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 RECURSOS HÍDRICOS

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

Colaboradores: Matías Taucare
María Jesús Ugarte

SANTIAGO, CHILE
3 de octubre de 2022

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: 3 de octubre de 2022
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 RECURSOS HÍDRICOS

La crisis climática y el calentamiento global están estresando seriamente los ciclos hidrológicos en toda su extensión. El agua cubre el 70 % de la superficie del planeta, sin embargo, se estima que solo un 0,62 % de ella es apta para el consumo humano.

Chile cuenta con numerosas zonas bajo un extenso estrés hídrico y muchas de ellas incluso con escasez de agua. La gestión sostenible de los cada vez más escasos recursos hídricos presenta dificultades técnicas para su realización.

Este trabajo presenta el diseño e implementación de una plataforma integrada de fácil uso para el procesamiento, análisis y visualización de datos y la detección de anomalías en calidad de agua para apoyar la gestión de los recursos hídricos. Se analiza la integración de métodos de procesamiento de datos, clasificación y detección de anomalías para generar alertas de posibles problemas a partir de las mediciones fisicoquímicas.

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 optimizando una métrica de desempeño configurada por el usuario para la detección de anomalías en datos de calidad de agua. Este trabajo tiene el potencial de ser una herramienta útil y atingente y sus aplicaciones podrían ir más allá del agua.

Para ustedes, mamá, papá.

Los amo.

Agradecimientos

FRANCISCO JARAMILLO, PROFESORA DORIS SÁEZ, MATÍAS TAUCARE

Este trabajo de Tesis fue posible gracias a todo el equipo del proyecto FONDEF Idea 2019 (ID19I10363) a cargo de la Agencia Nacional de Investigación y Desarrollo. Mención especial también al programa Enlace VID perteneciente a la Vicerrectoría de Investigación y Desarrollo de la Universidad de Chile ya que de él ha nacido esta propuesta.

Tabla de Contenidos

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

La crisis climática en el mundo avanza a pasos agigantados y está trayendo repercusiones medioambientales de gran magnitud. Según un último reporte del panel intergubernamental sobre cambio climático [1], los efectos de la crisis están ocurriendo más rápido y de forma más intensa de lo que se había previsto. Entre los puntos más preocupantes es que este proceso está afectando a todo el ciclo del agua [2] y perjudicando directamente a las fuentes de aguas subterráneas, que corresponden a la mayor reserva de agua fresca a nivel mundial [3].

Los pozos subterráneos actualmente suministran agua potable a billones de personas alrededor del mundo, sin embargo, estudios recientes muestran que millones de pozos están en riesgo de secarse [4], producto principalmente de que el nivel de las aguas subterráneas disponibles disminuye cada día más [5]. Esta alarmante noticia no es algo nuevo en el caso de Chile, que posee zonas declaradas en "crisis hídrica" y que no cuentan con un suministro constante de agua potable no solo producto del cambio climático, sino también de una insuficiente gestión de los recursos hídricos [6].

La escasez de agua en Chile se ha agravado aún más en la última década producto de factores como el profundo déficit de precipitaciones [7], el considerable avance de la desertificación y la megasequía que afecta a gran parte del territorio nacional [8]. Estos fenómenos están trayendo repercusiones muy negativas para la vegetación, la fauna y han escalado en crisis humanitarias y económicas en los territorios [6]. La situación empeora año a año y entre los factores más influyentes de tipo antropogénico, a nivel local, se encuentran la escasa regulación del uso de suelos y recursos hídricos [9], junto con la constante deforestación y el uso intensivo de agua por sectores productivos como la agricultura, que utiliza alrededor de un 70 % del agua extraída [10].

Ante este escenario tan complejo, poder asegurar la disponibilidad y la calidad de los recursos hídricos al que tienen acceso las comunidades en zonas donde la sequía está golpeando fuertemente, toma un rol fundamental y urgente. El actual código de aguas en Chile define que quien posea *derechos de aprovechamiento de agua* , puede explotar los recursos a perpetuidad y tranzar los derechos de aprovechamiento como un bien económico [11]. Lo anterior está provocando problemas como el acaparamiento y la mala distribución del recurso, que se ha convertido en un negocio, creando conflictos territoriales en torno al agua [6].

Los derechos de agua permiten a privados extraer una cantidad de agua determinada mediante un caudal máximo regulado a través de sistemas de telemetría para monitorear su cumplimiento. Según la resolución 1238 de la *Dirección General de Aguas* (DGA) [12], los componentes para los sistemas de *Mediciones de Extracción Efectiva* (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. También se cuenta con una red de control de calidad de aguas superficiales que envían información periódica del estado trófico de cuerpos de agua superficial ubicados en las macrozonas centro y sur.

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.

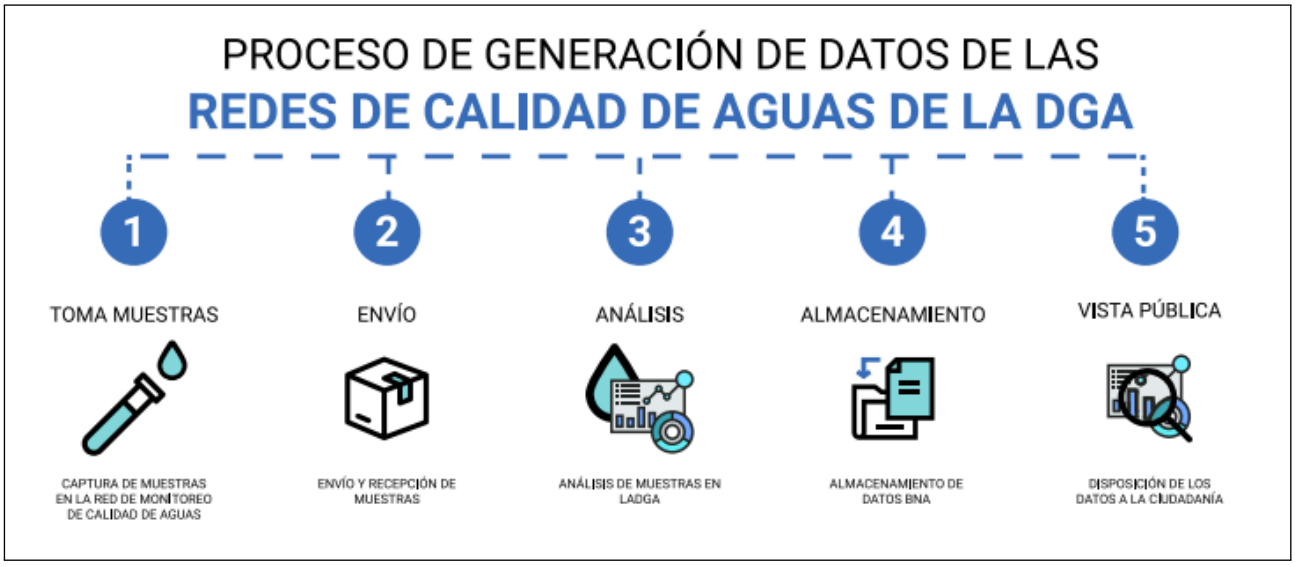


Figura 1.1: Estructura de adquisición y procesamiento de los datos provenientes de las redes de calidad de aguas de la DGA. [13]

La actual Red de Calidad de Aguas de la DGA considera un extenso y amplio trabajo multidisciplinario tanto a nivel regional como a nivel nacional [15], sin embargo, 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 muy

complejo 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.

1.2. Monitoreo de acuíferos y calidad de aguas

El monitoreo tanto de la calidad como de la cantidad de agua disponible de forma superficial y subterránea en diferentes acuíferos propone una oportunidad clave para poder gestionar este recurso vital con una mirada en la sustentabilidad y en el aseguramiento del acceso al agua como un derecho fundamental. Por tanto, se plantea que mediante un sistema de monitoreo inteligente es posible obtener información respecto al funcionamiento y el estado de los sistemas hidrogeológicos analizados, y de esta forma apoyar la toma de decisiones estratégicas y alertar sobre eventos de interés asociados a los recursos hídricos.

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.

En la actualidad, la DGA requiere un sistema de medición de extracciones que monitorea el nivel, además del flujo de extracción para el caso de aguas subterráneas [16]. Y 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 [17]. 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.

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) [18],[14] titulada “Norma Calidad Agua Potable”, establece qué se entiende por Agua Potable en Chile y plantea 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 tanto de parámetros físicos, como inorgánicos y orgánicos, estos parámetros pueden apreciarse en la Tabla 1.1.

Existe también la Norma Chilena 1333 [19] 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 para poder, por

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.			

ejemplo, detectar eventos de contaminación o de peligro para el consumo.

Estas normas tienen como objetivo principal proponer un marco normativo tecnificado en cuanto a los tipos de mediciones y la frecuencia con que deben realizarse estas, sin embargo, no abordan en su mayoría una etapa de disponibilización de los datos para los usuarios y que estos puedan realizar gestiones informadas.

1.2.2. Variables monitoreadas

1. **Conductividad eléctrica:** representa a la capacidad del agua de conducir corriente eléctrica, la cual depende de la cantidad de sólidos disueltos, tales como la sal. Esta variable es importante para medir calidad del agua, ya que representa la ionización del agua, y con ello, el contenido disuelto de iones y aniones. A partir del contenido de Conductividad Eléctrica, es posible determinar eventos antropogénicos o de causa natural, como lluvias o intrusión marina. Según la Organización Mundial de la Salud el rango de calidad para este parámetro, en uso potable, es en torno a 500 [µS/cm] a 1000 [µS/cm].

2. **Temperatura:** es importante medir la temperatura dado que afecta la capacidad de los microorganismos para resistir los contaminantes del agua. Por otra parte, está relacionada con el oxígeno que se encuentra en el agua, lo que regula la vida acuática. Su unidad de media es grados Celsius [°C].

3. **pH:** El 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, donde el 7 indica neutralidad. Según la Organización Mundial de la Salud el rango de calidad para este parámetro,

en uso potable, es en torno a 6,5 pH a 8,5 pH. Una variación en el pH puede indicar contaminación química en el agua.

4. **Nivel de agua:** esta variable se mide comúnmente a través de un piezómetro e indica la cantidad de agua que contiene el acuífero. Por lo general al tratarse del nivel, se relaciona directamente con las mediciones de presión del sensor inserto en el acuífero. Su unidad de medida son los pascales [Pa].

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 asociados al agua. 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. Sin embargo, una de las grandes falencias es la escasa información de monitoreo disponible asociado a la calidad de agua.

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 se 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 SQM [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 wambiental que autoriza el proyecto y establece que las mediciones se pueden distanciar en el tiempo hasta en 90 días [20].

Las falencias detectadas dentro de las políticas de monitoreo de calidad de aguas plantean una oportunidad de desarrollo tecnológico importante y pertinente en la actualidad. Así nace el proyecto Fondef ID19I10363, que abordó en profundidad la problemática con un desarrollo multidisciplinario y tanto de hardware como software que busca ser una solución efectiva y de bajo costo para ampliar y mejorar las redes de monitoreo de calidad de agua en Chile.

