El código de la Figura 21 describe el método de clase para entrenar el clasificador Random Forest.

Figura 21

Código de entrenamiento para el algoritmo de Random Forest desarrollado en Colab.

```
#
@classmethod
def generar_randomforest_modelo(cls):
    # data_etiquetas=cls.leer_para_etiquetas() #Lista de objetos Category1A2
    # print(data_etiquetas)
    # if data_etiquetas==None:
    #     return None
    X, labels=cls.cargar_datos()
    print("Lista de factores ",X)
    print("Lista de etiquetas ",labels)
    # Codificar etiquetas de clase
    label_encoder = LabelEncoder()
    Y = label_encoder.fit_transform(labels)
    # Codificar las etiquetas de texto a valores numéricos
    # Entrenar el clasificador RandomForest
    clf_rnd = RandomForestClassifier(random_state=42, n_jobs=-1) # -1 indica que usará todo el poder de cómputo para el modelo
    grid_search = GridSearchCV(clf_rnd, cls.params_grid_random_forest, cv=5, scoring="fl_weighted", return_train_score=True)
    grid_search.fit(X, Y)

    print(grid_search.best_params_)
    print(grid_search.best_estimator_)

    # clf_rnd.fit(X, Y)

    if not os.path.exists(f'./{cls.__name__}'):
        os.makedirs(f'./{cls.__name__}')
        print(f'Directorio '{f'./{cls.__name__}'}' creado.")

    modelo="modelo_rnd_forest.pkl"
    # joblib.dump(clf_rnd, f'./{cls.__name__}/{modelo}.')
    joblib.dump(grid_search.best_estimator_, f'./{cls.__name__}/{modelo}.')
    print(f'Modelo exportado como '{f'./{cls.__name__}/{modelo}.')')
    exito = cls.esperar_modelo(clase_nombre=f'{cls.__name__}', archivo=f'{modelo}.')
    if exito:
        print("Archivo encontrado y listo para su uso.")
    else:
        print("El archivo no fue encontrado dentro del tiempo de espera especificado.")
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

Este proceso es parte de la clasificación de parámetros de agua utilizando Machine Learning, y está estructurado para manejar la optimización de parámetros y la

exportación del modelo.

El método intenta cargar datos utilizando la función `leer_para_etiquetado()`, que retorna una lista de objetos `Category1A2` con los datos necesarios. Si no se encuentran datos, el método retorna `None` y termina prematuramente. El método `ecaListFromRndDataset(1600)` se utiliza para extraer características (X) y etiquetas (labels) de los datos, seleccionando un subconjunto aleatorio o específico de los datos disponibles. Utiliza `LabelEncoder` para transformar las etiquetas de texto a etiquetas numéricas, una práctica estándar para adaptar etiquetas categóricas para uso en modelos de aprendizaje automático. Luego se inicializa una instancia de `RandomForestClassifier` con `random_state=42` y `n_jobs=-1`. El parámetro `n_jobs=-1` permite que el clasificador utilice todos los núcleos del procesador disponibles para acelerar el entrenamiento.

Se configura `GridSearchCV` para optimizar los hiperparámetros del clasificador, utilizando validación cruzada de 5 pliegues (`cv=5`) y evaluando los modelos basados en la métrica "`f1_weighted`". Esto último es particularmente útil para tratar con conjuntos de datos desbalanceados. La función `grid_search.fit(X, Y)` entrena el modelo Random Forest sobre los datos codificados y encuentra el mejor conjunto de parámetros. Los mejores parámetros y el mejor modelo se imprimen después del entrenamiento, proporcionando información valiosa sobre el rendimiento del modelo y la configuración óptima encontrada.

Se verifica si existe un directorio específico para almacenar el modelo; si no existe, se crea. El modelo entrenado y su configuración óptima se guardan usando `joblib.dump`, lo que permite una fácil reutilización del modelo sin necesidad de reentrenamiento.

Este enfoque para entrenar y exportar un modelo de Random Forest es robusto y adecuado para las clasificaciones de parámetros de agua de forma precisa y reproducible, la optimización de hiperparámetros mediante `GridSearchCV` asegura que el modelo está bien ajustado para proporcionar el mejor rendimiento posible según los datos y las etiquetas proporcionadas (ver Figura 22).

Figura 22

Función principal del algoritmo de entrenamiento para SVM, Random Forest y KNN

desarrollado en Colab.

```
def trainCategory1A2() -> None:
    #Genera SVM
    Category1A2.generar_SVM_multiclass()
    ### lee directamente de la carpeta Category1A2

    #Genera random forest
    Category1A2.generar_randomForest_multiclass()
    ### lee directamente de la carpeta Category1A2

    #-----Generar KNN -----
    Category1A2.generar_KNN_multiclass()

def trainCategoryAnimalDrinkingWater() -> None:

    #Genera SVM
    AnimalDrinkingWater.generar_SVM_multiclass()
    ### lee directamente de la carpeta AnimalDrinkingWater

    #Genera random forest
    AnimalDrinkingWater.generar_randomForest_multiclass()
    ### lee directamente de la carpeta AnimalDrinkingWater

    #-----Generar KNN -----
    AnimalDrinkingWater.generar_KNN_multiclass()
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

Estas funciones están diseñadas para entrenar diferentes modelos para el cálculo del índice de calidad de agua.

1. Función trainCategory1A2:

- Esta función entrena tres tipos de modelos diferentes para un conjunto de datos que parece estar en la carpeta Category1A2:
- SVM (Support Vector Machine): Llama a la función `generar_SVM_multiclass()` que genera un modelo SVM multiclase.
- Random Forest: Llama a `generar_randomForest_multiclass()` para generar un modelo de Random Forest multiclase.
- KNN (K-Nearest Neighbors): Llama a `generar_KNN_multiclass()` para generar un modelo KNN multiclase.

Cada modelo se entrena utilizando los datos directamente desde la carpeta Category1A2.

2. Función trainCategoryAnimalDrinkingWater:

- Esta función realiza exactamente los mismos pasos que `trainCategory1A2`, pero con un conjunto de datos diferente ubicado en la carpeta `AnimalDrinkingWater`:

- SVM: Genera el modelo SVM multiclase.
- Random Forest: Genera el modelo Random Forest multiclase.
- KNN: Genera el modelo KNN multiclase.

En resumen, ambas funciones están diseñadas para entrenar tres tipos de clasificadores (SVM, Random Forest y KNN) en diferentes conjuntos de datos, utilizando las funciones correspondientes para cada algoritmo.

Para los resultados del algoritmo entrenado, se cargan diversos datos de entrada. En este caso el modelo de datos de entrada esta definido como `TrainKNN`, que es una tabla en formato Excel que permite buscar los datos por columnas, clasificar los datos, y mostrar de forma gráfica los resultados. La implementación de un clasificador KNN para clasificación multiclase se muestra a continuación, dividido en dos partes principales, el entrenamiento y la clasificación:

- **Entrenamiento del Clasificador KNN**

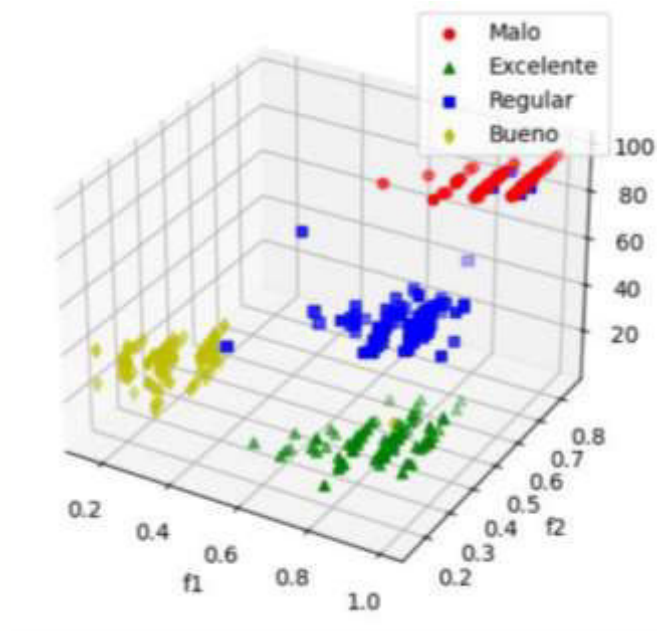
1. **Corte del Dataset:** Selecciona un subconjunto de datos con 1600 elementos usando el método `cls.cutDataset(1600)`.
2. **Extracción de Características y Etiquetas:**
 - X almacena las características extraídas y transpuestas (f1, f2, f3) del DataFrame.
 - C almacena las etiquetas.
3. **Codificación de Etiquetas:** Las etiquetas se convierten en valores numéricos mediante un mapeo.
4. **Visualización:** Se grafican los datos en un gráfico 3D, utilizando diferentes colores y marcadores para cada etiqueta.
5. **Entrenamiento del Modelo KNN:**
 - Se instancia un modelo KNN con $k=8$.
 - Se entrena el modelo con las características y las etiquetas codificadas.
6. **Serialización:** (Comentado en el código)

- Se propone crear un directorio específico para guardar el modelo y el codificador de etiquetas.
- Se guardan ambos objetos usando joblib.
- **Clasificación Usando el Modelo KNN Entrenado**
 1. **Definición de Rutas:** Define las rutas para cargar el modelo KNN y el codificador de etiquetas.
 2. **Carga de Modelo y Codificador:** Se cargan el modelo y el codificador de etiquetas desde sus respectivos archivos.
 3. **Clasificación de Nuevos Datos:**
 - Se transforman los datos de entrada a formato NumPy.
 - Se realiza la clasificación utilizando el método `clasificacion` del modelo KNN cargado.
 - Se convierten los índices clasificados de vuelta a etiquetas originales usando el mapeo almacenado.

Este código proporciona tanto la funcionalidad de entrenamiento como de clasificación en un entorno de prueba. En la Figura 23 se muestra la distribución de los datos clasificados en cuatro categorías de calidad del agua utilizando el modelo entrenado. En la Figura 24, se observa la diferenciación más marcada entre las categorías tras el ajuste de parámetros.

Figura 23

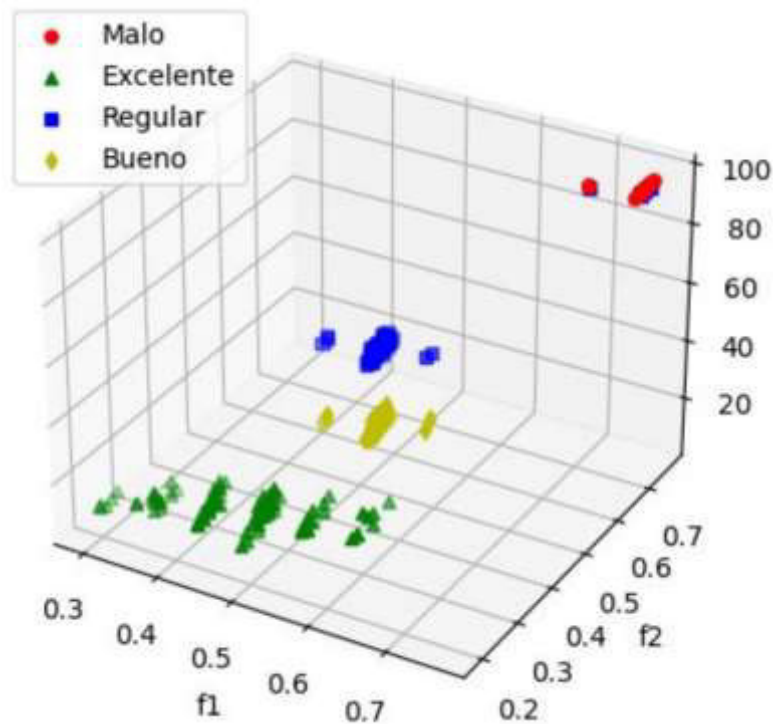
Resultados del entrenamiento del algoritmo K vecinos más cercanos (KNN) en los datos para la categoría 1-A2 de consumo humano.



Nota. Resultados obtenidos a partir del entrenamiento del algoritmo KNN en Google Colab, desarrollado por Jean Paul Pajuelo Leon.

Figura 24

Resultados del entrenamiento del algoritmo K vecinos más cercanos (KNN) en los datos para la categoría de bebida animal 3-D2.



Nota. Resultados obtenidos a partir del entrenamiento del algoritmo KNN en Google

Colab, desarrollado por Jean Paul Pajuelo Leon.

Las tres dimensiones del gráfico están etiquetadas como "f1", "f2" y una tercera dimensión vertical que es la variable que refleja un puntaje o valor en una escala del 0 al 100 del ICA representando f3.

- El eje horizontal "f1" representa una característica o métrica cuantitativa, con valores que oscilan entre 0 y 1.
- El eje "f2" también mide una característica o métrica, y sus valores van desde 0 hasta aproximadamente 0.7.
- El eje vertical representa una variable de resultado o evaluación, con valores que suben de 0 a 100.

Los datos están etiquetados con cuatro categorías diferentes:

- **Rojo (Círculos):** Representa la categoría "Malo", con puntuaciones altas (cercanas a 100) en el eje vertical y ubicadas en una zona particular de los ejes "f1" y "f2".
- **Verde (Triángulos):** Representa la categoría "Excelente", con datos distribuidos principalmente en la parte baja del gráfico (puntuaciones bajas en el eje vertical) y concentrados en un rango específico de "f1" y "f2".

- **Azul (Cuadrados):** Representa la categoría "Regular", con puntos principalmente en la parte media de la gráfica (eje vertical alrededor de 50) y con valores intermedios en "f1" y "f2".
- **Amarillo (Rombos):** Representa la categoría "Bueno", con valores distribuidos principalmente en el cuadrante izquierdo del gráfico.

Para justificar el uso del algoritmo K vecinos más cercanos (KNN) en la clasificación de parámetros de agua según el Índice de Calidad de Agua (ICA), es importante analizar el desempeño del modelo utilizando métricas clave de evaluación en el contexto de clasificación. La validación cruzada es una técnica que divide los datos en varios subconjuntos para entrenar y probar el modelo en diferentes particiones, asegurando que el rendimiento del modelo no dependa de un solo conjunto de datos (Laura Ochoa, 2019). La validación cruzada es una técnica utilizada en el campo de la inteligencia artificial y el aprendizaje automático para medir el rendimiento de un modelo. Su propósito principal es evitar que el modelo se ajuste en exceso a los datos de entrenamiento (overfitting), ofreciendo una evaluación más sólida y confiable (Fernández, 2022). En este caso, los datos de entrenamiento provinieron de la recolección y generación de datos, y se validaron constantemente con los ICAs calculados a partir de los datos obtenidos de la ANA, junto con los registros de los años 2020, 2021 y 2023. Así, en cada iteración, los datos de entrenamiento se comprobaron utilizando datos independientes de estos. Para manejar overfitting, se ajustaron los hiperparámetros correctamente y se implementaron técnicas como la regularización y el control de la profundidad de los árboles. Se evitó el underfitting seleccionando modelos suficientemente complejos, ajustando adecuadamente los parámetros del KNN y la configuración de los árboles en Random Forest. Además, la validación cruzada permitió una evaluación robusta del modelo y evitó que las conclusiones se basaran en un solo conjunto de datos, proporcionando una visión más generalizada del rendimiento.

4.5.3 Estructura de contenido del código a implementar

Un índice bien organizado permite a los desarrolladores y otros usuarios del código

encontrar rápidamente las secciones o módulos que necesitan revisar o modifica. En la Figura 25 se muestra la estructura del código para la implementación del algoritmo principal,

Figura 25

Estructura del código para la implementación del algoritmo principal para la clasificación de agua mediante el ICA-Perú en la plataforma Colab.



