UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

PLATAFORMA ABIERTA DE DETECCIÓN DE ANOMALÍAS Y APRENDIZAJE AUTOMÁTICO PARA APOYO A LA TOMA DE DECISIONES EN LA GESTIÓN DE AGUAS SUBTERRÁNEAS

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

MAXIMILIANO TOMÁS JONES HERRERA

PROFESORA GUÍA:
DORIS SÁEZ HUEICHAPAN

PROFESOR CO-GUÍA:
FRANCISCO JARAMILLO MONTOYA

MIEMBROS DE LA COMISIÓN:
MATÍAS TAUCARE TORO
CONSTANZA AHUMADA SANHUEZA

Este trabajo fue financiado por:
FONDEF IDEA ID19I10363, ID UChile ENL08/21,
ANID/FONDAP/1522A0006 y ANID/FONDECYT 1220507.

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
Y AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: MAXIMILIANO TOMÁS JONES HERRERA
FECHA: 2023
PROF. GUÍA: DORIS SÁEZ HUEICHAPAN

PLATAFORMA ABIERTA DE DETECCIÓN DE ANOMALÍAS Y APRENDIZAJE AUTOMÁTICO PARA APOYO A LA TOMA DE DECISIONES EN LA GESTIÓN DE AGUAS SUBTERRÁNEAS

Según informes recientes de las Naciones Unidas (2022), las cuencas principales de Chile se encuentran bajo un estrés hídrico extremadamente alto, lo que amenaza tanto a la población como al crecimiento económico. De ellas, los acuíferos subterráneos poseen un bajo nivel de monitoreo además de complejidades técnicas y físicas asociadas a su gestión.

Este trabajo presenta el diseño e implementación de una plataforma que integra un sistema experto para el procesamiento, análisis y la detección de anomalías en datos de monitoreo de acuíferos subterráneos. Se propone una metodología de selección automática del mejor modelo de detección de anomalías para generar alertas de posibles problemas a partir de sus parámetros fisicoquímicos. La metodología incluye el análisis de múltiples modelos incluyendo modelos de apilamiento y ensamble para una solución robusta ante las particularidades propias de cada acuífero.

Se analiza que la integración de las distintas herramientas dentro de la plataforma desarrollada logra facilitar el procesamiento de los datos de series de tiempo hidrológicas, para la detección de anomalías. Se logra detectar automáticamente hasta un 85 % de las anomalías en los datos de calidad de agua. Las potenciales aplicaciones de este trabajo podrían ir extenderse a la detección de anomalías en diferentes casos de uso.

Para ustedes, mamá, papá.

Los amo.

Agradecimientos

Agradezco el apoyo de los siguientes proyectos para el desarrollo este trabajo de Tesis: FONDEF ID19I10363 “Sistema abierto experto para apoyar la gestión de recursos hídricos mediante monitoreo de bajo costo en tiempo real de aguas superficiales y subterráneas”, Instituto Sistemas Complejos de Ingeniería (ISCI) ANID PIA/PUENTE AFB220003, Solar Energy Research Center SERC-Chile ANID/FONDAP/1522A0006, ANID/FONDECYT 1220507 “Distributed Predictive Control Strategies based on Evolving Prediction Intervals for Energy-Water Microgrids” y Project VID UChile ENL08/21 “Design of Prediction Intervals based on Computational Intelligence and Evolving Systems for Modeling and Control of a Water and Energy Management System”.

Luego quiero agradecer encarecidamente a mi familia, sin duda todo esto ha sido posible gracias a su apoyo constante. A mi padres, Yolanda y Mario por todo el amor y esos cariños, a veces a distancia, que nunca faltaron.

Me gustaría agregar a mis profesores guías, Dra. Dóris Sáez Hueichapan y Dr. Francisco Jaramillo Montoya por su constante ayuda, confianza y paciencia en este proceso. Aquí tampoco puedo dejar de agradecer a Dr. Matías Taucare, por su apoyo fundamental en temas ajenos a mi área de especialización.

A Vanessa, mi compañera, que me soportó estos últimos años y toda la etapa del Magíster, gracias por todo el amor y la paciencia.

A todos mi amigos de universidad que conocí, tanto al principio en plan común, como durante la especialidad en eléctrica por su compañía y apoyo incondicional durante estos años.

Finalmente no puedo olvidarme de todos mis amigos de la vida, los que están y los que ya no, por su amistad y mantenerse en contacto pese al tiempo y la distancia.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

Asegurar el suministro de agua potable en medio de la actual crisis climática se ha convertido en una preocupación mundial de gran magnitud. Esta tarea, sin embargo, se vuelve cada vez más compleja. Según el último informe del Panel Intergubernamental sobre Cambio Climático [1], se prevé que la temperatura global aumente más de 2°C, superando el límite de 1,5°C establecido para el crecimiento desde la era preindustrial.

Una de las consecuencias más preocupantes de esta crisis es su impacto en todo el ciclo del agua [2], particularmente en las fuentes de agua subterránea [3], que representan la mayor reserva de agua dulce a nivel mundial [4]. Actualmente, los pozos subterráneos abastecen de agua potable a miles de millones de personas en todo el mundo [5]. Sin embargo, estudios recientes revelan que millones de estos pozos corren el riesgo de quedarse sin agua [6], principalmente debido a la constante disminución del nivel de las aguas subterráneas disponibles [7].

Esta alarmante noticia no es nada nuevo en el caso de Chile, que posee zonas declaradas en "crisis hídrica" ya que no cuentan con un suministro constante de agua potable. Se cree que esta situación se debe no solo al cambio climático, sino también a una insuficiente gestión de los recursos hídricos [8]. A lo largo de los años, se ha evidenciado una falta de planificación y coordinación efectiva en la administración del agua, lo que ha llevado a un uso ineficiente y desigual de este recurso vital además de su sobreexplotación.

En la última década, la escasez de agua en Chile ha empeorado significativamente debido a la megasequía que afecta a gran parte del territorio nacional [9]. Esta sequía se manifiesta en una disminución de hasta un 40 % y un 90 % de las precipitaciones y los caudales de los ríos, respectivamente. Estos fenómenos tienen graves repercusiones tanto para los ecosistemas como para la vida humana y las actividades productivas.

La disminución de las precipitaciones y de caudales de los ríos está afectando directamente a la disponibilidad de aguas subterráneas [6]. Dentro de los factores locales más influyentes de tipo antropogénico se encuentran la escasa regulación del uso de suelos y el otorgamiento de derechos de agua [10]. Lo anterior se evidencia en una constante deforestación y el uso intensivo de agua por sectores productivos como la agricultura, que utiliza alrededor de un 70 % del agua extraída [11].

En este contexto, es esencial establecer mecanismos efectivos de monitoreo y control de los recursos hídricos, tanto subterráneos como superficiales, con el fin de asegurar su disponibilidad y calidad para toda la población. En este sentido, una herramienta clave para regular la extracción de agua son soluciones que ya se implementan como el Sistema de *Monitoreo de Extracción Efectiva* (MEE) que busca hacer cumplir los límites de extracción de los derechos de agua perteneciente a privados.

1.2. Monitoreo de acuíferos y calidad de aguas en Chile

En Chile la autoridad competente del monitoreo y gestión de los recursos hídricos es la *Dirección General de Aguas* (DGA), organismo administrativo dependiente del Ministerio de Obras Públicas (MOP) que tiene como objetivo verificar, gestionar y difundir la información hídrica del País, en especial la referente a cantidad y calidad de los recursos hídricos extraídos, además de los permisos de extracción y aprovechamiento.

Según la resolución 1238 de la DGA [12], los componentes para los sistemas de MEE deben contar como mínimo con: flujómetro o caudalímetro, el cual puede ser de tipo electromagnético, de ultrasonido o mecánico. En caso de que el volumen de agua extraída sea igual o mayor a los estándares medio o mayores, el titular del derecho de extracción debe contar también en su sistema con: sensor de niveles freáticos, sistema de almacenamiento y registro de datos para ser cargados en la plataforma web de MEE de la DGA.

Es importante destacar que el monitoreo continuo de la calidad de aguas mediante parámetros fisicoquímicos no es obligatorio en la actualidad. La DGA cuenta con una Red de Monitoreo de Calidad de Aguas [13]. Esta red posee múltiples estaciones de monitoreo desde donde se toman muestras manualmente y se envían a un laboratorio ambiental que las procesa y realiza los análisis, sin embargo, pasa mucho tiempo desde que se toma una muestra hasta que se tienen los resultados.

Para el caso de las aguas superficiales se cuenta con una red de control de calidad de aguas que envían información periódica, con una frecuencia del orden de meses, del estado trófico de algunos de los principales cuerpos de agua superficial ubicados en las macrozonas centro y sur. Esta frecuencia es insuficiente para poder hacer una gestión o detectar posibles eventos de contaminación o dinámicas propias de ríos y lagos.

La actual estrategia de monitoreo de las redes de calidad de aguas se realiza a través de captura *in situ* y envío de muestras al laboratorio centralizado de la DGA, como puede apreciarse en la Figura 1.1. Los requerimientos concretos de frecuencia y metodología de muestreo de calidad de aguas puede encontrarse en la Norma Chilena 409/2 [14], más información sobre los requerimientos pueden encontrarse en la Sección 1.2.1.

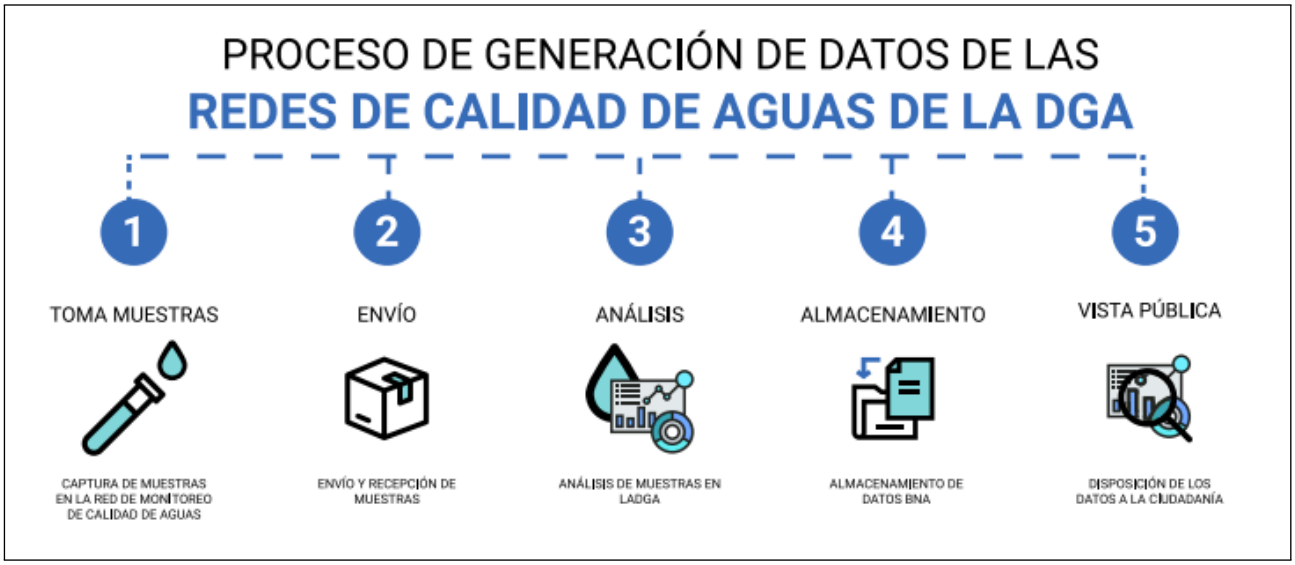


Figura 1.1: Estructura de adquisición y procesamiento de los datos provenientes de las redes de calidad de aguas de la DGA (fuente: [13])

La Red de Calidad de Aguas de la DGA considera un extenso y amplio trabajo multidisciplinario tanto a nivel regional como a nivel nacional [15]. Esto supone un desafío complejo para los sistemas de información disponibles de localidades en donde no existen en la actualidad pozos o acuíferos monitoreados por esta entidad, siendo difícil para privados poder anexarse a esta red de monitoreo. También existen problemas asociados a que el impacto de esta red de monitoreo es bajo pues los datos generados en ella en ocasiones no llegan hasta los usuarios y por tanto no influyen en las decisiones para la gestión de los recursos hídricos.

La DGA exige un sistema de medición de extracciones que monitorea el nivel, además del flujo de extracción para el caso de aguas subterráneas [12]. Para la extracción de aguas superficiales se requiere reportar mediciones del caudal desviado para aprovechamiento más el caudal restablecido para los casos que se sobrepasa los derechos de extracción [16]. En ambos casos estos sistemas deben almacenar la información histórica de al menos tres años en su memoria y el usuario debe enviar de forma periódica la información al servidor de la DGA para informar estos parámetros y monitorear.

El estado actual del monitoreo de calidad y cantidad de agua disponible en acuíferos subterráneos propone una oportunidad para poder incluir no solo parámetros de calidad a los sistemas de MEE, sino también herramientas de procesamiento de datos para la supervisión de los datos en línea. Con esto es posible procesar datos de forma automática a través de una plataforma intuitiva que, mediante visualizaciones, puede ayudar a la toma de decisiones y alertar sobre eventos de interés asociados a los recursos hídricos.

Se presentan a continuación los principales marcos normativos locales asociados al monitoreo de acuíferos y sus requerimientos:

1.2.1. Parámetros de Calidad de agua

La Norma Chilena 409 (Nch.409) [17],[14] titulada "Norma Calidad Agua Potable", establece qué se entiende por Agua Potable en Chile y fija los límites para la presencia de distintos elementos o sustancias químicas de importancia para la salud. Plantea también los límites además de las características organolépticas que debe tener el agua potable asociado a parámetros físicos, tanto inorgánicos como orgánicos. Estos parámetros pueden apreciarse en la Tabla 1.1.

Tabla 1.1: Parámetros relativos a características organolépticas para el agua potable

Parámetros	**Expresado como**	**Unidad**	**Límite máximo**
Físicos:			
Color verdadero	-	Unidad PT-Co	20
Olor	-	-	inodora
Sabor	-	-	insípida
Inorgánicos:			
Amoníaco	NH_3	mg/L	1,5
Cloruro	Cl^-	mg/L	400(1)
pH	-	-	6,5 <pH <8,5
Sulfato	SO_4^{-2}	mg/L	500(1)
Sólidos disueltos	-	mg/L	1500
Orgánicos: **Compuestos fenólicos**	Fenol	µg/L	2
(1) La Autoridad Competente, de acuerdo con las instrucciones impartidas por el Ministerio de Salud, podrá autorizar valores superiores a los límites máximos señalados en esta tabla, conforme a la reglamentación sanitaria vigente.			

Existe también la Norma Chilena 1333 [18] que plantea una flexibilización de algunos de los parámetros planteados anteriormente pero asociado a los requisitos de uso de agua para otros fines. Entre estos otros fines se destaca la utilización para cultivos agrícolas, recreacional, estético y el sustento de la vida acuática. En cuanto a los parámetros relativos de calidad de agua muchos de ellos se obtienen con una frecuencia menor de la deseable[1] para poder, por ejemplo, detectar eventos de contaminación o de peligro para el consumo.

El principal objetivo de estas normas es establecer un marco normativo tecnificado que especifique los tipos de mediciones y la frecuencia con la que deben llevarse a cabo. Sin embargo, estas normas no abordan la problemática de cómo poner a disposición esta información para los usuarios, de manera que puedan tomar decisiones y realizar gestiones en base a su análisis.

Este trabajo representa una contribución al procesamiento de datos de monitoreo de acuíferos, centrándose en la captura de datos a través de nodos sensores de bajo costo. El objetivo fundamental es fomentar la adopción generalizada de estas tecnologías en todo el país. Con

[1] Una frecuencia que permitiera hacer gestión incluiría varias mediciones por día, idealmente en el orden de minutos.

esto, se espera poder contar en el futuro con mediciones basadas en parámetros fisicoquímicos capturados mediante sensores de bajo costo. En este estudio, se aborda un subconjunto específico de parámetros fisicoquímicos, los cuales se detallan en la Sección 1.2.2.

1.2.2. Variables a considerar en monitoreo de aguas subterráneas

En este trabajo se analizan los siguientes parámetros fisicoquímicos basado en la disponibilidad y confiabilidad de sensores con los que ya se contaba un registro de varios años de monitoreo. Estas variables son:

1. **Conductividad eléctrica:** representa a la capacidad del agua de conducir corriente eléctrica, la cual depende de la cantidad de iones disueltos (tales como Cl-, SO42-, HCO3-, Na+, K+, Ca2+, Mg2+). Esta variable es importante, pues a partir de su valor, es posible inferir cambios en la calidad de aguas, los cuales pueden asociarse a eventos como lluvias o intrusión marina. Según la Organización Mundial de la Salud (OMS) el rango de calidad para este parámetro, en uso potable, es en torno a 50-800 [$\mu S/cm$].

2. **Temperatura:** es importante medir la temperatura dado que afecta la capacidad de los microorganismos para resistir los contaminantes del agua. Su unidad de medida es grados Celsius [$°C$].

3. **pH:** es la medida relativa de la concentración de iones de hidrógeno e hidroxilos en el agua. Sus mediciones pueden variar entre 0 y 14, de 0 a 7 aumenta la acidez (H+), mientras que de 7 a 14 aumenta la alcalinidad (OH-). Según la OMS el rango de calidad para este parámetro, en uso potable, es aceptable entre 6,5 a 8,5 pH. Una variación en el pH puede indicar contaminación química en el agua, como por ejemplo la asociada a infiltración de drenajes con fertilizantes o de aguas residuales de procesos industriales como la minería [19].

4. **Nivel de agua:** se refiere a la altura o profundidad a la que se encuentra el agua en un determinado punto de medición. Es una medida que indica la cantidad de agua presente en un sistema, como un acuífero, un cuerpo de agua superficial, un pozo o un embalse. Swe mide comúnmente a través de un piezómetro y se relaciona con la cantidad de agua que contiene el acuífero si se conocen bien sus propiedades geométricas e hidráulicas y, si no, representa una referencia de almacenamiento. Cuando se toma la presión su unidad de medida son los Pascales [Pa], aunque también puede representarse por metros de columna de agua o [mH_2O] haciendo referencia a la presión que ejerce el fluido sobre el sensor.

1.2.3. Plataformas de monitoreo

En Chile existen iniciativas tanto públicas como privadas que han desarrollado plataformas [20, 21, 22] con un foco en el monitoreo de parámetros ambientales y algunos presentando datos de calidad de aguas. Estas plataformas tienen en común que permiten visualizar mediante mapas la disponibilidad de información de múltiples fuentes, ubicar distintas estaciones de monitoreo y obtener visualizaciones dinámicas en forma de series de tiempo de los datos meteorológicos disponibles.

Por ejemplo, el observatorio georreferenciado de la DGA [20] entrega información de las extracciones efectivas y opciones para visualizar las zonas que actualmente se encuentran declaradas en emergencia hídrica o similares, pero no muestra información sobre parámetros fisicoquímicos de acuíferos. Dependiente del Centro de ciencia del Clima y la Resiliencia (CR)2 se encuentra el Explorador climático CR2 [21] que entrega información climática de precipitaciones y temperaturas en distintas estaciones además de incorporar datos desde la DGA. Finalmente, la plataforma de Soquimich [22] es una plataforma de monitoreo de parámetros fisicoquímicos de múltiples pozos y su desarrollo se encuentra condicionada al cumplimiento de una resolución de calificación ambiental que autoriza el proyecto y establece que las mediciones se pueden distanciar en el tiempo hasta en 90 días [20].

Algo en común de las plataformas analizadas es la escasa información de monitoreo de disponible asociado a la calidad de agua y en el caso que existe, una baja frecuencia y nulo análisis de los datos. Estas falencias detectadas dentro de las políticas de monitoreo de calidad de aguas plantean una oportunidad de desarrollo tecnológico importante y pertinente en la actualidad. De la necesidad de poder generar nodos sensores de bajo costo para ampliar y mejorar las redes de monitoreo de calidad de agua en Chile nace el proyecto FONDEF ID19I10363. Este proyecto abordó la problemática con un desarrollo multidisciplinario y tanto de hardware como software para el procesamiento mediante un Sistema Experto de los datos de monitoreo de acuíferos

Así este trabajo se enmarca dentro del proyecto titulado "Sistema abierto experto para apoyar la gestión de recursos hídricos mediante monitoreo de bajo costo en tiempo real de aguas superficiales y subterráneas" se desarrolla un prototipo de unidad experimental para medir de forma continua niveles tanto dinámicos como estáticos de acuíferos, además de parámetros de calidad de agua como el pH, la conductividad eléctrica (CE) y la temperatura.

La contribución de este trabajo al proyecto es el desarrollo de un sistema de información abierto, titulado *Sistema Experto* (SE), para ayudar en la toma de decisiones oportunas por parte de gestores de recursos hídricos. El SE se encarga de analizar e interpretar los datos de calidad de aguas subterráneas para detectar *anomalías* en las variables fisicoquímicas monitoreadas. Un ejemplo de esta interacción y en donde interactúa el usuario, a través de una aplicación web, puede apreciarse en la Figura 1.2.