1.3. Proyecto Fondef Sistema experto para el monitoreo de acuíferos

En el marco del proyecto FONDEF ID19I10363 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. El objetivo general del proyecto es implementar un sistema de información abierto disponible para ayudar en la toma de decisiones oportunas por parte de gestores de recursos hídricos. Se considera así la implementación de un *Sistema Experto* (SE) que analiza e interpreta los datos capturados por los nodos sensores desarrollados para detectar *anomalías* en las varia-

bles fisicoquímicas monitoreadas.

El proyecto busca resolver un problema muy frecuente en el país, que es la obtención y el procesamiento de datos sobre recursos hídricos con una periodicidad y precisión suficiente que permita detectar de forma temprana problemas o alteraciones de las aguas. Este problema no ocurre por una falta de tecnologías disponibles, sino a que en su mayoría los equipos de medición especializados de calidad de aguas, con sistemas de almacenamiento de datos, poseen un alto costo e incluso a veces solo permiten realizar mediciones puntuales. Lo anterior, sumado a que en ocasiones las unidades de sensado se instalan en lugares de complejo acceso, dificulta aún más la obtención de la información, pues se deben poder acceder al dispositivo para descargar los datos de forma manual.

El avance de nuevas tecnologías de comunicación de bajo requerimiento energético, junto con la miniaturización de sensores y microcontroladores proponen una nueva alternativa para poder conseguir mediciones de múltiples parámetros de calidad de agua de forma continua y con capacidad de transmitir la información en tiempo real de forma remota. Sin embargo, soluciones disponibles en el mercado que incorporan estas nuevas tecnologías, tienen un costo aún mayor a las unidades de mediciones puntuales y esto supone un obstáculo para que empresas locales, que no se encuentran obligadas a realizar mediciones con mucha frecuencia, las incorporen en sus operaciones.

El objetivo de desarrollo del proyecto y la solución final propuesta es un sistema de monitoreo automático compuesto por una plataforma abierta (de libre acceso) con componentes de hardware y software, que permiten la recolección a bajo costo de mediciones de nivel y calidad de aguas superficiales y subterráneas, y que ofrece el acceso a información enriquecida para apoyar la toma de decisiones de los actores involucrados en la gestión de los recursos hídricos monitoreados y que ofrece el acceso a información enriquecida para apoyar la toma de decisiones de los actores involucrados en la gestión de los recursos hídricos monitoreados [23].

En conjunto con el desarrollo del nodo sensor multiparámetro, que mide Conductividad Eléctrica, pH, Temperatura, Nivel freático y Turbiedad el proyecto consideró también una parte de desarrollo de software que consiste en el Sistema Experto basado en el procesamiento de información en la Nube. Aquí se incorpora el conocimiento de expertos y expertas junto a herramientas de procesamiento y visualización mediante un tablero de control, el cual se alimenta de la información recolectada por los sensores de bajo costo conectados vía Internet al sistema. Un ejemplo de esta interacción y en donde interactúa el usuario, a través del tablero de control, puede apreciarse en la Figura 1.2.

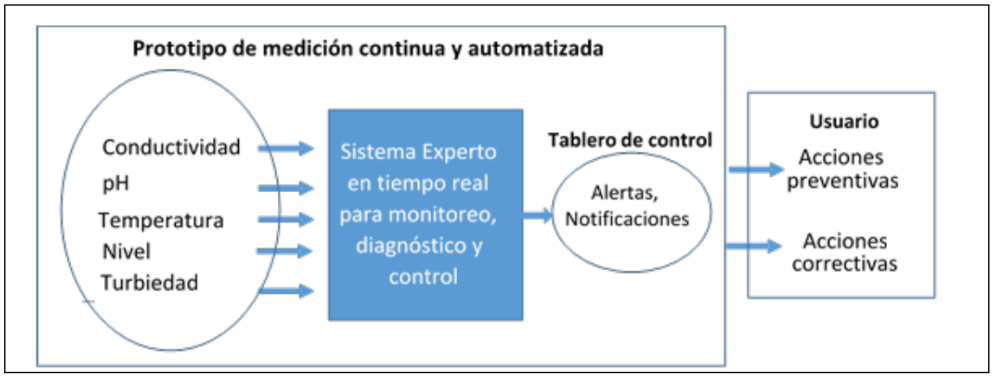


Figura 1.2: Esquema de información, comunicación y procesamiento del sistema propuesto en el proyecto.

El objetivo de desarrollo del proyecto y la solución final propuesta es un sistema de monitoreo automático compuesto por una plataforma abierta (de libre acceso) con componentes de hardware y software, que permiten la recolección a bajo costo de mediciones de nivel y calidad de aguas superficiales y subterráneas, y que ofrece el acceso a información enriquecida para apoyar la toma de decisiones de los actores involucrados en la gestión de los recursos hídricos monitoreados y que ofrece el acceso a información enriquecida para apoyar la toma de decisiones de los actores involucrados en la gestión de los recursos hídricos monitoreados [23].

1.3.1. Sistema experto

La principal componente de producción de la etapa de desarrollo de software y la propuesta de este trabajo de tesis es el SE y el panel de datos o plataforma. Este sistema busca poder combinar el conocimiento de un experto en sistemas dinámicos y complejos, como es el caso del monitoreo de acuíferos, junto con un programa computacional que analiza las series de tiempo, provenientes de los sensores, de manera análoga a como lo haría un experto 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.

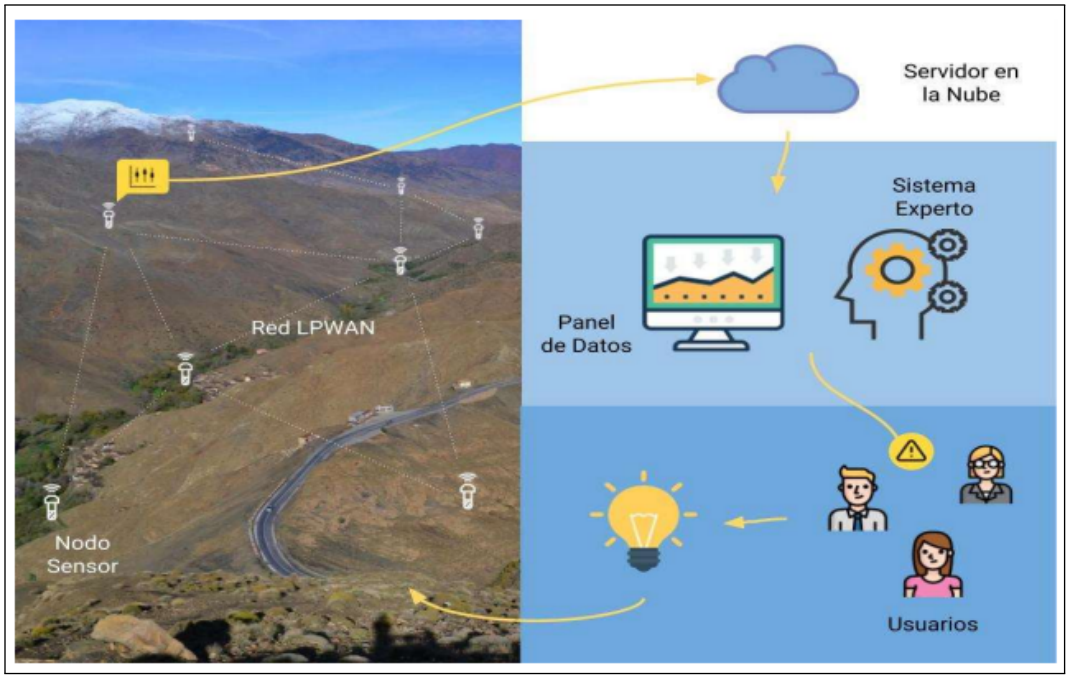


Figura 1.3: Vista general del proyecto.

Los datos de monitoreo de acuíferos presentan comportamientos dinámicos complejos de caracterizar en el tiempo para lógicas de control tradicional. Adicionalmente, cada acuífero puede presentar una dinámica totalmente diferente a otro cercano. Así, cada acuífero podría asociarse a un estado libre de riesgos y caracterizarse 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* que 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.

Algunas de las posibles situaciones que podrían detectarse con estos sistemas son la georreferenciación de puntos que aportan salinidad a partir de los cambios 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 quizás extracciones o descargas de terceros a partir de cambios en nivel o la temperatura. Con información como esta los gestores de los recursos podrían evitar la explotación insostenible de ciertos pozos, el cambio a otros en mejor estado para no interrumpir el suministro o levantar una alerta a los servicios medioambientales por posible contaminación.

Considerando los objetivos de desarrollo del proyecto y la importancia de obtener información valiosa de parámetros de calidad de agua y el estado de los datos monitoreados de acuíferos es que se caracteriza un nicho de potenciales usuarios sin los conocimientos técnicos necesarios y que podrían beneficiarse de una herramienta analítica integrada que se alimente de los datos y realice un procesamiento automático de ellos. Adicionalmente se ha encontrado que la aplicación de herramientas asociadas al Machine Learning (ML) proponen la robustez suficiente para la tarea de detectar anomalías y levantar alarmas a través de una plataforma de visualización.

1.4. Hipótesis

- La capacidad de poder generar, entrenar y comparar, de forma simple, múltiples modelos basados en aprendizaje de máquinas para la tarea de detección de anomalías facilitará el acceso a herramientas de procesamiento de alto nivel en un tiempo reducido a usuarios sin un conocimiento técnico abundante en la materia.
- 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 monitoreo de los recursos hídricos.

1.5. Objetivos

El objetivo principal y los objetivos específicos del trabajo se definen tomando en consideración tanto los requerimientos que nacen desde el proyecto mismo como los que provienen de una revisión bibliográfica y una caracterización del nicho para abordar el desafío del monitoreo automático de la disponibilidad y calidad de aguas de acuíferos subterráneos en Chile.

1.5.1. Objetivo general

Diseñar y desarrollar una plataforma web para el procesamiento en línea y visualización de datos hidrogeológicos con la finalidad de detectar anomalías basadas en ML para generar alarmas considerando una selección automática de los modelos con mejor desempeño; permitiendo realizar un diagnóstico oportuno del estado del acuífero monitoreado y de la calidad del agua.

1.5.2. Objetivos específicos

- Diseñar una estrategia de pre-procesamiento, curado de datos automático y una selección de características configurable por el usuario.
- Integrar modelos de detección de anomalías basadas en series de tiempo hidrogeológicas a través de modelos de ML para emular la capacidad de generalización de un conjunto de expertos.
- Diseñar y desarrollar una plataforma web para el procesamiento automático de los datos y el monitoreo a través de visualizaciones, de alarmas de anomalías detectadas y análisis mediante métricas de desempeño.

1.6. Contribuciones

- Revisión bibliográfica en estrategias detección de anomalías supervisadas aplicadas a series de tiempo hidrogeológicas y sistemas multivariables. Evaluación de eficacia de aplicación de estrategias de ensamble y apilamiento de modelos para mejorar la detección de distintos tipos de anomalías.