Nota. Índice desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

- Algoritmos ICA-PE incluye una breve descripción de lo que contienen los módulos en general para la elaboración de la codificación del algoritmo principal.
- Importaciones es el apartado para mapear y distinguir todas las librerías usadas en la implementación del algoritmo y el manejo matemático de las fórmulas para la clasificación de las categorías del ICA.
- Se toma como título del apartado de la estructura a Estados iniciales, el cual

contiene las variables de los parámetros de agua basados en el reglamento de máximos y mínimos aceptados dadas por el ICA establecidas de Perú.

- Para las Clases, se toma al algoritmo KNN como el líder principal de la función final, así como también el cálculo del ICA para validar la categoría de estados en el que se maneje el algoritmo principal. Se tiene también las clases de entrenamiento final para los algoritmos de clasificación tomadas como SVM y RF.
- Se tiene el apartado de Algoritmos, esto con el fin de unir en una estructura de código final la función que llamará al resto de operaciones para la salida de resultados en el manejo de la clasificación de la categorización de Riego Animal y Consumo Humano establecidos por el índice de calidad de agua, así como también el detalle para mostrar de forma gráfica los resultados y mapear de forma concisa la clasificación final de los datos de entrada.

4.5.4 Desarrollo del código principal a implementar

Se empieza la estructura del algoritmo con el uso de las librerías que Python nos brinda. Colab permite importar fácilmente librerías de Python como TensorFlow, Keras o PyTorch, que son fundamentales para el desarrollo de algoritmos de clasificación. Esto simplifica la implementación de modelos complejos sin necesidad de configuraciones adicionales en el entorno de trabajo (ver Figura 26).

Figura 26

Importación de librerías en la elaboración del algoritmo de clasificación en Colab.

C

```
#Importaciones
import os
import pandas as pd
from typing import List, Dict, Union, Callable, TypeVar, Type
import math
from abc import ABC, abstractmethod
import numpy as np
from sklearn import svm
from sklearn.preprocessing import LabelEncoder
import joblib
import matplotlib.pyplot as plt
import random
from collections import Counter
from mpl_toolkits.mplot3d import Axes3D
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

Es importante el bloque de código con importaciones de librerías y módulos en Python para facilitar el manejo de datos y mejorar las líneas de código.

Librerías estándar y de utilidad:

- os: Para interactuar con el sistema operativo.
- pandas as pd: Utilizado para manipulación y análisis de datos.
- math: Proporciona acceso a funciones matemáticas básicas.
- numpy as np: Fundamental para operaciones matemáticas y lógicas en arrays.
- random: Generación de números aleatorios.
- joblib: Utilizado para serializar objetos Python.
- collections: Contiene estructuras de datos alternativas (como Counter).

Tipado y abstracciones:

- from typing import List, Dict, Union, Callable, TypeVar, Type: Facilita el tipado estático en Python.

Machine Learning y preprocesamiento:

- from sklearn import svm: Módulo de máquinas de vectores de soporte para algoritmos de clasificación y regresión.
- from sklearn.preprocessing import LabelEncoder: Herramienta para codificar etiquetas de texto a valores numéricos.
- from sklearn.ensemble import RandomForestClassifier: Clasificador

basado en el método de ensamble de árboles de decisión.

- `from sklearn.model_selection import GridSearchCV`: Herramienta para optimizar los parámetros de los modelos.

Visualización:

- `import matplotlib.pyplot as plt`: Para la creación de gráficos.
- `from mpl_toolkits.mplot3d import Axes3D`: Soporte para gráficos en 3D con matplotlib.

Abstracciones y patrones de diseño:

- `from abc import ABC, abstractmethod`: Para definir clases y métodos abstractos, que son una forma de imponer estructuras y patrones en el diseño del software.

Los módulos de los estados iniciales en código Python define las operaciones de comparación para una clase relacionada al cálculo del índice de calidad de agua. Estas operaciones son utilizadas en la clase "Eca" para realizar comparaciones entre valores numéricos, relacionados con los parámetros de agua. Estos métodos proporcionan una base para lógicas condicionales que se requieren en la evaluación de estándares de calidad del agua (ver Figura 27).

Figura 27

Estados iniciales para las clases ECA en la elaboración del algoritmo de clasificación en Colab.

▼ Estados iniciales Clase Eca

```
[ ] Operation = Callable[[float, float], bool]

EcaType = TypeVar('EcaType', bound='Eca')

compare_operations : Dict[str, Operation] = {
    'greater_than': lambda x, y: x > y,
    'less_than': lambda x, y: x < y,
    'equal_to': lambda x, y: x == y,
    'greater_or_equal': lambda x, y: x >= y,
    'less_or_equal': lambda x, y: x <= y
}
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

Se define un diccionario de valores nominales para una categoría de calidad de agua, etiquetada como "Category1A2" sobre la categoría de Consum Humano 1-A2. Este diccionario lista los valores para diferentes parámetros químicos y microbiológicos que son esenciales en la evaluación de la calidad del agua (ver Figura 28).

Figura 28

Estados iniciales de la clase ECA para la categoría de Consumo Humano 1-A2 del algoritmo de clasificación en Colab.



```
category1A2PropertyNumber = 13

category1A2Nominals: Dict[str, Union[float, int]] = {
    'O': 5.0,
    'DBO': 5.0,
    'AS': 0.01,
    'CD': 0.005,
    'CU': 2.0,
    'CR': 0.05,
    'FE': 1.0,
    'MN': 0.4,
    'PB': 0.05,
    'HS': 0.002,
    'ZN': 5.0,
    'MIN_PH': 5.5,
    'MAX_PH': 9.0,
    'MAX_NPP_COLIFORMES': 2000
}
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 29 muestra lo que se define como un conjunto de valores nominales para distintos parámetros de muestras de agua en una categoría etiquetada como "Category3" con respecto a la categoría de Bebida de Animales 3-D2. El rango de pH y los niveles máximos de coliformes son especialmente indicativos de la potabilidad del agua para uso animal.

Figura 29

Estados iniciales de la clase ECA para la categoría de Bebida de Animales 3-D2 del algoritmo de clasificación en Colab.

```
Estados iniciales Clase Category3

[ ] Category3Type = TypeVar('Category3Type', bound='Category3')

category3PropertyNumber = 17

animalDrinkingWaterNominals: Dict[str, Union[float, int]] = {
    'CONDUCTIVIDAD': 2500,
    'DBO': 15,
    'O': 5.0,
    'AL': 5.0,
    'AS_': 0.2,
    'B': 5.0,
    'CD': 0.05,
    'CU': 0.05,
    'MN': 0.2,
    'HG': 0.01,
    'PB': 0.05,
    'ZN': 24.0,
    'MIN_PH': 0.5,
    'MAX_PH': 8.4,
    'MAX_NMP_COLIFORMES': 1000
}
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 30 muestra una implementación de la clase de algoritmo K-Nearest Neighbors (KNN) en Python, diseñada para clasificar datos basándose en características similares. Este algoritmo es útil para clasificar nuevos datos basados en su similitud con datos conocidos, donde las características del agua pueden clasificarse en diferentes niveles de calidad basados en muestras previamente clasificadas.

Figura 30

Clases y funciones principales del módulo de Clase KNN para el algoritmo de clasificación en Colab.

Clase KNN

```
[1] class KNN:
    def __init__(self, k=3):
        self.k = k

    @staticmethod
    def EUCLIDIANA(x, y):
        return np.sqrt(np.sum((x - y) ** 2))

    def aprendizaje(self, X, C):
        self.X = X # matriz de vectores de características
        self.c = C # clases asociadas a cada vector x(n)
        self.n_muestras = X.shape[1] # cantidad de muestras

    def clasificacion(self, Y):
        clases = []
        for i in range(Y.shape[1]): # por cada vector y(n) a clasificar
            distancias = np.empty(self.n_muestras)
            for n in range(self.n_muestras): # por cada vector x(n) de características
                distancias[n] = self.EUCLIDIANA(self.X[:, n], Y[:, i])

            # distancias más cercanas:
            k_distancias = np.argsort(distancias)
            # identificar las k distancias - clases
            k_etiqueta = self.c[k_distancias[:self.k]]
            # votación
            c = Counter(k_etiqueta).most_common(1) # (5,0)
            clases.append(c[0][0]) # almacenamos la clase asignada al vector y(n)
        return clases
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 31 muestra la definición de una clase Python llamada “Eca”, que se utiliza para representar y manejar los valores considerados en la categoría de Consumo Animal 1-A2 y la categoría de Bebida de animales 3-D2. La clase está diseñada para encapsular todas las propiedades químicas y microbiológicas necesarias para evaluar la calidad del agua según las normativas aplicables para cada categoría. La clase Eca sirve como un modelo para representar y manipular los parámetros críticos en la evaluación de la calidad del agua, facilitando la implementación de cálculos del ICA de manera estructurada y consistente en el algoritmo.

Figura 31

Clases y funciones principales del módulo de Clase ECA para el algoritmo de clasificación en Colab.

▼ Clase Eca

```
[ ] class Eca(ABC):

    """
    Representa los valores a considerar para el cálculo del ICA para agua Categoría 1-A2

    Atributos:
    - _o : Oxígeno disuelto (valor mínimo) mg/L (float),
    - _dbo : Demanda Bioquímica de Oxígeno (DBO5) mg/L (float),
    - _as_ : Arsénico mg/L (float),
    - _cd : Cadmio mg/L (float),
    - _cu : Cobre mg/L (float),
    - _cr : Cromo total mg/L (float),
    - _fe : Hierro mg/L (float),
    - _mn : Manganeso mg/L (float),
    - _pb : Plomo mg/L (float),
    - _hg : Mercurio mg/L (float),
    - _zn : Zinc mg/L (float),
    - _ph : Potencial de Hidrógeno (pH) Unid. de pH (float),
    - _coliformes : Coliformes Termotolerantes (44,5°C) MPN/ 100 ml (int)

    """

    _o : float = 0
    _dbo : float = 0
    _as_ : float = 0
    _cd : float = 0
    _cu : float = 0
    _cr : float = 0
    _fe : float = 0
    _mn : float = 0
    _pb : float = 0
    _hg : float = 0
    _zn : float = 0
    _ph : float = 0
    _al : float = 0
    _b : float = 0
    _coliformes : float = 0
    _cloruros : float = 0
    _conductividad : float = 0
    _huevos_de_helminos : float = 0

    property number : int = 0

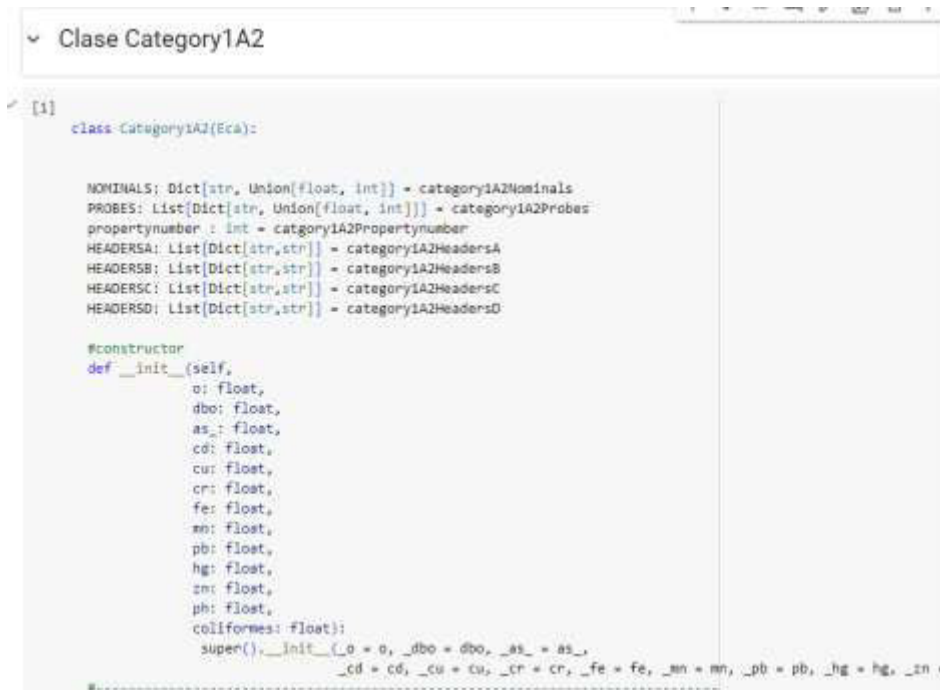
    NOMINALS: Dict[str, Union[float, int]] = {}
    COMPARES : Dict[str, Operation] = compare_operations
    PROBS: List[Dict[str, Union[float, int]]] = []
    HEADERSA: List[Dict[str, str]] = []
    HEADERSB: List[Dict[str, str]] = []
    HEADERSC: List[Dict[str, str]] = []
    HEADERSD: List[Dict[str, str]] = []
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 32 muestra la definición de una clase Python denominada “Category1A2”, que hereda la clase base llamada “Eca”. La estructura y los atributos está específicamente diseñada para manejar datos relacionados con la categoría 1-A2 de Consumo Humano, con puntos específicos de evaluación y límites en los parámetros que difieren de otras categorías. Esta clase se usa para manejar y evaluar la calidad del agua dentro de marcos regulativos y estándares específicos, proporcionando un mecanismo claro y orientado a objetos para trabajar con las muestras del monitoreo cargado.

Figura 32

Clases y funciones principales del módulo de Clase Category1A2 para el algoritmo de clasificación en Colab.



```
class Category1A2(Eca):

    NOMINALS: Dict[str, Union[float, int]] = category1A2Nominals
    PROBES: List[Dict[str, Union[float, int]]] = category1A2Probes
    propertynumber: int = category1A2Propertynumber
    HEADERSA: List[Dict[str, str]] = category1A2HeadersA
    HEADERSB: List[Dict[str, str]] = category1A2HeadersB
    HEADERSC: List[Dict[str, str]] = category1A2HeadersC
    HEADERSD: List[Dict[str, str]] = category1A2HeadersD

    #constructor
    def __init__(self,
                  o: float,
                  dbo: float,
                  as_: float,
                  cd: float,
                  cu: float,
                  cr: float,
                  fe: float,
                  mn: float,
                  pb: float,
                  hg: float,
                  zn: float,
                  ph: float,
                  coliformes: float):
        super().__init__(o = o, _dbo = dbo, _as_ = as_,
                        _cd = cd, _cu = cu, _cr = cr, _fe = fe, _mn = mn, _pb = pb, _hg = hg, _zn = zn)
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 33 muestra la definición de una clase Python llamada Category3 representando a la categoría de Bebida de Animales 3-D2, que hereda de una clase base Eca. Esta clase está diseñada para manejar la configuración y evaluación de parámetros específicos de calidad del agua para una categoría denominada "Category3".

Figura 33

Clases y funciones principales del módulo de Clase Category3 para el algoritmo de clasificación en Colab.

Clase Category3

```
[ ] class Category3(Eca):
    param_grid_random_forest=[
        # n_estimators Probará con 100,500,1000 arboles de decisión en el randomforest
        # usando todos los datos para cada randomforest
        # con un maximo de ramas de 16,32,64
        # max_depth profundidad de los arboles
        # max_features maximo de características
        {"n_estimators":[100,500,1000],"bootstrap":[False,],"max_depth":[16,32,64],"max_features":[2,3,4]},
    ]

    propertynumber : int = category3Propertynumber
    PROBES: List[Dict[str, Union[float, int]]] = category3Probes

    HEADERSA: List[Dict[str, str]] = category3HeadersA
    HEADERSB: List[Dict[str, str]] = category3HeadersB
    HEADERSC: List[Dict[str, str]] = category3HeadersC
    HEADERSD: List[Dict[str, str]] = category3HeadersD

    @constructor
    @abstractmethod
    def __init__(
        self,
        cloruros: float,
        conductividad: float,
        dbp: float,
        o: float,
        ph: float,
        al: float,
        as_: float,
        bi: float,
        cd: float,
        cu: float,
        fe: float,
        mn: float,
        hg: float,
        pb: float,
        zn: float,
        coliformes: float,
        huevos_de_hesintos: float
    ):
        super().__init__(
            _cloruros = cloruros, _conductividad = conductividad, _dbp = dbp, _o = o, _ph = ph
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 34 muestra el código de una clase Python llamada “AnimalDrinkingWater”, que hereda de Category3. Esta clase está diseñada específicamente para gestionar las muestras de consumo animal. La funcionalidad de createFromDF permite integrar datos de calidad de agua desde un formato estructurado (es este caso el Template ICA con formato Excel) directamente en el modelo de datos de la aplicación, facilitando los resultados del análisis de calidad del agua.