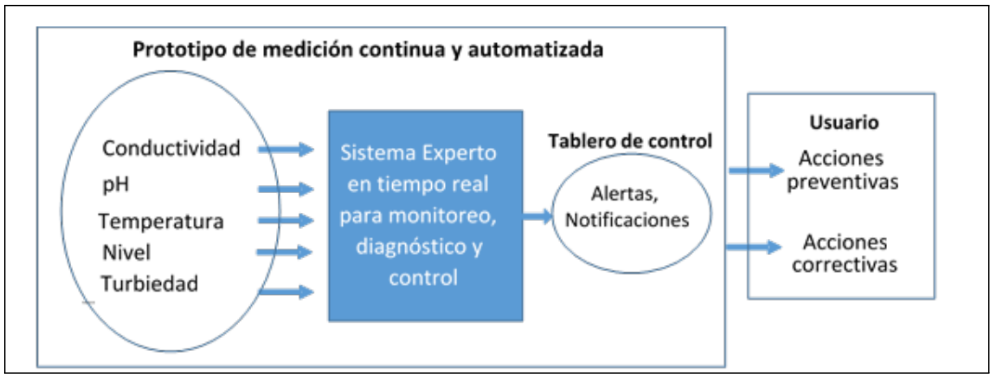


Figura 1.2: Esquema de información, comunicación y procesamiento del sistema propuesto en el proyecto

1.2.4. Sistema experto

La principal componente de producción de la etapa de desarrollo de software y la propuesta de este trabajo de tesis es el sistema de detección de anomalías de datos de calidad de agua y el panel de datos o plataforma web de visualización. Este sistema conjunto busca poder combinar el conocimiento de un experto en sistemas hidrológicos, como es el caso del monitoreo de acuíferos, junto con un procesamiento computacional que analiza las series de tiempo, provenientes de los sensores, de manera análoga a como lo haría un/a experto/a en busca de anomalías o riesgos. Este sistema supone una capa adicional de monitoreo automático a través de la generación de *alertas* y *notificaciones* a través de un tablero de control interactivo. Un esquema informativo general del proyecto y la interacción del SE se aprecia en la Figura 1.3.

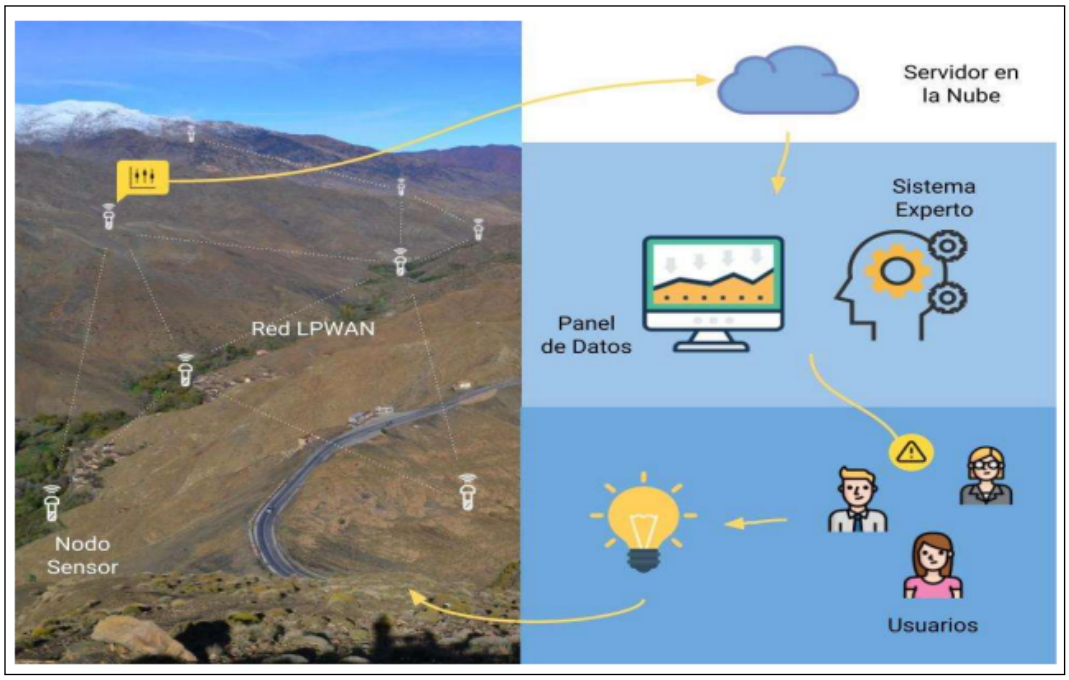


Figura 1.3: Vista general del proyecto.

Los datos de monitoreo de acuíferos presentan comportamientos dinámicos complejos de caracterizar dentro de parámetros normales o no normales en el tiempo pues esto dependerá de las particularidades de cada pozo más la utilización del mismo. Adicionalmente, diferentes pozos asociados a un mismo acuífero pueden presentar una dinámica totalmente diferente a otro cercano que puede relacionarse con su geometría o hidráulica particular.

Así, cada acuífero podría asociarse a un estado dentro de un rango y comportamiento *normal*, pues los parámetros fisicoquímicos están en orden o por el contrario, clasificarse dentro de un estado de *anomalía*. Una anomalía supone una alerta para los usuarios que gestionan el recurso y para los usuarios dependientes, pues puede significar un riesgo para su salud o del sistema en sí o incluso algo positivo como una recarga en tiempos de sequía. Aquí el conocimiento experto es fundamental en el etiquetado de los datos inicialmente para entrenar los modelos de detección basados en su conocimiento y las particularidades de cada pozo.

Algunas de las posibles situaciones que podrían encontrarse con la detección de anomalías son:

- La identificación de puntos que aportan salinidad a partir de cambios inusuales en la conductividad eléctrica.
- A través de los cambios de pH o conductividad detectar posibles derrames químicos o infiltración de agua contaminada al acuífero.
- También sería posible detectar puntos de recarga y/o quizás extracciones o descargas de terceros a partir de cambios en nivel o la temperatura.

Dadas las particularidades de cada pozo, incluso en un mismo acuífero, es que se necesitan herramientas lo suficientemente robustas para poder diferenciar de qué datos corresponden o no a una anomalía basado en un etiquetado inicial de datos por un/a experto/a. Las distintas singularidades de cada pozo también son un factor a considerar a la hora de establecer que una herramienta pueda servir en distintos casos de uso, por lo tanto, se considera que una herramienta capaz de auto-ajustarse a cada acuífero es un requisito fundamental. Aquí es donde las metodologías de ensamble de modelos de detección de anomalías proponen una buena generalización a variaciones en los datos y a escalar de buena forma a medida que se cuenta con más datos [23].

Para este trabajo se analiza un caso de estudio correspondiente a un pozo monitoreado en la localidad de Horcón, Región de Coquimbo. Este pozo contiene mediciones fisicoquímicas de un período de casi cuatro años con un etiquetado experto de todos los datos para un análisis cuantitativo de la estrategia de detección de anomalía basada en modelos de ensamble y entrenamiento automático. También se analizan las visualizaciones de sus datos y la integración de la metodología dentro del Sistema Experto, que incluye también a la plataforma para el entrenamiento y selección del mejor modelo.

Uno de los objetivos de este trabajo es desarrollar una herramienta útil para los usuarios finales, de modo que puedan cargar sus propios datos y beneficiarse de una herramienta analítica integrada que se entrene a partir de esos datos y realice un procesamiento automático de los mismos. Se ha encontrado que la aplicación de herramientas asociadas al Aprendizaje Automático (Machine Learning, ML) propone una solución factible, y se buscará analizar si son lo suficientemente robustas para la tarea de entrenar y seleccionar el mejor modelo o conjunto de modelos para detectar anomalías y generar alarmas a través de un Sistema Experto.

1.2.5. Revisión de estado del arte

En la literatura el problema de detección de anomalías en calidad de agua se ha abordado en su mayoría con un foco en el monitoreo de los sistemas de agua potable y de causes de ríos. En [24] se propone una metodología completa de detección de contaminantes en ríos con el objetivo de poder generar alertas tempranas de eventos de contaminación a través de un análisis multiparámetro. Otros estudios, como [25], analizan la utilización de diferentes modelos basados en ML para la predicción de indicadores de calidad de agua en fuentes de

agua superficiales.

Para el caso de acuíferos subterráneos, en [26] se analiza la detección mediante ML de proliferación de algas nocivas para la salud. Se comparan distintos modelos para la detección de anomalías en análisis de fluorescencia de ficocianina en sustitución de un análisis de laboratorio. En [19] se analizan dos modelos de detección de anomalías en datos de napas subterráneas qué tenían riesgo de contaminación por fertilizantes al encontrarse un una región rodeada de cultivos industriales.

En el análisis de la literatura se identifica una oportunidad importante de desarrollo en cuanto a que los acuíferos subterráneos suponen presentan un estado que es invisible a los ojos pero de suma importancia para la vida misma. Así la capacidad de integración de herramientas de ML y de detección de anomalías en datos de acuíferos subterráneos propone una oportunidad para el desarrollo.

1.3. Hipótesis

- La capacidad de generación, entrenamiento y comparación, de forma simple, de múltiples modelos basados en la selección del mejor modelo individual o ensamble de modelos, para la tarea de detección de anomalías, supone una solución robusta ante la variabilidad de los datos de pozos monitoreados.
- El desarrollo de una plataforma que integre un análisis estadístico de la datos históricos, además de poder caracterizar anomalías de calidad de agua de forma automática, permitirá levantar alarmas sobre eventos de interés para los usuarios y así apoyar a la gestión a través del monitoreo de los recursos hídricos.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar una plataforma web que facilite tanto el procesamiento en línea como la visualización de datos hidrogeológicos. Además, se busca implementar técnicas de aprendizaje automático para detectar anomalías en dichos datos y generar alarmas de manera automatizada. La plataforma permitirá la selección automática del modelo con mejor desempeño para optimizar la detección de anomalías basada en los datos.

1.4.2. Objetivos específicos

- Diseñar una estrategia de pre-procesamiento, manejo de datos perdidos automático y una selección de las variables monitoreadas a procesar configurable por el usuario.
- Integrar modelos de detección de anomalías basadas en series de tiempo hidrogeológicas a través de aprendizaje automático y el ensamble de modelos para emular la capacidad de generalización de un conjunto de expertos.
- Diseñar y desarrollar un sistema experto que integre una metodología de procesamiento, entrenamiento y selección automática del mejor modelo dependiendo de los datos

cargados dentro de una plataforma web. Permitiendo potenciar el monitoreo a través de visualizaciones de alarmas de anomalías detectadas y análisis mediante métricas de desempeño en una interfaz interactiva y de fácil uso.

1.5. Contribuciones

Este trabajo considera varias contribuciones asociadas a la tarea de detección de anomalías en monitoreo de acuíferos subterráneos.

- Revisión bibliográfica en estrategias detección de anomalías supervisadas aplicadas a series de tiempo hidrogeológicas y sistemas multivariables.
- Propuesta de estrategia de automatización de los procesos de análisis exploratorio de los datos, la etapa de pre-procesamiento, entrenamiento distintos modelos y selección del mejor modelo individual o ensamble de modelos para la tarea de detección de anomalías en series de tiempo.
- Integración de un sistema experto a través de una plataforma web que implementa la estructura de automatización de la estrategia de entrenamiento y selección de modelos de detección de anomalías para el diagnóstico de acuíferos subterráneos a partir de la data histórica.

1.6. Estructura del documento

Este trabajo se organiza en 7 capítulos, estructurados de la siguiente forma: en el Capítulo 2 se presenta un marco teórico en donde se definen los principales conceptos sobre detección de anomalías y procesamiento automático de series de tiempo hidrogeológicas asociados a este trabajo de Tesis. Se presentan también aquí los principales avances del estado del arte asociado a estas temáticas y se analiza la pertinencia de las distintas alternativas y modelos, además de sus aplicaciones al problema de detección de anomalías.

En el Capítulo 3, se expone la metodología utilizada, abordando las estructuras de funcionamiento y las propuestas de automatización para incorporar la estrategia de entrenamiento y selección de modelos aplicadas a la detección de anomalías en el monitoreo de acuíferos. Además, se presentan las principales métricas de desempeño de los distintos clasificadores asociados al caso de estudio ubicado en la localidad de Horcón, Región de Coquimbo

El caso de estudio y los datos utilizados para este trabajo se presentan en el Capítulo 4 junto a una descripción y un breve análisis exploratorio de las particularidades de los datos de monitoreo de acuíferos además de los escenarios a analizar.

Los resultados tanto de la metodología de detección de anomalías como también de la aplicación de modelos de ensamble y apilamiento se encuentran en el Capítulo 5, junto con la discusión de sus desempeños en ambos escenarios propuestos a partir del caso de estudio.

En el Capítulo 6 se muestra el desarrollo de la plataforma junto a una descripción y justificación de los principales requerimientos propuestos para su diseño y su desarrollo. Aquí se

detallan las distintas funcionalidades del sistema experto en cuanto al procesamiento de datos y su integración como aplicación web. Se presentan también los gráficos que corresponden a la aplicación de la estrategia de detección de anomalías dentro de la plataforma.

Finalmente, en el Capítulo 7 se presentan las principales conclusiones en torno a las funcionalidades y el cumplimiento de los objetivos de la investigación y de este trabajo de Tesis.

Capítulo 2

Marco teórico

En la literatura se presentan diversas aplicaciones de aprendizaje de máquinas e inteligencia artificial para el procesamiento y la detección de anomalías en series de tiempo, además de distintas metodologías para el pre-procesamiento y visualización de los datos.

En este capítulo se aborda una recopilación de los principales antecedentes, definiciones, conceptos y marcos de conocimiento necesarios para comprender el trabajo realizado. Se presentan también antecedentes relativos al monitoreo y análisis de datos de acuíferos en Chile según exigencias normativas, y por último, los conceptos claves acerca de detección de anomalías en series de tiempo, aplicadas al análisis de calidad de aguas.

2.1. Anomalías y outliers

Existen muchas definiciones distintas de qué representa un dato o un conjunto de datos anómalos en la literatura, y varía según el campo de estudio en el que se aborda. Algunos autores utilizan indistintamente definiciones como "anomalía" y valores atípicos, comúnmente estudiados como "outliers" para referirse a una medición que no se condice con el resto de las mediciones. Pueden encontrarse también en la literatura el uso de definiciones como anormalidades, discordancias o desviaciones en las muestras [27].

Hay autores que diferencian estos dos conceptos añadiendo a la definición de valores atípicos un concepto más amplio, pues, plantean que puede representar ruido o información corrupta [28]. Por otra parte, una anomalía podría representar puntos irregulares pero que siguen cierto patrón de desviación. Una de las definiciones más comunes de anomalías es la dada por Hawkins en [29] :

"Un valor atípico o outlier es una observación que se desvía considerablemente del resto de observaciones, tanto que despierta sospechas que ha sido generada por un mecanismo diferente. "

En este trabajo se consideran los conceptos de anomalías y outliers de forma indistinta. Y es necesario destacar que una definición común de anomalías incluye que sus distribuciones se desvían considerablemente de la distribución característica de los datos. Además de esto, las anomalías representan una pequeña fracción del conjunto de datos y es debido a que en su mayoría representan datos normales (no anómalos).

En adelante se entenderá anomalía según la definición 2.1.

Definición 2.1 *Una anomalía es una observación o una secuencia de observaciones que se desvían considerablemente de la distribución común de los datos. El conjunto de anomalías representa una pequeña parte de la base de datos.*

Por ejemplo, un conjunto de datos anómalos podría representar desde un fraude de tarjetas de crédito [30] hasta un defecto estructural en un edificio [31] o una condición médica desconocida detectada a través del procesamiento de imágenes [32]. Así, en el área del aprendizaje de máquinas, la detección de anomalías en series de tiempo es una tarea de vital importancia y su estudio ha alcanzado mayor relevancia de la mano con los avances de infraestructuras y modelos de almacenamiento y procesamiento de datos provenientes de las más variadas fuentes.

Tipos de anomalías

Las anomalías pueden presentarse en diversas formas [33], en específico se clasificarán en tres grandes grupos:

1. **Anomalía puntual:** Un punto que se desvía considerablemente del resto de los datos, es considerado una anomalía puntual. Un ejemplo de esto podría ser un cambio drástico en la conductividad eléctrica en el agua subterránea de un pozo monitoreado.

2. **Conjunto de anomalías:** En algunas ocasiones un dato desviado levemente por si solo no representa una anomalía, sin embargo, un conjunto de estos datos representan una anomalía, en su conjunto. Aquí también podrían incluirse una tendencia anómala que se aprecia en una ventana de tiempo más grande y que podría representar un estado que se desvía de forma continua (o no) en el tiempo. Un ejemplo de esto podría ser un seguidilla de retiros de dinero en un cajero automático por un monto considerable pero no muy distinto de lo habitual por sí solo.

3. **Anomalías contextuales:** Representan información que puede ser correcta en cierto contexto, pero se detecta como anomalías en otro. Un ejemplo sería que una temperatura elevada de 28° podría ser normal en verano, sin embargo, en invierno esto representa una anomalía.

La diferenciación de los tipos de anomalía es clave para poder comprender las singularidades de los datos estudiados y los resultados obtenidos de los distintos modelos aplicados en la detección.

2.2. Detección de anomalías

Cuando un proceso como un sistema de monitoreo entrega datos que no concuerdan con lo que se espera del modelo de operación "normal", estos datos pueden entenderse como anomalías. Identificar la aparición de estos datos y como se relacionan su generación en un sistema con un evento de interés o un suceso de importancia es uno de los objetivos principales de

modelos de detección de anomalías en series de tiempo.

De forma general los modelos de detección de anomalías pueden entregar dos tipos de salidas [34]:

- **Puntaje de anomalía:** Muchos de los algoritmos para detección de anomalías entregan como una salida una cuantificación del nivel de "atipicidad" o qué tan anómalo una medición se considera. Es una forma general de salida de estos modelos que permite ordenar las mediciones a partir de su desviación de los modelos normales, sin embargo, no entrega una etiqueta cualitativa pues no tiene asociado un umbral de decisión en sí.
- **Etiqueta binaria:** Corresponde a una etiqueta asignada que indica si un dato corresponde a una anomalía o no. Algunos modelos entregan esta etiqueta directamente y otros la generan a partir de un umbral fijado a partir de un puntaje de anomalía. Aporta menos información que el puntaje de anomalía, sin embargo muchos procesos asociados a la toma de decisiones requieren que las etiquetas sean binarias.

La detección de *outliers* aplicada a datos de sensores de monitoreo incluye en algunos casos la detección de desviaciones leves o también conocidas como mediciones ruidosas. Encontrar el espacio de separación óptimo entre qué se considera una medición ruidosa de una anomalía se basa en qué entenderá el analista por una "desviación considerable" de una medición, es decir, el *groundtruth* con el cual se entrenarán los modelos. Así, es posible definir dos grandes grupos de metodologías de entrenamiento para detección o clasificación de anomalías, los métodos que requieren de entrenamiento no supervisado y los que son supervisados.

2.2.1. Aprendizaje no supervisado

Cuando los datos a analizar no cuentan con una etiqueta previa de si corresponden a datos "normales" o a "anomalías" los métodos disponibles para detectar estas anomalías están limitados a modelos no supervisados. En este caso los métodos de detección de anomalías no son capaces de encontrar por sí solos o aprender cuál es el mejor modelo pues no conoce qué datos son anómalos y cuales no. En estos modelos la validación del desempeño debe realizarla el analista en base a prueba y error.

Una aplicación común de métodos no supervisados es detectar mediciones discordantes en una red de monitoreo que podrían entenderse como anomalías pues se diferencian considerablemente (estadísticamente) al resto de muestras. También tienen aplicación en detectar mediciones ruidosas, es decir, que presentan una menor tasa de discordancia, pero que no llegan a ser un evento de interés y podría afectar a métodos que se aplicarán posteriormente sobre los datos. En la aplicación de estos métodos una forma de discriminar entre una y otra puede ser a través del *puntaje de anomalía* de cada muestra [27].

2.2.2. Aprendizaje supervisado

En un esquema de aprendizaje supervisado, los datos cuentan con una etiqueta binaria previa de cuando ocurren anomalías así como también de los ejemplos de datos "normales".

Estos datos se etiquetan generalmente de forma manual por un experto, para entrenar modelos de clasificación binaria que diferencian entre una muestra "normal" (etiqueta negativa o 0) y una muestra "anómala" (positiva o 1).

Los métodos que se entrenan de forma supervisada utilizan la información contenida en las etiquetas para crear modelos que relacionan y caracterizan las diferencias entre los datos normales y anomalías. Es posible evaluar los métodos supervisados en base a distintas métricas asociadas a la detección de las anomalías, estas pueden encontrarse en la sección 2.2.3. Un ejemplo de aprendizaje supervisado es la detección de fallas o mediciones anómalas dentro de una red de monitoreo utilizando la información de eventos del mismo tipo ocurridos y etiquetados con anterioridad

2.2.3. Métricas en modelos de detección supervisados

Para medir el desempeño de los distintos modelos que pueden ser usados para clasificar las anomalías en datos de series de tiempo hidrológicas, es necesario definir las principales métricas con la que se evaluarán los resultados obtenidos en el caso de los modelos supervisados, es decir, que utilizan datos previamente etiquetados que se considerarán como *groundtruth*.

Para definir las distintas componentes de las métricas de clasificación en un problema de detección de anomalías es posible realizar un paralelo con un problema de clasificación binario, en donde se debe decidir entre dos etiquetas. Así una medición multivariable de calidad de agua puede clasificarse como una anomalía y significaría un caso positivo (1) o ser un dato normal, negativo (0) tras procesarse mediante un modelo de detección de anomalías [35].