- Estructura de automatización de análisis exploratorio de la data, 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 controlable 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 data histórica.

1.7. Estructura del documento

Este trabajo se organiza en capítulos como sigue: 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 esta Tesis. Se presenta también aquí los principales avances del estado del arte asociado a estas temáticas y en donde se analiza la pertinencia de las distintos alternativas y modelos, además de sus respectivos funcionamientos y aplicaciones al problema de detección de anomalías.

En el Capítulo 3 se presenta la metodología del trabajo y se abordan los temas relacionados a las estructuras de funcionamiento y las propuesta de automatización para la incorporación de estrategia de entrenamiento y selección de modelos aplicadas a la detección de anomalías en el monitoreo de acuíferos.

En el Capítulo 4 se muestra el desarrollo de la plataforma junto a una descripci y justificación de los principales requerimientos propuestos para el diseño y desarrollo de la plataforma. Aquí se detallan las distintas funcionalidades de la plataforma en cuanto al procesamiento de datos, entrenamiento y selección de múltiples modelos y su integración como aplicación WEB. Se presenta también el caso de estudio, la obtención de los datos experimentales reales y otros casos sintéticos, con el objetivo de validar la capacidad de generalización de la estrategia propuesta.

Dentro del Capítulo 5 se presentan los resultados estudiados aplicados al caso de estudio específico y acotado que corresponde a un pozo monitoreado en la localidad de Horcón, Región de Valparaíso. La base de datos contiene mediciones de parámetros fisicoquímicos del agua del pozo registradas de forma continua con una frecuencia de muestreo de una hora desde marzo de 2013 a febrero de 2017. Se analiza así la pertinencia de la aplicación y de los resultados obtenidos para la detección de anomalías en el caso de estudio.

Finalmente, en el Capítulo 6 se presentan las discusiones en torno las funcionalidades y los cumplimientos de los objetivos junto a las principales conclusiones« de la investigación y del trabajo de tesis .

Capítulo 2

Marco teórico

En la literatura se presentan de manera extensa 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 y 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 [24].

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 [25]. 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 [26] :

"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.*

Un conjunto de datos anómalos podría representar desde un fraude de tarjetas de crédito [27] hasta un defecto estructural en un edificio [28] o una condición médica desconocida detectada a través del procesamiento de imágenes [29]. 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 [30], 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 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 podría representa 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 35° 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 [31]:

- **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 que 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 supervisado y los que son no 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 cual 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.

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". Los métodos que se entrenan de forma supervisada utilizan la información contenida en las

etiquetas para crear modelos numéricos de las características que definen y diferencian los datos normales de las 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 con anterioridad y etiquetados, generalmente de forma manual por une experto, para entrenar modelos de clasificación binaria que diferencian entre una muestra *normal* (etiqueta negativa o 0) y una muestra anómala (positiva o 1).

2.2.3. Métricas en modelos de detección supervisados

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

Un ejemplo para definir las distintas componentes de las métricas de clasificación en un problema de detección de anomalías es posible realizar un símil con un problema de clasificación binario. Así una medición multivariable de calidad de agua puede dar positivo (1) o negativo (0) utilizando un modelo de detección de anomalía supervisado.[32]

Así, se definen los siguientes conceptos en base a si el clasificador 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):*** *Dada 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.*

En resumen, si el modelo clasifica eficazmente como positivo es un Verdadero positivo o negativo y si no, es un falso positivo o negativo según corresponda. Ahora con estas definiciones es posible definir diferentes métricas que relacionan el desempeño relativo a la hora de identificar correctamente los datos anómalos y normales. Una forma gráfica de poder comparar la composición final de la clasificación es a través de la matriz de confusión se ilustra en la Figura 2.1

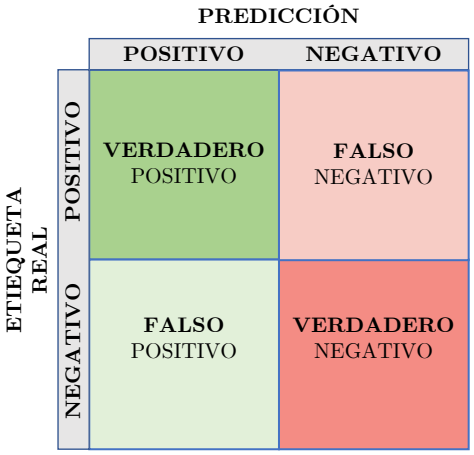


Figura 2.1: Matriz de confusión.

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 F1

También conocido como *F1 Score*, es una métrica que combina en un valor tanto la precisión (P), como la exhaustividad (R) y toma el promedio armónico de la forma:

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

o equivalentemente:

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

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 su 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. 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.

2.3.1. Series de tiempo

En el mundo real 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 [33]. 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.

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

Para formalizar aún más el concepto una definición dada por Morris en [34]:

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 de la serie de tiempo.*

Las series de tiempo pueden tener una forma regular o irregular dependiendo de la frecuencia en la que se muestren u obtiene 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.

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 general en este reporte se trabajará con series de tiempo multivariable pero los métodos de análisis se analizarán en algunos casos de forma univariable para resolver el problema de detección de anomalías [31].

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. Otras de sus múltiples aplicaciones consiste como el procesamiento de señales para el diagnóstico y pronóstico de fallas. [35].

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.

2.4. Métodos de Ensamble

Los métodos de ensamble, a diferencia de los métodos clásicos utilizados para problemas de clasificación o regresión, no se basan en solo un único modelo con una alta eficacia, sino que, utilizan un conjunto de clasificadores o regresores base, en algunos casos de menor eficacia, para así obtener un modelo compuesto que utiliza la salida de los distintos modelos individuales con el objetivo de mejorar el desempeño de los clasificadores individuales utilizados en el ensamble, un ejemplo de esto se puede apreciar en 2.2.

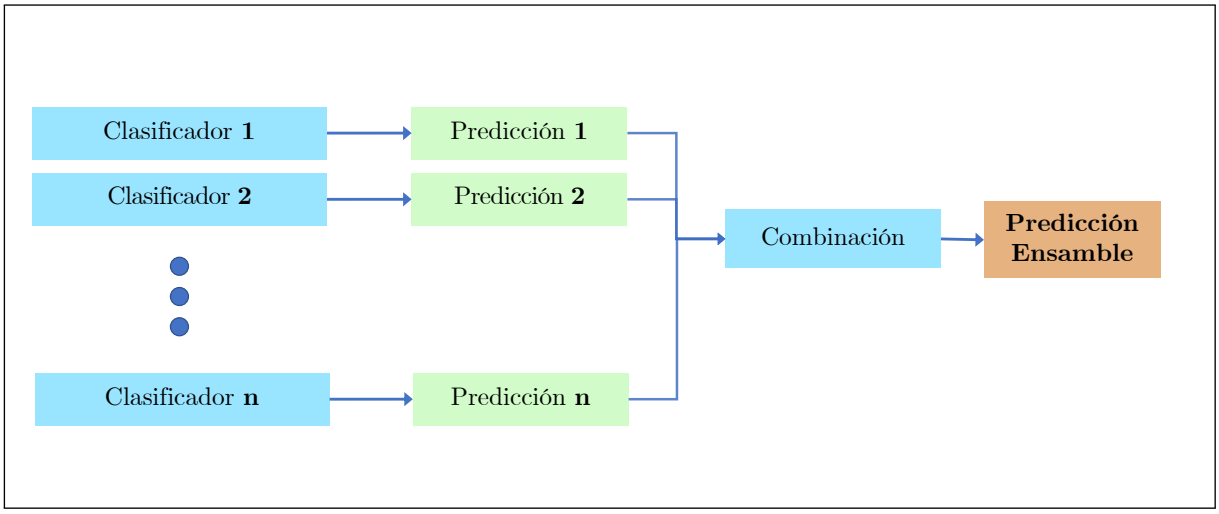


Figura 2.2: Modelo de ensamble de n clasificadores mediante agregación o combinación en su salida, en la práctica suele ser mediante votación de mayoría.

Una idea común de 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 clasificadores bases. 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 tipo de clasificador.

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

- **Bagging:** Corresponden a técnicas basadas en el entrenamiento mediante agregaciones

de bootstrap [36] 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 predictores y combinar sus salidas en **paralelo** [37]. 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* [38].

- **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 aquennas muestras con un error más significativo para alimentar un siguiente estimador. En cada iteración se almacena el desempeño general de cada estimador para la ponderación final. La estrategia general es un tipo de *Adaptive Boosting* [39]. Se desarrollan luego también las basadas en *Gradient Boosting*[40] que optimizan una función de pérdida de forma secuencial y que desde el principio mejora el desempeño en cada iteración.

- **Stacking**: También se le conoce como Stacked Generalization o apilamiento de modelos [41] y es una técnica que aborda la agregación como un proceso de aprendizaje en si. Su funcionamiento consiste en entrenar un *meta-modelo* 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 [41, 42, 37]. 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* y los basados en *gradient boosting* como lo son *XG-Boost*, *CatBoost* y *LightGBM*.

2.4.1. Modelos basados en Boosting

Son algortimos que utilizan un ensamble de múltiples clasificadores, en este caso árboles de decisión, que se añaden de forma constante para mejorar el desempeño en cada iteración.

2.4.1.1. XGBoost

Clasificaciones basados en boosting son una alternativa a la utilización de redes neuronales en casos en donde no se tiene un gran número de etiquetas en los datos. 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 para estos clasificadores débiles [43].

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 un desempeño

de que supera a otros algoritmos basados en Gradient Boosting Decision Trees.˙

Su funcionamiento se basa en que el ensamble de árboles se construyen de forma secuencial en donde cada árbol subsecuente apunta a reducir el error del árbol anterior, aquí es donde se aplica el método del descenso del gradiente. Cada árbol aprende de sus predecesores y actualiza sus errores residuales en cada iteración, por lo tanto, cada árbol que crezca en la siguiente secuencia aprende de los estados actualizados de los residuos.

De forma muy general, XGBoost utiliza el método del descenceso del gradiente en una función de pérdida desde los pasos anteriores. La novedad de XGBoost, es que se modifica el algoritmo de gradiente para los métodos boosting tradicionales y poder funcionar con cualquier función de pérdida que sea diferenciable.

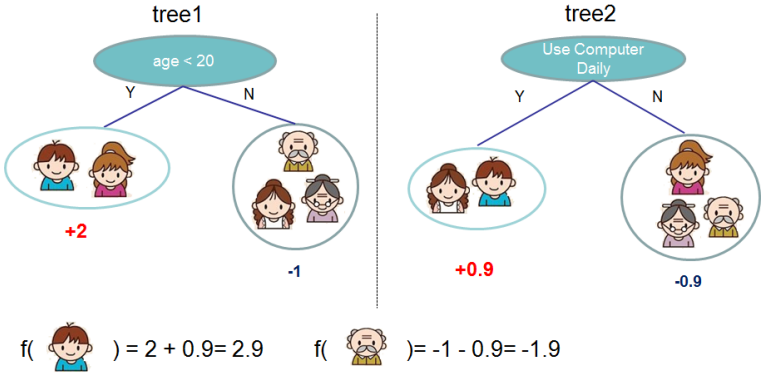


Figura 2.3: Modelo de emsable de árboles de decisión. El valor para la predicción final para cada muestra es la suma de las predicciones de cada árbol [43]

La figura 2.3 es un ejemplo de la salida para la predicción de de por ejemplo si determinada persona le gustaría jugar un videojuego X. Así una predicción de salida de un método de ensamble está compuesta por la suma de todas las puntuación o salidas de cada árbol individual. Otro factor importante es que estos dos árboles se complementar. Matemáticamente esto se puede escribir de la siguiente forma:

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

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 último la función a minimizar queda de la forma:

$$\text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.2}$$

Es necesario destacar que una implementación como XGBoost es similar a Random Forest en su funcionamiento mediante ensamble y solo cambia la forma en la que se entrenan los modelos y se ajustan los parámetros en la ejecución.