Figura 34

Clases y funciones principales del módulo de Clase AnimalDrinkingWater para el algoritmo de clasificación en Colab.

▼ Clase AnimalDrinkingWater

```
[ ] class AnimalDrinkingWater(Category3):

    NOMINALS: Dict[str, Union[float, int]] = animalDrinkingWaterNominals
    PROBES: List[Dict[str, Union[float, int]]] = animalDrinkingWaterProbes

    def __init__(
        self, conductividad: float, dbo: float, o: float, ph: float, al: float, as_: float, b: float, cd: float,
    ):
        super().__init__(
            cloruros=0, conductividad=conductividad, dbo=dbo, o=o, ph=ph, al=al,
            as_=as_, b=b, cd=cd, cu=cu, fe=0, mn=mn, hg=hg, pb=pb, zn=zn, coliformes=coliformes,
            huevos_de_hemintos=0
        )

    @classmethod
    def createFromDF(
        cls : Type[Category3Type],
        df : pd.DataFrame) -> List[Category3Type] :

        """
        Crea una lista de instancias de Category1A2 a partir de un DataFrame.

        Itera sobre cada fila del DataFrame y crea una instancia de Category1A2
        para cada una, añadiéndola a una lista.

        Args:
            df (pd.DataFrame): DataFrame que contiene los datos para crear instancias de Category1A2.

        Returns:
            List[Category1A2]: Una lista de objetos Category1A2 creados a partir de los datos del DataFrame.

        Raises:
            KeyError: Si alguna columna esperada falta en el DataFrame.
            Exception: Para cualquier otro error imprevisto durante el procesamiento.
        """
        categories = []
        try:
            for _, row in df.iterrows():
                category = cls(
                    conductividad = row['conductividad'], dbo = row['dbo'], o = row['o'], ph = row['ph'], al = row['al'],
                )
                categories.append(category)
            except KeyError as e:
                print(f"Error de clave: {e}")
            except Exception as e:
                print(f"Ocurrió un error: {e}")
            return categories
```

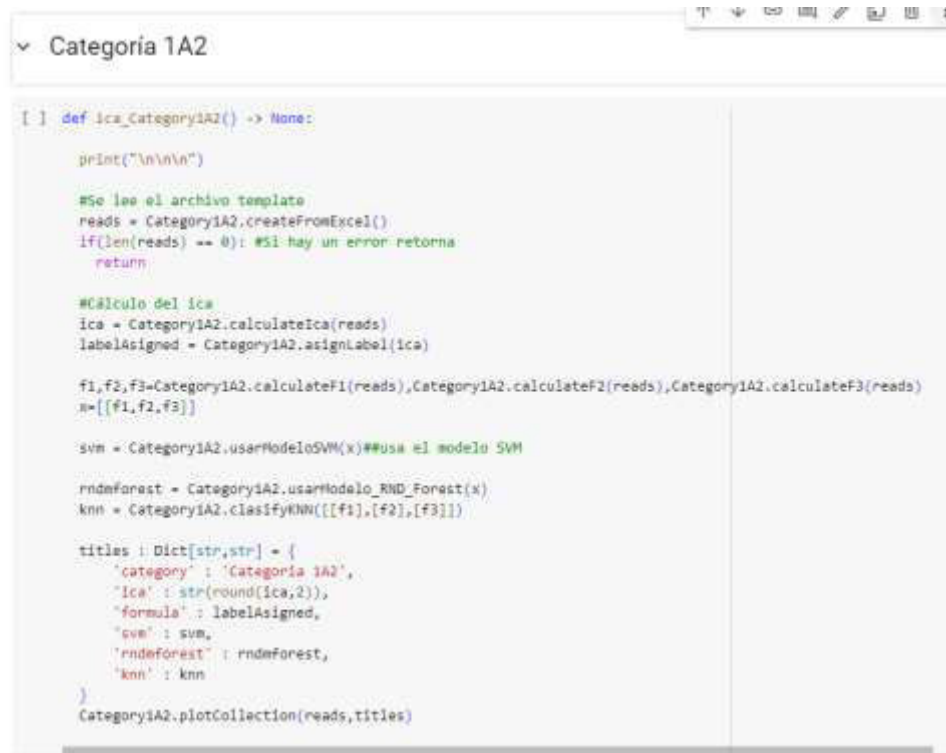
Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 35 muestra el código de una función en Python, `def_ica_Category1A2()`, sobre la categoría de Consumo Humano 1-A2. Este código encapsula el flujo desde la lectura de datos, cálculo de ICA, aplicación de modelos de clasificación, hasta la presentación de los resultados, haciendo uso de

técnicas de machine learning con una mejor precisión y utilidad del análisis.

Figura 35

Módulo de la categoría de Consumo Humano para su clasificación en Colab.



```
[ ] def ica_CategoryIA2() -> None:

    print("\n\n\n")

    #Se lee el archivo template
    reads = CategoryIA2.createFromExcel()
    if(len(reads) == 0): #Si hay un error retorna
        return

    #Cálculo del ica
    ica = CategoryIA2.calculateIca(reads)
    labelAssigned = CategoryIA2.assignLabel(ica)

    f1,f2,f3=CategoryIA2.calculateF1(reads),CategoryIA2.calculateF2(reads),CategoryIA2.calculateF3(reads)
    x=[[f1,f2,f3]]

    svm = CategoryIA2.usarModeloSVM(x)##usa el modelo SVM

    rndmforest = CategoryIA2.usarModelo_RND_Forest(x)
    knn = CategoryIA2.classifyKNN([[f1],[f2],[f3]])

    titles : Dict[str,str] = {
        'category' : 'Categoría 1A2',
        'ica' : str(round(ica,2)),
        'formula' : labelAssigned,
        'svm' : svm,
        'rndmforest' : rndmforest,
        'knn' : knn
    }

    CategoryIA2.plotCollection(reads,titles)
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

La Figura 36 muestra una función Python denominada `ica_AnimalDrinkingWater()`, sobre la categoría "Categoría 3 Bebida de animales". El código encapsula un flujo completo para el cálculo y clasificación del ICA. Este código facilita un enfoque automatizado y sistemático para evaluar la calidad del agua destinada a animales, aplicando métodos de data science para mejorar la precisión de las evaluaciones y la toma de decisiones basada en datos concretos.

Figura 36

Módulo de la categoría de Bebida de Animales para su clasificación en Colab.

▼ Categoría 3

▼ Bebida de animales

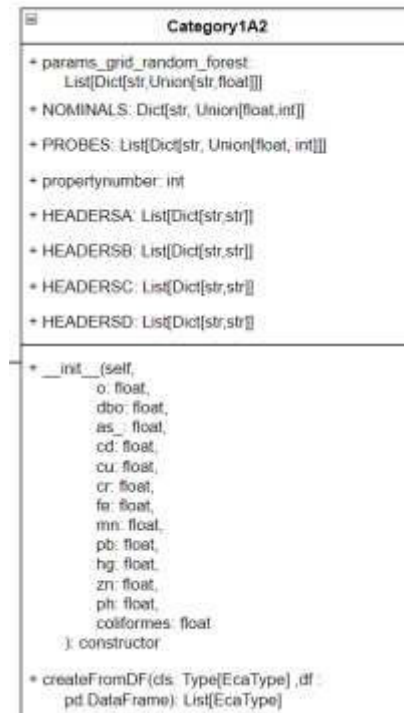
```
[ ] def iia_AnimalDrinkingwater() -> Name :  
    print("Verano")  
    #Se lee el archivo template  
    reads = AnimalDrinkingwater.createfromexcel()  
    if(len(reads) == 0): #Si hay un error retorna  
        return  
  
    ica = AnimalDrinkingwater.calculateIca(reads)  
    labelAssigned = AnimalDrinkingwater.assignLabel(ica)  
  
    F1,F2,F3=AnimalDrinkingwater.calculateF1(reads),AnimalDrinkingwater.calculateF2(reads),AnimalDrinkingwater.calculateF3(reads)  
    m=[F1,F2,F3]  
  
    svm = AnimalDrinkingwater.esetModelosVet(x)##usa el modelo SVM  
    randomforest = AnimalDrinkingwater.marModelo_RF1_forest(x)##usa el  
    knn = AnimalDrinkingwater.classifyKnn([F1],[F2],[F3])  
  
    titles : Dict[str,str] = {  
        'category' : 'Categoría 3 Bebida de animales',  
        'ica' : str(round(ica,F1)),  
        'formula' : labelAssigned,  
        'svm' : svm,  
        'randomforest' : randomforest,  
        'knn' : knn  
    }  
  
    AnimalDrinkingwater.plotcollection(reads,titles)
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

El diagrama de flujos muestra una descripción estructurada de una clase de Python denominada Category1A2 de la categoría de Consumo Humano 1-A2 (Ver Figura 37). La descripción de esta clase indica que está diseñada para manejar y procesar datos DE agua, para la categorización y evaluación según normativas específicas. El uso de modelos como Random Forest y la estructura de datos como DataFrames que se integra con métodos de análisis de datos permiten un manejo eficiente y efectivo de los datos de calidad del agua de los monitoreos cargados.

Figura 37

Diagrama de flujo del algoritmo para la Categoría de Consumo Humano 1-A2.



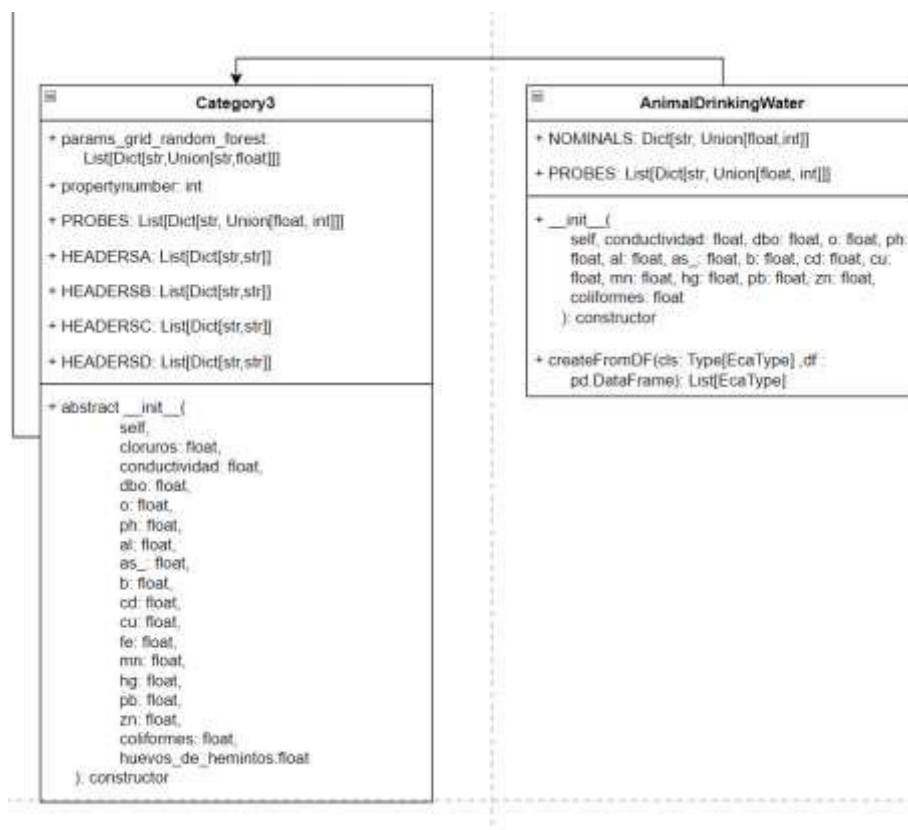
Nota. Elaboración propia utilizando *draw.io* como herramienta de diagramación.

El diagrama de flujos muestra las dos clases Python sobre la categoría de Bebida de Animales 3-D2 definidas como *Category3* y *AnimalDrinkingWater*, ambas diseñadas para las muestras de agua cargadas (Ver Figura 38). Ambas clases están claramente enfocadas en la gestión de datos de calidad del agua, con *Category3* orientada a una categorización más general y *AnimalDrinkingWater* específicamente destinada al agua de consumo animal.

La implementación de métodos como *createFromDF* en *AnimalDrinkingWater* facilita la integración de datos de diversas fuentes, convirtiéndolos en formatos útiles para análisis y gestión continuos.

Figura 38

Diagrama de flujo del algoritmo para la Categoría de Bebida de Animales 3-D2.



Nota. Elaboración propia utilizando *draw.io* como herramienta de diagramación.

La Figura 39 muestra el diagrama que resume el diseño para manejar datos relacionados con el ICA de la clasificación de la categoría de Consumo humano 1-A2 y la categoría de Bebida de animales 3-D2 establecido por la Autoridad Nacional de Agua (ANA) en la clasificación de los parámetros de las muestras de las cuencas monitoreadas. También tiene herramientas para configurar modelos de Random Forest (a través de `params_grid_random_forest`) y convertir datos tabulares (DataFrames) en instancias de esta clase, lo que facilita su uso en tareas de modelado o análisis.

Figura 39

Función principal del algoritmo para mostrar los resultados de la clasificación de datos de muestras de agua de las cuencas de Huarmey.

```
[ ] def __CALCULATE_ICA__():
    ica_CategoryIA2()
    ica_AnimalDrinkingWater()
    __CALCULATE_ICA__()
```

Nota. Código desarrollado por Jean Paul Pajuelo Leon, ejecutado en Google Colab.

4.5.5 Proceso de la Función generar_KNN_multiclass

1. Cargar y preparar el dataset:

- La función inicia cargando los datos de entrenamiento mediante un método cutDataset, que selecciona un subconjunto del dataset según un número total de lecturas (totalReads).
- Extrae las características f1, f2, y f3, que son utilizadas como vectores de entrada X, y las etiquetas labels se transforman en una matriz C.

2. Codificación de etiquetas:

- Se convierte cada etiqueta de clase en un valor numérico utilizando un diccionario etiquetas_map que asigna un índice a cada etiqueta única.

3. Configuración de validación cruzada:

- Implementa validación cruzada con KFold de la biblioteca scikit-learn. Se utiliza un esquema de 5 divisiones (n_splits=5) con barajado de datos (shuffle=True) y una semilla fija para reproducibilidad (random_state=42).