Así, se definen los siguientes conceptos en base a si el modelo acierta o no en función de cada caso como:

Definición 2.2 ***Verdadero Positivo (VP):*** *Dado el problema, se clasifica como* ***positivo*** *cuando realmente* ***corresponde*** *a un caso positivo.*

Definición 2.3 ***Falso Negativo (FN):*** *Dado el problema, se clasifica como* ***negativo*** *cuando realmente* ***no corresponde*** *a un caso negativo, sino a uno positivo.*

Definición 2.4 ***Verdadero Negativo (VN):*** *Dado el problema, se clasifica como* ***negativo*** *cuando realmente* ***corresponde*** *a un caso negativo.*

Definición 2.5 ***Falso Positivo (FP):*** *Dado el problema, se clasifica como* ***positivo*** *cuando realmente* ***no corresponde*** *a un caso positivo, sino a uno negativo.*

Una forma gráfica de poder comparar la composición final de la clasificación es a través de la matriz de confusión, la cual se se ilustra en la Figura 2.1. A partir de esta base de definiciones es posible definir también diferentes métricas más especializadas que relacionan el desempeño relativo a la hora de identificar correctamente los datos anómalos y normales.

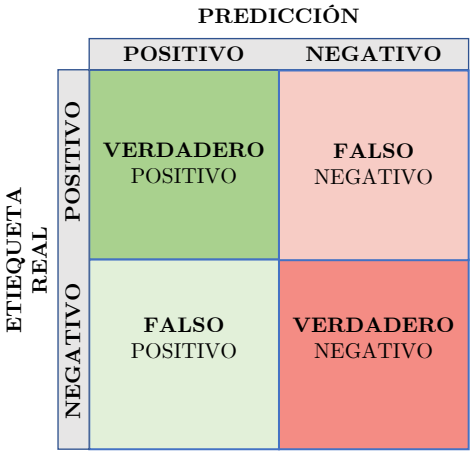


Figura 2.1: Matriz de confusión

Exactitud

Es una de las métricas más utilizadas en machine learning, se conoce también como *accuracy*. Define que tan exacto un modelo de clasificación binaria es en relación al total de muestras. Su valor numérico está dado por:

$$Accuracy\ score = \frac{VP + VN}{VP + VN + FP + FN}$$

Precisión

En un set de datos sesgado o con un desbalance de clases, métricas como el *accuracy* pueden esconder algunos problemas importantes de clasificación. Así, la precisión considera que proporción de verdaderos positivos se detectó en relación a todos los casos positivos detectados, es decir:

$$Precision\ score = \frac{VP}{VP + FP}$$

Exhaustividad

Ahora la exhaustividad o *Recall* representa si el modelo de clasificación es capaz de detectar bien los casos positivos sobre el total de casos positivos en las muestras, es decir:

$$Recall\ score = \frac{VP}{VP + FN}$$

Valor F_β

También conocido como *F1 Score* en el caso cuando $\beta = 1$, es una métrica que combina un valor tanto la precisión (P), como la exhaustividad o *recall* (R) y toma el promedio armónico de la forma:

$$F1\ Score = \frac{2PR}{P+R}$$

o equivalentemente:

$$F1\ Score = \frac{2VP}{2VP+FP+FN}$$

En el caso general su calcula como:

$$F_\beta\ Score = (1+\beta^2)\frac{PR}{(\beta^2 P)+R}$$

Para este trabajo se analiza la métrica F_β con un valor de $\beta = 2$, en adelante llamado F_2 pues concentra un esfuerzo mayor de optimización en minimizar falsos negativos que en minimizar falsos positivos. Es decir, se busca poder detectar la mayor cantidad de anomalías posible pues su potencial costo de no detección es mayor a un falso positivo.

Es importante destacar que el análisis de métricas como la *precision* y el *recall*, además de sus interacciones en los valores F_1 y F_2, toman mayor importancia en casos de datos desbalanceados, como es el caso de los problemas de detección de anomalías. Así métricas como el *accuracy* pueden esconder malos desempeños en la detección de anomalías cuando el desbalance es muy grande, es decir, cuando las anomalías son poco frecuentes.

El análisis de distintas métricas y su pertinencia en determinados casos o aplicaciones es clave pues permite comparar cuantitativamente distintos modelos en relación a su desempeño considerando distintos componentes en sus cálculos. Luego, para una determinada aplicación o caso de uso, una métrica puede ser más informativa que otra.

2.3. Detección de anomalías basadas en series de tiempo

Se analizan modelos basados tanto en aprendizaje supervisado como no supervisado, que pueden presentarse también como clasificadores. Esto debido a que se espera que los datos que se procesarán podrán estar o no etiquetados previamente por un experto. Es necesario también entregar algunas definiciones claves y conceptos para entender el tipo de datos con el que se trabaja y se entrenarán los modelos.

El estado de sistemas complejos puede representarse mediante múltiples mediciones de sensores de forma repetitiva y espaciada en el tiempo, es decir, mediante series de datos ordenados cronológicamente [36]. Así, una serie de tiempo representa una colección de datos agregados sobre un eje temporal en donde se refleja su evolución y eso es lo que las diferencia de otros tipos de series, desde un punto de vista matemático.

$$X = \{X_1, X_2, \dots X_t\}$$

Para formalizar aún más el concepto una definición dada por Morris en [37]:

Definición 2.6 *Una serie de tiempo es una secuencia de observaciones medidas de forma continua en el tiempo. Por lo general estas observaciones son tomadas en intervalos de tiempo*

equidistantes: $T = (t_0^d, t_1^d, ..., t_t^d)$, $d \in \mathbb{N}_+$, $t \in \mathbb{N}$, *donde d representa la dimensión y t el tiempo.*

Las series de tiempo pueden tener una forma regular o irregular dependiendo de la frecuencia en la que se muestrea u obtenga la información. Una medición instrumental de un equipo de forma periódica, sería una serie de tiempo regular. En cambio una serie de tiempo irregular puede ser una basada en eventos gatillados por agentes internos o externos del sistema como el paso de un automóvil por un pórtico de peaje o mediciones manuales sin una frecuencia fija.

Una serie de tiempo proveniente de datos obtenidos de un sensor es del tipo univariable $d = 1$. En el caso de datos provenientes de múltiples sensores, se genera una serie de tiempo multivariable $d > 1$. En este trabajo se consideran series de tiempo multivariable para resolver el problema de detección de anomalías [34].

El análisis de las series de tiempo toma una vital importancia en muchas áreas del conocimiento como las estadísticas, la econometría, la finanza, sismología, meteorología, geofísica y la **hidrogeología** para realizar tareas como pronósticos, diagnósticos y detección de anomalías para poder caracterizar eventos o situaciones de riesgo [38].

Otras aplicaciones que han tomado relevancia años es su utilización en el aprendizaje de máquinas y en la minería de datos para poder obtener información escondida dentro de grandes cantidades de datos. Existen también aplicaciones consiste como el procesamiento de señales para el diagnóstico y pronóstico de fallas. [39].

El problema de detección de anomalías puede entenderse también como un problema de clasificación binario, en donde se clasifica entre dos variables, por lo que los pueden utilizarse modelos de clasificación de distinto tipo en la tarea de detección de anomalías.

2.4. Métodos de Ensamble

Los métodos de ensamble, a diferencia de los métodos clásicos como la regresión logística, se utilizan en problemas de clasificación o regresión. En lugar de basarse en un único modelo altamente eficaz, emplean un conjunto de clasificadores o regresores base. Aunque algunos de estos clasificadores individuales pueden tener menor eficacia, el objetivo es obtener un modelo compuesto que aproveche las salidas de los diferentes modelos individuales para mejorar el rendimiento general. Un ejemplo de esto es la aplicación de una técnica de combinación simple, como la "votación por mayoría", que se ilustra en la Figura 2.2.

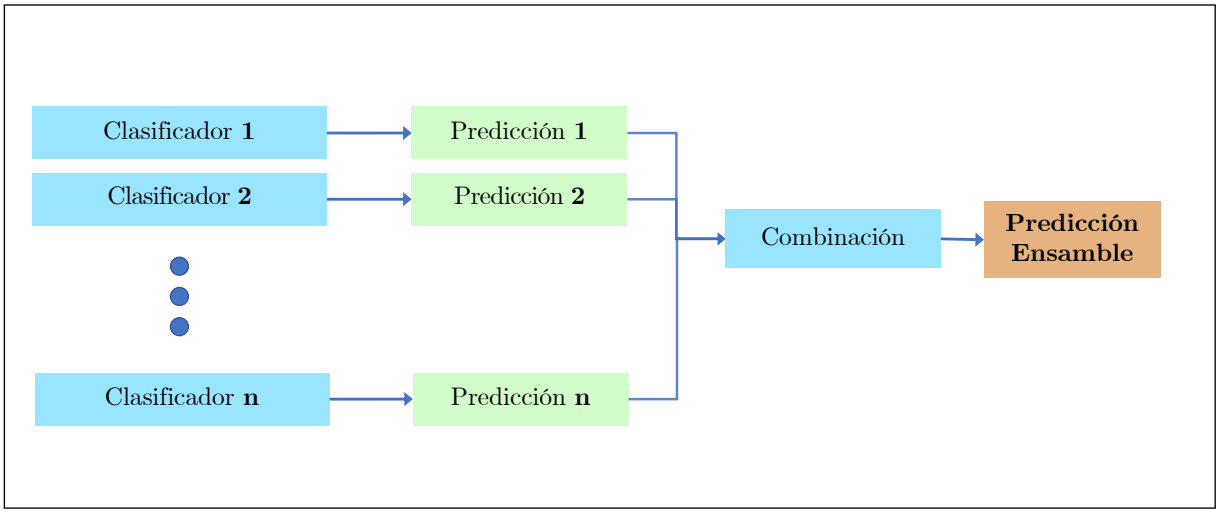


Figura 2.2: Modelo de ensamble de n clasificadores.

Una idea común de los métodos de ensamble es que están formados por múltiples clasificadores y las distintas variantes de ensambles difieren en la metodología con la que se entrenan los clasificadores bases y en la estrategia con la que se integran las salidas de cada uno de estos. Otra noción importante respecto a los ensambles es que en la práctica buscan poder generalizar de mejor manera y reducir la incertidumbre asociada a los sesgos individuales de un clasificador o de un tipo de clasificador.

De forma general los diferentes métodos de ensambles que se abordan en este trabajo pueden agruparse dentro de las siguientes categorías:

- **Boosting**: Estos algoritmos se basan en el entrenamiento secuencial de un modelo inicial de baja complejidad, generalmente árboles de decisión, que genera predicciones en todas las muestras. Se asignan ponderaciones de entrenamiento, inicialmente iguales a cada muestra, y luego va aumentando la ponderación de aquellas muestras con un error mayor para retroalimentar un siguiente clasificador. En cada iteración se almacena el desempeño general de cada clasificador para la ponderación final. Una de sus primeras aplicaciones consiste en las de tipo *Adaptive Boosting* [40]. Se desarrollan luego también las basadas en *Gradient Boosting* [41] que optimizan una función de pérdida de forma secuencial y que busca mejorar el desempeño en cada iteración.

- **Bagging:** Corresponden a técnicas basadas en el entrenamiento mediante agregaciones de *bootstrap* [42] y es un *meta-modelo* que busca generalizar de mejor manera entrenando distintas variaciones de un algoritmo inicial al submuestrear datos de entrenamiento para provocar inestabilidad en los clasificadores y combinar sus salidas en paralelo [43]. La etapa de agregación de los estimadores se basa por lo general en votación por mayoría o un promedio de ellas. Un modelo que utiliza esta técnica para entrenar es *Random Forest* [44].

- **Stacking**: También se le conoce como *Stacked Generalization* o apilamiento de modelos [23] y es una técnica que aborda la agregación como un proceso de aprendizaje en si. Su funcionamiento consiste en entrenar un *meta-modelo*, que es un modelo que recibe la salida y se entrena con las salidas de los modelos que componen el ensamble, para

encontrar la mejor relación entre la salida de los modelos que componen el ensamble y la etiqueta que se busca predecir.

Los primeros estudios sobre metodologías de ensamble de clasificadores datan de los años 90 [23, 45, 43]. Desde entonces la investigación de este tipo de modelos ha seguido evolucionando y en la actualidad algunas de estas herramientas entregan resultados sobresalientes en competencias de clasificación y detección de anomalías en comunidades asociadas al procesamiento de datos, algunos de estos modelos son basados en *bagging* como lo es *random forests* [44] y los basados en *gradient boosting* como lo son *XGBoost* [46], *CatBoost* y *LightGBM* [47].

2.4.1. Modelos basados en Boosting

Son algortimos que utilizan un ensamble de múltiples clasificadores débiles, que se añaden de forma constante para mejorar el desempeño en cada iteración. Esta metodología consiste en que el modelo aumenta la ponderación de los casos en que se equivoca en clasificar en una siguiente iteración hasta que logra obtener la salida correcta.

Clasificadores basados en boosting son una alternativa a la utilización de redes neuronales en casos en donde no se cuenta con muchos datos o no se tiene un gran número de datos etiquetados. Generalmente estos clasificadores de baja eficacia son llamados *weak learners* o "clasificadores débiles" y su principal característica es que no pueden aprender problemas complejos, sin embargo, debido a su simplicidad se entrenan muy rápido y clasifican en muy poco tiempo. Frecuentemente se utilizan árboles de decisión [46], como es el caso de XGBoost (Extreme Gradient Boosting).

XGBoost

XGBoost ha sido utilizando extensamente para resolver problemas de clasificación, predicción y detección de anomalías en múltiples aplicaciones y con datos proveniente de las más diversas disciplinas debido a una gran velocidad de entrenamiento y su capacidad de entrenar y escalar eficientemente en conjuntos de datos con millones de muestras [46]. En tiempos de entrenamiento supera incluso a otros algoritmos basados en *Gradient Boosting Decision Trees* o *Random Forest*.

Su funcionamiento se basa en que su salida combina las salidas de múltiples árboles de decisión mediante un ensamble que se construye de forma secuencial. Cada árbol mapea la muestra hasta una salida en sus nodos para luego compararla con la etiqueta real, luego a través de una minimización de una función de perdida actualiza sus errores residuales en cada iteración.

XGBoost se basa en el algoritmo de Gradient Boosting Machine (GBM), por lo que comparten algunos conceptos y principios básicos. Sin embargo, existen algunas diferencias importantes entre XGBoost y los modelos de GBM:

Regularización: incluye una regularización adicional en su algoritmo, lo que le permite controlar la complejidad del modelo y reducir el sobreajuste. Utiliza términos de regularización como la penalización L1 (lasso) y L2 (ridge) para evitar el crecimiento excesivo del árbol

y mejorar la generalización del modelo.

Manejo de valores faltantes: tiene un manejo integrado de valores faltantes en los datos de entrada. Puede aprender de manera automática cómo tratar los valores ausentes y hacer divisiones en los nodos correspondientes de los árboles de decisión.

Paralelización: está diseñado para aprovechar el paralelismo y el rendimiento eficiente. Puede utilizar múltiples núcleos de procesamiento para realizar el entrenamiento y la predicción más rápidamente.

Función objetivo personalizada: permite definir una función objetivo personalizada, lo que brinda flexibilidad para adaptar el modelo a problemas específicos y optimizar métricas específicas de acuerdo con las necesidades del usuario.

Optimización del rendimiento: utiliza técnicas avanzadas de optimización, como el muestreo aproximado y la poda temprana de árboles, para acelerar el proceso de entrenamiento y mejorar la eficiencia computacional.

En el ejemplo de la Figura 2.3 se muestra un problema de clasificación binaria en donde se busca predecir si a una persona le gustaría jugar un videojuego en función de sus edad y si utiliza o no computador. Así una predicción de salida de un método de ensamble está compuesta por la suma de todas las puntuaciones o salidas de cada árbol de decisión individual para cada persona. Otro factor importante es que todos los árboles de salida se complementan.

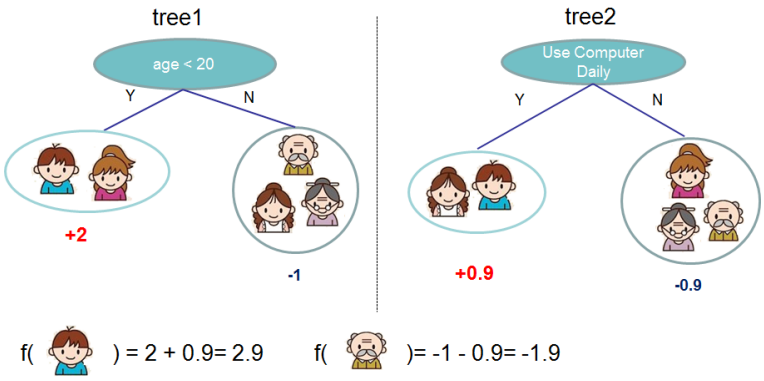


Figura 2.3: Modelo de emsable de árboles de decisión. El valor para la predicción final para cada muestra es la suma de las predicciones de cada árbol [46]

Matemáticamente esto se puede escribir de la siguiente forma: Con K el número de árboles, f es un funcional en el espacio de F, y F es el conjunto de todos los posibles Árboles de Clasificación y Regresión (CARTs, por sus siglas en inglés). l representa la función de pérdida o la diferencia entre la predicción y el *target* y Ω un término de regularización para evitar sobreajuste. Por último la función a minimizar queda de la forma:

$$\text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.1}$$

XGBoost ha sido utilizado en el análisis de calidad de agua abarcan desde predecir índices de calidad de agua [48] con desempeños sobre el 80 %. También se reportan aplicaciones híbridas, combinados con otros modelos para realizar predicciones a corto plazo de calidad de aguas en datos de ríos altamente contaminados [49]. En el campo de detección de intrusos o de ataques en redes [38], que puede considerarse un tipo de detección de anomalías, presenta desempeños con accuracys de 98.7 % con bajas tasas de error.

2.4.2. Modelos basados en Bagging

Random Forest

Corresponde a un modelo de clasificación supervisado basado en una combinación de árboles de decisión inicializados de forma aleatoria y entrenados en un submuestreo del set de datos de entrenamiento [44]. Su funcionamiento se basa en que el error de generalización asociado al bosque de árboles depende del desempeño de los árboles individuales que componen el bosque y sus relaciones, supliendo una de las principales falencias de los árboles de decisión, el sobreajuste.

Para el entrenamiento este método genera un submuestreo con una cantidad de datos inferior al conjunto de entrenamientos para cada árbol de decisión a través del método de *boostrapping* [42]. En el caso de un Random Forest, se inicializa una cantidad de estimadores igual al número de árboles configurado y cada uno de ellos se entrenará con un subconjunto diferente de los datos, adquiriendo la capacidad de generalizar y evitando sobre-ajustarse a los datos. Un ejemplo de submuestreo con el que se entrena cada árbol se aprecia en la Figura 2.4.

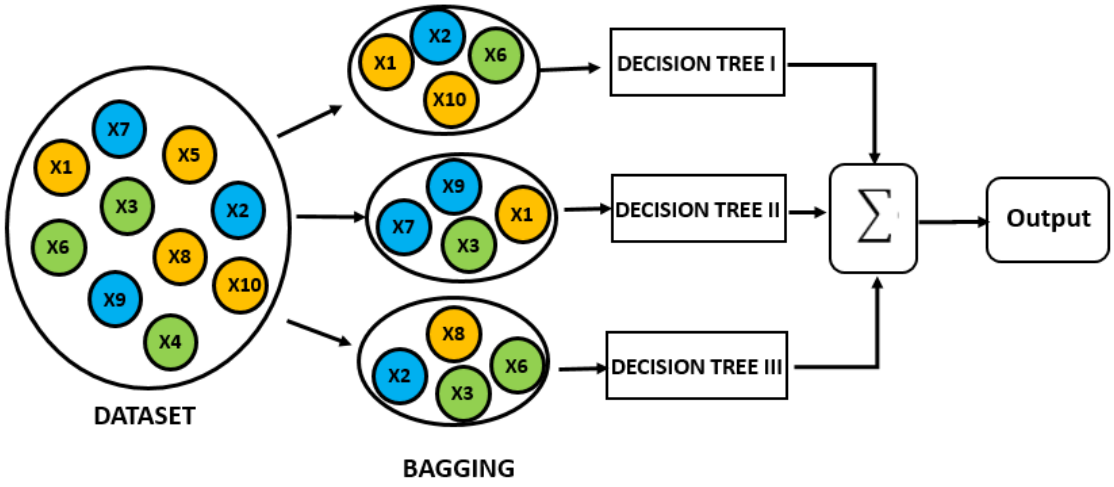


Figura 2.4: Ejemplo de Random Forest y metodología de submuestro *Boostrapping* (también conocida como Bagging) y la integración de las salidas.

Luego del entrenamiento de los múltiples clasificadores se ensamblan todos los árboles de decisión a través de la técnica de bagging. En este ensamble se consideran ahora pasar todos los datos del conjunto de entrenamiento a través de todos los árboles y sus salidas se combinan a través de una votación simple, que no es más que una suma ponderada de los árboles

que entreguen resultados más seguros, en otros casos se utiliza "Votación por mayoría" o *Soft Voting* al comparar sus salidas en probabilidades versus un umbral fijado.

2.4.3. Métodos de apilamiento de modelos (Stacking)

Desde la exploración de los ensambles a través de técnicas como las basadas en *bagging* o *boosting* [50] y más recientemente la gran atención de variantes basadas en *gradient boosting* [51], se ha ampliado la exploración a las distintas maneras de poder ensamblar estas técnicas para aprovechar las ventajas de cada una de ellas y suplir sus falencias, incluso se han combinando técnicas basadas en aprendizaje supervisado y no supervisado [52].

Las técnicas de stacking o de apilamiento se han estudiado desde la década de los 90's [23] y buscan reducir la varianza de otros estimadores, incluso basados en ensambles. Existen diferentes variantes también en torno al entrenamiento de modelos apilados debido a qué pueden utilizar distintas formas de elegir a los modelos bases idóneos, además de distintas maneras de combinar sus salidas. Entre las más destacadas se encuentra el uso de un *metamodelo*, que a su vez se entrena utilizando como conjunto de datos las salidas de los clasificadores bases que genera el ensamble, un ejemplo de esto en un caso aplicado puede verse en la Figura 2.5.