Diferentes aplicaciones de XGBoost se han reportado en el análisis de calidad de agua, en [44] se analiza su poder de predicción a la hora de predecir índices de calidad de agua 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 [45]. En el campo de detección de intrusos o de ataques en redes [46] presenta excelentes desempeños con accuracys de 98.7 % con tasas de error bajas.

2.4.2. Modelos basados en Bagging

2.4.2.1. 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 [38]. 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* [36]. 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.

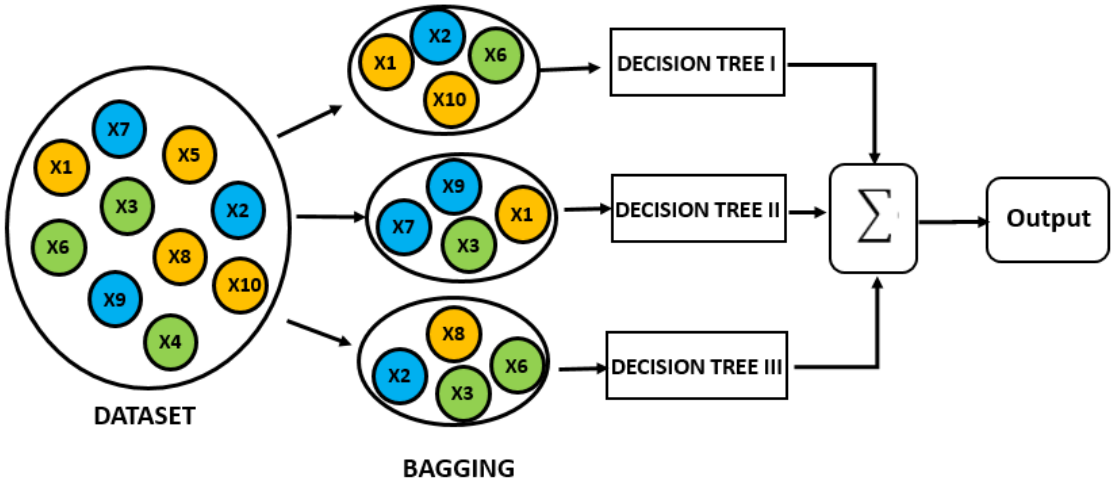


Figura 2.4: Ejemplo de Random Forest y metodología Bagging y la integración de las alidas. En el caso de Random Forest en Clasificación se suele utilizar *Votación por mayoría*.

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.

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* [47] y más recientemente la gran atención de variantes basadas en *gradient boosting* [48], 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 [49].

Las técnicas de stacking o de apilamiento llevan estudiándose desde hace bastante tiempo [41] 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 utilizado como set 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.

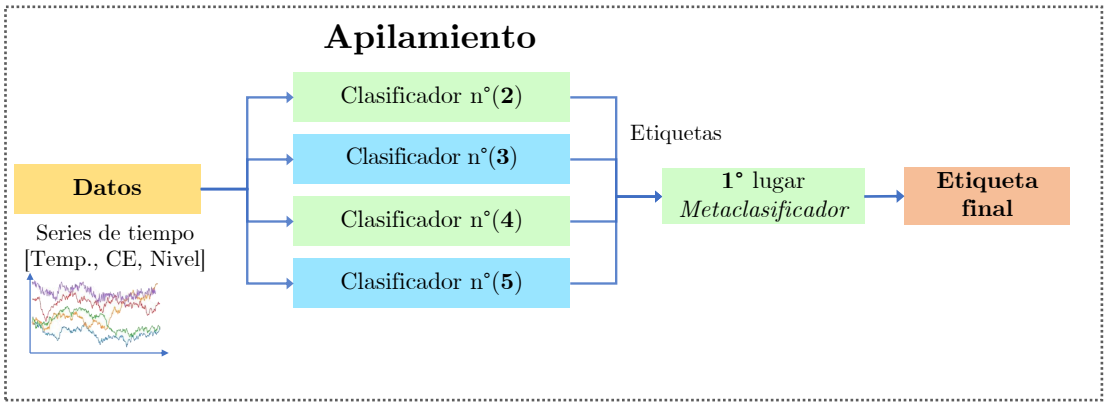


Figura 2.5: Ejemplo de estructura de apilamiento de modelos.

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 simplemente votación por mayoría para decidir 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.

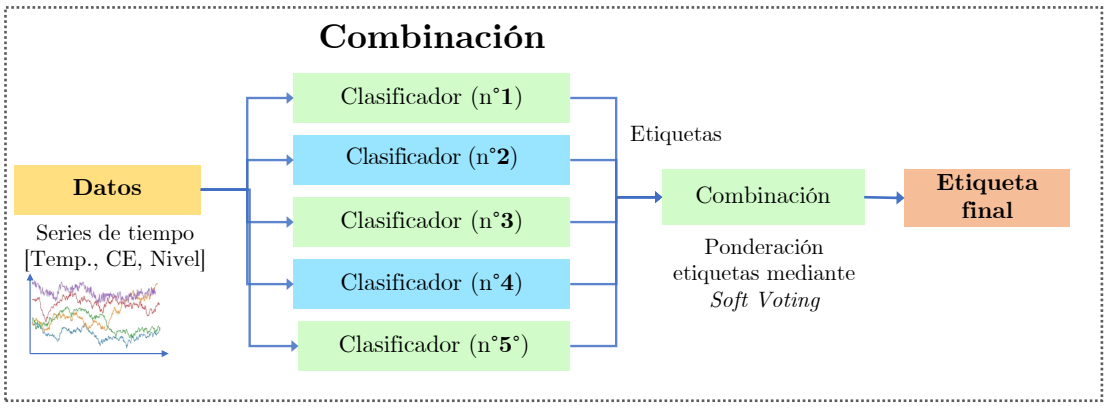


Figura 2.6: Ejemplo de estructura de apilamiento de modelos.

Algorithm 1 Apilamiento de modelos (Stacking)

Entrada: Set de datos de entrenamiento: $\mathbb{D}$, Clasificadores base: $\mathbb{C}_1, \mathbb{C}_2, ..., \mathbb{C}_n$
Salida: Modelo de Ensamble por apilamiento entrenado: $\hat{\mathbb{C}}$

Inicialización:

1: **Paso 1**: Se entrenan los clasificadores individuales $\mathbb{C}_i$ en el set de datos $\mathbb{D}$ y se ajustan sus hiperparámetros.
2: **Paso 2**: Se seleccionan los k mejores clasificadores y se ajustan mediante optimización bayesiana de hiperparámetros.
3: **Paso 3**: Se selecciona el mejor modelo como el meta modelo o meta clasificador $\mathbb{C}_0$ que debe reentrenarse a partir de la salida del resto de clasificadores para apilar todo esto en una salida dada por el modelo final.
4: **return** $\hat{\mathbb{C}}$

2.5. Aprendizaje de máquinas automatizado

Desde los comienzos de la computación nació el interés en la posibilidad de que los computadores pudieran aprender tareas complejas y realizarlas de forma autónoma, es así como surgió el nicho de investigación del Aprendizaje de máquinas o *machine learning* (ML). Los primeros científicos en plantear esta problemática estudiaban formas de modelar matemáticamente el funcionamiento de redes neuronales para imitar el comportamiento humano. En la actualidad estas técnicas pueden encontrarse en los más diversos campos de conocimiento y ya superan la habilidad humana para resolver ciertas tareas como la detección de objetos o para jugar ajedrez [50].

El proceso iterativo asociado al entrenamiento de modelos de ML conlleva un período de limpieza inicial de los datos, un proceso de exploración de ellos y luego un proceso iterativo de experimentación con distintos modelos e hiperparámetros para analizar sus desempeño dependiendo del caso de uso y del tipo de datos con los que se trabaje. Estos requerimientos del proceso son frecuentemente una piedra de tope para el desarrollo de nuevas aplicaciones

para ciudadanos y principiantes [51].

El aprendizaje de máquinas automatizados plantea que una automatización del proceso iterativo, de diseño y de experimentación en aplicaciones de ML pueden ahorrar tiempo [52] recursos y acercar tecnologías de alto nivel a nuevos casos de uso [53]. En la actualidad, existen diversos proyectos que buscan en automatizar el proceso completo, sin embargo, es algo que avanza de forma gradual y hasta cierto punto se requerirá siempre de una supervisión y entendimiento de la materia, por lo que es posible afirmar que es un proceso semiautónomo en el que se busca liberar liberar de forma gradual al científico o ingeniero a cargo de estas tareas para priorizar otras tareas.

Capítulo 3

Detección de anomalías y aprendizaje automático

3.1. Metodología de detección de anomalías

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. Edel trabajo propuestos en la Sección 1.5. Respecto al diseño y desarrollo de la plataforma se abordarán en concreto en el capítulo 4. En cuanto a la integración de modelos de detección de anomalías en series de tiempo hidrogeológicas mediante el apilamiento y ensamble se realizan pruebas cuantitativas en el lenguaje de programación *Python* para comparar la mejor versión de cada modelo con su respectiva optimización de hiperparámetros además de un apilamiento de ellos. Una vista general de las distintas etapas del trabajo se aprecia en la figura 3.1.

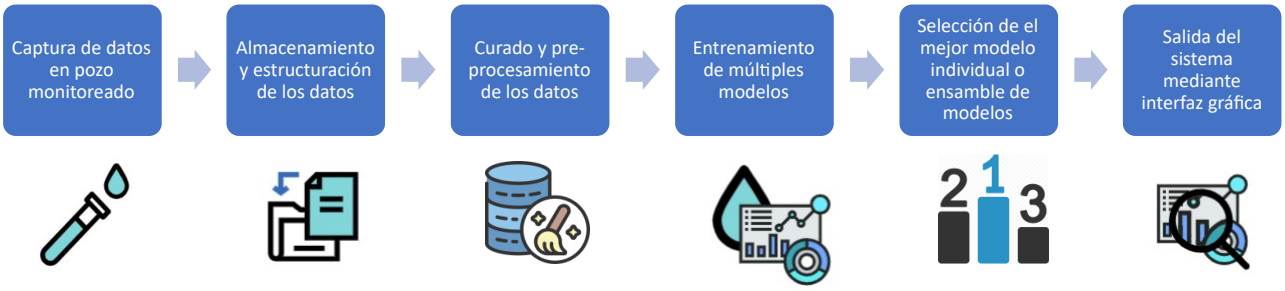


Figura 3.1: Estructura del flujo de información y procesamiento de datos desde la recolección por parte del usuario hasta el procesamiento en la plataforma.