4. Entrenamiento y evaluación por pliegues:

- Para cada pliegue generado por KFold, se realiza lo siguiente:
 - A. Se dividen los datos en conjuntos de entrenamiento y prueba.
 - B. Se grafican las distribuciones de los datos de entrenamiento y prueba en 3D mediante la función plot3D.
 - C. Se entrena un modelo KNN con k=12 usando el método aprendizaje.
 - D. Se evalúa el modelo calculando el puntaje R^2 para las predicciones del conjunto de prueba.
 - E. Se almacena el puntaje R^2 de cada pliegue.

5. Selección del mejor modelo:

- Se identifica el modelo con el puntaje R^2 más alto entre los pliegues y se guarda como el mejor clasificador.

6. Visualización de resultados:

- Se grafica la evolución del puntaje R^2 a lo largo de los pliegues para observar el desempeño del modelo.

7. Exportación del modelo:

- El modelo con el mejor puntaje R^2 se guarda en un archivo utilizando `joblib.dump`.

Módulos principales utilizados:

- `KFold`: Divide los datos en varios pliegues para realizar validación cruzada.
- `r2_score`: Calcula el puntaje R^2 para evaluar la precisión del modelo.
- `plot3D`: Visualiza los datos en un espacio tridimensional para análisis.

El modelo KNN se optimiza utilizando validación cruzada, y se selecciona el modelo con el mejor desempeño basado en el puntaje R^2 . Los resultados de cada pliegue son visualizados, lo que permite identificar tendencias y evaluar la estabilidad del modelo.

4.5.6 Evitar el Overfitting con Validación Cruzada

1. Definición del Proceso de Validación Cruzada:

- Se utiliza validación cruzada de k-pliegues (en este caso, $k=5$) para dividir los datos en subconjuntos de entrenamiento y prueba. Esto asegura que cada parte del dataset sea usada tanto para entrenar como para evaluar el modelo, proporcionando una evaluación robusta del rendimiento.
- El uso de validación cruzada evita el sobreajuste al verificar que el modelo no dependa de un conjunto específico de datos.

2. Configuración de los Datos:

- Los datos se dividen aleatoria y uniformemente en categorías según el área de cada monitoreo.

- Se asegura una distribución homogénea, lo que mejora la capacidad del modelo para generalizar.

3. **Ajuste de Hiperparámetros:**

- A través de técnicas como GridSearchCV, el algoritmo ajusta el valor de k (número de vecinos cercanos) para encontrar la configuración que maximice el rendimiento.
- Se evalúan métricas como `f1_weighted`, que son sensibles a datos desbalanceados.

4.5.7 Módulo del Algoritmo para R^2 y su Uso

1. **Cálculo del R^2 :**

- Durante la validación cruzada, el algoritmo evalúa el rendimiento del modelo en los conjuntos de prueba utilizando R^2 , que mide qué tan bien las predicciones del modelo se ajustan a los datos reales.
- Un valor de R^2 cercano a 1 indica un ajuste excelente, mientras que valores cercanos a 0 sugieren bajo rendimiento.

2. **Optimización del Valor de k :**

- Se prueba una gama de valores de k para el modelo KNN.
- El valor óptimo de k es aquel que maximiza el R^2 y minimiza errores en la validación cruzada.
- En el análisis realizado, el valor de $k=8$ fue seleccionado como el más adecuado, después de observar su desempeño durante las pruebas.

4.5.8 Impacto del Enfoque en el Cálculo del ICA

1. **Ventajas Frente a Métodos Tradicionales:**

- A diferencia del uso de macros en Excel, este enfoque asegura que el modelo no esté sobreajustado ni subajustado, proporcionando predicciones más precisas.
- La automatización del ajuste de k mediante validación cruzada evita errores humanos.

2. **Beneficio para la Clasificación del Agua:**

- La combinación de validación cruzada y métricas como R^2 permite clasificar las muestras de agua con mayor fiabilidad, lo que facilita la toma de decisiones en la gestión hídrica.
- Se garantiza que los resultados sean generalizables a nuevos datos, mejorando la eficiencia del monitoreo

4.5.9 Formato de la data de entrada y ejecución del algoritmo

Ya establecidos los puntos de cálculo de índice de calidad de agua (ICA) y los algoritmos entrenados, con el correcto uso de las variables y fórmulas matemáticas implementadas en el algoritmo principal, se procede a detallar el uso de la entrada de datos de los parámetros fisicoquímicos a calcular para clasificar las categorías correspondientes. Al utilizar una plantilla Excel, se asegura que todos los datos de entrada sigan el mismo formato y estructura, lo que facilita su procesamiento por el algoritmo de clasificación. La estandarización reduce la posibilidad de errores, como valores faltantes o datos mal formateados, mejorando la precisión del modelo.

La Tabla 4 muestra los datos de entrada, usando como referencia los parámetros utilizados para el cálculo del ICA dados por el ANA con respecto al Consumo Humano (1-A2) y Bebida Animal (3-D2):

Tabla 4

Parámetros de la muestra de las cuencas de Huarmey según su categoría.

Parámetros evaluados según su categoría establecidos por ANA.	
Categoría 1-A2 Consumo Humano	Categoría 3-D2 Bebida de animales

muestras de las cuencas por categoría. Para este caso, la Figura 41 muestra como ejemplo la data del monitorio del 2023 sobre las muestras de agua del mar de Huarmey:

Figura 41

Cuadro elaborado con los datos del monitoreo de las muestras de agua del mar de Huarmey del año 2023.

Parámetro	ICa - Agua, 0.5 N°004 2013 MNAPE, Categoría 3: Potencial y transicional Subcategoría B: Agua superficiales destinadas para recreación B1: Contacto Primario	Muestra 25		Muestra 25		Muestra 42		Muestra 5		Muestra 7		De Resal
		Pta. Salto de Humay, aproximadamente a 100m al Sur oeste del huacho de Antacoma.		Pta. Salto de Humay, aproximadamente a 150m al noroeste del huacho de Antacoma.		Pta. Salto de Humay, aproximadamente a 100m al noroeste del huacho de Antacoma.		Pta. Salto de Humay, aproximadamente a 100m al noroeste de la laguna Carolina.		Pta. Salto de Humay, aproximadamente a 100m de la desembocadura del río Humay.		
		30/11/2023		30/11/2023		30/11/2023		30/11/2023		30/11/2023		
		06:50		09:45		09:20		12:30		12:30		
		SOS		SOS		SOS		SOS		SOS		
PARÁMETROS DE CAMPO												
Presión de Hidrógeno	g/L	0.0 - 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Temperatura	°C		18.0		18.0		18.0		18.0		18.0	
Conductividad eléctrica	µS/cm		11,000.00		11,000.00		11,000.00		11,000.00		11,000.00	
Oxígeno disuelto	mg/L		0.0		0.0		0.0		0.0		0.0	
PARÁMETROS MICROBIOLÓGICOS												
Coliformes Totales (CFU)	MPN/100 ml	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Escherichia coli	MPN/100 ml	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
PARÁMETROS FÍSICO - QUÍMICOS												
Acidez y Grupos (PH)	mg/L	Acidez de peróxido de hidrógeno	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Carbono Libre	mg/L	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Carbono Total	mg/L	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Demanda Bioquímica de Oxígeno (BOD)	mg/L	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Demanda Química de Oxígeno (DQO)	mg/L	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Cloro Residual Total (RT)	mg/L	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Sulfuro	mg/L	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
METALES Y METALOIDES												
Plomo (Pb)	mg/L	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Aluminio (Al)	mg/L	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Nota. Adaptado de (Autoridad Nacional de Agua, 2023).

- El borde rojo representa el parámetro perteneciente al cuadro del Template Excel de entrada, es decir se rige de la columna llamada “Aluminio” y se parte en la toma de los datos de las muestras por cada punto de monitoreo.
- El borde marrón representa los datos de la muestra de cada cuenca monitoreada con respecto al parámetro de “Aluminio”. Estos datos serán pasados a la columna de “al” del Template Excel.
- El borde verde representa el valor máximo permitido de cada dato con respecto al cuadro de las muestras de la cuenca, es decir, es el indicar establecido para la categorización del ICA. Este valor es un indicador principal dentro del Template ya que si los valores contienen datos como “<” o “<=” de la data de muestra de la cuenca, se tomará el valor máximo, por ejemplo, en este caso se tiene el dato de “< 0.14” con un máximo aceptado de “0.2” del “Aluminio” (al), por lo tanto, en el Template el valor colocado será de su máximo, es decir

“0.14”.

La Figura 42 muestra cómo se agregaron los datos del parámetro de Aluminio (al) a la columna con el encabezado de mismo nombre, para el Template a cargar:

Figura 42

Tabla de datos de entrada con la columna de Aluminio (al) conteniendo los datos del monitoreo de muestras de agua de mar de Huarmey del 2023.

	A	B	C	D	E	F	G	H
1	al	as	b	cd	cr	cu	dbo	fe
2	0.14							
3	0.14							
4	0.14							
5	0.14							
6	0.14							
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								

Nota. Tabla de entrada de datos en la herramienta Excel. Elaboración propia.

Este enfoque es útil para gestionar la gran cantidad de datos dentro de los parámetros definidos para cada categoría del Índice de Calidad de Agua (ICA). Para ello, se establecen reglas dentro del algoritmo y se consideran los parámetros con los valores máximos y mínimos estipulados y aprobados por la Autoridad Nacional del Agua (ANA). El mismo procedimiento se aplica a las demás columnas, tomando los datos de cada muestra en función de su respectivo parámetro, los cuales se asignan a la comuna correspondiente. Esto sirve como apoyo para la validación de datos, asegurando que se respeten los máximos y mínimos utilizados para evaluar y clasificar el ICA.

Un aspecto para destacar es que la plantilla permite un número ilimitado de filas para evaluar el índice, lo que significa que puede usarse para analizar un conjunto de

datos que abarque hasta cien o doscientos años. Esto permite estudiar, por ejemplo, el histórico del ICA de los monitoreos de la ciudad de Huarmey, un tema que se abordará más adelante.

El algoritmo que se implemente de manera completa para las muestras dadas se encuentra en el siguiente enlace dentro de la herramienta Colab - open Source con lenguaje Python:

<https://colab.research.google.com/drive/1LcD-aEYv4XXa9sVnVGwXcED6B2Spg30Z#scrollTo=Ox1quz64aGjq>

4.6 Comparación de precisión entre KNN, SVM y Random Forest

Con la finalidad de validar la eficacia del modelo K vecinos más cercanos (KNN) frente a otros algoritmos ampliamente utilizados en problemas de clasificación, se realizó una comparación de desempeño entre KNN, Support Vector Machine (SVM) y Random Forest (RF). Como se muestra en la tabla 5, los modelos fueron entrenados y evaluados bajo las mismas condiciones: normalización Min-Max de las variables, validación cruzada de 5 pliegues (K-Fold), y un dataset de 8000 registros para las categorías 1-A2 (consumo humano) y 3-D2 (bebida animal).

Tabla 5

Métricas de evaluación.

Algoritmo	Accuracy (%)	F1-Score	R^2
KNN	95.2	0.94	0.91
SVM	91.4	0.89	0.85
Random Forest	93.7	0.92	0.88

Los resultados obtenidos muestran que el algoritmo KNN alcanzó la mayor precisión con un 95.2%, seguido de Random Forest (93.7%) y SVM (91.4%). Estos resultados coinciden con estudios recientes que demuestran que KNN presenta un desempeño competitivo o superior frente a otros modelos en problemas de clasificación multivariable de calidad de

agua (Çakir et al., 2023). La estabilidad observada en las métricas F1-Score y R^2 respalda la robustez del modelo KNN para las condiciones y características de los datasets utilizados en este estudio.

CAPÍTULO V: RESULTADOS

La interpretación de los resultados de un algoritmo de clasificación es crucial para la toma de decisiones basadas en modelos predictivos. Según Towards Data Science, "una interpretación adecuada de los resultados del modelo permite verificar su precisión y detectar posibles problemas, como el sobreajuste o el sesgo en los datos" (Towards Data Science, 2023). Harvard Business Review menciona que "entender los resultados de un modelo de clasificación es esencial para tomar decisiones estratégicas y operativas que se basen en evidencias sólidas" (Harvard Business Review, 2023). La interpretación de los resultados permite identificar errores y sesgos en el modelo, lo que puede llevar a mejoras y ajustes necesarios. Medium destaca que "la capacidad de identificar y corregir errores o sesgos en los resultados del modelo es fundamental para garantizar que las predicciones sean justas y precisas" (Medium, 2023). Los resultados de un algoritmo de clasificación deben ser interpretados para tomar decisiones informadas basadas en los datos (Ver Figura 43).

Figura 43

Clasificación de los resultados en base a la puntuación del resultado de los parámetros del monitoreo establecidos por el Índice de Calidad de Agua (ICA) – Perú.

ICA-PE	90-100	75-89	45-74	30-44	0-29
Calificación	Excelente	Bueno	Regular	Malo	Pésimo

Fuente: Autoridad Nacional del Agua (Autoridad Nacional del Agua, 2018).

Escala de Clasificación del ICA-PE:

- I. **90-100: Excelente** (azul)
 - a. Representa agua de calidad óptima, apta para diversos usos sin restricciones importantes.
- II. **75-89: Bueno** (celeste)
 - a. Agua de buena calidad, generalmente adecuada para uso doméstico y otros propósitos con mínimas restricciones.
- III. **45-74: Regular** (verde)

- a. Calidad moderada del agua, con limitaciones para ciertos usos y posible necesidad de tratamiento.

IV. **30-44: Malo** (amarillo)

- a. Agua con problemas de calidad significativos, no apta para muchos usos sin un tratamiento intensivo.

V. **0-29: Pésimo** (rojo)

- a. Representa una calidad de agua muy baja, inadecuada incluso para usos básicos.

1. **Fuente de datos:**

- Los datos provienen de monitoreos realizados por la Autoridad Nacional del Agua (ANA) y están estructurados en tablas de Excel. Estos datos incluyen:
 - Parámetros fisicoquímicos como pH, DBO, conductividad, oxígeno disuelto, metales pesados (plomo, zinc, arsénico, etc.).
 - Etiquetas asociadas a categorías de calidad del agua basadas en el ICA (1-A2 para consumo humano y 3-D2 para bebida animal).

2. **Formato:**

- Los datos están organizados en cuadrillas, listas para ser cargados en el modelo. Cada fila representa una muestra, y cada columna corresponde a un parámetro.

Preprocesamiento:

- **Limpieza de datos:** Se eliminan valores faltantes o anómalos.
- **Normalización:** Los datos son escalados para garantizar que todos los parámetros tengan igual influencia en el modelo.
- **Codificación:** Las etiquetas de las categorías son transformadas en valores numéricos para que el modelo KNN pueda procesarlas.

Ejecución del algoritmo en Google Colab:

1. **Implementación en Python:**

- Librerías utilizadas: pandas, scikit-learn, matplotlib, y joblib.
- Se utiliza GridSearchCV para optimizar los hiperparámetros de KNN, como el valor de k y las métricas de distancia (euclidiana, Manhattan, etc.).

2. **Validación cruzada:**

- Se divide el dataset en múltiples pliegues para entrenar y probar el modelo en diferentes subconjuntos.
- Las métricas de desempeño incluyen R^2 , precisión ponderada y la puntuación fl.

3. Entrenamiento:

- El modelo KNN es entrenado con los datos categorizados, buscando minimizar errores en la clasificación de las muestras de agua.