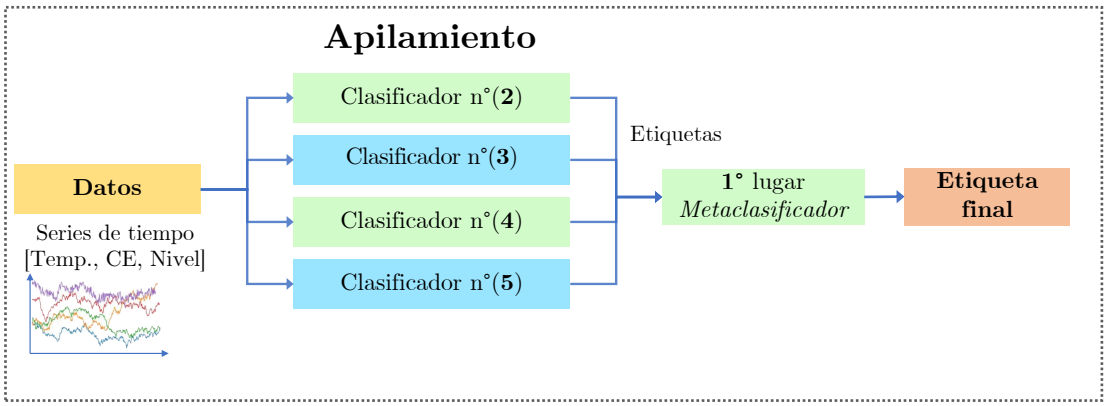


Figura 2.5: Ejemplo de estructura de apilamiento de modelos.

Las salidas de estos modelos de ensamble también pueden ser mezclados por reglas más simples como las basadas en *blending* o combinación directa, que significa que pueden utilizar reglas como una ponderación de probabilidades o votación por mayoría (también conocida como *Soft Voting*) para asignar una etiqueta de clasificación, como en el caso de la tarea de detección de anomalías. Esto puede apreciarse con un ejemplo de uso en la Figura 2.6.

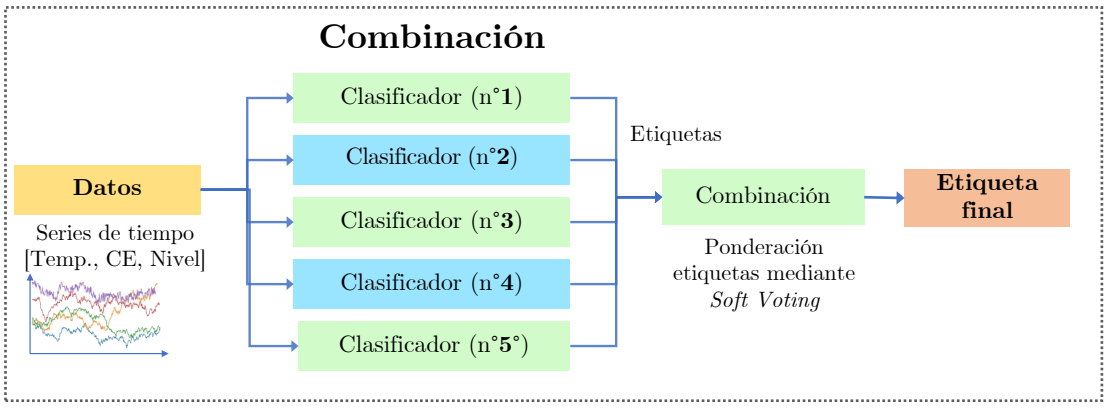


Figura 2.6: Ejemplo de estructura de apilamiento de modelos.

Capítulo 3

Metodología de detección de anomalías en acuíferos basado en ensambles

A continuación se describe la estructura y la propuesta metodológica para la integración de múltiples modelos de detección de anomalías basadas en series de tiempo para calidad de agua y el cumplimiento de los objetivos del trabajo propuestos en la Sección 1.4.

De forma general el trabajo comienza con el pre-procesamiento y limpieza de los datos para continuar luego con el entrenamiento del conjunto inicial de modelos de detección de anomalías en series de tiempo hidrogeológicas para luego realizar la selección del mejor modelo individual o conjunto de modelos, mediante apilamiento.

Con la selección del modelo con mejor desempeño, la salida de los datos de validación se pueden analizar en términos de métricas individuales, una matriz de confusión y resultados gráficos. Para todo el procesamiento de los datos, entrenamiento y visualizaciones se trabaja con el lenguaje de programación *Python* y utilizando la librería de Aprendizaje de máquinas automatizado o (*AutoML*) de código abierto llamada *PyCaret*.

Una vista general del flujo de la información y el procesamiento de los datos presentados en este trabajo se presenta en la Figura 3.1.

Los resultados de este trabajo asociados al desarrollo de la plataforma web interactiva y la presentación del sistema mediante interfaz gráfica se aborda junto a la integración de esta metodología en el Capítulo 6.

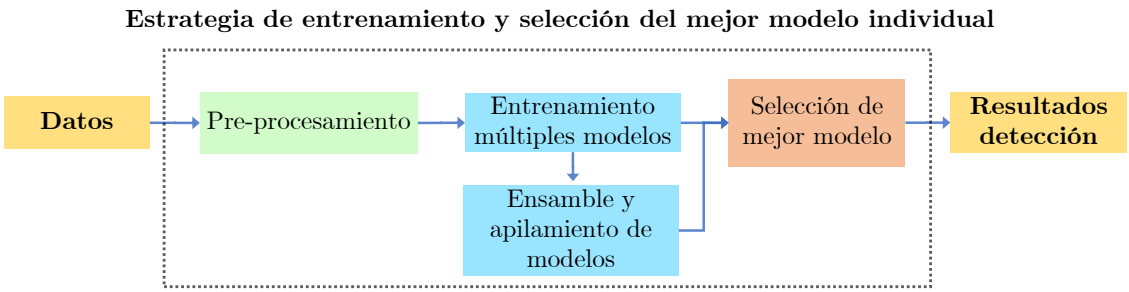


Figura 3.1: Estructura del flujo de información y procesamiento de datos.

3.1. Pre-procesamiento

Los modelos de detección de anomalías basados en *Machine Learning* necesitan, en algunos casos, recibir los datos con una estructura común, para que al momento de entrenar varios modelos los resultados sean comparables. En este caso, se realizan los siguientes pasos:

1. Se identifican la columna de objetivo o la etiqueta de anomalía y las columnas correspondientes a las características.
2. Se buscan datos perdidos, nulos (NaNs) o erróneos y se reemplazan con un valor diferente dado el tipo de dato. Para datos numéricos se reemplazan por el promedio y en el caso de datos categóricos se mantiene el último valor constante.
3. Se normalizan y escalan las características numéricas utilizando el Z-Score [53], que transforma los datos respecto a su desviación estándar de la media.
4. Se realiza un *Shuffle* o reordenamiento de los datos de forma aleatoria.
5. Se separan los datos en un 70 % para el conjunto de entrenamiento y un 30 % para validación.
6. Se crean 10 sub muestras a partir de los datos de entrenamiento para realizar una validación cruzada o *Cross Validation* mediante la estrategia denominada *KFold* [54].

3.2. Entrenamiento de modelos

Para el entrenamiento del conjunto de modelos en este trabajo se consideran los presentados en la sección 2.3, sin embargo, al ser un sistema modular podrían agregarse más o cambiarse otros en futuras iteraciones de este trabajo. Con esto para cada modelo a entrenar se realizan los siguientes pasos:

1. Entrenamiento del modelo base con parámetros por defecto y utilizando la estrategia de *Cross Validation* para calcular las métricas de las 10 submuestras para cada clasificador. Esto se realiza entrenando los datos con 9 de las 10 muestras y calculando las métricas con la que se dejó afuera y así suscesitvamente para las 10 posibles combinaciones. Un ejemplo puede encontrarse en el Anexo A.1.

2. Se promedian las métricas de desempeño calculada para cada uno de las 10-submuestras y se considera el desempeño general del modelo como el promedio de las métricas para cada sub-muestra.

Al recorrer toda la lista de los modelos se genera una grilla de búsqueda que contiene las métricas para cada uno y se almacena en una tabla. Un ejemplo de la salida de esta etapa puede apreciarse en la Tabla 3.1.

Tabla 3.1: Ejemplo de grilla de búsqueda de modelos

Index	Model Name	Accuracy	Recall	Prec.	F1
0	Logistic Regression	0.8729	0.2080	1.0000	0.3433
3	Decision Tree Classifier	0.9333	0.7907	0.7931	0.7916
4	SVM - Linear Kernel	0.8698	0.1887	1.0000	0.3113
6	Random Forest Classifier	0.9473	0.7420	0.9138	0.8186
8	Ada Boost Classifier	0.9069	0.4647	0.9134	0.6153
9	Gradient Boosting Classifier	0.9349	0.6213	0.9581	0.7532
11	Extra Trees Classifier	0.9448	0.7167	0.9227	0.8063
12	Extreme Gradient Boosting	0.9480	0.7633	0.8984	0.8249
13	Light Gradient Boosting Machine	0.9500	0.7567	0.9188	0.8288
14	CatBoost Classifier	0.9517	0.7600	0.9261	0.8344

3.3. Selección de modelos

La validación del desempeño de los distintos modelos aplicados a la tarea de detección de anomalías se lleva a cabo de forma iterativa. Inicialmente, se busca identificar los modelos con el mejor rendimiento utilizando un conjunto de hiperparámetros predeterminados. Luego, se realiza una búsqueda y optimización bayesiana [55] de los hiperparámetros basada en métricas de desempeño, en particular, se utiliza el *valor F2*. Este proceso se aplica a los tres modelos más prometedores mediante la librería *scikit-optimize*. La validación de desempeño de los distintos modelos aplicados a la tarea de detección de anomalías se realiza de forma iterativa. Se comienza la búsqueda de los modelos con mejor desempeño en un conjunto de hiperparámetros por defecto y luego mediante una búsqueda y optimización bayesiana de hiperparámetros por métricas de desempeño [55], en este caso del *valor F2*, de los 3 mejores modelos a través de la librería *scikit-optimize* .

Con la selección de los modelos con mejor desempeño se entrena el modelo de apilamiento, utilizando el modelo con mejor desempeño como el meta-modelo ensamblador y los otros dos como modelos débiles como se muestra en el Capítulo 2. Con esto realizado se analiza la grilla

de búsqueda de modelos por el mejor modelo apilado o el mejor modelo individual basados en la comparación de la métrica *F2*.

La estructura final de la metodología y su integración dentro de la plataforma puede apreciarse a través en el esquema de la Figura 3.2.

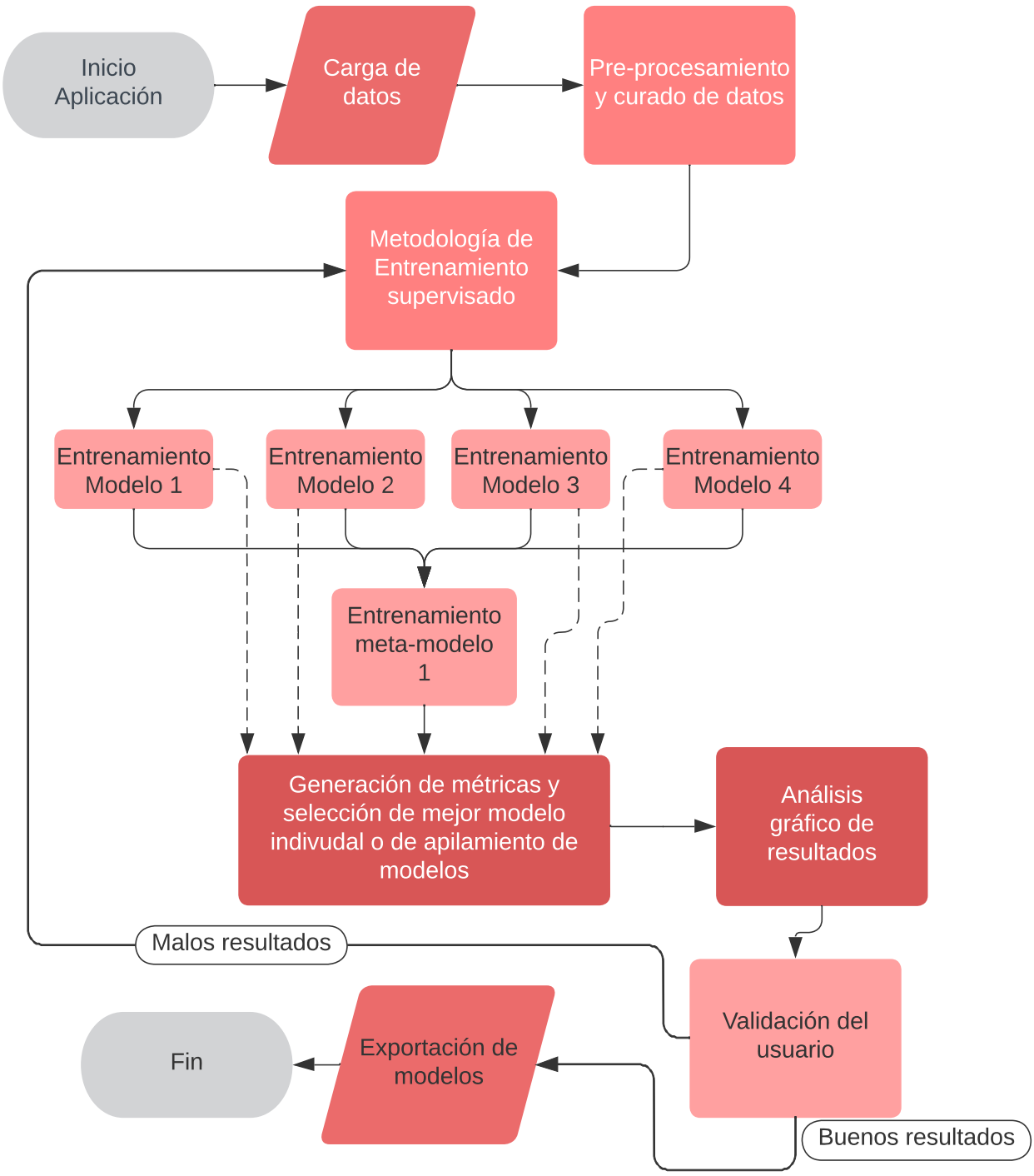


Figura 3.2: Esquema integración plataforma y sistema automático de detección de anomalías.

Capítulo 4

Caso de estudio: Pozo monitoreado en localidad de Horcón

Para el desarrollo de los diferentes modelos a estudiar se trabaja con una base de datos proporcionada por la *Hydrosciences de la Université de Montpellier* para el desarrollo del proyecto. Estos datos contienen la información de monitoreo de un pozo ubicado en la localidad de Horcón, Región de Coquimbo. La base de datos contiene mediciones de parámetros fisicoquímicos del agua registrados con una frecuencia de muestreo de una hora desde **marzo de 2013 a febrero de 2017**.

Esta base de datos es presentada como una serie de tiempo tabulada por parámetros, que contiene también tres columnas adicionales de etiquetas, una para cada variable, en donde se indican si un experto considera que esas mediciones corresponden a una anomalía puntual (valor de 1) o a un dato de operación normal (0). A continuación, se presentan algunas de sus principales características y la visualización de sus parámetros monitoreados se encuentran en la Figura 4.1.

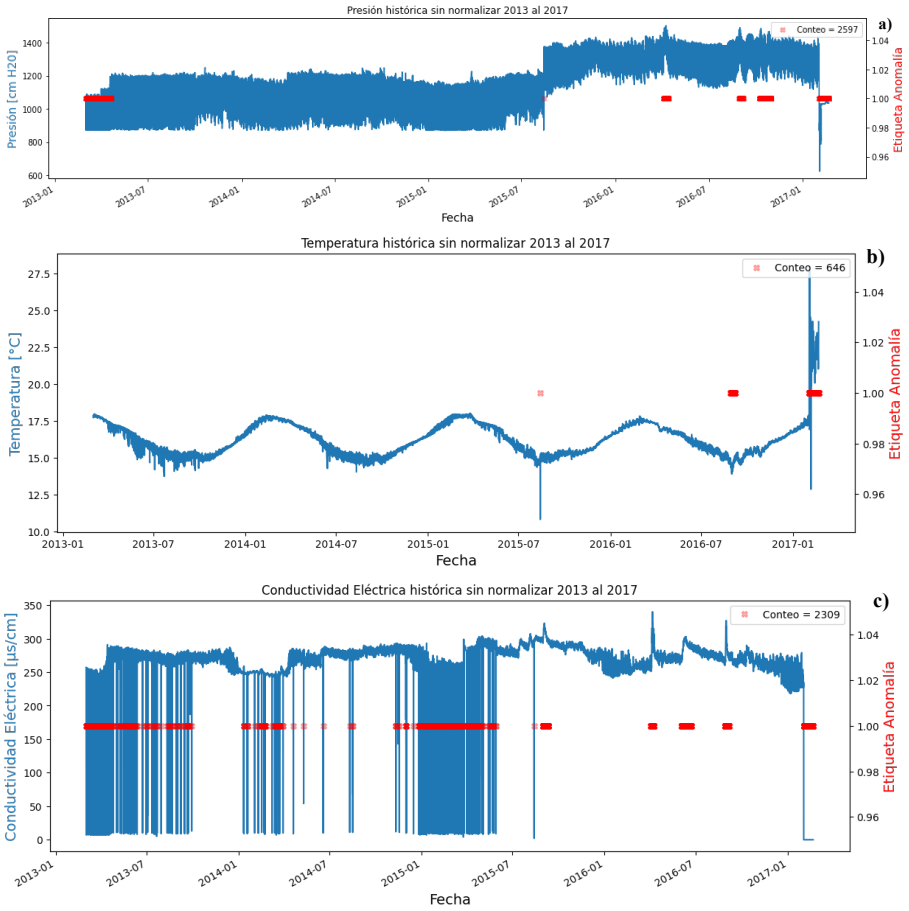


Figura 4.1: Base de datos de Horcón etiquetado por experto, en rojo una superposición de las marcas temporales en donde se ha etiquetado un dato anómalo. a) Presión; b) Temperatura ; c) Conductividad eléctrica

4.1. Datos disponibles

El set de datos está compuesto por 34.827 mediciones sin mediciones perdidas y ningún NaN o dato nulo, debido a que ya ha pasado por una etapa de pre-procesamiento. Una descripción estadística más detallada de cada una de las variables fisicoquímicas monitoreadas, sin considerar las columnas correspondientes a las etiquetas de anomalías, se presentan en la Tabla 4.1.

Tabla 4.1: Descripción estadística de la base de datos y sus variables monitoreadas

	Presión [cm H2O]	Temperatura [°C]	EC [µs/cm]
conteo	34827.000000	34827.000000	34827.000000
media	1155.082798	16.247972	262.208057
std	155.703183	1.173972	51.467876
min	622.600000	10.800000	0.000000
25 %	1030.700000	15.310000	260.000000
50 %	1158.900000	16.100000	273.000000
75 %	1284.800000	17.070000	281.000000
max	1501.900000	27.990000	340.000000

La altura de columna de agua con unidad de medida de presión, varía desde 620 cm hasta 1500 *cm* y con una media de 1150 *cm*. Es posible apreciar, al analizar la información diaria que se presentan varias oscilaciones en el valor de la columna de agua del pozo como puede apreciarse en la Figura 4.2, en este día se tiene una media de 275 cm de columna de agua y una desviación estándar de 0.83. Esto es muy probable que se deba a la extracción con bombas de forma intermitente que se realiza en estos pozos. Además, es posible apreciar que en el período aproximado de septiembre-octubre de 2015 se presenta un cambio considerable en el nivel promedio de la presión en el tiempo y se reduce por otro lado la desviación estándar de sus valores de forma definitiva marcando dos períodos bien diferenciados.

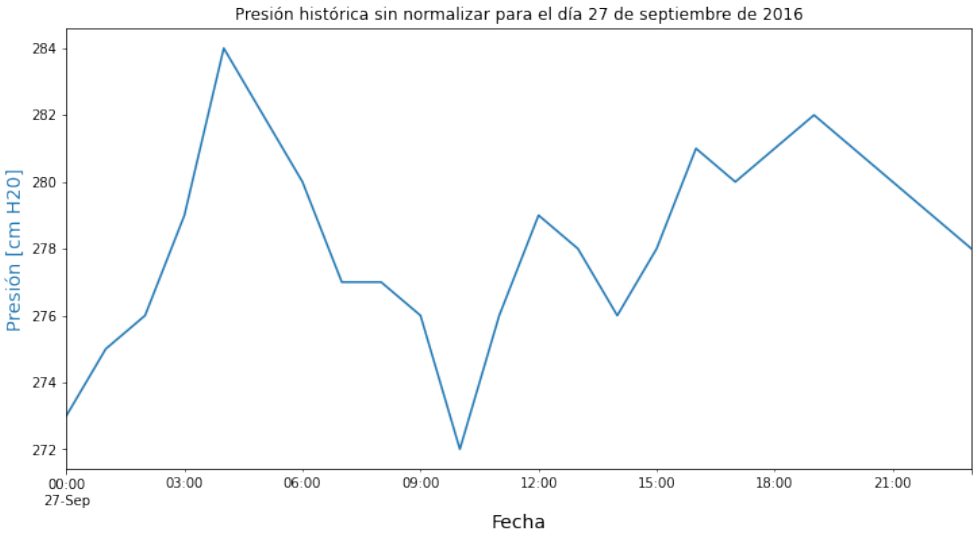


Figura 4.2: Cambios en la presión en un día

Por otro lado la temperatura presenta variaciones menores durante el día y es debido a que al estar bajo tierra tiene menos incidencia la temperatura ambiental. Sin embargo si se presentan variaciones de tipo estacional, pues las temperaturas disminuyen desde el período de abril a septiembre y luego aumentan desde octubre a marzo, es bastante directo ver la forma oscilatoria periódica marcada por esta estacionalidad en la Figura 4.1b.