Para las simulaciones de diferentes casos de uso asociados a la cantidad de información histórica o datos disponibles es que se dividió el set de datos original en cuatro, en donde todos son subconjuntos del original abarcando desde solo el primer año de datos históricos, luego los dos primeros años y así hasta llegar a los cuatro años disponibles de información. Esto con la finalidad de analizar el funcionamiento de la estrategia y su robustez a la falta de datos para el entrenamiento y los cambios en las dinámicas estacionales.

El manejo de datos perdidos, que corresponde a la etapa de "curado y pre-procesamiento", se realiza mediante una imputación mediante el valor medio de la categoría. Aquí también se procesa un archivo de valores separados por comas (.csv) para transformarlo en un *Data-Frame* y trabajar con ellos.

Para el entrenamiento de múltiples modelos se consideraron los presentados en la sección 2.3, sin embargo, al ser un sistema modular podrían agregarse más o cambiarse otros en futuros trabajos. La validación de desempeño de los distintos modelos aplicados a la tarea de detección de anomalías se realiza de forma iterativa mediante una búsqueda inicial de los modelos con mejor desempeño en un conjunto de hiperparámetros por defecto y luego en una búsqueda y optimización bayesiana de hiperparámetros [54] de los n mejores modelos a través de la librería *scikit-optimize*.

Los análisis cuantitativos se realizan a través de comparaciones según las métricas: *valor F1*, por lo que a su vez se consideran para ello la *precisión* y la *exhaustividad* de los métodos, pues la tarea de detección de anomalías suele presentar un desbalance de clases etiquetadas considerable. Se analizó la capacidad de generalización de los distintos modelos individuales y en los diferentes tipos de anomalías en los datos del caso de estudio y en las distintas variaciones sintéticas generadas a partir de agregar nuevos atributos.

La estructura final de la metodología cuantitativa en relación a la integración dentro de la plataforma puede apreciarse a través de un esquema en la figura 3.2.

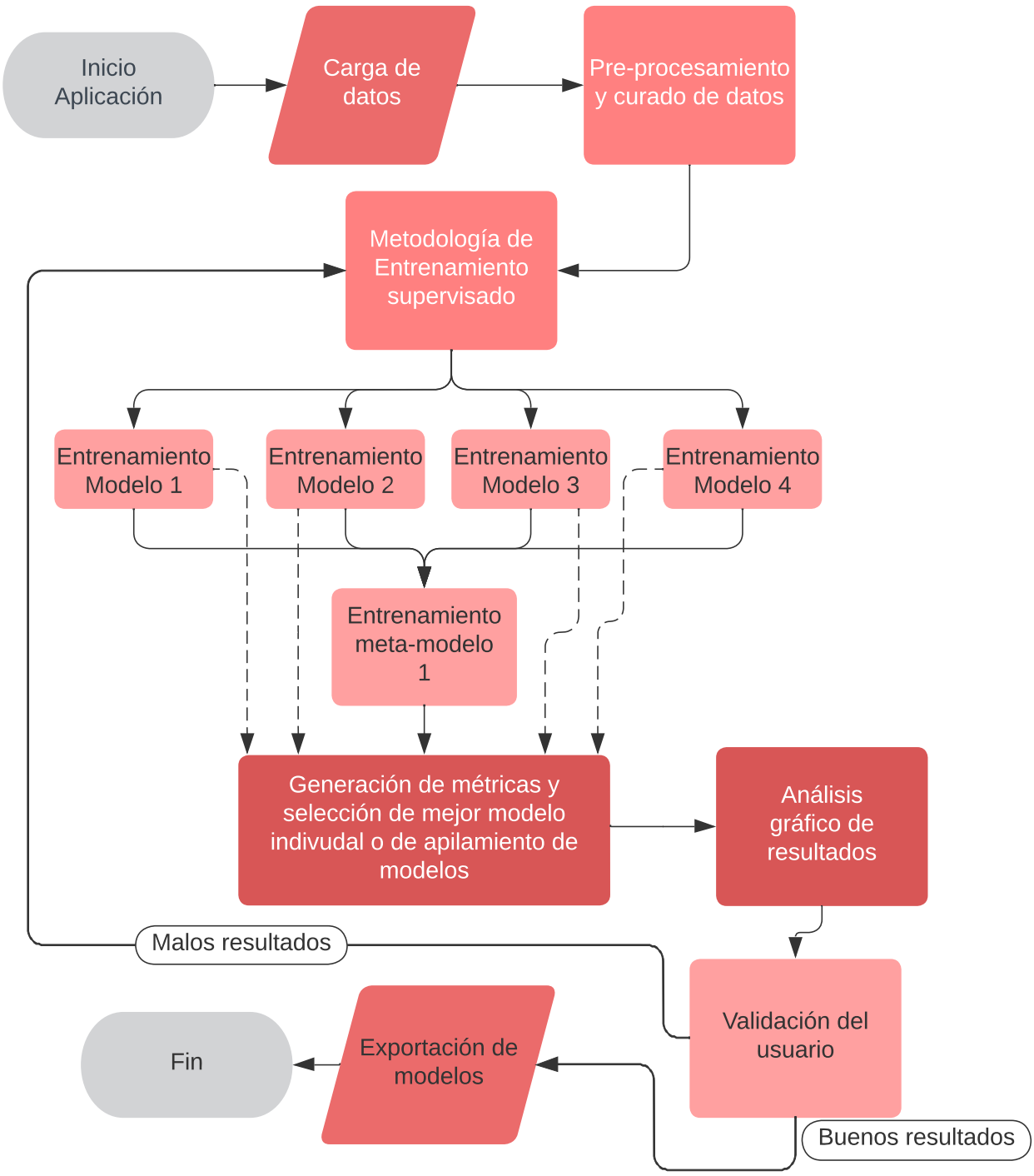


Figura 3.2: Esquema integración plataforma y sistema automático de detección de anomalías.

3.2. Casos de estudio: Pozo monitoreado en localidad de Horcón - Universidad de Chile

Para el desarrollo de los diferentes modelos a estudiar se trabajará con una base de datos proporcionada por la Universidad de Chile para el desarrollo del proyecto. Estos datos contienen la información de una estación de monitoreo de un pozo ubicada en la localidad de Horcón, Región de Valparaíso. La base de datos contiene mediciones de parámetros del agua de forma continua que fueron 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 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.

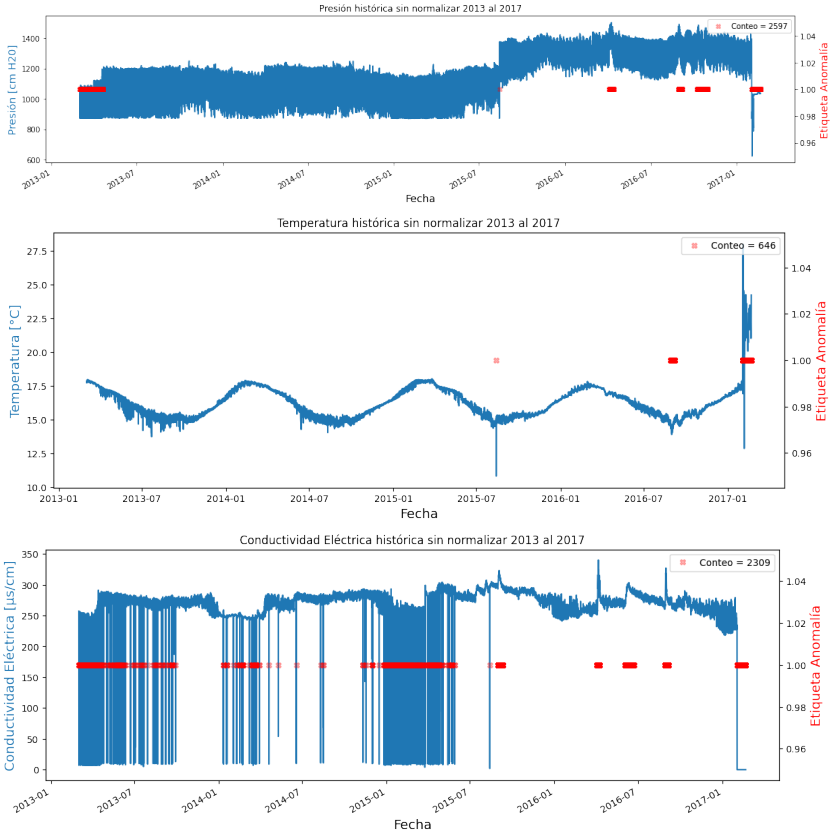


Figura 3.3: Base de datos de Horcón etiquetado por experto, en rojo una superposición de las marcas temporales en donde se ha etiquetado un dato anómalo. a) Presión; b) Temperatura ; c) Conductividad eléctrica

3.2.1. Datos disponibles

El dataset está compuesto por 34.827 mediciones sin mediciones perdidas y ningún NaN o dato nulo, por lo que ya ha pasado por una etapa de pre-procesamiento. Una descripción esta-

dística más detallada de cada una de las variables fisicoquímicas monitoreadas, sin considerar las columnas correspondientes a las etiquetas de anomalías, se presentan en la siguiente tabla:

Tabla 3.1: 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 3.4, 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.

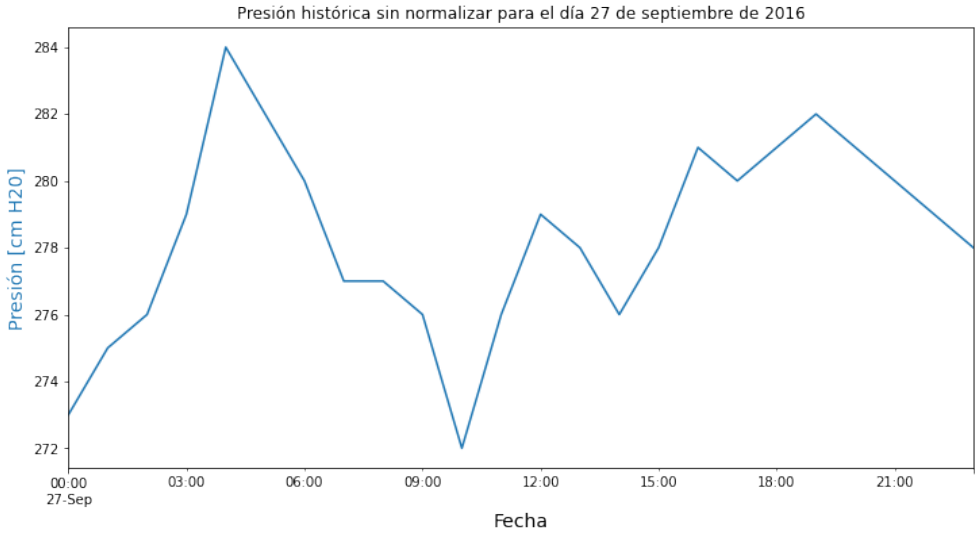


Figura 3.4: Cambios en la presión en un día

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, estas varia. De forma estacional, las temperaturas disnminuyen desde el período de abril a septiembre y lue-

go aumentan desde octubre a marzo, es bastante directo ver la forma oscilatoria periódica marcada por esta estacionalidad en la figura 3.3b.