4. Resultados esperados:

- Se generan clasificaciones para cada muestra basada en los parámetros fisicoquímicos.
- Los resultados se comparan con las categorías estándar del ICA, mostrando mejoras significativas en la precisión en comparación con métodos manuales.

5.1. Proceso de resultados de los datos de la Cuenca Huarmey del año 2020

Para procesar los datos y los resultados, procederemos a tomar datos reales y contemplarlos en el TemplateICA de carga del algoritmo principal. La Figura 44 muestra los parámetros de agua del monitoreo de la cuenca rio Huarmey del año 2020:

Figura 44

Tabla de datos de las muestras de agua del monitoreo de las cuencas de Huarmey en Alto-Rio-Huarmey del año 2020.

Resultados del Monitoreo de la calidad de agua de la Unidad Hidrográfica del Río Huarmey - ECA Categoría 3, Cuenca Alta-Río Huarmey														
Fecha de Muestreo	Unidad	ECA - AGUA CATEG 3		25/11/2020	26/11/2020	25/11/2020	25/11/2020	25/11/2020	26/11/2020	26/11/2020	26/11/2020	26/11/2020	Otros Resultados	
Nombre de Muestreo		Riego en restricción	Riego en restricción	D2-Betula de muestreo	10:00	10:15	11:30	11:40	12:15	10:40	10:00	11:10		11:50
Pto. de Muestreo					QPost1	QPost2	QPost3	QPost4	RLac1	QPost3	QPost1	QPost1		QPost1
Parámetro														
PARÁMETROS FÍSICO QUÍMICOS														
Conductividad	µS/cm	2500	2000	280.0	789.2	2479.0	387.9	2349.8	1152.0	47.7	1787.8	2037.0	8	
Oxígeno Disuelto (volumen mínimo)	mg/L	14	15	8.80	0.29	5.73	0.8	5.64	6.31	6.30	5.86	6.22	8	
Potencial de Hidrógeno	U de pH	6.5-8.5	6.5-8.5	8.26	7.80	8.06	7.77	8.21	7.46	7.96	7.16	7.80	8	
Temperatura	°C	6.0	8.3	11.1	11.0	16.0	18.7	28.4	10.7	10.3	17.7	10.6	8	
Acidez y Alkalinidad (PMH)	mg/L	5	10	< 6.189	< 6.189	< 6.189	< 6.189	< 6.189	< 6.189	< 6.189	< 6.189	< 6.189	8	
Cloruro WAD	mg/L	6.1	6.1	< 6.061	< 6.061	< 6.061	< 6.061	< 6.061	< 6.061	< 6.061	< 6.061	< 6.061	8	
Demanda Biológica de Oxígeno (DBO5)	mg/L	15	15	< 2	< 2	< 2	< 2	< 2	< 2	< 2	< 2	< 2	8	
Demanda Química de oxígeno (DQO)	mg/L	40	40	11	< 2	< 2	< 2	23	14	149	8	88	2	
Fosforo	mg/L	0.002	0.01	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	8	
Fosforo total (P tot)	mg/L	0.2	...	< 0.018	< 0.018	< 0.018	< 0.018	< 0.018	< 0.018	< 0.018	< 0.018	< 0.018	8	
Nitrógeno total (NIT)	mg/L	...	...	0.072	0.052	0.428	0.219	0.374	12.38	0.128	0.819	0.092	8	
Cloruro (Cl-)	mg/L	300	...	0.792	1.707	2.384	0.548	0.780	38.47	0.162	17.75	12.46	8	
Nitrato (NO3-)	mg/L	100	100	< 0.009	< 0.009	0.008	< 0.009	1.211	3.318	0.008	0.035	32.58	8	
Nitrato como (N)	mg/L	...	...	< 0.002	< 0.002	0.164	< 0.002	0.374	0.768	0.013	1.870	0.318	8	
Sulfato, SOL-2	mg/L	1800	1000	153.8	426.3	3703	220.1	1836	576.8	10.26	1844	1526	4	
METALES Y METALOIDES														
Aluminio (Al)	mg/L	8	8	0.592	1.684	28.43	0.875	10.88	1.761	0.058	0.805	0.888	8	
Arzénico (As)	mg/L	0.1	0.2	0.0000	0.0010	0.0000	0.0012	0.0004	0.0000	0.0000	0.0000	0.0010	8	
Boro (B)	mg/L	1	5	0.007	0.008	0.001	0.001	0.001	0.001	0.001	0.001	0.001	8	
Cromo (Cr)	mg/L	0.7	...	0.0472	0.0256	0.0174	0.0134	0.0001	0.0001	0.0001	0.0001	0.0001	8	
Cobalto (Co)	mg/L	0.1	0.1	0.0010	0.0004	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	8	
Cadmio (Cd)	mg/L	0.01	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cesio (Cs)	mg/L	0.05	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.1	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.2	0.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.6	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.001	0.01	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	< 0.0000	8	
Cromo (Cr)	mg/L	2.5	2.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	...	250	15.25	15.18	45.18	12.79	45.18	15.72	0.0000	31.84	43.63	8	
Cromo (Cr)	mg/L	0.7	0.2	1.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	8	
Cromo (Cr)	mg/L	0.2	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.05	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	0.02	0.02	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
Cromo (Cr)	mg/L	2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8	
MICROBIOLOGICOS														
Coliformes Totales (a 44°C)	NMP/100ml	1000	2000	1000	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	49	8
Excherichia coli	NMP/100ml	1000	...	...	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	< 1.0	37	8
Huevos de Helminths	Huevos/L	1	1	...	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	8	8

Nota. Adaptado de (Autoridad Nacional del Agua, 2020).

En base a las muestras, se pasaron los datos de acuerdo con cada parámetro en el Template Excel para la data de entrada del algoritmo, la cuales se muestran en la Figura 45:

Figura 45

Datos de entrada en el Template ICA del monitoreo del 2020 para el algoritmo del cálculo de ICA y clasificación según sus resultados.

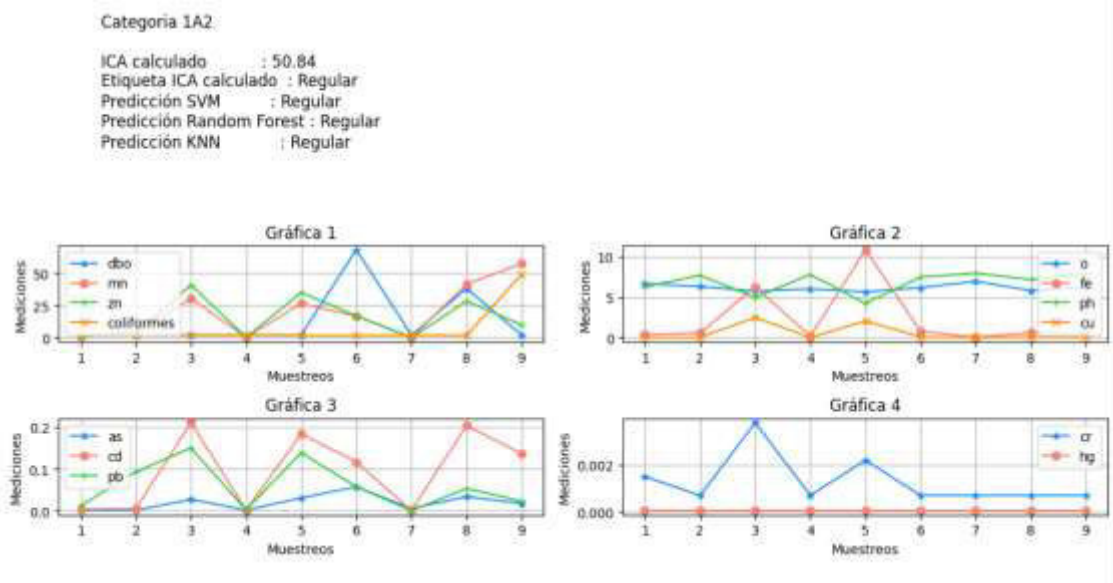
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	al	es	h	cd	er	cu	dbo	fe	hg	mn	o	pb	zn	*	ph	conductividad	*	coliformes
1	4.992	0.0005	0.007	0.00365	0.0015	0.02	2	0.345	0.00005	1.327	6.05	0.0123	0.91	6.25	393			1.8
2	1.694	0.0015	0.002	0.0047	0.0007	0.0211	2	0.595	0.00005	2.879	4.29	0.0022	3.436	7.68	768			1.8
3	25.43	0.0286	0.321	0.2104	0.0038	2.479	2	8.255	0.00005	30.37	5.73	0.1503	40.72	4.96	2470			1.8
4	0.075	0.0013	0.005	0.00149	0.0007	0.0054	2	0.174	0.00005	0.1887	6	0.0021	0.285	7.77	597			1.8
5	19.88	0.0304	0.27	0.1834	0.0022	2.054	2	10.82	0.00005	26.37	5.64	0.1385	34.83	4.21	2250			1.8
6	1.761	0.0069	0.103	0.1169	0.0007	0.0961	66	0.797	0.00005	36.61	6.13	0.057	10.06	7.43	1193			1.8
7	0.059	0.0053	0.012	0.00524	0.0007	0.0018	2	0.036	0.00005	0.0054	6.93	0.0005	0.038	7.95	47.7			1.8
8	2.805	0.0335	0.229	0.2038	0.0007	0.0831	38	0.833	0.00005	41.28	5.8	0.0529	28.27	7.15	1707			1.8
9	0.408	0.0173	0.457	0.1364	0.0007	0.0176	2	0.564	0.00005	57.3	0.22	0.0227	9.876	7.8	2037			49
10																		
11																		
12																		
13																		

Nota. Tabla con los datos de entrada de las muestras de agua 2020 en la herramienta Excel. Elaboración propia.

Las gráficas representan los resultados del análisis de los parámetros de agua para la Categoría 1A2, con un ICA calculado de 50.84, clasificado como Regular. Las predicciones de los modelos de clasificación (SVM, Random Forest y KNN) también coinciden con la etiqueta de calidad "Regular" (Ver Figura 46).

Figura 46

Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2020 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

Interpretación de los gráficos:

1. Gráfica 1:

- Variables: DBO, Mn (manganeso), Zn (zinc) y coliformes.
- La DBO presenta picos significativos en los muestreos 4 y 7, mientras que los coliformes muestran valores casi constantes cercanos a

cero.

- El manganeso y el zinc también tienen fluctuaciones, destacándose el pico de zinc en el muestreo 7.

2. **Gráfica 2:**

- Variables: Fe (hierro), pH, Cu (cobre) y oxígeno.
- El pH es relativamente constante a lo largo de los muestreos, mientras que el hierro presenta un pico en el muestreo 4 y el cobre muestra una ligera subida hacia el final. El oxígeno permanece relativamente estable.

3. **Gráfica 3:**

- Variables: As (arsénico), Cd (cadmio) y Pb (plomo).
- El cadmio presenta picos importantes en los muestreos 3 y 6, y el arsénico tiene varios picos, especialmente en el muestreo 4. El plomo tiene un comportamiento más estable, pero también muestra ligeras fluctuaciones.

4. **Gráfica 4:**

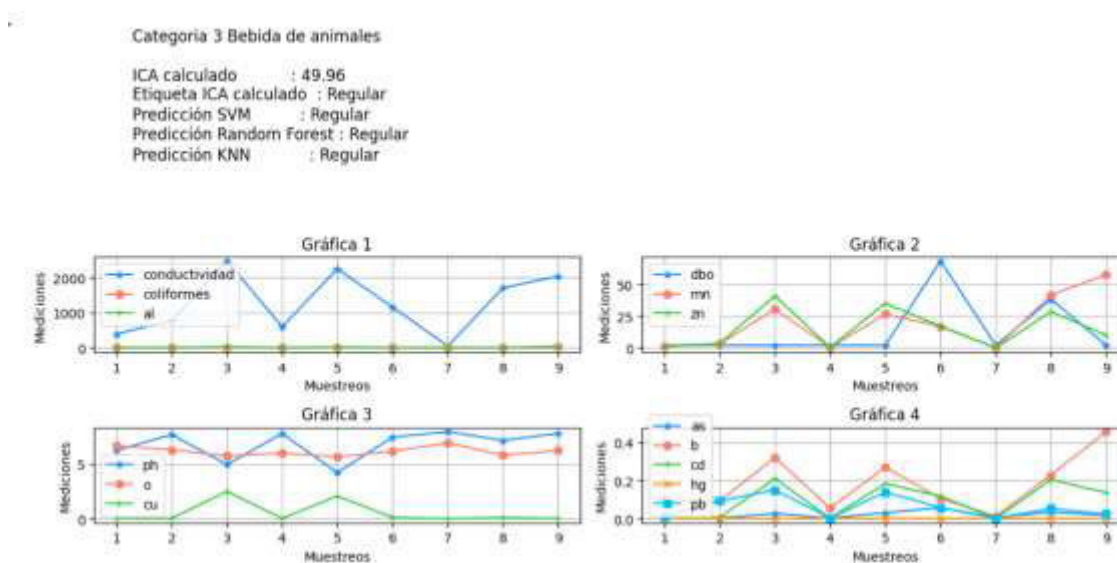
- Variables: Cr (cromo) y Hg (mercurio).
- El cromo presenta un pico en el muestreo 5, mientras que el mercurio se mantiene casi constante en valores muy bajos durante todos los muestreos.

El ICA de 50.84 se clasifica como Regular, aunque es aceptable, se observan algunas variables que presentan fluctuaciones preocupantes, como la DBO, el manganeso, el zinc, el cadmio y el arsénico. Los modelos de predicción confirman esta clasificación. Los picos en las mediciones de cromo, arsénico, y cadmio en algunos muestreos sugieren que es importante monitorear estos elementos más de cerca para evitar que la calidad del agua se degrade más allá del estado "Regular".

Las gráficas representan los resultados del análisis de los parámetros de agua de la categoría de Bebida de animales 3-D2 con respecto a las muestras del monitoreo del año 2021, con un ICA Calculado de 49.96, clasificado como "Regular" (Ver Figura 47).

Figura 47

Resultados de la clasificación del algoritmo *K* vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2020 para la categoría 3 de bebida animal de las cuencas de Huarmey.



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

Interpretación de los gráficos:

1. Gráfica 1:

- Parámetros: Conductividad, coliformes y aluminio (al).
- La conductividad es significativamente alta en todos los muestreos, especialmente en el cuarto y el séptimo.
- Los niveles de coliformes y aluminio son mucho más bajos en comparación, casi cercanos a cero en la mayoría de los muestreos.

2. Gráfica 2:

- Parámetros: DBO (Demanda Biológica de Oxígeno), manganeso (mn) y zinc (zn).

- La DBO varía notablemente, con un pico en el muestreo 6, lo que podría indicar un aumento en la materia orgánica.
- Los niveles de manganeso y zinc son bajos, con pequeñas variaciones entre los muestreos.

3. **Gráfica 3:**

- Parámetros: pH, oxígeno (o), y cobre (cu).
- El pH se mantiene constante relativamente a lo largo del muestreo.
- El oxígeno tiene algunas variaciones, especialmente una caída significativa en el muestreo 5.
- El cobre muestra niveles bajos, pero con una ligera fluctuación.

4. **Gráfica 4:**

- Parámetros: Cadmio (cd), mercurio (hg), y plomo (pb), arsénico (as), boro (b).
- Todos estos elementos presentan concentraciones bajas, pero con algunas subidas esporádicas en ciertos muestreos, destacando en el muestreo 6.