Para esta base de datos la conductividad eléctrica (CE) (Figura 4.1c) presenta múltiples oscilaciones en sus mediciones diarias con valores que en ocasiones son muy cercanos a 0 o incluso 0, es posible apreciar una correlación directa de estos valores con un bajo nivel de columna de agua del pozo en los momentos en que se obtiene esta baja medición de conductividad eléctrica. Esto puede apreciarse de mejor manera en la Figura 4.3, analizando el período comprendido entre marzo 2013 a septiembre 2015 prácticamente todos los datos que presentan una anomalía asociada al parámetro de CE se realizan cuando el pozo presenta un nivel de columna de agua constante que además representa el mínimo de ese período. Se interpreta que el sensor de conductividad eléctrica estuviera realizando mediciones en el aire debido a que el nivel del pozo no alcanzaba a cubrir el sensor. Esto podría explicar el por qué del drástico cambio de niveles con el período siguiente asociado a una manipulación de los instrumentos de medición al interior del pozo o una reubicación.

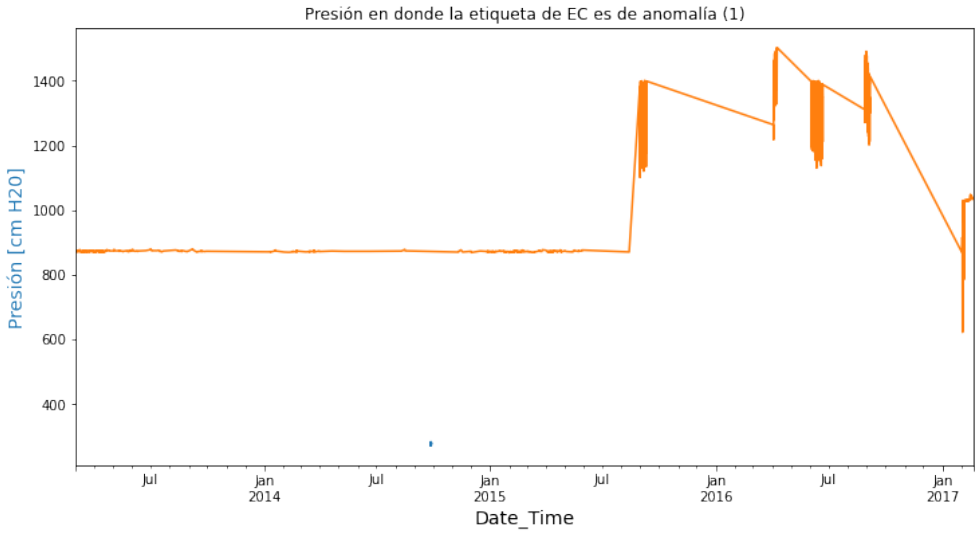


Figura 4.3: Nivel de columna de agua cuando la etiqueta asociada a la variable de conductividad eléctrica indica anomalía (1)

Es posible visualizar estos datos en relación a sus valores diarios mediante su media y la desviación estándar utilizando una ventana deslizante de tiempo. Esto se realiza considerando los datos agrupados cada 24 horas y calculando su media y su desviación estándar, comparándose gráficamente con el valor en bruto en la Figura 4.4. Además se presentan también los gráficos de pares variables en la Figura 4.5a) y un mapa de calor para analizar correlaciones de forma gráfica en la Figura 4.5b).

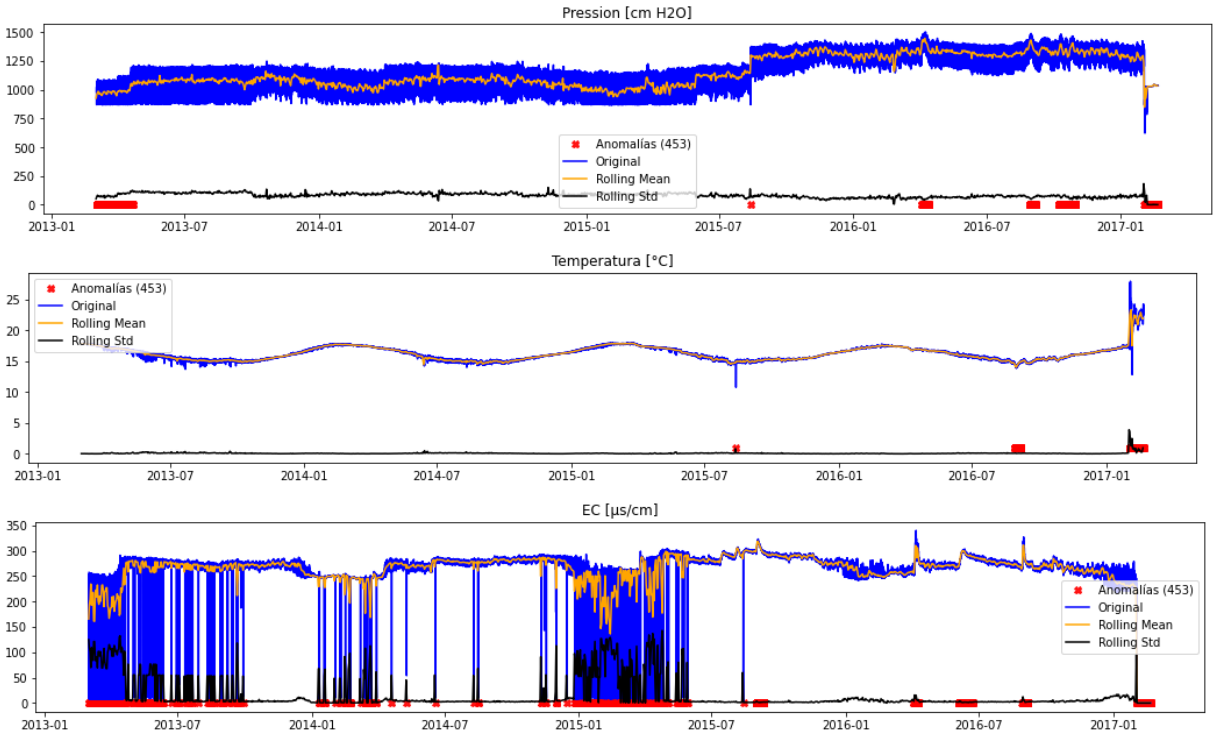


Figura 4.4: Análisis por ventana deslizante de 24 horas. Original en azul, promedio deslizante en amarillo, desviación estándar deslizante en negro, etiquetas de anomalías en rojo.

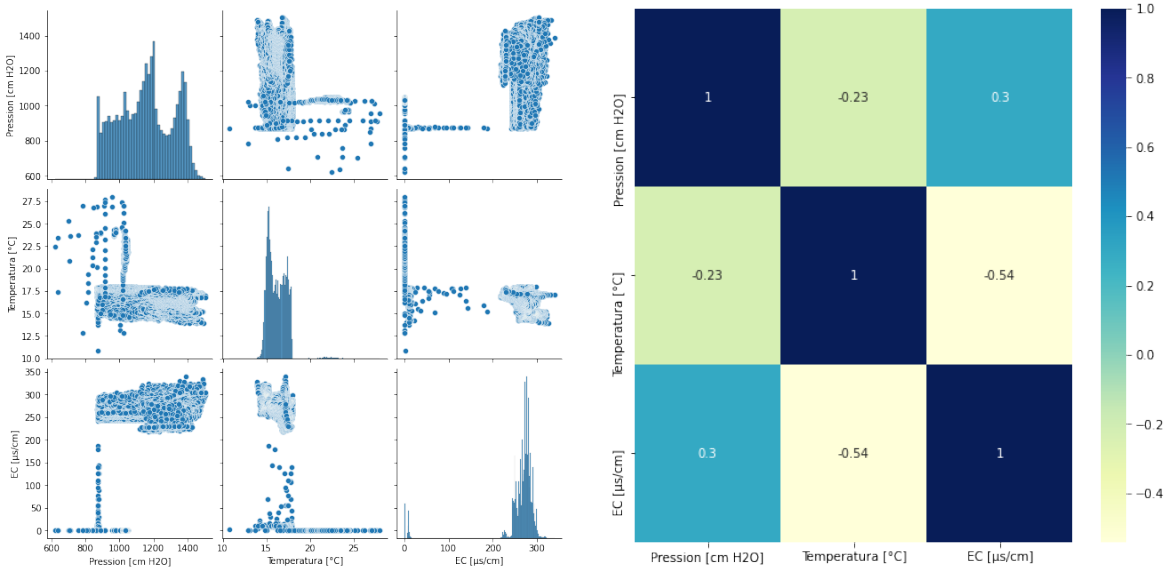


Figura 4.5: Relaciones entre características. a) Gráficos de pares de variables b) Mapa de calor

Es importante destacar aquí que para el caso del análisis con ventanas deslizantes es posible ver, en el caso de la Figura 4.4 c) que la mayoría de las anomalías se condicen directamente con una alta desviación estándar y por tanto un cambio notorio también en el promedio diario. Esto ocurre también, pero en menor magnitud, en la temperatura Figura 4.4 b) y no se aprecia una correlación tan directa para el caso de la presión en la Figura 4.4 a).

Luego a través de los plots por pares y sus histogramas, que se encuentran en la diagonal, es posible ver primero como sus distribuciones no se asemejan demasiado a una distribución normal. Luego es posible ver que hay algunas correlaciones entre variables, que podrían explicarse por algún fenómeno como el encontrado anteriormente para el caso de la conductividad eléctrica y el nivel o simplemente por su dependencia física.

Capítulo 5

Resultados detección de anomalías

Los experimentos para analizar el desempeño de la metodología de detección de anomalías se procesan los datos del caso de estudio de forma local, y no a través de plataforma web. Luego para generar variabilidad en los experimentos y analizar la capacidad de generalizar se separa en dos escenarios distintos el set de datos original en el Capítulo 4. Además se prueba también la metodología con todos los datos. Una muestra de la división de los datos analizados y las ventanas de tiempo que generan, en este caso para el parámetro de presión, se muestra en la Figura 5.1.

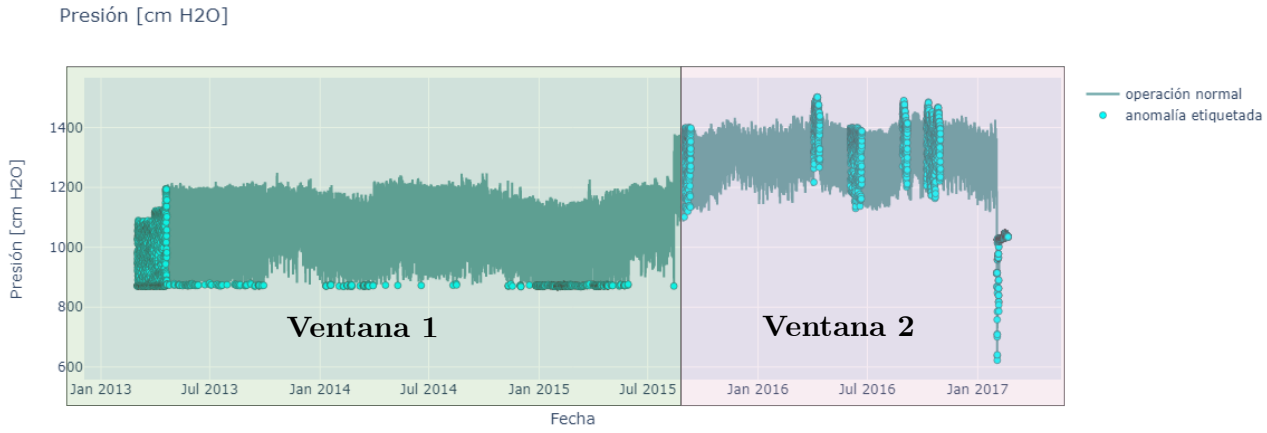


Figura 5.1: Separación del set de datos original en dos ventanas de tiempo.

Los análisis cuantitativos se realizan a través de comparaciones según la métrica *Valor F2* y se analizan también el valor *F1*, la *precisión* y la *exhaustividad* de los métodos, pues la tarea de detección de anomalías suele presentar un desbalance de clases etiquetadas considerable como se encuentra en la exploración y se ve gráficamente en la Figura 6.5. Se analiza la capacidad de generalización de los distintos modelos individuales con un énfasis en tener una mayor tasa de detección de anomalías, es decir ponderando más la exhaustividad que la precisión pues se calcula sobre el total de casos positivos reales.

5.1. Primer Caso

El primer subconjunto de datos toma los datos desde el 2 de marzo de 2013 hasta el 13 de agosto de 2015, para hacer 894 días en total. En este periodo se tienen 21470 muestras de datos con frecuencia de una hora para presión o columna de agua, temperatura y conductividad eléctrica. Los datos pueden apreciarse en la Figura 5.2. Este conjunto de datos tiene una proporción de 8,11 % de anomalías etiquetadas en su totalidad.

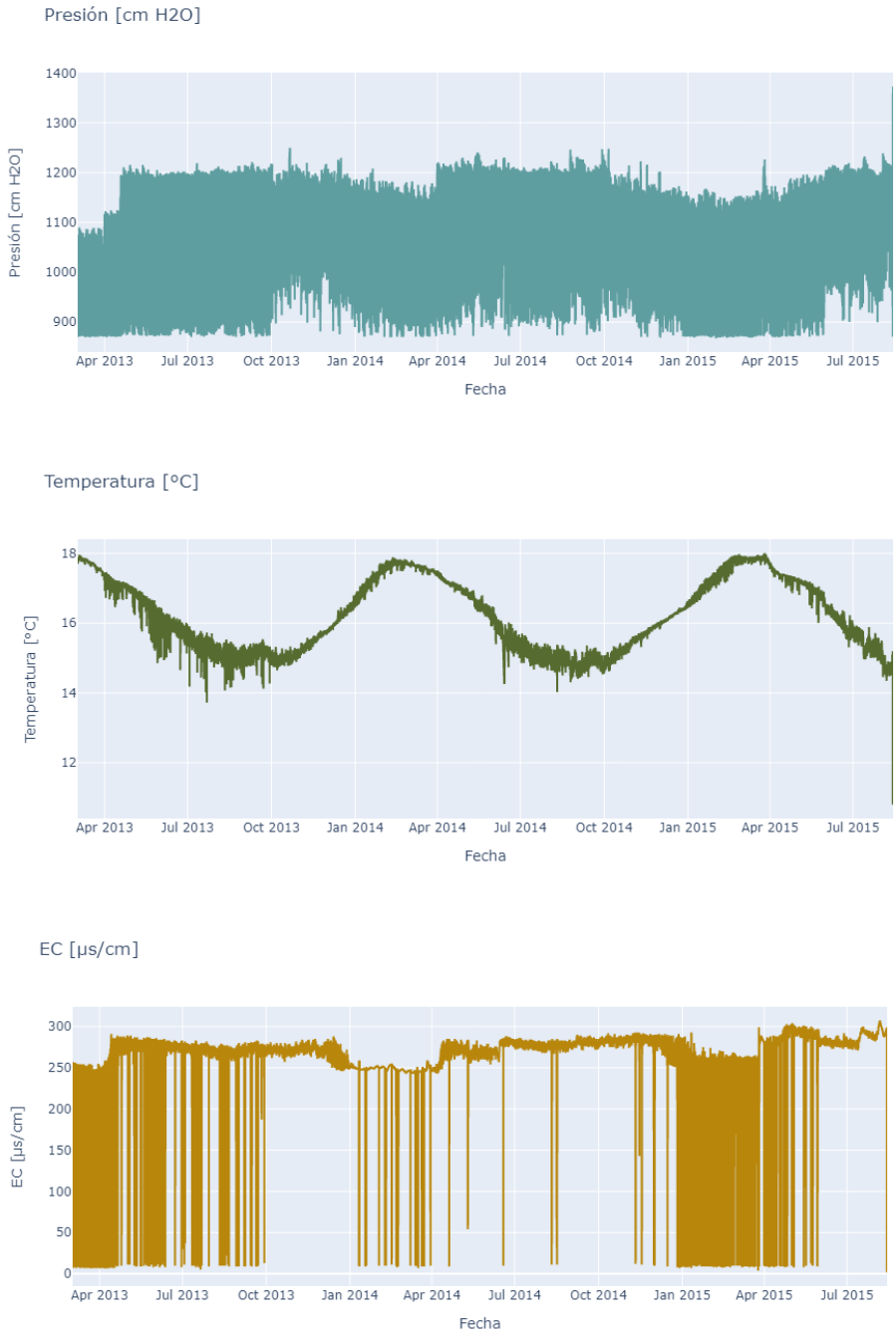


Figura 5.2: Datos asociados al primer caso analizado 2013-2015.

5.1.1. Entrenamiento y selección de modelos

Como se detalló en el Capítulo 3 el entrenamiento comienza luego de la etapa de preprocesamiento común y en ella se genera una grilla de búsqueda de modelos inicial para los datos de entrenamiento cargados. En este caso la grilla de búsqueda de modelos y sus métricas puede apreciarse en la Figura 5.4. El conjunto inicial de modelos son los que presentan índice 1 al 15, aquí se incluyen una lista inicial de modelos que no tienen un buen desempeño y por tanto no se trabaja con ellos en etapas posteriores. La escala de calor utilizada para analizar de mejor manera las métricas se aprecia en la Figura 5.3.

Figura 5.3: Escala de color.

Index	Model Name	Accuracy	Recall	Prec.	F1	f2score
14	CatBoost Classifier	0.976600	0.842900	0.867600	0.854500	0.847400
12	Extreme Gradient Boosting	0.976000	0.835500	0.865500	0.849800	0.841100
13	Light Gradient Boosting Machine	0.975700	0.836300	0.862000	0.848300	0.840900
3	Decision Tree Classifier	0.971700	0.827300	0.827100	0.826300	0.826700
6	Random Forest Classifier	0.975500	0.805900	0.884000	0.842300	0.819900
11	Extra Trees Classifier	0.974000	0.773900	0.892500	0.828300	0.794600
9	Gradient Boosting Classifier	0.970100	0.725300	0.887000	0.797200	0.752300
8	Ada Boost Classifier	0.960800	0.657900	0.822800	0.730800	0.685200
1	K Neighbors Classifier	0.956400	0.612700	0.805200	0.694800	0.642900
4	SVM - Linear Kernel	0.925700	0.542100	0.858200	0.580600	0.529000
0	Logistic Regression	0.955700	0.453100	1.000000	0.623000	0.508500
2	Naive Bayes	0.949200	0.462200	0.838500	0.595300	0.507500
7	Quadratic Discriminant Analysis	0.954600	0.453100	0.970500	0.617100	0.507000
10	Linear Discriminant Analysis	0.955600	0.451500	1.000000	0.621400	0.506900
5	Ridge Classifier	0.955500	0.450700	1.000000	0.620700	0.506100
15	Dummy Classifier	0.919100	0.000000	0.000000	0.000000	0.000000

Figura 5.4: Grilla de búsqueda inicial para primera ventana de tiempo.

Luego de esta búsqueda se realiza el proceso de optimización de los mejores cuatro modelos ordenados por el *F2 Score*. En este caso particular los modelos seleccionados para la optimización de hiperparámetros y posteriormente pasar a ser parte de un ensamble por votación y un modelo apilado son CatBoost como modelo ensamblador y con XGBoost, LGBM y Decision Tree como clasificadores. Aquí se prueban la eficacia de los modelos Stacking Classifier y Voting Classifier que corresponden al modelo apilado y al ensamble por votación simple.

Ambos modelos se prueban en dos configuraciones cada uno, una es utilizando el top 4 de modelos previo a la optimización y otro ya considerando los modelos con sus hiperparámetros optimizados. La tabla resumen con las métricas calculadas para los datos de entrenamiento y aplicando la estrategia de validación dada por *10-fold cross validation* se muestra en la Figura 5.5.

Index	Model Name	Accuracy	Recall	Prec.	F1	F2score
17	Extreme Gradient Boosting Optimized	0.973500	0.875000	0.815500	0.843300	0.861800
14	CatBoost Classifier	0.976600	0.842900	0.867600	0.854500	0.847400
21	Voting Classifier	0.976200	0.840500	0.864300	0.851500	0.844700
12	Extreme Gradient Boosting	0.976000	0.835500	0.865500	0.849800	0.841100
13	Light Gradient Boosting Machine	0.975700	0.836300	0.862000	0.848300	0.840900
22	Stacking Classifier Optimized	0.975100	0.833900	0.857500	0.844900	0.838100
20	Stacking Classifier	0.974400	0.828900	0.854600	0.840500	0.833300
23	Voting Classifier Optimized	0.976500	0.814100	0.889000	0.849300	0.827700
3	Decision Tree Classifier	0.971700	0.827300	0.827100	0.826300	0.826700
16	CatBoost Classifier Optimized	0.975500	0.809200	0.881400	0.843000	0.822200
6	Random Forest Classifier	0.975500	0.805900	0.884000	0.842300	0.819900
19	Decision Tree Classifier Optimized	0.971900	0.801000	0.846200	0.821900	0.809000
11	Extra Trees Classifier	0.974000	0.773900	0.892500	0.828300	0.794600
9	Gradient Boosting Classifier	0.970100	0.725300	0.887000	0.797200	0.752300
18	Light Gradient Boosting Machine Optimized	0.968100	0.720400	0.863800	0.784900	0.744800
8	Ada Boost Classifier	0.960800	0.657900	0.822800	0.730800	0.685200
1	K Neighbors Classifier	0.956400	0.612700	0.805200	0.694800	0.642900

Figura 5.5: Grilla de búsqueda optimizada y con ensambles para primera ventana de tiempo.