Para esta base de datos la conductividad eléctrica (CE) (figura 3.3c) 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 3.5, 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 cree que es muy probable 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.

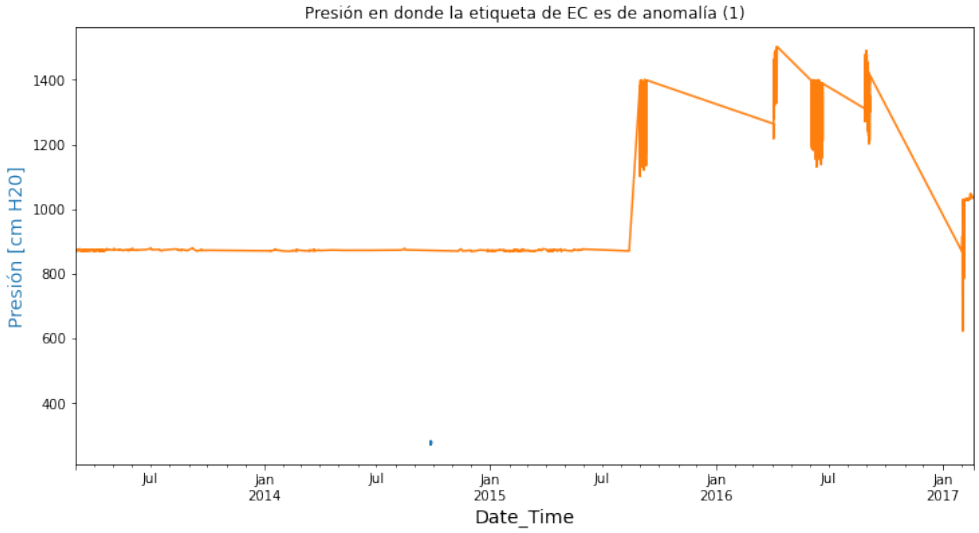


Figura 3.5: Nivel de columna de agua cuando la etiqueta asociada a la variable de conductividad eléctrica indica anomalía (1)

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 estándard, comparándose gráficamente con el valor en bruto en la figura 3.6. Además se presentan también los gráficos de pares variables en la figura 3.7a) y un mapa de calor para analizar correlaciones de forma gráfica en la figura 3.7b).

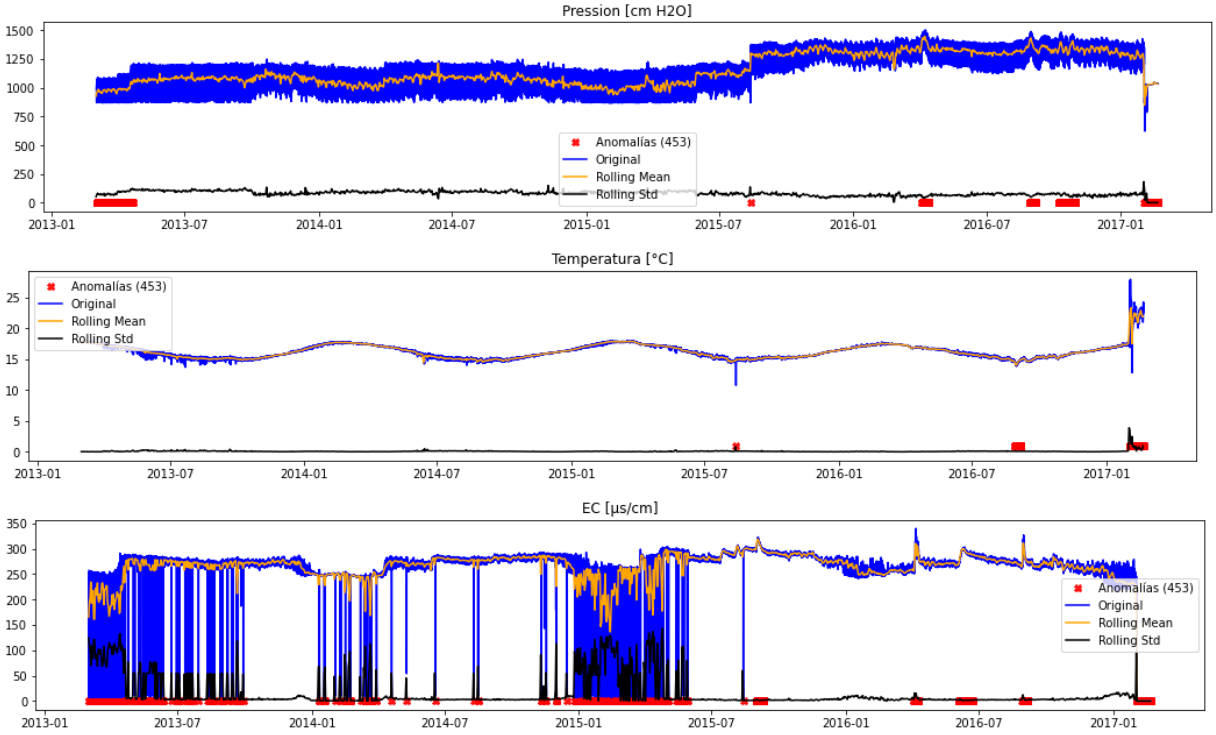


Figura 3.6: Análisis por ventana deslizante de 24 horas. Original en azul, promedio deslizante en amarillo, desviación estándar deslizante en negro, etiquetas de anomalías en rojo.

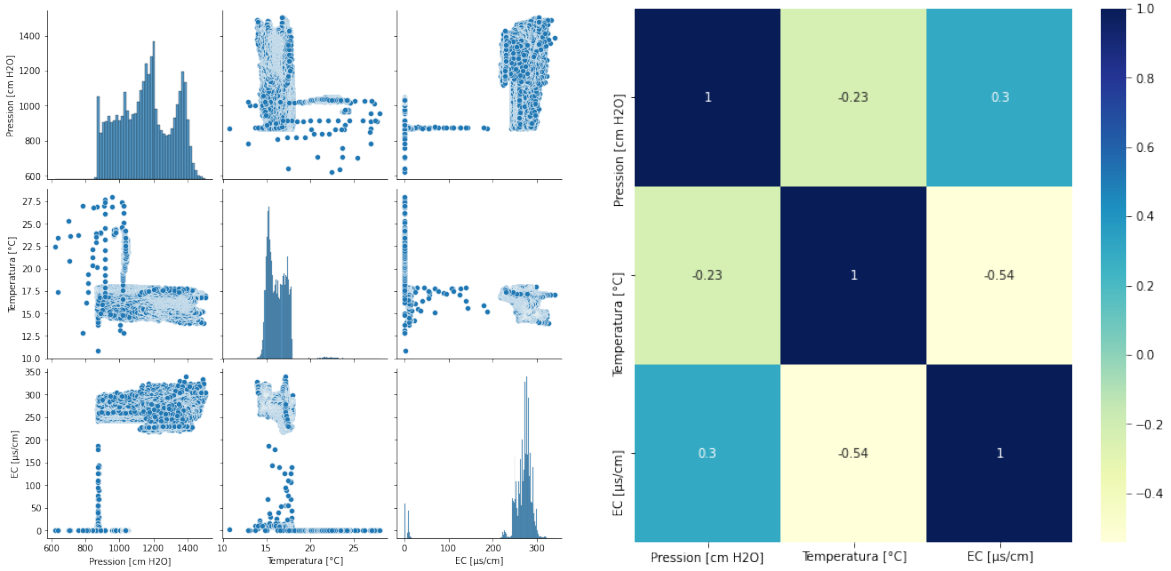


Figura 3.7: Relaciones entre características. a) Gráficos de pares de variables b) Mapa de calor

Es importante destacar aquí que para el caso del análisis con ventanas deslizantes es posible ver, en el caso de la figura 3.6 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 3.6 b) y no se aprecia una correlación tan directa para el caso de la presión en la figura 3.6 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.

3.3. Desarrollo de la estrategia de detección de anomalías

Para el desarrollo de la estrategia de detección de anomalías y el entrenamiento automático se trabajó con el lenguaje de programación *Python* en su versión 3.9.13. Además se trabajó con las siguientes librerías, disponibles también en el documento "requirements.txt" del repositorio del proyecto.

El desarrollo se

- pycaret==2.3.10
- streamlit==1.12.0
- pandas==1.3.5
- pytz==2022.1
- xgboost==1.6.1
- lightgbm==3.3.2
- bokeh==2.4.3
- scikit-optimize==0.9.0
- scikit-learn==0.23.2

3.4. Entrenamiento y selección de modelos

Como se detalló en la metodología el entrenamiento comienza luego de la etapa de preprocesamiento y en ella se genera una grilla de búsqueda de modelos para los datos de entrenamiento cargados. En el caso de los datos de un año la grilla de búsqueda puede apreciarse en la Figura 3.8

Grilla de búsqueda de modelos:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.961100	0.990700	0.966100	0.988100	0.976900	0.853600	0.856300	0.094000
catboost	CatBoost Classifier	0.960300	0.992300	0.965500	0.987700	0.976400	0.850600	0.853500	0.883000
lightgbm	Light Gradient Boosting Machine	0.959300	0.991800	0.966800	0.985200	0.975900	0.845400	0.847700	0.027000
rf	Random Forest Classifier	0.958700	0.991100	0.965900	0.985400	0.975500	0.843600	0.845900	0.117000
xgboost	Extreme Gradient Boosting	0.958200	0.991400	0.969500	0.981300	0.975300	0.838400	0.839600	0.123000
dt	Decision Tree Classifier	0.952800	0.917500	0.967600	0.976800	0.972200	0.816900	0.817700	0.008000
gbc	Gradient Boosting Classifier	0.946800	0.988300	0.944600	0.992600	0.967900	0.811300	0.820600	0.101000
ada	Ada Boost Classifier	0.930300	0.983300	0.926400	0.991200	0.957700	0.760800	0.775200	0.045000
qda	Quadratic Discriminant Analysis	0.879000	0.931000	0.941200	0.918800	0.929800	0.489700	0.492300	0.006000
nb	Naive Bayes	0.869500	0.922900	0.885800	0.958000	0.920400	0.560700	0.574400	0.005000

Figura 3.8: Grilla de búsqueda de modelos para el primer año de datos.