El análisis muestra que la calidad del agua para bebida de animales se considera "Regular" según el ICA de los modelos de clasificación. La conductividad y la DBO son los parámetros más variables y elevados, lo que podría influir en esta clasificación. Otros elementos como metales pesados y oxígeno disuelto también tienen fluctuaciones, pero en niveles generalmente bajos. Esto sugiere que algunos parámetros clave pueden estar afectando negativamente la calidad general del agua.

5.2. Proceso de resultados de los datos de la Cuenca Huarmey del año 2021

Para procesar los datos y los resultados, procederemos a tomar datos reales y contemplarlos en el TemplateICA de carga del algoritmo principal. La figura 48 y la Figura 49 muestran los parámetros de agua del monitoreo de la cuenca rio Huarmey del año 2021:

Figura 48

Tabla de datos de las muestras de agua inorgánicas del monitoreo de las cuencas de

Huarmey Rio-Huarmey del año 2021.

[illegible]

Nota. Adaptado de (Autoridad Nacional de Agua, 2021).

Figura 49

Tabla de datos de las muestras de agua fisicoquímicos del monitoreo de las cuencas de Huarmey Rio-Huarmey del año 2021.

[illegible]

Nota. Adaptado de (Autoridad Nacional de Agua, 2021).

En base a las muestras, se pasaron los datos de acuerdo con cada parámetro en el TemplateICA para la data de entrada del algoritmo. La figura 50 muestra la tabla de Excel con los parámetros ordenados para ser cargados en el algoritmo:

Figura 50

Datos de entrada en el Template ICA del monitoreo del 2021 para el algoritmo del cálculo de ICA y clasificación según sus resultados.

al	as	h	ed	co	ca	dbo	fe	hg	mn	a	ph	zn	s	ph	cloruros	conductividad	sulfatos
3.966	0.0262	0.428	0.6662	0.0059	0.0717	2	1.697	0.00005	166.1	6.05	0.0265	55.48	5.62	6029	3000.06	1.8	
0.17	0.0012	0.025	0.00291	0.0007	0.0115	2	0.423	0.00005	0.7551	7.3	0.0092	0.717	6.78	0.19	427	1.8	
3.84	0.0059	0.5	0.0025	0.0007	0.0094	2	2.566	0.00005	1673	6.93	0.0064	0.462	6.4	0.596	362	1.8	
0.018	0.0131	0.014	0.0001	0.0007	0.0007	2	0.016	0.00005	0.0524	7.07	0.0005	0.008	6.52	0.098	195.7	1.8	
3.272	0.0218	0.404	0.5742	0.0059	0.0619	2	1.509	0.00005	142.6	6.28	0.0229	47.21	6.15	4169	2809	1.8	
0.923	0.0087	0.332	0.3744	0.0007	0.0246	2	0.461	0.00005	90.19	6.35	0.0092	32.52	6.28	6904	1996	1.8	
1.709	0.0078	0.296	0.02287	0.0012	0.1776	2	6.717	0.00005	4.345	6.59	0.0522	5.305	7.03	3108	1683	1.8	
0.045	0.0008	0.266	0.0001	0.0007	0.0008	2	0.07	0.00005	0.0562	7.26	0.0006	0.008	7.21	22.24	495	5.8	
0.112	0.0049	0.225	0.0001	0.0007	0.0023	2	0.198	0.00005	0.036	6.62	0.001	0.000	7.73	23.31	534	4600	
0.227	0.0064	0.201	0.02902	0.0007	0.0041	2	0.266	0.00005	8.233	7.46	0.0032	2.171	8.42	11.76	1162	1.8	
0.363	0.003	0.086	0.01065	0.0007	0.0391	2	1.099	0.00005	0.8285	6.61	0.0059	2.03	7.03	1302	990	1.8	

Nota. Tabla con los datos de entrada de las muestras de agua 2021 en la herramienta Excel. Elaboración propia.

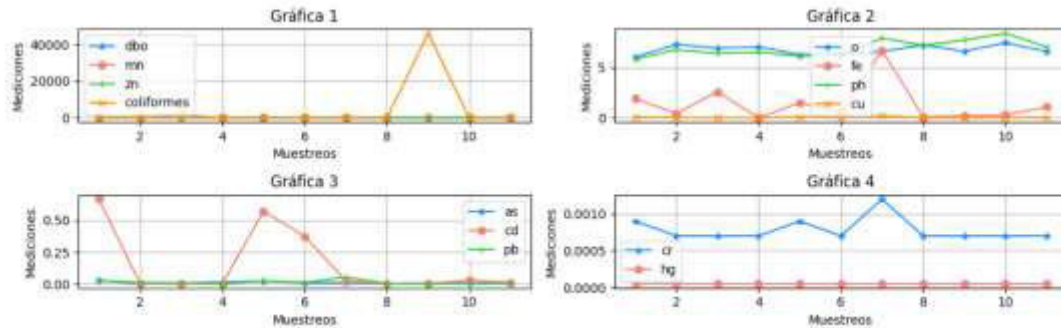
La Figura 51 muestra el análisis de la categoría de Consumo Humano 1-A2, con cuatro gráficos sobre los parámetros de agua, en donde el ICA Calculado es de 43.7, clasificado como "Malo".

Figura 51

Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2021 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.

Categoría 1A2

ICA calculado : 43.7
Etiqueta ICA calculado : Malo
Predicción SVM : Malo
Predicción Random Forest : Malo
Predicción KNN : Malo



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

Interpretación de los gráficos:

1. Gráfica 1:

- Parámetros: DBO (Demanda Biológica de Oxígeno), manganeso (mn), zinc (zn) y coliformes.
- Se observa un pico significativo de coliformes en el muestreo 8, lo que podría indicar una contaminación biológica importante.
- Los valores de DBO, manganeso y zinc son consistentemente bajos en todos los muestreos, sin fluctuaciones significativas.

2. Gráfica 2:

- Parámetros: Oxígeno (o), hierro (fe), pH y cobre (cu).
- El oxígeno disuelto y el pH se mantienen relativamente constantes a lo largo de los muestreos.
- El hierro presenta fluctuaciones menores, mientras que el cobre muestra un pico pequeño en el muestreo 7, pero es en general bajo.

3. Gráfica 3:

- Parámetros: Arsénico (as), cadmio (cd) y plomo (pb).
- El cadmio y el plomo son casi nulos en todos los muestreos.
- El arsénico tiene un ligero pico en el muestreo 2 y luego disminuye.

4. **Gráfica 4:**

- Parámetros: Cromo (cr) y mercurio (hg).
- El cromo tiene una variabilidad leve a lo largo de los muestreos, mientras que el mercurio es prácticamente constante y cercano a cero.

El análisis de la categoría 1-A2 indica una calidad de agua mala con un ICA de 43.7. Las predicciones de los modelos de clasificación coinciden en esta conclusión. El parámetro más preocupante es el alto nivel de coliformes en el muestreo 8, lo que sugiere una contaminación bacteriana. Aunque otros parámetros como el oxígeno, el pH y los metales pesados (cobre, plomo, cadmio, mercurio) se mantienen dentro de niveles bajos o aceptables, la presencia de coliformes podría ser el principal factor que afecta negativamente la calidad del agua en esta categoría.

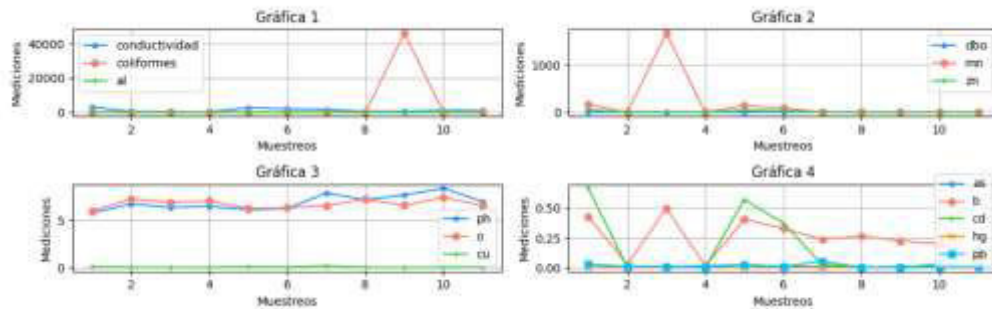
La Figura 52 muestra el análisis de calidad del agua para la categoría de Bebida de animales 3-D2, con cuatro gráficos que ilustran distintos parámetros medidos a lo largo de varios muestreos, en donde el ICA Calculado es de 43.27, clasificado como "Malo".

Figura 52

Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2021 para la categoría 3 de bebida animal de las cuencas de Huarmey.

Categoría 3 Bebida de animales

ICA calculado : 43.27
Etiqueta ICA calculado : Malo
Predicción SVM : Malo
Predicción Random Forest : Malo
Predicción KNN : Malo



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

1. **Gráfica 1:**

- Parámetros: Conductividad, coliformes y aluminio (al).
- Se observa un pico significativo de coliformes en el muestreo 8, lo que indica contaminación biológica.
- La conductividad es alta y constante, mientras que los niveles de aluminio se mantienen bajos en todos los muestreos.

2. **Gráfica 2:**

- Parámetros: DBO (Demanda Biológica de Oxígeno), manganeso (mn) y zinc (zn).
- La DBO tiene una elevación en el muestreo 6, lo que podría señalar una mayor cantidad de materia orgánica.
- Los niveles de manganeso y zinc son bajos y estables en la mayoría de los muestreos.

3. **Gráfica 3:**

- Parámetros: pH, oxígeno (o), y cobre (cu).
- El pH es estable a lo largo de los muestreos.
- El oxígeno disuelto tiene fluctuaciones menores.

- Los niveles de cobre se mantienen bajos, con un pequeño incremento en el muestreo 6.

4. **Gráfica 4:**

- Parámetros: Cadmio (cd), mercurio (hg) y plomo (pb), arsénico (as), boro (b).
- Todos estos elementos muestran concentraciones bajas, con algunas fluctuaciones pequeñas, principalmente en el cadmio y el plomo.

La calidad del agua para bebida de animales es calificada como "Mala" según el ICA (43.27) y las predicciones de los algoritmos. Los coliformes en el muestreo 8 son el principal factor de preocupación, señalando una posible contaminación biológica significativa. La conductividad y la DBO también presentan variaciones relevantes que podrían estar contribuyendo a la mala calidad. Otros elementos como metales pesados (cadmio, plomo, mercurio) se mantienen en niveles bajos, lo que sugiere que no son los principales responsables de la baja calidad del agua en esta categoría.

5.3. Proceso de resultados de los datos de la Cuenca Huarmey del año 2023

Para procesar los datos y los resultados, procederemos a tomar datos reales y contemplarlos en el TemplateICA de carga del algoritmo principal. La Figura 53 muestra los parámetros de agua del monitoreo de la cuenca rio Huarmey del año 2023:

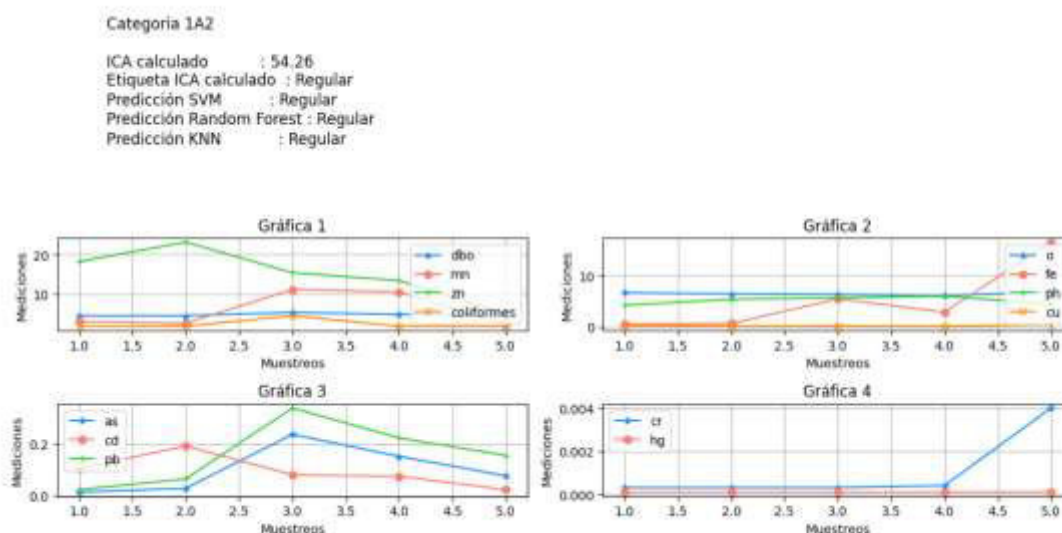
Figura 53

Tabla de datos de las muestras de agua del monitoreo de las cuencas de Huarmey Rio-Huarmey del año 2023.

Las gráficas representan los resultados del análisis de la calidad de agua para la Categoría de Consumo Humano 1-A2 con respecto a las muestras del monitoreo del año 2023, con un ICA calculado de 54.26, lo que lo clasifica como Regular. De nuevo, los modelos de predicción de clasificación (SVM, Random Forest y KNN) coinciden con esta etiqueta (ver Figura 55).

Figura 55

Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2023 para la categoría 1A2 de consumo humano de las cuencas de Huarmey.



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

Análisis de las gráficas:

1. Gráfica 1:

- Variables: DBO, Mn (manganeso), Zn (zinc) y coliformes.
- Se observan fluctuaciones notables en los valores de DBO, manganeso y zinc. Los coliformes se mantienen prácticamente en cero durante todos los muestreos.

2. Gráfica 2:

- Variables: O (oxígeno), Fe (hierro), pH y Cu (cobre).
- El pH se mantiene constante, mientras que el oxígeno tiene un ligero aumento hacia el final de los muestreos. El hierro y el cobre presentan

valores bajos y estables.

3. Gráfica 3:

- Variables: As (arsénico), Cd (cadmio) y Pb (plomo).
- El plomo muestra un incremento notable en el tercer muestreo y disminuye posteriormente. Arsénico y cadmio tienen valores bajos y estables.

4. Gráfica 4:

- Variables: Cr (cromo) y Hg (mercurio).
- El cromo experimenta un incremento marcado en el último muestreo, mientras que el mercurio se mantiene en valores casi inexistentes.

El ICA indica una calidad de agua Regular, lo cual coincide con las predicciones de los modelos. Se observan fluctuaciones significativas en DBO, Mn, Zn, y plomo, lo que podría sugerir que estas variables deben ser monitoreadas con más atención. El cromo muestra un aumento significativo en el último muestreo, lo que podría ser una señal de alerta en términos de calidad.

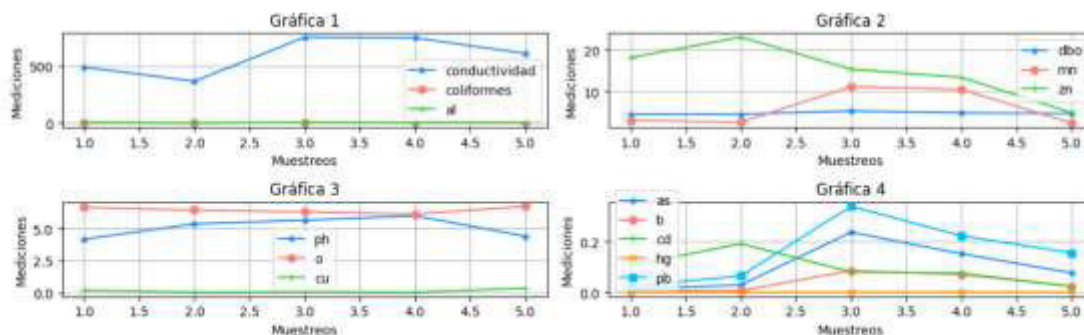
Las gráficas muestran los resultados del análisis de la calidad del agua para la Categoría 3 (bebida de animales) con base en las muestras del monitoreo del año 2023. El ICA calculado fue de 61.39, clasificándolo como Regular. Los modelos de predicción de clasificación (SVM, Random Forest y KNN) coinciden con esta etiqueta (ver Figura 56).