Una vez generados y entrenados todos los modelos se selecciona el modelo final basado en el mejor *F2 Score*, en este caso el mejor modelo corresponde a XGBoost Optimized, es decir, XGBoost tras optimizar sus hiperparámetros. Para profundizar en el desempeño final del mejor modelo seleccionado se muestran a continuación la matriz de confusión en la Figura 5.6 y el reporte de clasificación que entrega las principales métricas analizadas, recall, precisión y F1 en cada una de las clases en la figura 5.7. Cabe destacar que las métricas se calculan en el set de datos de validación, no visto por los modelos a la hora de entrenar ni de optimizar sus respectivos parámetros.

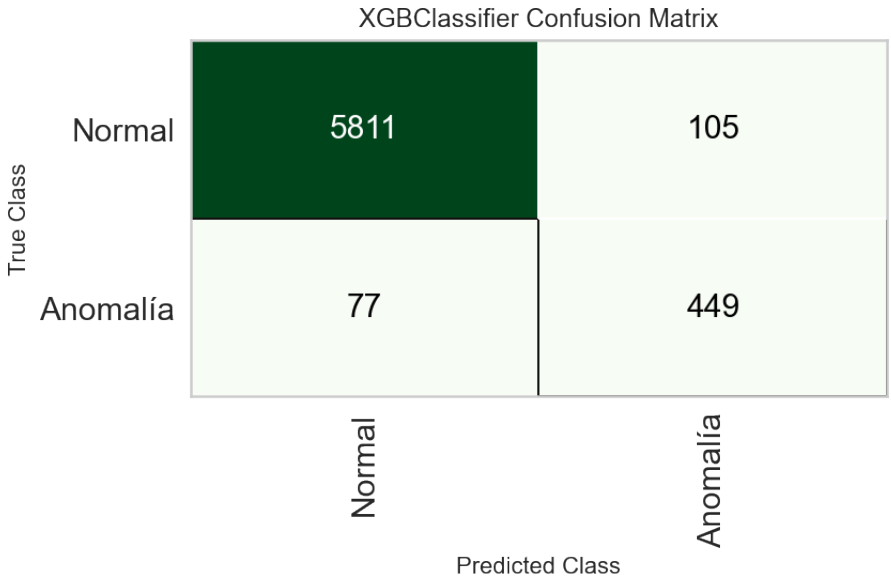


Figura 5.6: Matriz de confusión del mejor modelo seleccionado, XGBoost Optimizado.

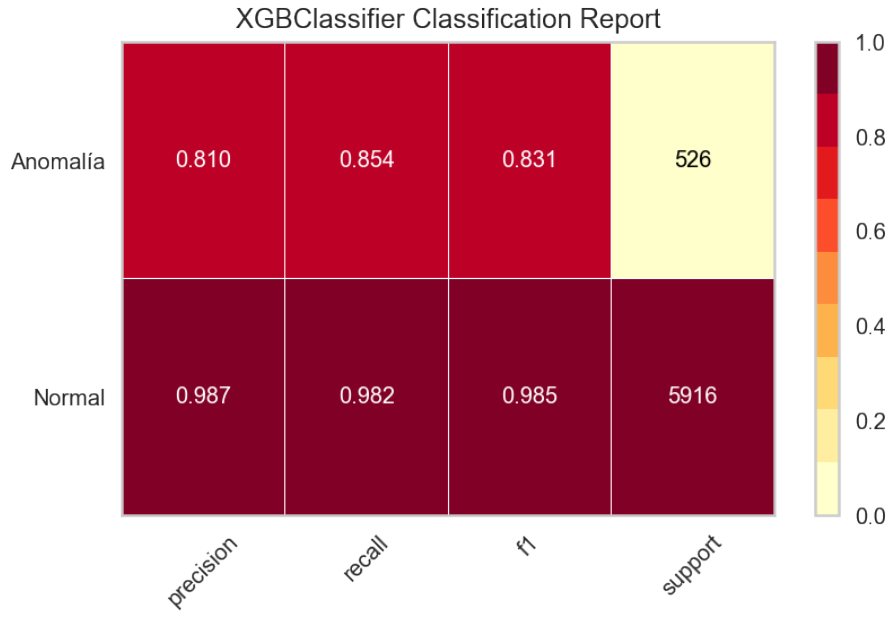


Figura 5.7: Reporte de clasificación para XGboost.

5.2. Segundo Caso

Le ventana de tiempo dos abarca las fechas desde el 14 de agosto de 2015 al 20 de febrero de 2017, 556 días en total, conteniendo las mismas variables monitoreadas que en el caso anterior para las 13357 muestras. Los datos se muestran en la Figura 5.8. Es necesario destacar que en este subconjunto de los datos se encuentra una mayor proporción de anomalías etiquetadas, de 16,1 %, prácticamente el doble en el caso anterior .

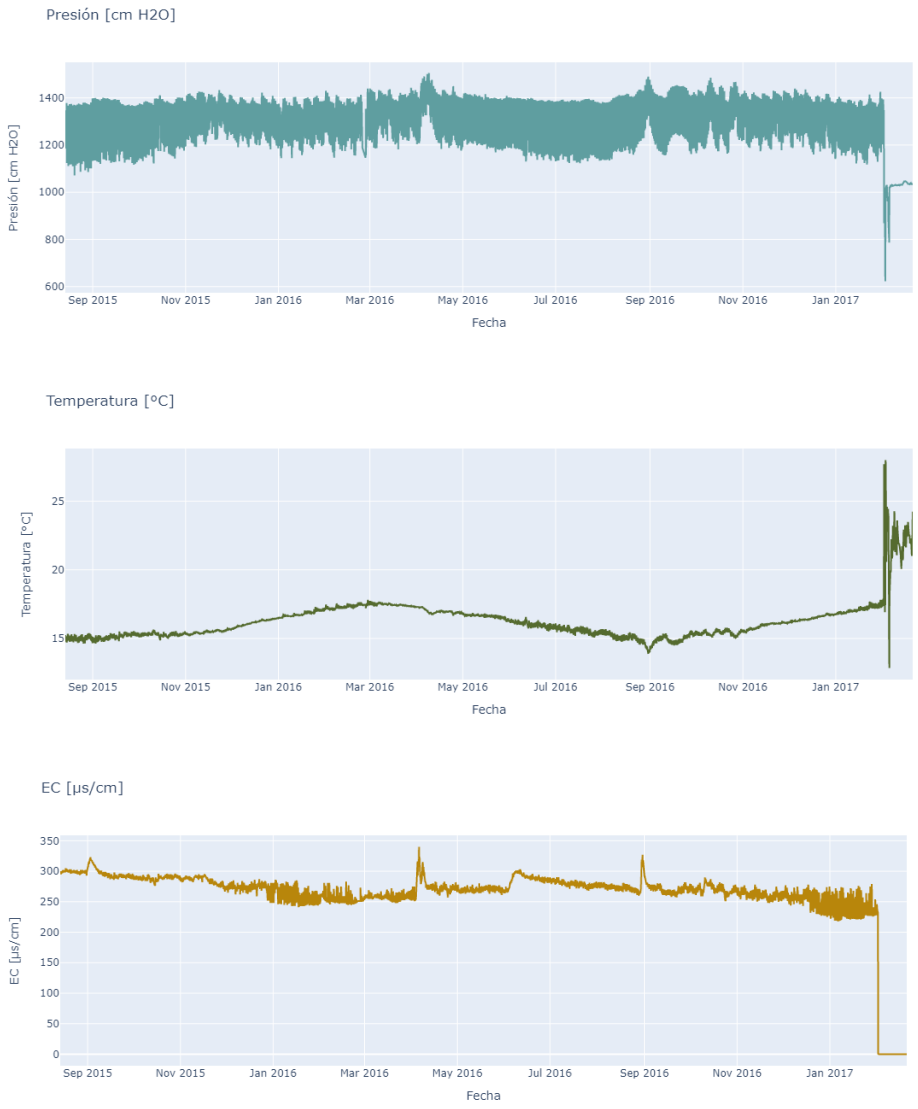


Figura 5.8: Datos asociados al segundo caso analizado 2015-2017.

5.2.1. Entrenamiento y selección de modelos

Este experimento es análogo a la ventana de tiempo del primer caso y el primer resultado analizado es la grilla de búsqueda de modelos para el conjunto de entrenamiento. Los resultados asociados se presentan en la Figura 5.9, al igual que en el Caso 1, los modelos que agregan *Optimizado* son los 4 que han pasado por este proceso y por tanto los que presentaron un mejor desempeño individual. En este caso son CatBoost, LightGBM,XGboost y Random Forest, utilizando el primero como modelo ensamblador del resto. También se presentan las versiones de ensamble por apilamiento y por votación previo a optimizar y una vez optimizados.

Index	Model Name	Accuracy	Recall	Prec.	F1	F2score
22	Stacking Classifier Optimized	0.950400	0.778800	0.896000	0.832500	0.799300
20	Stacking Classifier	0.951100	0.774200	0.905000	0.833800	0.796800
18	Extreme Gradient Boosting Optimized	0.835500	0.938000	0.490500	0.644100	0.793100
21	Voting Classifier	0.949100	0.764700	0.899700	0.825700	0.787800
19	CatBoost Classifier Optimized	0.952200	0.758600	0.927500	0.834100	0.787000
13	Light Gradient Boosting Machine	0.951100	0.758600	0.919800	0.830800	0.785800
3	Decision Tree Classifier	0.932400	0.784900	0.788600	0.786300	0.785300
16	Light Gradient Boosting Machine Optimized	0.947900	0.755200	0.900800	0.821400	0.780300
12	Extreme Gradient Boosting	0.949600	0.752500	0.915200	0.825200	0.779900
14	CatBoost Classifier	0.951200	0.748500	0.930900	0.828500	0.778300
6	Random Forest Classifier	0.948900	0.739100	0.923800	0.820400	0.769400
11	Extra Trees Classifier	0.946500	0.718800	0.929400	0.809500	0.752400
23	Voting Classifier Optimized	0.949300	0.711400	0.958600	0.815800	0.749700
9	Gradient Boosting Classifier	0.935000	0.621000	0.953500	0.751200	0.667100
17	Decision Tree Classifier Optimized	0.920700	0.516500	0.970200	0.673700	0.569600

Figura 5.9: Grilla de búsqueda de modelos para la segunda ventana de tiempo.

Para el modelo de ensamble, en este caso el mejor modelo seleccionado automáticamente por la estrategia es el de Apilamiento de los 4 mejores modelos optimizados mencionados en el párrafo anterior. Este presenta el mayor *Score F2* del conjunto y seguido en segundo lugar por su la variante que utiliza los modelos sin optimizar para el apilamiento. Los resultados del mejor modelo seleccionado se presentan en las Figuras 5.10 y 5.11.

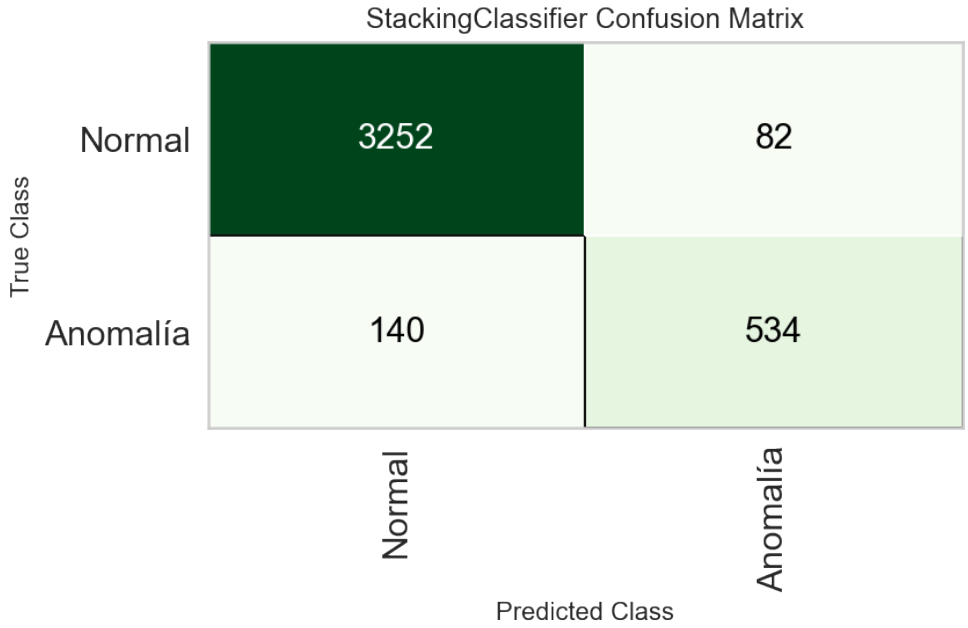


Figura 5.10: Matriz de confusión modelo Apilado.

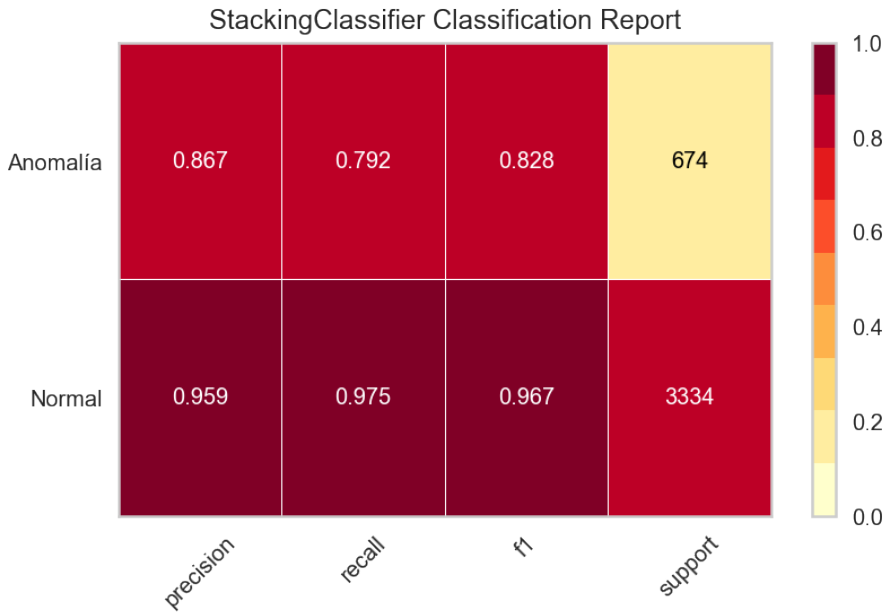


Figura 5.11: Reporte de clasificación mejor modelo seleccionado.

5.3. Discusión

Desde los primeros resultados obtenidos la estrategia de búsqueda de modelos genera múltiples clasificadores con resultados de eficacia en donde todas sus métricas están cercanas o sobre al 80 %. Los buenos resultados individuales de los múltiples clasificadores entrenados para los casos de estudio, aplicando la estrategia de pre-procesamiento, da sustento a que el desarrollo de una herramienta automática puede entregar resultados similares a los que

puede entregar una herramienta especializada entrenada con un experto realizando los experimentos con un mínimo esfuerzo en el manejo de los datos.

Es importante destacar que en todos los experimentos la métrica de exactitud o *Accuracy* presenta un sesgo al no reflejar un mal desempeño detectando anomalías debido al desbalance de las clases. Este problema con la métrica queda claro al analizar la columna *Accuracy* de todas las tablas presentadas en este capítulo. Tomando el caso presentado en la Figura 5.4, en donde tomando un clasificador que ignora por completo las variables de entrada, como lo es el *Dummy Classifier*, aún así presenta un *accuracy* en torno al 91,9 % sin embargo las métricas como *precision* y *recall* son 0 al igual que el Score F1 y F2 que se calcular a partir de ellos pues considera todos los casos negativos.

En el caso de la ventana de tiempo 1, como se muestra en la Figura 5.6, el modelo final seleccionado presenta 449 Verdaderos positivos y 77 Falsos Negativos, esto significa que se detecta un 85,5 % de los casos positivos totales correctamente además de un 98,2 % de los casos negativos correctamente, como se aprecia en el reporte de clasificación en la Figura 5.7. En este caso, se encuentra puede apreciar la influencia de un mayor *recall* por sobre la precisión entre el modelo n° 17 y el 14 pues el segundo presenta mayor *Score F1* por tener una precisión mayor sin embargo detecta menos casos positivos y por tanto tiene menor *recall*, esto si se ve reflejado en la fórmula de cálculo del *Score F2*.

Para la ventana 2, que presenta una mayor proporción de anomalías etiquetadas en los datos, como se esperaba, cambia el orden en el que quedan los modelos tras el entrenamiento y el mejor modelo seleccionado es el dado por la estrategia de Apilamiento de los cuatro mejores modelos tras optimizar sus hiperarámetros seguido por su variante utilizando los cuatro modelos sin optimizar. Este resultado se condice con lo esperado respecto a que los modelos de ensamble y apilamiento pueden llegar a tener una mayor capacidad de generalización. En este caso el modelo de apilamiento presenta un buen resultado detectando un 79,2 % de las anomalías totales considerando que tenía aproximadamente un 40 % menos de datos que la ventana 1.

Respecto a los resultados de la ventana 2, Figuras 5.9, 5.10 y 5.11, es necesario destacar la importancia de elegir una métrica adecuada para potenciar el tipo de métrica que potencie la necesidad de detección y el trade-off entre especificidad y exhaustitividad que se necesita. En el caso de la detección de eventos anómalos en calidad de agua, la no detección de un posible evento es un riesgo inminente para la salud de la población, por lo que una mayor detección de anomalías pero considerando poder tener más falsos positivos es una posible solución para involucrar a los gestores de los recursos solo cuando es necesario en comparación a un monitoreo exhaustivo y manual de los datos.

En este ejemplo en particular, es necesario destacar que todos los clasificadores presentan un buen desempeño y se debe a que los datos en particular de este ejemplo de la metodología de detección utiliza una ventana de tiempo en donde las etiquetas son pocas, sin embargo, muchas de ellas corresponden a valores atípicos muy notorios incluso gráficamente, por lo que la estrategia se considera exitosa para la comparación de desempeño con un etiquetado experto previo como el que contienen estos datos.

Capítulo 6

Plataforma abierta de detección de anomalías

Para el desarrollo de la plataforma que integra la estrategia de procesamiento de datos, la detección de anomalías y la interacción con el usuario se trabaja también en *Python* con un *framework* de desarrollo de aplicaciones web de código abierto denominado Streamlit, el cual se especializa en el desarrollo y despliegue de aplicaciones para trabajo con datos de forma sencilla y con la capacidad de ejecutarse tanto de forma local como en la nube. En específico se utilizan las siguientes versiones de librerías en el desarrollo de la aplicación web.

- python==3.9.13
- streamlit==1.12.0
- pandas==1.3.5
- plotly==5.10.0
- pandas-profiling==3.2.0

Un diagrama de la estructura y el flujo de información se presenta en la Figura 6.1, se muestra desde la carga de datos, el análisis exploratorio y el pre-procesamiento hasta la implementación dentro de la plataforma la metodología de detección de anomalías y la visualización los resultados del mejor modelo o ensamble de modelo seleccionado para realizar la tarea.

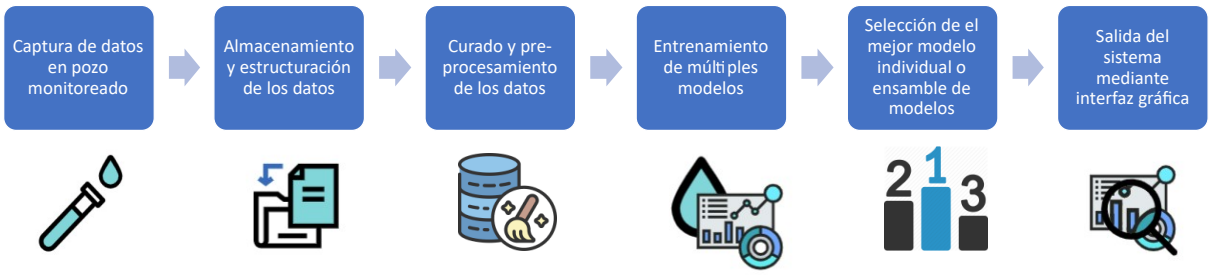


Figura 6.1: Estructura de procesamiento de datos desde la recolección por parte del usuario hasta el procesamiento a través de la plataforma.

6.1. Consideraciones de diseño

El primer requerimiento considera poder analizar las series de tiempo históricas en bruto para un análisis exploratorio previo al entrenamiento. También, se deben mostrar las salidas del sistema, que corresponden a las alarmas en donde se etiqueta una anomalía para que el usuario pueda ubicarlas en el tiempo y relacionar las variables involucradas.

Algunas de estos requerimientos ya se encuentran disponibles en el "Observatorio Georreferenciado" de la DGA [20] que permite, a través de un mapa, acceder a los registros de derechos de aguas, solicitudes de derechos de agua y el monitoreo de extracciones efectivas. Otra cualidad interesante es que permite también poder agregar o filtrar diferentes fuentes de aguas como lo son superficiales y subterráneas además de sobreponer en el mapa áreas de prohibiciones o donde se ha decretado una escasez o declarado agotamiento de agua. Su interfaz puede apreciarse en la Figura B.1.

Otra plataforma abierta disponible es el "Explorador climático" [21] perteneciente al Centro de Ciencias del Clima y la Resiliencia (CR) 2. Este explorador permite acceder a datos climáticos históricos de distintos sectores mediante un mapa georreferenciado además de un panel de control con múltiples, opciones como puede apreciarse en la Figura B.3. Se destaca que permite realizar un refinamiento por variable y se modifica en el mapa todas las estaciones que presentan monitoreo y una escala de color con un promedio diario . Es posible seleccionar desde una lista desplegable o desde el mapa la estación a monitorear y muestra la serie de tiempo en la parte inferior. Presenta una sección de detección de anomalías en donde se puede ajustar un parámetro que mide cuanto se alejan del promedio histórico de los datos y fijar un umbral.