Luego se genera el modelo de ensamble por apilamiento considerando, para este caso, a Extra Trees classifier como el *metaclasificador* y a Catboost, LightGBM, Random Forest y XGBoost como los modelos que componen el ensamble como se explicaba en la Figura 2.5. La matriz de confusión final de este modelo se aprecia a continuación en la Figura 3.9

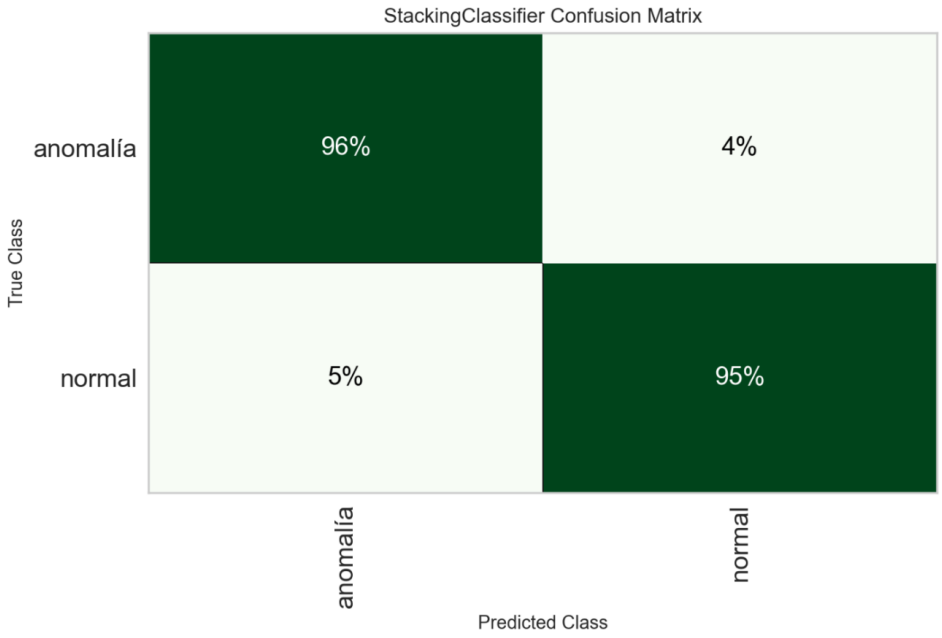


Figura 3.9: Matriz de confusión modelo de apilamiento para el primer año de datos.

3.5. Discusión

Es posible apreciar de manera concreta que la estrategia de búsqueda de modelos se genera

Capítulo 4

Desarrollo de plataforma propuesta

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 trabajó 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.

- pycaret==2.3.10
- streamlit==1.12.0
- pandas==1.3.5
- plotly==5.10.0
- pandas-profiling==3.2.0

4.1. 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 súmamente 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. Aquí 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 **??** Comparativa de atributos de distintas plataformas abiertas disponibles.

Tabla 4.1: 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

4.2. Requerimientos de la plataforma

Los requerimientos principales comprometidos para este resultado 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.

4.3. 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[1]. 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.

Entre los atributos principales que se consideraron para el diseño de la aplicación se destaca la incorporación de un mapa para georreferenciar la estación a monitorear, esta información es obtenida mediante la conexión a la base de datos del proyecto que tiene la meta-data de cada estación. 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 las etiquetas de la data que previamente el experto ha completado al analizar la data. Es posible también visualizar las series de tiempo en crudo mediante gráficos interactivos.

La última consideración de diseño de la que dependen los dos principales resultados de producción y requerimientos del hito es la detección de anomalías. Este módulo de detección de anomalías carga de forma automática un modelo pre-entrenado de un detector de anomalías mediante boosting, específicamente utilizando un modelo Light GBM (Anexo **??**) que presentó, utilizando los datos de validación, todas sus métricas sobre 75 % (Anexo **??**) y el tiempo que demora en procesar un dataset de varios años de data con frecuencia horaria es de escasos 3 segundos en la plataforma.

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

1. **Adquisición de data** En esta primera sección de la aplicación es posible realizar las consultas de información adicional de la estación con la que el sistema esta implementada directamente desde la base de datos además de una estación de prueba para validar

[1] Repositorio en GitHub de la aplicación Sistema Experto APP

el funcionamiento de la consulta dinámica desde una lista desplegable.

Luego de esto hay un cuadro de archivo en donde se carga el archivo .csv con la data histórica previamente formateada para un sitio de monitoreo acotado, en este caso, este sitio es el correspondiente a Horcón y fue previamente analizado en el caso de estudio y la data que se presentan en la sección **??**.

Figura 4.1: 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.

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 estandard, mínimos, máximos, promedio y cuartiles.

	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 4.2: 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, sin procesamiento. Seguido de

los gráficos por variable es posible también desplegar un panel de información adicional que muestra algunas estadísticas básicas y análisis exploratorio generados automática a partir de la data.

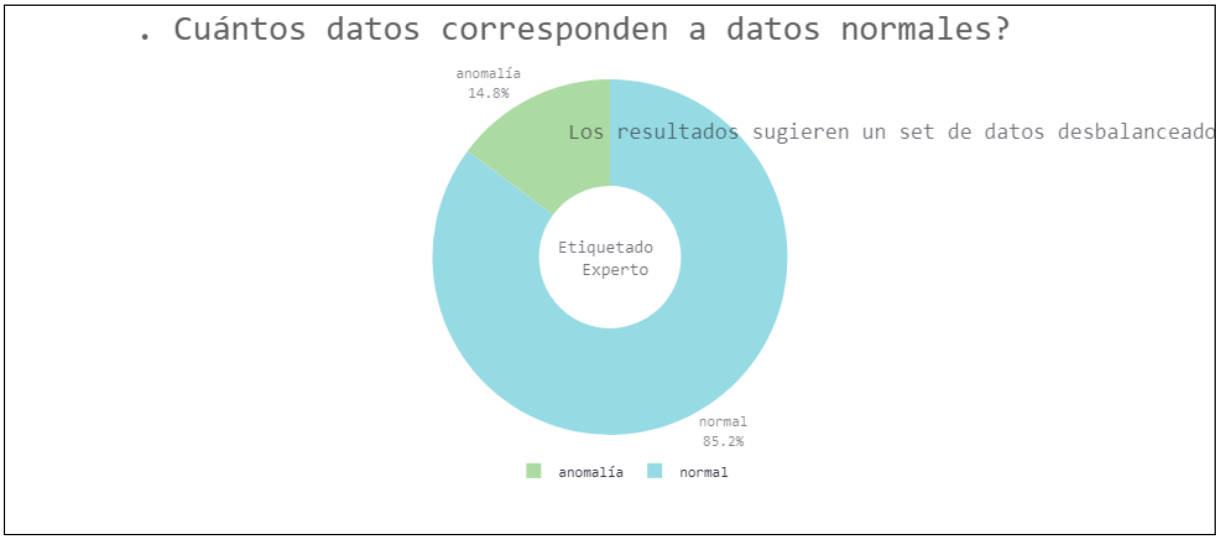


Figura 4.3: Sección de visualización de información del sitio de medición

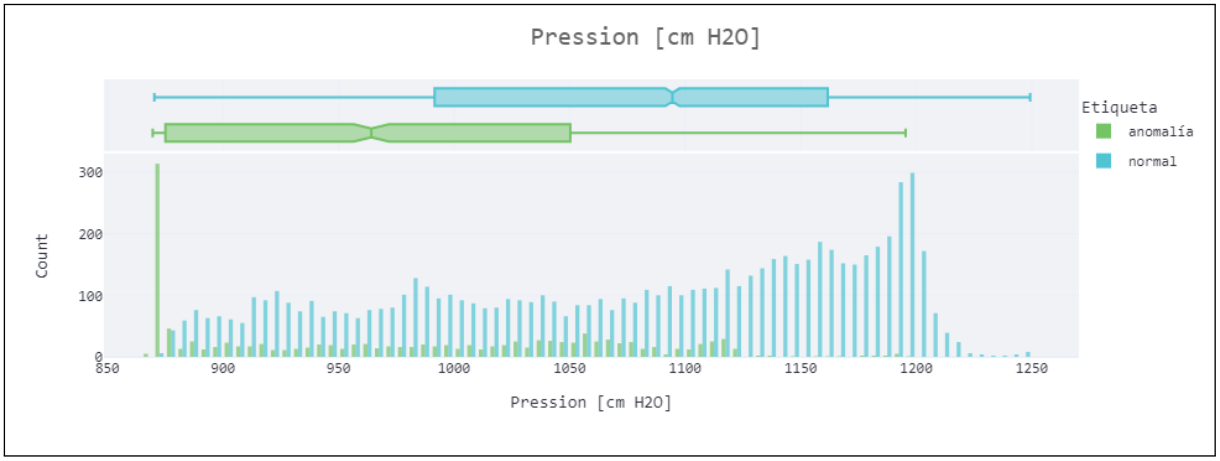


Figura 4.4: Análisis estadístico gráfico de información cargada.

3. **Procesamiento y detección de anomalías** Así se llega a la sección principal de la aplicación, la correspondiente a detección de anomalías. En esta sección es posible encontrar primero la tabla 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 4.5: Grilla de búsqueda de modelos