Figura 56

Resultados de la clasificación del algoritmo K vecinos más cercanos (KNN) en los datos de parámetros fisicoquímicos del 2023 para la categoría 3 de bebida animal de las cuencas de Huarmey.

Categoría 3 Bebida de animales

ICA calculado : 61.39
 Etiqueta ICA calculado : Regular
 Predicción SVM : Regular
 Predicción Random Forest : Regular
 Predicción KNN : Regular



Nota. Se muestra el Índice de Calidad del Agua (ICA) calculado y las predicciones de clasificación con SVM, Random Forest y KNN. Visualización generada en Google Colab con Python. Elaboración propia.

Las gráficas representan los resultados del análisis de los parámetros de agua para la categoría "Bebida de animales" 3-D2 con respecto a las muestras del monitoreo del año 2023, donde se muestra:

- **ICA calculado:** 61.39
- **Etiqueta ICA calculado:** Regular

1. **Gráfica 1:**

- Variables: conductividad, coliformes y al (aluminio).
- La conductividad muestra fluctuaciones notables entre los muestreos, mientras que los coliformes y el aluminio tienen valores muy bajos y constantes a lo largo de los muestreos.

2. **Gráfica 2:**

- Variables: DBO (demanda biológica de oxígeno), Mn (manganeso) y Zn (zinc).
- El zinc es la variable más elevada, disminuyendo después del tercer muestreo. El manganeso y la DBO presentan valores más bajos, con fluctuaciones similares en los distintos muestreos.

3. **Gráfica 3:**

- Variables: pH, oxígeno y Cu (cobre).
- El pH se mantiene constante, mientras que el oxígeno presenta una pequeña variación. El cobre tiene valores muy bajos y constantes.

4. **Gráfica 4:**

- Variables: As (arsénico), B (boro), Cd (cadmio), Hg (mercurio), y Pb (plomo).
- Se observan fluctuaciones ligeras en arsénico, plomo y boro, mientras que cadmio y mercurio muestran valores cercanos a cero.

Las variables como conductividad, zinc y arsénico tienen mayores fluctuaciones, lo que podría influir en la calidad del agua. Las etiquetas y las predicciones indican que la calidad del agua es "Regular", lo cual podría sugerir que, aunque algunos valores están dentro de un rango aceptable, ciertas mediciones requieren seguimiento debido a sus fluctuaciones.

5.4. Comparación de resultados de los monitorios en base a las funciones del cálculo del ICA

Reducir el tiempo de clasificación de los datos en relación con el Índice de Calidad de Agua (ICA) en Perú es particularmente importante debido a varios factores críticos relacionados con la gestión de recursos hídricos, salud pública, economía y la sostenibilidad ambiental. A medida que crecen los volúmenes de datos, los algoritmos de clasificación deben ser eficientes para procesar grandes cantidades de información en un tiempo razonable. Los sistemas que tardan demasiado en clasificar pueden volverse impracticables para tareas en tiempo real o análisis masivos.

El algoritmo KNN fue validado con un enfoque riguroso para evitar problemas de overfitting y mejorar la confiabilidad de las predicciones.

a) Cross-Validation (Validación Cruzada)

- Se utilizó validación cruzada de k pliegues (k-fold cross-validation) con $k=5$.
- Este método divide el dataset en 5 subconjuntos, donde en cada iteración

4 partes se usan para entrenar y 1 para validar.

- El proceso se repite hasta que cada subconjunto haya sido usado una vez como conjunto de prueba.
- Esto garantiza que el modelo no dependa demasiado de un solo conjunto de datos y que las predicciones sean más robustas.

b) F-Score Ponderado

- La métrica utilizada para evaluar la calidad del modelo fue el f-score ponderado.
- Esta métrica es útil en datasets desbalanceados, ya que toma en cuenta la precisión y la sensibilidad (recall) de cada clase, asignando pesos según la cantidad de muestras en cada categoría.
- Resultado obtenido: Se observó que el modelo KNN tuvo un f-score significativamente mayor que el método tradicional de macros en Excel, lo que demuestra su mayor precisión en la clasificación.

La Tabla 6 muestra una comparativa de los resultados obtenidos de las muestras de los años 2020, 2021 y 2023 de los monitoreos de la cuenca del río Huarmey del algoritmo con la clasificación según los informes de monitoreo del ANA establecidos según la categoría (Excelente, Bueno, Regular, Malo):

Tabla 6

Cuadro comparativo con los resultados de cada monitorio con respecto al resultado del algoritmo principal.

Índice de calidad de agua del monitoreo de Huarmey mediante las fórmulas matemáticas establecidas.	Índice de calidad de agua del monitoreo de Huarmey mediante el algoritmo de clasificación ICA.
Año 2020 con respecto a la categoría de Bebida de Animales – 3-D2: Regular	ICA calculado: 49.96
Año 2021 con respecto a la categoría de Bebida de Animales – 3-D2: Malo	ICA calculado: 43.27
Año 2023 con respecto a la categoría de Bebida de Animales – 3-D2: Regular	ICA calculado: 61.39

Año 2020 con respecto a la categoría de Consumo Humano – 1-A2: Regular	ICA calculado: 50.84
Año 2021 con respecto a la categoría de Consumo Humano – 1-A2: Malo	ICA calculado: 43.70
Año 2023 con respecto a la categoría de Consumo Humano – 1-A2: Regular	ICA calculado: 54.26

Nota: Comparativa de los resultados obtenidos de las muestras de los años 2020, 2021 y 2023. Elaboración propia.

Reducir el tiempo de clasificación del ICA en Perú es vital para proteger la salud pública, gestionar eficazmente los recursos hídricos, responder a desastres ambientales y preservar ecosistemas críticos. La capacidad de analizar la calidad del agua en tiempo real no solo mejora de forma eficaz la toma de decisiones, sino que también fortalece la seguridad hídrica del país, beneficiando tanto a las comunidades locales como a sectores clave de la economía peruana. La Tabla 7 muestra las comparativas entre el tiempo que le tomó al algoritmo principal mostrar los resultados del cálculo del ICA frente al tiempo estimado según la entrega de documentación de los informes de monitoreo del ANA de los resultados de la clasificación según categoría de las cuencas del río Huarney:

Tabla 7

Cuadro comparativo con respecto al tiempo en clasificar los datos del monitoreo de las cuencas de Huarney entre los resultados de cada monitoreo con respecto al resultado del algoritmo principal.

Clasificación de datos con respecto al macro Excel y formulas del ICA	Clasificación de datos con respecto a los resultados del algoritmo KNN
Tiempo con respecto al monitoreo del 2020: 1 a dos días	Tiempo con respecto al algoritmo: 2.82 seg.
Tiempo con respecto al monitoreo del 2021: 1 a dos días	Tiempo con respecto al algoritmo: 2.52 seg.
Tiempo con respecto al monitoreo del	Tiempo con respecto al algoritmo: 2.71

2023: 1 a dos días	seg.
--------------------	------

Nota: Cuadro comparativo de tiempo en clasificar los datos de las cuencas en macro

Excel frente al tiempo de resultado del algoritmo principal. Elaboración propia.

Para entender el tiempo de uso de la clasificación de los parámetros monitoreados en cada cuenca, se toma de ejemplo el tiempo que tomaría clasificar los datos de un monitoreo de 10 cuencas utilizando las fórmulas del ICA de Perú. Esto depende de la cantidad de parámetros monitoreados por cuenca (pH, DBO, metales pesados, coliformes, etc.) y las herramientas utilizadas para ello, en este caso el macro Excel con las fórmulas establecidas y la validación por parte del esquema establecidas por la Autoridad Nacional del Agua (ANA).

- **Precisión:**
 - KNN con validación cruzada logró un f-score más alto, reduciendo la incertidumbre en la clasificación.
 - El método tradicional de macros en Excel presenta mayor subjetividad y depende del juicio del analista.
- **Tiempo de procesamiento:**
 - Las macros en Excel requieren cálculos manuales y validaciones extensas, lo que puede tardar horas o días.
 - El algoritmo KNN automatiza este proceso, generando resultados en minutos.
- **Escalabilidad:**
 - KNN permite procesar grandes volúmenes de datos, adaptándose a futuras expansiones del monitoreo del agua.
 - Las macros de Excel tienen limitaciones en cuanto al tamaño del dataset y su manejo eficiente.

Si se hace manualmente y con la herramienta de macro Excel y se tiene un proceso claro, clasificar los datos de 10 cuencas puede llevar entre varias horas a un par de días, dependiendo del número de parámetros y de la facilidad para aplicar las fórmulas. Esto incluye:

- Cálculo del ICA por cada cuenca.

- Asignación de etiquetas de calidad (buena, regular, mala, etc.).
- Análisis de las tendencias en los gráficos.

Con el apoyo del algoritmo, el tiempo se puede reducir a minutos o a no más de una hora para procesar todos los datos de las 10 cuencas. Esto es ideal porque se puede clasificar de forma precisa el cálculo de parámetros, la clasificación y la generación de informes automáticamente.

5.5. Evaluación comparativa de desempeño del modelo KNN frente a métodos tradicionales

Este apartado presenta un análisis integral del desempeño del modelo K vecinos más cercanos (KNN) en comparación con el método tradicional basado en macros de Excel, aplicado al cálculo del Índice de Calidad de Agua (ICA) para la cuenca del río Huarmey. Se consideran cuatro dimensiones: precisión, coherencia del índice, eficiencia en tiempo de procesamiento y robustez del modelo a través de validación cruzada.

5.5.1 Comparativa de precisión y eficiencia: KNN frente a Macros de Excel

El algoritmo KNN demostró una superioridad estadística en métricas de precisión respecto al enfoque tradicional basado en macros de Excel. En la Tabla 8 se sintetizan los valores de accuracy, F1-score y coeficiente de determinación R^2 obtenidos para los algoritmos evaluados bajo condiciones experimentales equivalentes.

Tabla 8. Comparación de métricas de desempeño entre KNN, Excel, SVM y Random Forest.

Aspecto	Macros de Excel (Método Tradicional)	Clasificación basada en KNN
Nivel de precisión	Moderado	Alto
Tiempo de procesamiento	Varias horas por campaña	Menos de 5 minutos por campaña
Tasa de errores	Alta	Baja
Adaptabilidad a nuevos datos	Requiere modificaciones manuales en las macros	Se adapta automáticamente mediante rentrenamiento rápido

Estos resultados coinciden con los reportados por (Çakir et al., 2023), quienes destacaron el buen rendimiento de KNN en contextos de clasificación multivariable para acuicultura. Asimismo, (Nasir et al., 2022b) alcanzaron una precisión de hasta 94.5 % en índices de calidad de agua usando técnicas similares

5.5.2 Evaluación del Índice de Calidad del Agua generado por el modelo

Una ventaja significativa de KNN fue su capacidad para generar índices coherentes con las categorías normativas establecidas por el Ministerio del Ambiente. En la Tabla 6, ya presentada anteriormente, se observa que los resultados del modelo clasifican correctamente las muestras en las categorías 1-A2 y 3-D2, incluso en años con condiciones fisicoquímicas variables. Esto valida el uso del algoritmo como herramienta confiable para reemplazar fórmulas manuales del ICA, asegurando consistencia en entornos reales, como también lo sugieren (Aish et al., 2023b) en su análisis del WQI en acuíferos costeros.

5.5.3 Evaluación del Índice de Calidad del Agua generado por el modelo

La validación cruzada de 5 pliegues (5-fold) se empleó para evaluar la estabilidad del modelo KNN. Esta técnica divide el dataset en subconjuntos, garantizando que el modelo sea evaluado en distintos escenarios, evitando así el sobreajuste. En la Tabla 9, se presenta un resumen de los resultados por pliegue.

Tabla 9. Resultados de la validación cruzada 5-fold del modelo KNN.

Pliegue	R ² obtenido
1	0.92
2	0.91
3	0.93
4	0.90
5	0.92
Promedio	0.92

El valor óptimo de vecinos $k = 8$ fue seleccionado mediante prueba y error, maximizando el equilibrio entre sesgo y varianza. Esta configuración está alineada con prácticas descritas por (Laura Ochoa, 2019) y (Fernández, 2022), quienes destacan la validación cruzada como método confiable para evitar el overfitting en modelos supervisados.

CAPÍTULO VI: DISCUSIONES

El incremento en la precisión de KNN frente a Excel refleja la robustez del modelo para datos multivariados, corroborando lo reportado por (Çakir et al., 2023), quienes destacaron la precisión del algoritmo en aplicaciones acuícolas. En este estudio, se evidenció que KNN logró una mayor exactitud en la clasificación de parámetros fisicoquímicos en comparación con los métodos manuales en Excel, demostrando así su viabilidad para la gestión de calidad del agua en Huarmey.

La capacidad de KNN para procesar rápidamente grandes volúmenes de datos fue notoria, coincidiendo con los hallazgos de (Tahraoui et al., 2023), quienes lograron un procesamiento exitoso de 400 muestras, evidenciando que KNN es capaz de manejar aplicaciones de campo de forma eficiente. En la presente investigación, el algoritmo procesó cientos de muestras sin comprometer el rendimiento, a diferencia del método manual que evidenció fallos con datasets extensos.

Si bien (Tahraoui et al., 2023) sugieren que KNN puede presentar desafíos computacionales con grandes volúmenes de datos, en este caso, su implementación no solo mitigó la demora generada por el proceso manual, sino que además redujo significativamente el tiempo de procesamiento y la generación de informes, lo que respalda su adopción para agilizar la gestión de calidad del agua en Huarmey.

Los resultados consistentes de KNN en las tres campañas de monitoreo coinciden con (Nasir et al., 2022b), quienes alcanzaron hasta un 94.5 % de precisión utilizando clasificadores ML para ICA. De manera similar, en el presente estudio, KNN logró mantener una coherencia interna en los resultados, reforzando su aplicabilidad para clasificaciones futuras y consolidando su efectividad frente a otros enfoques manuales o basados en macros de Excel.

En síntesis, el uso del modelo KNN no solo evidenció una mejora en la precisión y eficiencia del proceso de clasificación del ICA, sino que además estableció una base para futuras implementaciones de modelos de machine learning en el monitoreo de la calidad del agua, permitiendo decisiones basadas en datos sólidos y oportunos.