Las dos plataformas descritas anteriormente tienen en común que funcionan centralizando información desde distintas fuentes. y con diferentes formatos para que puedan visualizarse y exportarse en el formato de archivos de conveniencia. Es necesario destacar aquí que todas ellas llegan hasta la etapa de visualización, sin embargo, no realizan un procesamiento de los datos disponibles.

Otro ejemplo de plataformas de monitoreo de datos es la implementada por la Sociedad Química y Minera de Chile (SQM) para abrir a la comunidad un seguimiento ambiental para el cumplimiento de la normativa ambiental y la resolución de calificación ambiental que permite su operación. Dentro de esta plataforma[22] se puede acceder a las distintas variables fisicoquímicas medidas en cada estación de monitoreo y ver sus evoluciones a través de series de tiempo. Es necesario destacar de esta plataforma, que se realizan mediciones por una empresa consultora con una frecuencia de muestreo de tres meses, pues es la frecuencia mínima que exige la legislación.

Un resumen de estas cualidades destacadas pueden analizarse resumidas en la tabla 6.1 Comparativa de atributos de distintas plataformas abiertas disponibles.

Tabla 6.1: Comparativa atributos de plataformas existentes.

Atributo	Observatorio Georreferenciado DGA	Explorador Climático CR2	Plataforma Seguimiento Ambiental SQM
Conexión con bases de datos	Sí Fuentes propias	Si Fuentes externas	Sí Fuentes propias
Data histórica/ series de tiempo disponibles	No	Sí Con visualizaciones y ejes modificables	Sí Con frecuencia de muestreo de 3 meses
Detección de anomalías	No	Sí Comparando la desviación con el promedio histórico	No
Muestra datos de calidad de agua	No	No	Sí Fuentes propias

6.2. Requerimientos de la plataforma

Los requerimientos principales comprometidos para este resultado, que son considerados como un herencia del proyecto original en donde se desarrolla este Trabajo de Tesis, son los dos atributos cuantificables que se describen a continuación.

1. Detección de eventos: Capacidad de detectar correctamente anomalías en base a la data de entrenamiento. Se requiere el sistema experto tenga una precisión mayor al 75 % y menor al 95 %.

2. Tiempo procesamiento del sistema: Tiempo que toma, tras recibir los datos, al sistema experto procesarlo y generar las notificaciones. Se requiere que no supere los 60 minutos en total.

6.3. Aplicación WEB de detección de anomalías

El desarrollo de la Plataforma del Sistema Experto se desarrolla mediante una aplicación web programada en “Python” y utilizando el framework de desarrollo de aplicaciones “Streamlit” por la facilidad de programación y de despliegue en la nube. Para el diseño de la interfaz y sus funcionalidades se toman en consideración primeramente los requerimientos dados por el proyecto además de analizar la pertinencia de las consideraciones de diseño de otras plataformas abiertas disponibles.

En este caso la adquisición de los datos se realiza mediante la carga de un archivo .csv en el formato requerido por el sistema para interpretar correctamente las variables monitoreadas y se seleccionan la columna correspondiente a las etiquetas que previamente el experto ha completado al analizar los data o parte de ellos. Es posible también visualizar las series de tiempo con el pre-procesamiento propuesto mediante gráficos interactivos en la primera

sección de la plataforma.

Con esto la aplicación web tiene, a grandes rasgos, tres secciones principales que se detallan a continuación.

1. **Adquisición de los datos**

   La primera funcionalidad que posee la aplicación web el cuadro de archivo en donde se carga el archivo *.csv* con la data histórica estructurada por columnas para un sitio de monitoreo acotado, en este caso, este sitio es el correspondiente a Horcón y su análisis se encuentra en el Capítulo 4. Este primer paso a completar se aprecia en la Figura 6.2.

Figura 6.2: Sección de carga y adquisición de datos.

La aplicación queda a la espera de que se carguen los datos y una vez cargados se despliegan las dos secciones siguientes. Es importante destacar aquí que hay un paso requerido antes de poder visualizar los datos y entrenar los modelos y es que se deben seleccionar las columnas que poseen características y la columna de etiquetas, mediante listas desplegables y selección múltiple, como puede apreciarse el Figura 6.3.

Figura 6.3: Selección de atributos y *target*.

Por último se aprecia un botón titulado “Comenzar!” que comienza el entrenamiento de los modelos de detección de anomalías.

2. **Visualización interactiva:** La primera visualización obtenida es la de descripción estadística del sitio de medición con información como la cantidad de variables, sus desviaciones estándar, mínimos, máximos, promedio y cuartiles, además de una vista de los datos guardados en su formato tabular. Un ejemplo de esto se muestra en la Figura 6.4.

	Pression [cm H2O]	Temperatura [°C]	EC [µ	Etiqu
2013-03-03T00:00:00-03:00	945.3000	17.8900	255	1
2013-03-03T01:00:00-03:00	1,036.2000	17.8500	255	1
2013-03-03T02:00:00-03:00	1,012.3000	17.8700	255	1
2013-03-03T03:00:00-03:00	980.1000	17.8800	255	1
2013-03-03T04:00:00-03:00	1,074.9000	17.8700	255	1
2013-03-03T05:00:00-03:00	988.8000	17.8800	255	1
2013-03-03T06:00:00-03:00	1,058.6000	17.8600	255	1
2013-03-03T07:00:00-03:00	890.9000	17.8800	254	1
2013-03-03T08:00:00-03:00	1,026.7000	17.8300	255	1
2013-03-03T09:00:00-03:00	905.2000	17.8700	254	1
2013-03-03T10:00:00-03:00	873.4000	17.8500	8	1

	count	mean	std	min	25%	50%	75%	max
Pression [cm H2O]	17507.000000	1058.285726	103.255030	868.200000	970.200000	1069.800000	1149.500000	1249.300000
Temperatura [°C]	17507.000000	16.160395	0.994024	13.720000	15.220000	16.010000	17.100000	17.960000
EC [µs/cm]	17507.000000	259.897869	50.013480	5.000000	258.000000	272.000000	279.000000	293.000000
Etiqueta	17507.000000	0.090592	0.287037	0.000000	0.000000	0.000000	0.000000	1.000000

Figura 6.4: Análisis estadístico descriptivo de información cargada.

Con los datos lo primero que se realiza es una visualización de la estructura de datos cargados mediante una tabla que se despliega si se requiere y los gráficos interactivos de las series de tiempo directamente desde la base de datos, como se muestra en las Figuras 6.4 y 6.6.

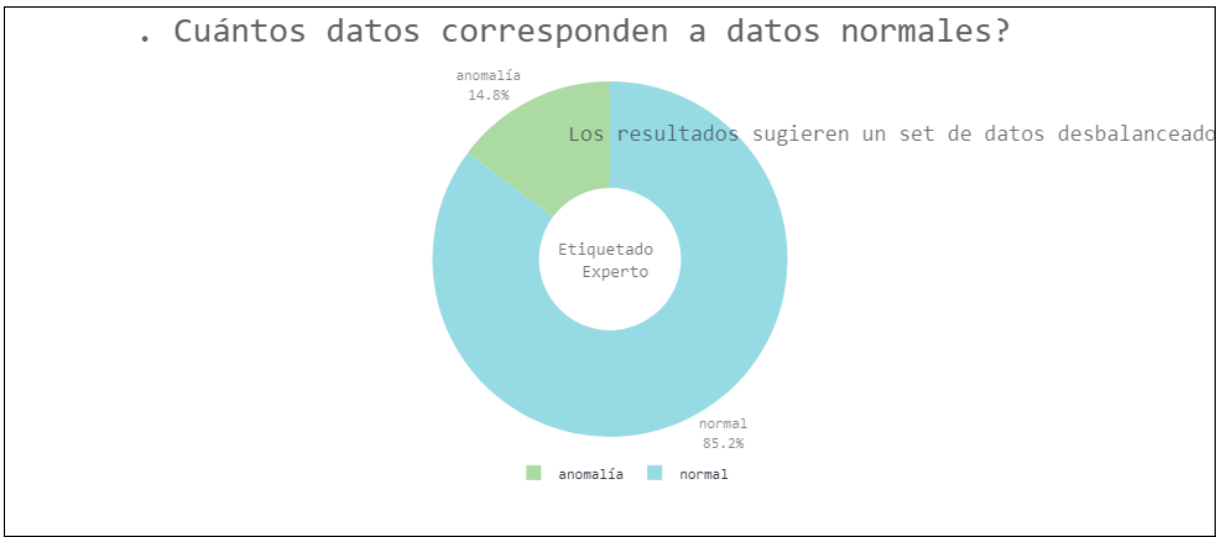


Figura 6.5: Sección de visualización de información del sitio de medición.

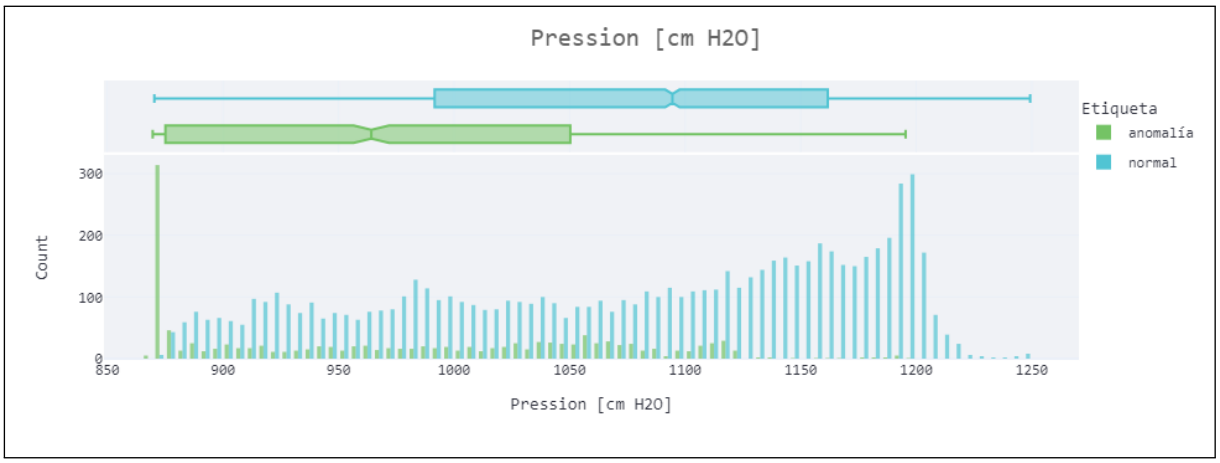


Figura 6.6: Análisis estadístico de la variable Presión de columna de agua según información cargada.

Adicional al análisis exploratorio mediante gráficos por variable es posible también desplegar un panel de información adicional que muestra algunas estadísticas adicionales y un reporte exploratorio generados automáticamente a partir de los datos cargados gracias a la librería *pandas-profilling*. Para realizar esta función se ubica un botón al final de esta sección de la plataforma titulado *Generar un reporte exploratorio más detallado.*

3. **Procesamiento y detección de anomalías** Así se llega a la sección principal de la aplicación, que corresponde a detección de anomalías. En esta sección es posible encontrar primero la tabla, Figura 6.7, que representa la grilla de búsqueda de modelos previo al ensamble y con el cual se genera el modelo apilado.

Los mejores clasificador fueron:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.967900	0.991900	0.976100	0.988400	0.982200	0.819300	0.821300	0.172000
catboost	CatBoost Classifier	0.966100	0.992100	0.971200	0.991300	0.981100	0.815300	0.819600	2.455000
rf	Random Forest Classifier	0.964500	0.989600	0.972000	0.988700	0.980300	0.803800	0.806900	0.194000
dt	Decision Tree Classifier	0.963500	0.910300	0.975600	0.984000	0.979800	0.791000	0.792100	0.011000
lightgbm	Light Gradient Boosting Machine	0.963400	0.992000	0.968200	0.991300	0.979600	0.802700	0.808100	0.033000

Figura 6.7: Grilla de búsqueda de modelos mostrando una selección de los mejores 5.

Luego se puede apreciar el cálculo de la matriz de confusión para el modelo apilado en la Figura 6.8.

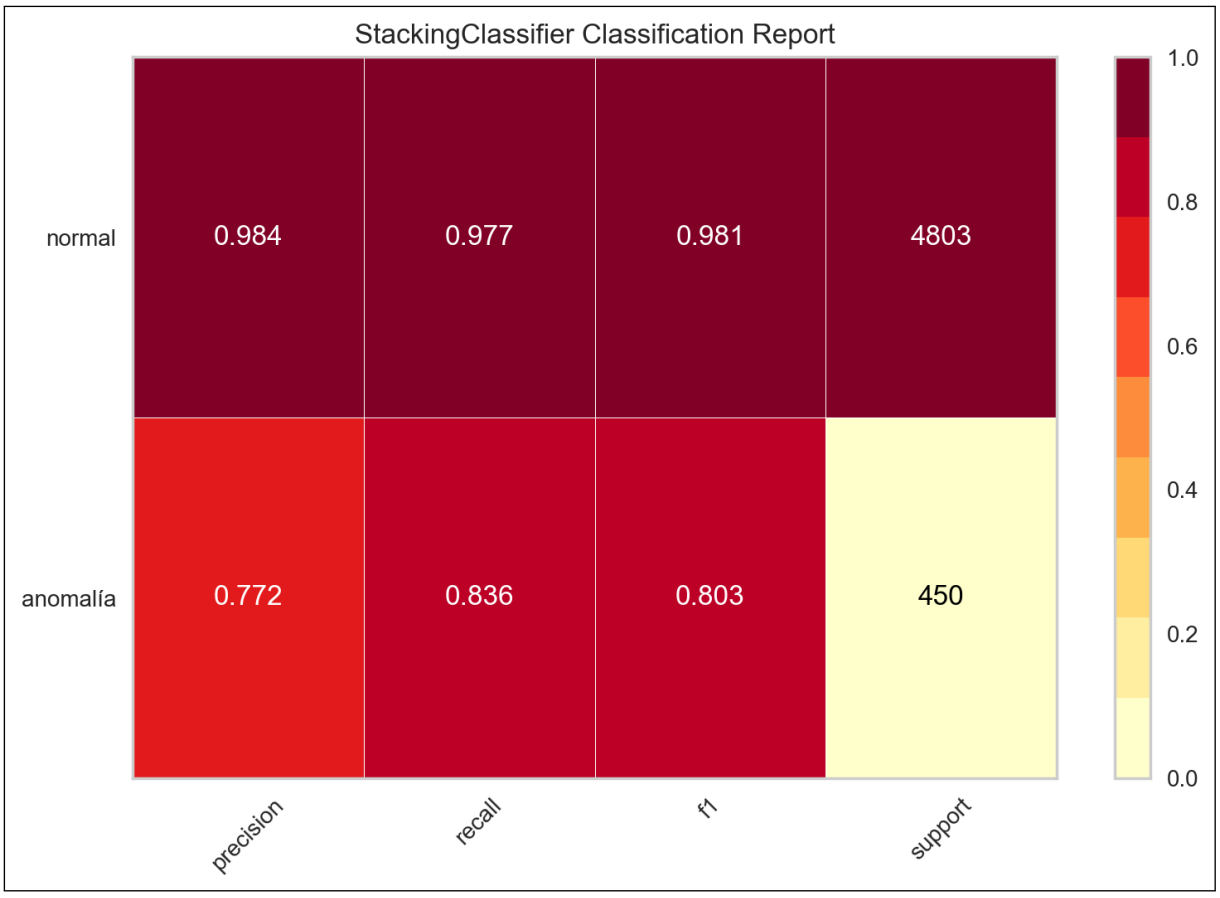
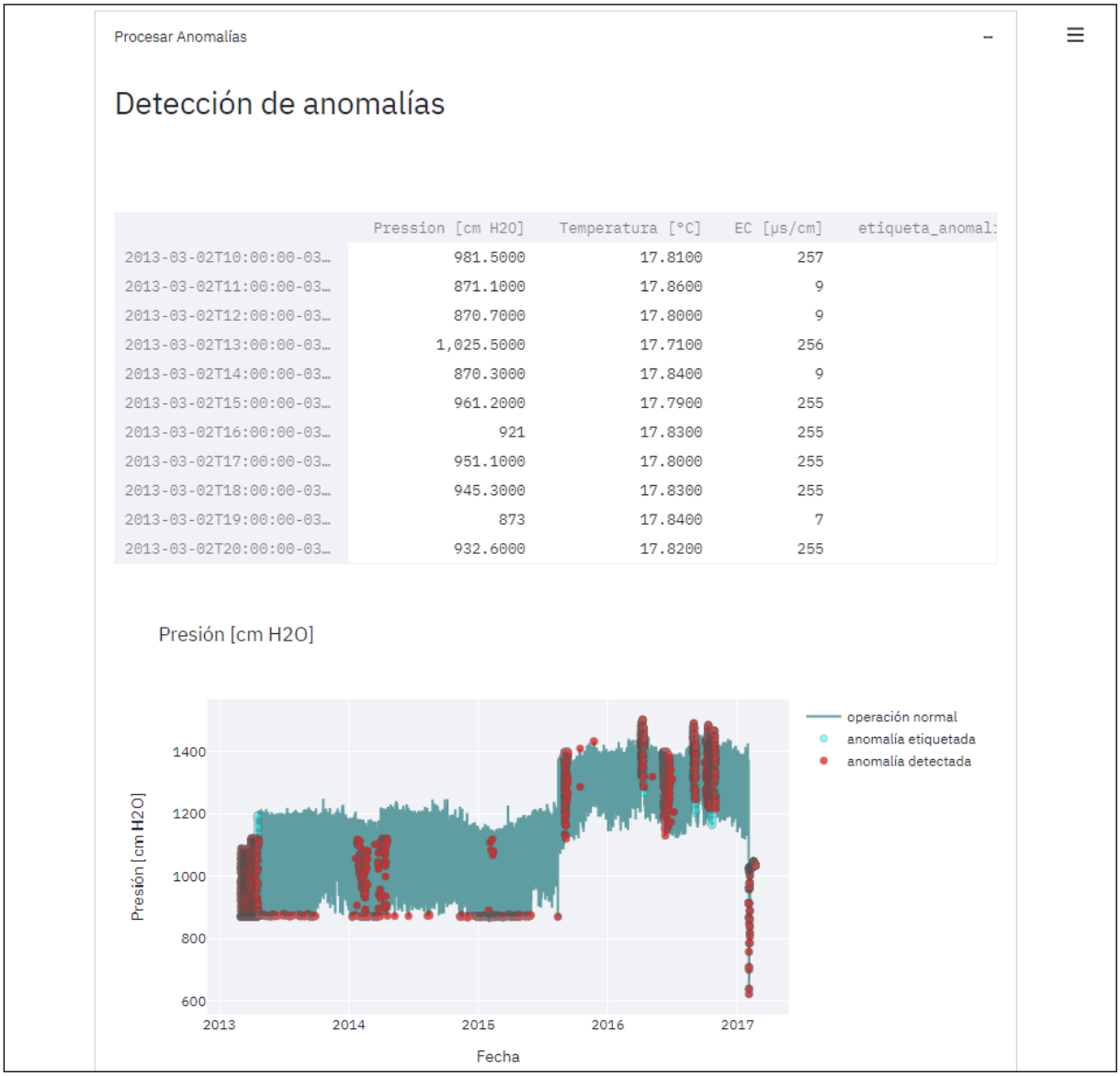


Figura 6.8: Matriz de confusión

Es necesario destacar aquí que los gráficos permiten refinar la ventana de tiempo a observar además de poder superponer o quitar las anomalías (puntos en rojo) de la serie de tiempo, además si se mueve el cursor sobre el punto entrega la información de la

marca de tiempo asociada a la anomalía etiquetada. Se superponen también el etiquetado experto para comparar el desempeño del modelo en casos específicos de forma gráfica.



	Pression [cm H2O]	Temperatura [°C]	EC [µs/cm]	etiqueta_anomal
2013-03-02T10:00:00-03…	981.5000	17.8100	257	
2013-03-02T11:00:00-03…	871.1000	17.8600	9	
2013-03-02T12:00:00-03…	870.7000	17.8000	9	
2013-03-02T13:00:00-03…	1,025.5000	17.7100	256	
2013-03-02T14:00:00-03…	870.3000	17.8400	9	
2013-03-02T15:00:00-03…	961.2000	17.7900	255	
2013-03-02T16:00:00-03…	921	17.8300	255	
2013-03-02T17:00:00-03…	951.1000	17.8000	255	
2013-03-02T18:00:00-03…	945.3000	17.8300	255	
2013-03-02T19:00:00-03…	873	17.8400	7	
2013-03-02T20:00:00-03…	932.6000	17.8200	255	

Figura 6.9: Sección de detección de anomalías

6.4. Discusión

Es necesario destacar que se logra satisfactoriamente cumplir con los requerimientos establecidos asociado a la obtención de notificaciones y alarmas en el sistema experto. Además, el sistema posee un tiempo de procesamiento del orden de segundos para cada modelo y un par de minutos para el entrenamiento de múltiples modelos, dependiendo de la cantidad de datos. Por tanto, los tiempos obtenidos para toda la integración de la estrategia de entrenamiento y selección de modelo toma un tiempo bastante lejano al máximo de 60 minutos.