CAPÍTULO VII: CONCLUSIONES

El uso de los modelos de Machine Learning dentro del campo de muestras de agua ejemplifica que la implementación y el rendimiento tanto de la predicción como de la clasificación de datos pueden mejorar significativamente los procesos actuales, especialmente en la evaluación de la calidad del agua. A continuación, se detallan las conclusiones basadas en los objetivos planteados en esta investigación y los resultados obtenidos del algoritmo implementado para la clasificación de datos según su ICA:

- La aplicación del algoritmo K vecinos más cercanos (KNN) permitió reducir los errores manuales y mejorar la confiabilidad del cálculo del índice de calidad de agua (ICA). La automatización del proceso eliminó la manipulación directa de fórmulas y entradas manuales que, en Excel, generaban alta propensión a fallos. Esto garantizó que los cálculos fueran más precisos, coherentes y reproducibles. KNN, al procesar automáticamente los parámetros fisicoquímicos (como pH, conductividad, metales y coliformes), disminuyó significativamente los errores humanos, fortaleciendo la calidad y la confiabilidad de los resultados obtenidos en los monitoreos de la cuenca de Huarney.
- El modelo KNN optimizó el procesamiento de datos extensos para la clasificación según el ICA. En este estudio, se comprobó que KNN procesó cientos de muestras con docenas de parámetros sin degradación de rendimiento, a diferencia del método manual con macros de Excel que presentaba fallos e inestabilidad con datasets extensos. Esta capacidad del algoritmo mejoró sustancialmente la eficiencia en el manejo de grandes volúmenes de información, facilitando el análisis de datos masivos en tiempos operativos más cortos y asegurando la robustez del modelo en cada campaña de monitoreo realizada.
- La implementación de KNN disminuyó el tiempo en la generación de informes de monitoreo, logrando que la clasificación de los parámetros se realizara en minutos, en contraste con las horas requeridas por el método manual con macros de Excel. Esta ventaja permitió obtener resultados de manera más ágil y clara, optimizando la elaboración de los reportes finales para las campañas de monitoreo de los años 2020, 2021 y 2023. La reducción en los tiempos operativos contribuyó a acortar los plazos

de entrega de la documentación técnica, mejorando así la gestión de los recursos hídricos y la toma de decisiones basada en datos actualizados.

- La aplicación del algoritmo KNN aumentó la precisión de la clasificación de los parámetros según el ICA, mostrando una coherencia interna superior en los resultados obtenidos durante las campañas de monitoreo. Los valores calculados coincidieron con las categorías normativas oficiales de forma más consistente que el método tradicional en Excel. Esta capacidad de KNN, sustentada en su enfoque no lineal y adaptable a la variabilidad geográfica y ambiental del Perú, optimizó el cálculo del índice de calidad de agua para las categorías de consumo humano y bebida animal. Esto permitió una clasificación más precisa y eficiente, beneficiando directamente la evaluación y monitoreo de la calidad del agua en la cuenca del río Huarmey.

Finalmente, la comparación de desempeño entre los tres algoritmos entrenados - KNN, SVM y Random Forest - demostró que KNN ofreció un balance superior entre precisión, tiempo de procesamiento y facilidad de implementación. KNN alcanzó una precisión promedio (R^2) de 0.92, superior a SVM (0.88) y Random Forest (0.90). Además, presentó un tiempo de entrenamiento rápido, una inferencia manejable y una implementación muy sencilla, lo cual facilita su adopción para la clasificación de calidad de agua. Estas ventajas consolidan a KNN como la opción más eficiente y efectiva dentro de las alternativas evaluadas para la optimización del cálculo del ICA en la cuenca del río Huarmey.

CAPÍTULO VIII: CONCLUSIONES

En función de los resultados obtenidos, del desarrollo metodológico aplicado y de los desafíos enfrentados durante el estudio, se propone las siguientes recomendaciones.

- Ampliar la cobertura del modelo hacia otras categorías del Índice de Calidad del Agua (ICA), como las relacionadas con usos recreacionales, riego o vida acuática. Aunque esta tesis se enfocó en las categorías 1-A2 (consumo humano) y 3-D2 (bebida animal), la metodología propuesta puede ser adaptada y escalada para cubrir una clasificación más integral del recurso hídrico.
- Incorporar técnicas de reducción de dimensionalidad y selección de variables, como el análisis de componentes principales (PCA) o algoritmos de selección basada en importancia, con el fin de optimizar el rendimiento del modelo en presencia de un mayor número de parámetros físicoquímicos o muestras futuras.
- Explorar la implementación de algoritmos complementarios, tales como XGBoost, LightGBM o redes neuronales, con el objetivo de comparar su desempeño frente a KNN en escenarios de mayor complejidad o presencia de ruido en los datos.

REFERENCIAS BIBLIOGRÁFICAS

- Ainapure, B., Baheti, N., Buch, J., Appasani, B., Jha, A. V., & Srinivasulu, A. (2023). Drinking water potability prediction using machine learning approaches: a case study of Indian rivers. *Water Practice and Technology*, 18(12), 3004–3020. <https://doi.org/10.2166/wpt.2023.202>
- Aish, A. M., Zaqoot, H. A., Sethar, W. A., & Aish, D. A. (2023a). Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques. *Water Practice and Technology*, 18(3), 501–521. <https://doi.org/10.2166/wpt.2023.028>
- Aish, A. M., Zaqoot, H. A., Sethar, W. A., & Aish, D. A. (2023b). Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques. *Water Practice and Technology*, 18(3), 501–521. <https://doi.org/10.2166/wpt.2023.028>
- Al-Sulttani, A. O., Al-Mukhtar, M., Roomi, A. B., Farooque, A. A., Khedher, K. M., & Yaseen, Z. M. (2021). Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction. *IEEE Access*, 9, 108527–108541. <https://doi.org/10.1109/ACCESS.2021.3100490>
- Ancash Noticias. (2021, March). *Preocupante: ANA constata la mala calidad de agua en la cuenca de agua del río de Huarmey - Ancash Noticias*. 3 de Enero de 2021. <https://ancashnoticias.com/2021/01/03/ancash-ana-constata-la-mala-calidad-de-la-cuenca-de-agua-del-rio-de-huarmey/>
- AquaChek. (2024). *Preguntas frecuentes: pruebas de piscinas y spas AquaChek*. <https://www.aquachek.com/faq/>
- Autoridad Nacional de Agua. (2021). *MONITOREO DE LA CALIDAD DEL AGUA DE LA CUENCA AIJA-HUARMEY 2021 (11 de agosto al 22 de setiembre del 2021) PERÚ*. Ministerio de Desarrollo Agrario y Riego Autoridad Nacional del Agua.
- Autoridad Nacional de Agua. (2023). *Monitoreo de parametros cuenca Huarmey_2023*.
- Autoridad Nacional del Agua. (2018, May 5). *Metodología para la determinación del índice de calidad de agua Ica-PE, aplicado a los cuerpos de agua continentales superficiales*. <https://repositorio.ana.gob.pe/handle/20.500.12543/2440>
- Autoridad Nacional del Agua. (2020). *Monitoreo de parametros de agua cuenca Huarmey_2020*.
- Banco Mundial. (2024). *Perú puede responder a las crecientes amenazas del cambio climático, la contaminación y la creciente demanda de agua*. <https://www.bancomundial.org/es/news/press-release/2023/07/06/per-puede-responder-a-las-crecientes-amenazas-del-cambio-clim-tico-la-contaminaci-n-y-la-creciente-demanda-de-agua>
- Çakir, M., Yilmaz, M., Oral, M. A., Kazanci, H. Ö., & Oral, O. (2023). Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. *Journal of King Saud University - Science*, 35(6), 102754. <https://doi.org/10.1016/J.JKSUS.2023.102754>
- Carbotecnia. (2021). *>> Bacterias coliformes en el agua potable ¿Cómo eliminarlas?* <https://www.carbotecnia.info/aprendizaje/desinfeccion/bacterias-coliformes-en-el-agua-potable/>
- Carbotecnia. (2023). *Metales pesados en el agua, efectos en la salud y eliminarlo - Carbotecnia*. <https://www.carbotecnia.info/aprendizaje/quimica-del-agua/los-metales-pesados-en-el-agua/>
- Chen, B., Mu, X., Chen, P., Wang, B., Choi, J., Park, H., Xu, S., Wu, Y., & Yang, H.

- (2021). Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data. *Ecological Indicators*, 133, 108434. <https://doi.org/10.1016/J.ECOLIND.2021.108434>
- Condorcuya, R. E. M., Cahuana, A. O. P., Ramos, E. O., & Quispe, A. H. (2025). Aplicación de Aprendizaje Automático para la Predicción de la Calidad del Agua Potable en la Universidad Nacional Micaela Bastidas de Apurímac. *Micaela Revista de Investigación - UNAMBA*, 6(1), 1–7. <https://doi.org/10.57166/MICAELA.V6.N1.2025.169>
- Defensoría del Pueblo. (2021). *Defensoria del Pueblo Peru*. <https://www.defensoria.gob.pe/defensoria-del-pueblo-gobiernos-locales-deben-garantizar-la-adecuada-cloracion-del-agua-en-huancavelica/>
- Dilmi, S. (2022). A combined water quality classification model based on kernel principal component analysis and machine learning techniques. *Desalination and Water Treatment*, 279, 61–67. <https://doi.org/10.5004/DWT.2022.29069>
- Dong, L., Zuo, X., & Xiong, Y. (2024a). Prediction of hydrological and water quality data based on granular-ball rough set and k-nearest neighbor analysis. *PLoS ONE*, 19(2 February). <https://doi.org/10.1371/journal.pone.0298664>
- Dong, L., Zuo, X., & Xiong, Y. (2024b). Prediction of hydrological and water quality data based on granular-ball rough set and k-nearest neighbor analysis. *PLoS ONE*, 19(2 February). <https://doi.org/10.1371/journal.pone.0298664>
- Fernández, L. H. H. S. *. (2022). Validación cruzada sobre una misma muestra: una práctica sin fundamento. *R.E.M.A. Revista Electrónica de Metodología Aplicada*, 24(1). <https://doi.org/https://doi.org/10.17811/rema.24.1.2022.38-40>
- Gestión sostenible del agua. (2024). *Sustainable Water Management*. <https://www.agry.purdue.edu/hydrology/projects/nexus-swm/es/Tools/WaterQualityCalculator.php>
- Hamzaoui, M., Aoueleiyine, M. O. E., & Bouallegue, R. (2023). A Hybrid Method of K-Nearest Neighbors with Decision Tree for Water Quality Classification in Aquaculture. *Communications in Computer and Information Science*, 1864 CCIS, 287–299. https://doi.org/10.1007/978-3-031-41774-0_23
- Hussein, E. E., Derdour, A., Zerouali, B., Almaliki, A., Wong, Y. J., Ballesta-de los Santos, M., Minh Ngoc, P., Hashim, M. A., & Elbeltagi, A. (2024). Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms. *Water (Switzerland)*, 16(2). <https://doi.org/10.3390/w16020264>
- Instituto Nacional de Estadística e Informática. (2023). *INEI: El 73,7% de peruanos no tiene acceso a agua de manera segura - Infobae*. <https://www.infobae.com/peru/2024/05/09/el-737-de-peruanos-no-tiene-acceso-a-agua-de-manera-segura-segun-inei/>
- Instituto Nacional del Cancer. (2015). *Cadmio*. <https://www.cancer.gov/espanol/cancer/causas-prevencion/riesgo/sustancias/cadmio>
- Kanyama, M. N., Bhunu Shava, F., Gamundani, A. M., & Hartmann, A. (2024). Machine learning applications for anomaly detection in Smart Water Metering Networks: A systematic review. *Physics and Chemistry of the Earth*, 134. <https://doi.org/10.1016/j.pce.2024.103558>
- La República. (2021). *Ancash: agua sin cloración consumen pobladores de ocho centros poblados de Huarmey | Sociedad | La República*. <https://larepublica.pe/sociedad/2021/11/17/ancash-agua-sin-cloracion-consumen-pobladores-de-ocho-centros-poblados-de-huarmey-lrnd>
- Laura Ochoa, L. (2019). Evaluation of Classification Algorithms using Cross Validation. *Industry, Innovation, And Infrastructure for Sustainable Cities and Communities*,

- 24–26. <https://doi.org/10.18687/LACCEI2019.1.1.471>
- López Lozano, L., Palazón Bru, I., Palazón Bru, A., Arroyo Fernández, M., & González-Estecha, M. (2015). Procedimiento de validación de un método para cuantificar cobalto en suero por espectroscopia de absorción atómica con atomización electrotérmica. *Revista Del Laboratorio Clínico*, 8(1), 46–51. <https://doi.org/10.1016/j.labcli.2014.12.001>
- Ministerio de medio ambiente. (2015). *APRENDE A PREVENIR LOS EFECTOS DEL MERCURIO MÓDULO 3: AGUA Y ALIMENTO*.
- Naghibi, S. A., Hashemi, H., Berndtsson, R., & Lee, S. (2020). Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *Journal of Hydrology*, 589, 125197. <https://doi.org/10.1016/J.JHYDROL.2020.125197>
- Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022a). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, 48, 102920. <https://doi.org/10.1016/J.JWPE.2022.102920>
- Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022b). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, 48. <https://doi.org/10.1016/j.jwpe.2022.102920>
- National Institutes of Health (NIH). (2024a). *Boro - Datos en español*. <https://ods.od.nih.gov/factsheets/Boron-DatosEnEspanol/>
- National Institutes of Health (NIH). (2024b). *Cromo - Datos en español*. National Institutes of Health (NIH). <https://ods.od.nih.gov/factsheets/Chromium-DatosEnEspanol/>
- Omer, N. H., & Omer, N. H. (2019). Water Quality Parameters. *Water Quality - Science, Assessments and Policy*. <https://doi.org/10.5772/INTECHOPEN.89657>
- Organización Mundial de la Salud. (2022). *Water, sanitation and hygiene links to health : facts and figures*. <https://iris.who.int/handle/10665/69489>
- Przybyła-Kasperek, M. (2022). Study of selected methods for balancing independent data sets in k-nearest neighbors classifiers with Pawlak conflict analysis. *Applied Soft Computing*, 129, 109612. <https://doi.org/10.1016/J.ASOC.2022.109612>
- Saddiqi, H. A., Javed, Z., Ali, Q. M., Ullah, A., & Ahmad, I. (2024). Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water. *Journal of Water Process Engineering*, 66, 106014. <https://doi.org/10.1016/J.JWPE.2024.106014>
- Seyedmohammadi, J., Zeinadini, A., Navidi, M. N., & McDowell, R. W. (2023). A new robust hybrid model based on support vector machine and firefly meta-heuristic algorithm to predict pistachio yields and select effective soil variables. *Ecological Informatics*, 74. <https://doi.org/10.1016/j.ecoinf.2023.102002>
- Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. *Chemosphere*, 276, 130265. <https://doi.org/10.1016/J.CHEMOSPHERE.2021.130265>
- Tahraoui, H., Toumi, S., Hassein-Bey, A. H., Bousselma, A., Sid, A. N. E. H., Belhadj, A. E., Triki, Z., Kebir, M., Amrane, A., Zhang, J., Assadi, A. A., Chebli, D., Bouguettoucha, A., & Mouni, L. (2023). Advancing Water Quality Research: K-Nearest Neighbor Coupled with the Improved Grey Wolf Optimizer Algorithm Model Unveils New Possibilities for Dry Residue Prediction. *Water (Switzerland)*, 15(14). <https://doi.org/10.3390/w15142631>
- Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023a). Performance analysis of the

- water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169, 808–828. <https://doi.org/10.1016/J.PSEP.2022.11.073>
- Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023b). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, 169, 808–828. <https://doi.org/10.1016/J.PSEP.2022.11.073>
- Umar, N., & Gray, A. (2023a). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, 15(8). <https://doi.org/10.3390/w15081519>
- Umar, N., & Gray, A. (2023b). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, 15(8). <https://doi.org/10.3390/w15081519>
- USGS Water Resources. (2024). *USGS Water-Quality Data for the Nation*. <https://waterdata.usgs.gov/nwis/qw>
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021a). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 598. <https://doi.org/10.1016/J.JHYDROL.2021.126266>
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021b). Ensemble machine learning paradigms in hydrology: A review. In *Journal of Hydrology* (Vol. 598). Elsevier B.V. <https://doi.org/10.1016/j.jhydrol.2021.126266>