En cuanto al diseño es necesario destacar que la versión mostrada es asociado a un caso

de uso específico, pues se considera para este Trabajo de Tesis el análisis de la data histórica local de un sitio específico y acotado incorporando el conocimiento experto y el procesamiento computacional para la detección de anomalías a través de la metodología propuesta. Sin embargo, se plantea que esta plataforma podría fácilmente aplicarse a otros sitios monitoreados y no solamente de acuíferos subterráneos, pues se adapta en función de los datos cargados y los modelos se entrenan a partir de la parámetros y configuraciones que ingresa el usuario.

La interfaz presentada incorpora una serie de observaciones y consideraciones provenientes tanto de potenciales usuarios como de un diseñador especializado en visualización durante desarrollo del proyecto, por lo que es un primer acercamiento a la validación de la metodología y el acercamiento de herramientas de ML a un mayor número de usuarios.

Capítulo 7

Conclusiones

Este trabajo de tesis se enfoca en el desarrollo de una Plataforma abierta de detección de anomalías para el monitoreo de acuíferos subterráneos, buscando ser una ayuda a la importante labor de preservación y sustentabilidad de los recursos hídricos en Chile y los potenciales riesgos para la salud asociados a su falta de atención. La metodología descrita abarca tanto la detección de anomalías como su integración en una plataforma web accesible, que permite a los usuarios realizar análisis exploratorios y estadísticos, y detectar anomalías en los datos de calidad de agua. Se presentan los principales resultados cuantitativos asociados al caso de estudio en la detección de anomalías, seguidos de una discusión sobre oportunidades de mejora

El objetivo principal de este trabajo es diseñar y desarrollar una plataforma web para el procesamiento en línea y visualización de datos hidrogeológicos, basada en Machine Learning para la detección de anomalías. La automatización del flujo de trabajo permite generar alarmas y facilita a los usuarios realizar un diagnóstico oportuno del estado del acuífero monitoreado. La metodología propuesta para el entrenamiento y selección automáticos del mejor modelo de detección de anomalías se integra en la plataforma web desarrollada.

La contribución principal es proporcionar una herramienta abierta que, a partir del conocimiento experto como input, automatiza el análisis de los datos de monitoreo de calidad de agua en acuíferos. Actualmente, este análisis se realiza manualmente, y en ocasiones los registros no se analizan debido a la falta de tecnologías y conocimientos asociados con su procesamiento. Es relevante destacar que tanto la metodología como el desarrollo de la plataforma pueden recibir conjuntos de datos de monitoreo de acuíferos de prácticamente cualquier estación de monitoreo, no estando limitados exclusivamente a datos de calidad de agua. La metodología propuesta incluye una capa de redundancia en el entrenamiento de modelos, lo que permite obtener un buen desempeño frente a escenarios que requieren una metodología robusta de selección de anomalías sin sacrificar la facilidad de uso..

El sistema implementado automatiza completamente el flujo de procesamiento de datos, análisis exploratorio, entrenamiento y selección del modelo con mejor desempeño, así como la generación automática de métricas de desempeño y gráficos. Por lo tanto, hay muchas opciones de configuración y oportunidades de mejora para futuras iteraciones. Se sugiere la implementación de un módulo de detección de anomalías no supervisado para abordar escenarios en los que no se cuenta con un etiquetado experto previo. Además, se considera que el análisis de una mayor cantidad de sitios o estaciones de monitoreo con datos etiquetados

permitiría profundizar en las ventajas y limitaciones del presente trabajo.

En resumen, la Plataforma desarrollada representa un avance significativo en la detección automatizada de anomalías en acuíferos, abriendo camino a futuras investigaciones y mejoras para una gestión más efectiva y segura de estos recursos vitales.

Bibliografía

[1] Intergovernmental Panel on Climate Change (IPCC), "Climate change widespread, rapid, and intensifying – IPCC," 2021.

[2] L. Yu, S. A. Josey, F. M. Bingham, and T. Lee, "Intensification of the global water cycle and evidence from ocean salinity: a synthesis review," *Annals of the New York Academy of Sciences*, vol. 1472, pp. 76–94, 7 2020.

[3] S. Jasechko and D. Perrone, "Global groundwater wells at risk of running dry," *Science*, vol. 372, pp. 418–421, 4 2021.

[4] R. G. Taylor, B. Scanlon, P. Döll, M. Rodell, R. van Beek, Y. Wada, L. Longuevergne, M. Leblanc, J. S. Famiglietti, M. Edmunds, L. Konikow, T. R. Green, J. Chen, M. Taniguchi, M. F. P. Bierkens, A. MacDonald, Y. Fan, R. M. Maxwell, Y. Yechieli, J. J. Gurdak, D. M. Allen, M. Shamsudduha, K. Hiscock, P. J.-F. Yeh, I. Holman, and H. Treidel, "Ground water and climate change," *Nature Climate Change*, vol. 3, pp. 322–329, 4 2013.

[5] N. Carrard, T. Foster, and J. Willetts, "Groundwater as a source of drinking water in southeast Asia and the Pacific: A multi-country review of current reliance and resource concerns," *Water (Switzerland)*, vol. 11, no. 8, 2019.

[6] J. S. Famiglietti and G. Ferguson, "The hidden crisis beneath our feet," *Science*, vol. 372, pp. 344–345, 4 2021.

[7] S. Jasechko and D. Perrone, "California's Central Valley Groundwater Wells Run Dry During Recent Drought," *Earth's Future*, vol. 8, 4 2020.

[8] A. A. Muñoz, K. Klock-Barría, C. Alvarez-Garreton, I. Aguilera-Betti, González-Reyes, J. A. Lastra, R. O. Chávez, P. Barría, D. Christie, M. Rojas-Badilla, and C. LeQuesne, "Water Crisis in Petorca Basin, Chile: The Combined Effects of a Mega-Drought and Water Management," *Water*, vol. 12, p. 648, 2 2020.

[9] R. D. Garreaud, J. P. Boisier, R. Rondanelli, A. Montecinos, H. H. Sepúlveda, and D. Veloso-Aguila, "The Central Chile Mega Drought (2010–2018): A climate dynamics perspective," *International Journal of Climatology*, vol. 40, pp. 421–439, 1 2020.

[10] C. Alvarez-Garreton, J. Boisier, M. Billi, I. Lefort, R. Marinao, and P. Barría, "Protecting environmental flows to achieve long-term water security," *Journal of Environmental Management*, vol. 328, p. 116914, 2 2023.

[11] P. Aceituno, J. P. Boisier, R. Garreaud, R. Rondanelli, and J. A. Rutllant, "Climate and Weather in Chile," in *Water Resources of Chile*, pp. 7–29, Springer International Publishing, 2021.

[12] Dirección General de Aguas, "Determina las condiciones técnicas y los plazos a nivel nacional para cumplir con obligaciones de instalar y mantener un sistema de monitoreo y transmisión de extracciones efectivas en las obras de captación subterráneas," 6 2019.

[13] A. Moreno, C. Leturia, O. C. Marfil, D. San, M. Cornejo, H. Moya, G. Daniela, F. Muños, and Y. Ulloa, "Atlas Calidad del Agua," 2020.

[14] Instituto Nacional de Normalización, "NCh 409/2: Agua potable - Parte 2: Muestreo," 2004.

[15] Departamento de Conservación y Protección de Recursos Hídricos (DCPRH), "Diagnóstico y desafíos de la Red de Calidad de Aguas Subterráneas de la DGA," 2017.

[16] Ministerio de Obras Públicas, "Aprueba Reglamento de Monitoreo de Extracciones Efectivas de Aguas Superficiales," 4 2020.

[17] Instituto Nacional de Normalización, "NCh 409/1: Agua potable - Parte 1: Requisitos," 2005.

[18] Instituto Nacional de Normalización, "NCh 1333: Requisitos de calidad del agua para diferentes usos," 1987.

[19] E. M. García-del Toro, S. García-Salgado, L. F. Mateo, M. Quijano, and M. I. Más-López, "Machine Learning as a Diagnosis Tool of Groundwater Quality in Zones with High Agricultural Activity (Region of Campo de Cartagena, Murcia, Spain)," *Agronomy*, vol. 12, p. 3076, 12 2022.

[20] D. G. de Aguas, "Observatorio georreferenciado." Accessed Nov, 2021 [Online]. Available: https://snia.mop.gob.cl/observatorio.

[21] C. de Ciencia del Clima y la Resiliencia (CR)2, "Explorador climático (CR)2." Accessed Nov, 2021 [Online]. Available: https://explorador.cr2.cl.

[22] S. Q. y Minera de Chile, "Plataforma seguimiento ambiental SQM." Accessed Nov, 2021 [Online]. Available: https://www.sqmsenlinea.com/data-source-type/2/show.

[23] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1 1992.

[24] J. Liu, P. Wang, D. Jiang, J. Nan, and W. Zhu, "An integrated data-driven framework for surface water quality anomaly detection and early warning," *Journal of Cleaner Production*, vol. 251, p. 119145, 2020.

[25] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, and H. Ren, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water Research*, vol. 171, p. 115454, 3 2020.

[26] H. Almuhtaram, A. Zamyadi, and R. Hofmann, "Machine learning for anomaly detection in cyanobacterial fluorescence signals," *Water Research*, vol. 197, 6 2021.

[27] C. Aggarwal, *An Introduction to Outlier Analysis*, pp. 1–34. 12 2017.

[28] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, *Noise Versus Outliers*, pp. 163–183. Cham: Springer International Publishing, 2016.

[29] D. M. Hawkins, *Identification of Outliers*. Dordrecht: Springer Netherlands, 1980.

[30] G. Moschini, R. Houssou, J. Bovay, and S. Robert-Nicoud, "Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model," in *The 7th International conference on Time Series and Forecasting*, vol. 5, (Basel Switzerland), p. 56, MDPI, 7 2021.

[31] S. Katipamula and M. R. Brambley, "Review article: Methods for fault detection, diagnostics, and prognostics for building systems—a review, part ii," *HVAC&R Research*, vol. 11, no. 2, pp. 169–187, 2005.

[32] M.-L. Antonie, O. R. Zaïane, and A. Coman, "Application of data mining techniques for medical image classification," in *Proceedings of the Second International Conference on Multimedia Data Mining*, MDMKDD'01, (Berlin, Heidelberg), p. 94–101, Springer-Verlag, 2001.

[33] M. Braei and S. Wagner, "Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art," *arXiv*, 4 2020.

[34] C. C. Aggarwal, *Outlier Analysis*. Cham: Springer International Publishing, 2017.

[35] A. Thakur, *Approaching (Almost) Any Machine Learning Problem*. 2020.

[36] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[37] C. W. J. Granger and M. J. Morris, "Time Series Modelling and Interpretation," *Journal of the Royal Statistical Society. Series A (General)*, vol. 139, no. 2, p. 246, 1976.

[38] S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective Intrusion Detection System Using XGBoost," *Information*, vol. 9, p. 149, 6 2018.

[39] R. Isermann and P. Ballé, "TRENDS IN THE APPLICATION OF MODEL BASED FAULT DETECTION AND DIAGNOSIS OF TECHNICAL PROCESSES," tech. rep., 1996.

[40] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 8 1997.

[41] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001.

[42] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, 1 1979.

[43] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 8 1996.

[44] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[45] Y. Freund, "Experiment with a new boosting algorithm," *Proc. Int. Conf. Mach. Learn.*, vol. 13, 01 1996.

[46] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, ACM, 8 2016.

[47] M. Gan, S. Pan, Y. Chen, C. Cheng, H. Pan, and X. Zhu, "Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia

River," *Journal of Marine Science and Engineering*, vol. 9, p. 496, 5 2021.

[48] K. Joslyn, "Water Quality Factor Prediction Using Supervised Water Quality Factor Prediction Using Supervised Machine Learning Machine Learning," *REU Final Reports. 6.*, 2018.

[49] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, 6 2020.

[50] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5991 LNAI, no. PART 2, pp. 340–350, 2010.

[51] M. H. D. M. Ribeiro and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Applied Soft Computing Journal*, vol. 86, 1 2020.

[52] P. Kalia, "Stacking Supervised and Unsupervised Learning Models for Better Performance," *International Research Journal of Engineering and Technology*, 2008.

[53] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.

[54] J. Rodriguez, A. Perez, and J. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 569–575, 3 2010.

[55] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," *Advances in neural information processing systems*, vol. 25, 6 2012.

Anexo A

Metodología

A.1. Cross Validation

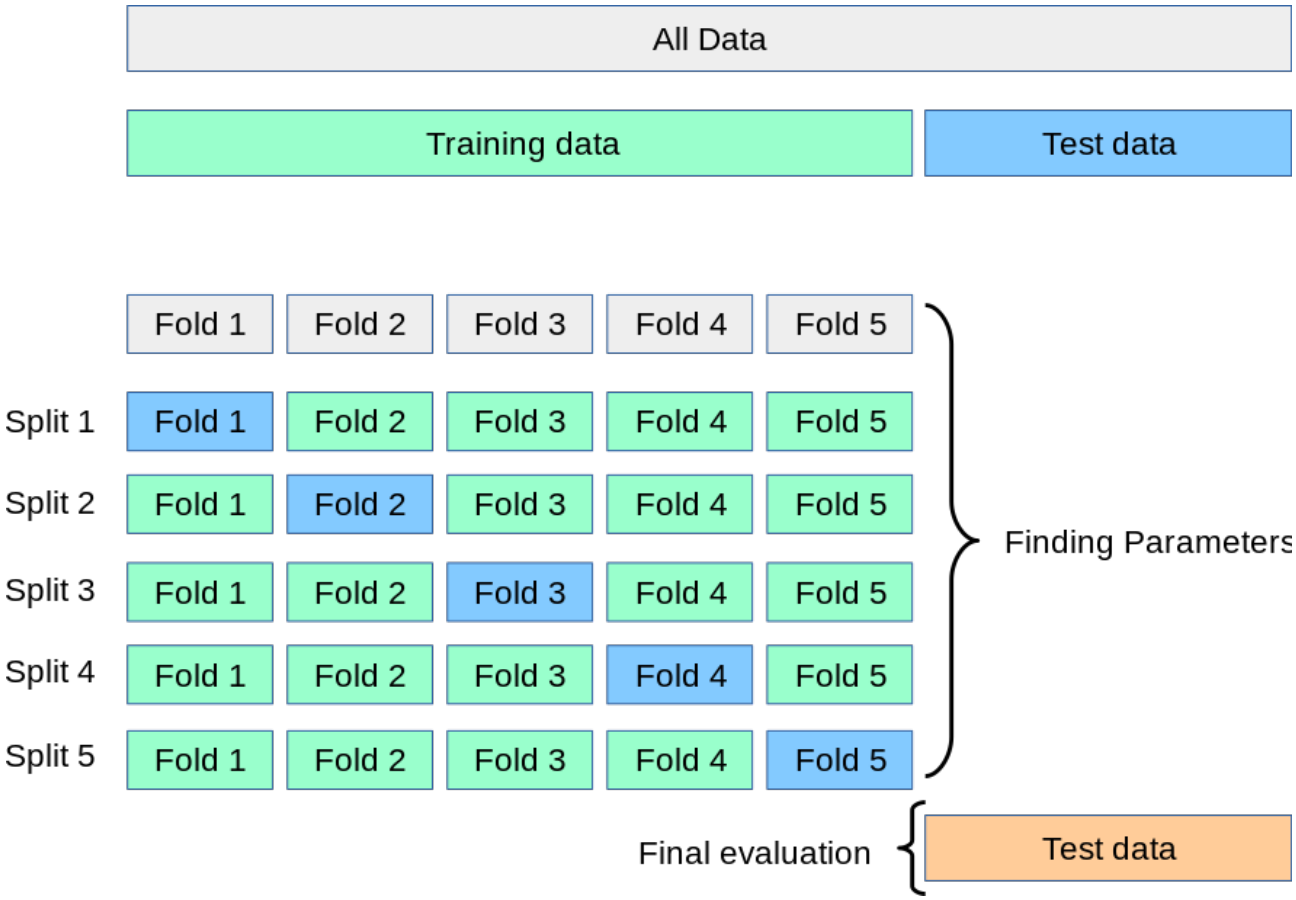


Figura A.1: Ejemplo de Cross Validation utilizando una estrategia de kfold con k=5.

Anexo B

Desarrollo plataforma

Análisis de otras plataformas disponibles y su interfaz visual.

B.1. Análisis otras plataformas

Figura B.1: Plataforma monitoreo DGA

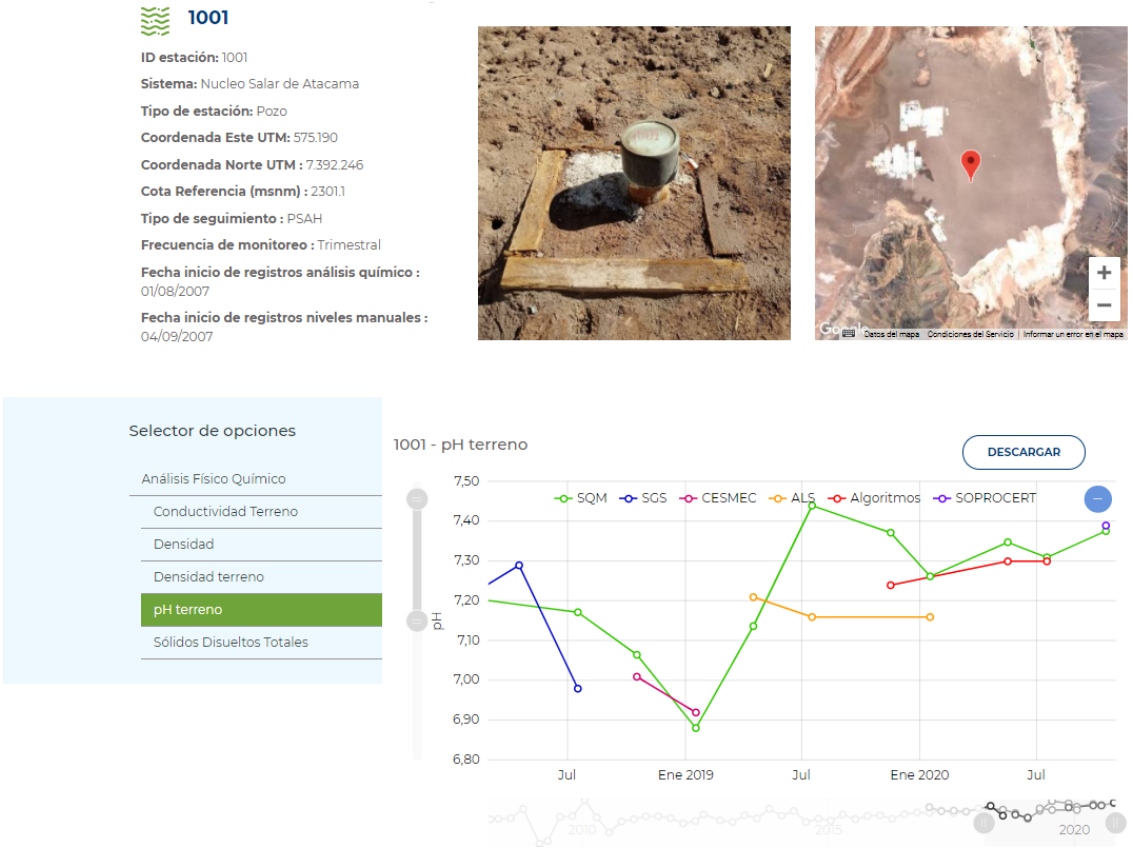


Figura B.2: Plataforma monitoreo SQM

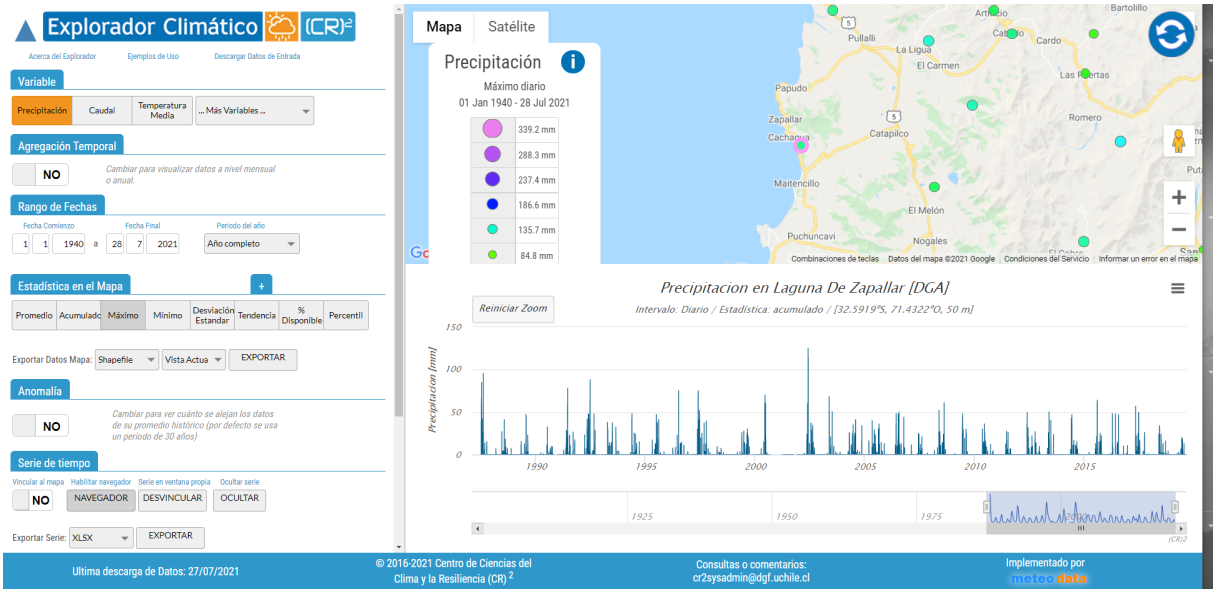


Figura B.3: Plataforma monitoreo CR2