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

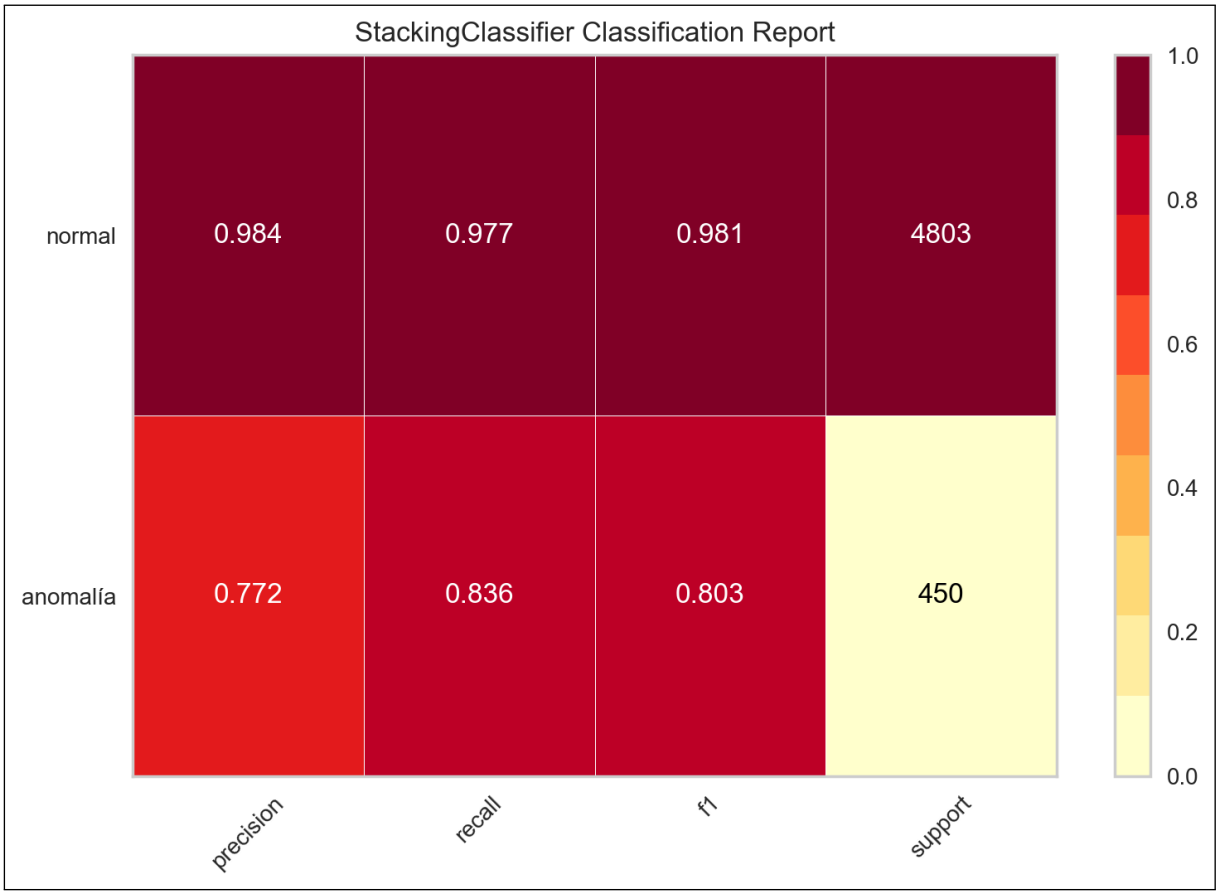
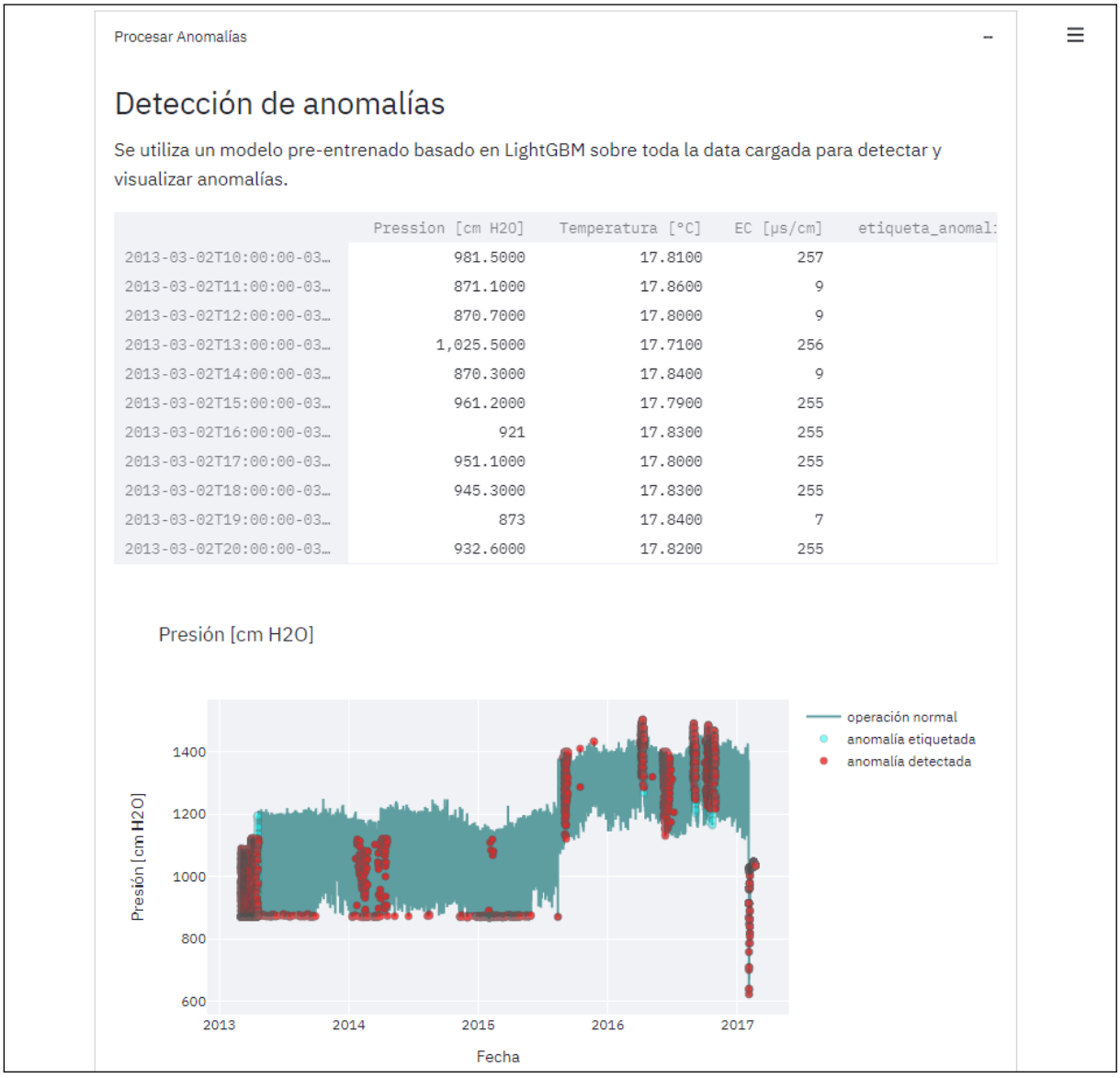


Figura 4.6: Matriz de confusión

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 4.7: Sección de detección de anomalías

4.4. 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 y muy lejano al máximo requerido de 60 minutos.

En cuanto al diseño es necesario destacar que la versión mostrada es asociado a un caso particular de datos cargados, 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.

La interfaz presentada incorpora una serie de observaciones y consideraciones provenientes

tanto de potenciales usuarios como de un diseñador especializado en plataformas WEB para visualización de data por lo que se puede concluir que puede llegar a ser una herramienta súmamente útil en acercar herramientas computacionales avanzadas a usuarios técnicos o expertos provenientes de otras áreas.

La plataforma se encuentra desplegada en línea en el siguiente enlace Sistema Experto APP.

Capítulo 5

Resultados

5.1. Aplicación al caso de datos de un año

5.2. Aplicación al caso de datos de dos años

5.3. Aplicación al caso de datos de tres año

5.4. Aplicación al caso de datos de cuatro años

Capítulo 6

Conclusiones

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

6.1. Trabajo futuro

Bibliografía

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

[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, 1 2020.

[3] 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. Bierkens, A. Macdonald, Y. Fan, R. M. Maxwell, Y. Yechieli, J. J. Gurdak, D. M. Allen, M. Shamsudduha, K. Hiscock, P. J. Yeh, I. Holman, and H. Treidel, "Ground water and climate change," 4 2013.

[4] S. Jasechko and D. Perrone, "Global groundwater wells at risk of running dry," tech. rep., 2021.

[5] S. Jasechko and D. Perrone, "California's Central Valley Groundwater Wells Run Dry During Recent Drought," 2020.

[6] 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," vol. 12, p. 648, 2020.

[7] A. Bustos, "Poca agua, altas temperaturas: el complejo escenario de sequía invernal que dejó en evidencia el mes de julio « Diario y Radio U Chile," 2021.

[8] 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.

[9] F. De la Vega, "Expertos de la u. de chile alertan sobre los riesgos de la desertificación en el país.," 2020.

[10] Escenarios hídricos 2030 Chile, "Radiografía del agua: brecha y riesgo hídrico en Chile," tech. rep., 2019.

[11] A. Panez-Pinto, P. Mansilla-Quiñones, and A. Moreira-Muñoz, "Agua, tierra y fractura sociometabólica del agronegocio. ActividAd frutícolA en PetorcA, chile," 2018.

[12] "res_1238_mee_nacional,"

[13] A. Moreno, C. Leturia, O. C. Marfil, D. San, M. Cornejo, H. Moya, G. Daniela, F. Muños, Y. Ulloa, Z. Colaboradores, C. Montecinos, M. Azócar, F. Aburto, and A. Bustos, "Atlas Calidad del Agua," tech. rep., MINISTERIO DE OBRAS PÚBLICAS Ministro de Obras

Públicas Subsecretario de Obras Públicas Área Desarrollo Ambiental Profesionales, 2020.

[14] INSTITUTO NACIONAL DE NORMALIZACION, "NORMA CHILENA 409/2," tech. rep., 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," tech. rep., DIRECCIÓN GENERAL DE AGUAS, Santiago, 3 2017.

[16] 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.

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

[18] INSTITUTO NACIONAL DE NORMALIZACION, "Agua potable-Parte 1-Requisitos," 2005.

[19] INSTITUTO NACIONAL DE NORMALIZACION, "NCh1333-1978_Mod-1987," 1987.

[20] "Observatorio georreferenciado DGA," *https://snia.mop.gob.cl/observatorio*.

[21] "Explorador climático (CR)2," *https://explorador.cr2.cl*.

[22] "Plataforma seguimiento ambiental SQM," *https://www.sqmsenlinea.com/data-source-type/2/show*.

[23] E. K. White, T. J. Peterson, J. Costelloe, A. W. Western, and E. Carrara, "Can we manage groundwater? A method to determine the quantitative testability of groundwater management plans,"

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

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

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

[27] G. Moschini, R. Houssou, J. Bovay, and S. Robert-Nicoud, "Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model," *Engineering Proceedings*, vol. 5, p. 56, 7 2021.

[28] 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.

[29] 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.

[30] M. Braei and S. Wagner, "Anomaly detection in univariate time-series: A survey on the state-of-the-art," *arXiv*, 2020.

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

[32] A. Thakur, *Approaching (Almost) Any Machine Learning Problem 1 Approaching (Al-*

most) Any Machine Learning Problem.

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

[34] C. W. J. Granger and M. J. Morris, "Time Series Modelling and Interpretation," Tech. Rep. 2, 1976.

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

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

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

[38] L. Breiman, "Random Forests," tech. rep., 2001.

[39] 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, no. 1, pp. 119–139, 1997.

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

[41] D. H. Wolpert, "Stacked Generalization," tech. rep., 1992.

[42] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," 1996.

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, Association for Computing Machinery, 8 2016.

[44] K. Joslyn, "Water Quality Factor Prediction Using Supervised Water Quality Factor Prediction Using Supervised Machine Learning Machine Learning," tech. rep., 2018.

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

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

[47] 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.

[48] 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.

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

[50] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," 2017.

[51] F. Hutter, L. Kotthoff, and J. Vanschoren, *The Springer Series on Challenges in Machine Learning Automated Machine Learning Methods, Systems, Challenges*. 2019.

[52] D. S. Liebeskind, C. S. Anderson, S. Wales, A. Vasileios-Arsenios Lioutas, Y.-H. Tsai, H.-L. Wang, W.-Y. Hsu, M.-H. Lee, H.-H. Weng, S.-W. Chang, and J.-T. Yang, "Automatic Machine-Learning-Based Outcome Prediction in Patients With Primary Intracerebral Hemorrhage," *Frontiers in Neurology | www.frontiersin.org*, vol. 1, p. 910, 2019.

[53] 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, 5 2021.

[54] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms,"

Anexo A

Cálculos realizados

Anexo B

Desarrollo plataforma

Análisis de otras plataformas disponibles y su interfaz visual.

B.1. Análisis otras plataformas

Figura B.1: Plataforma monitoreo DGA

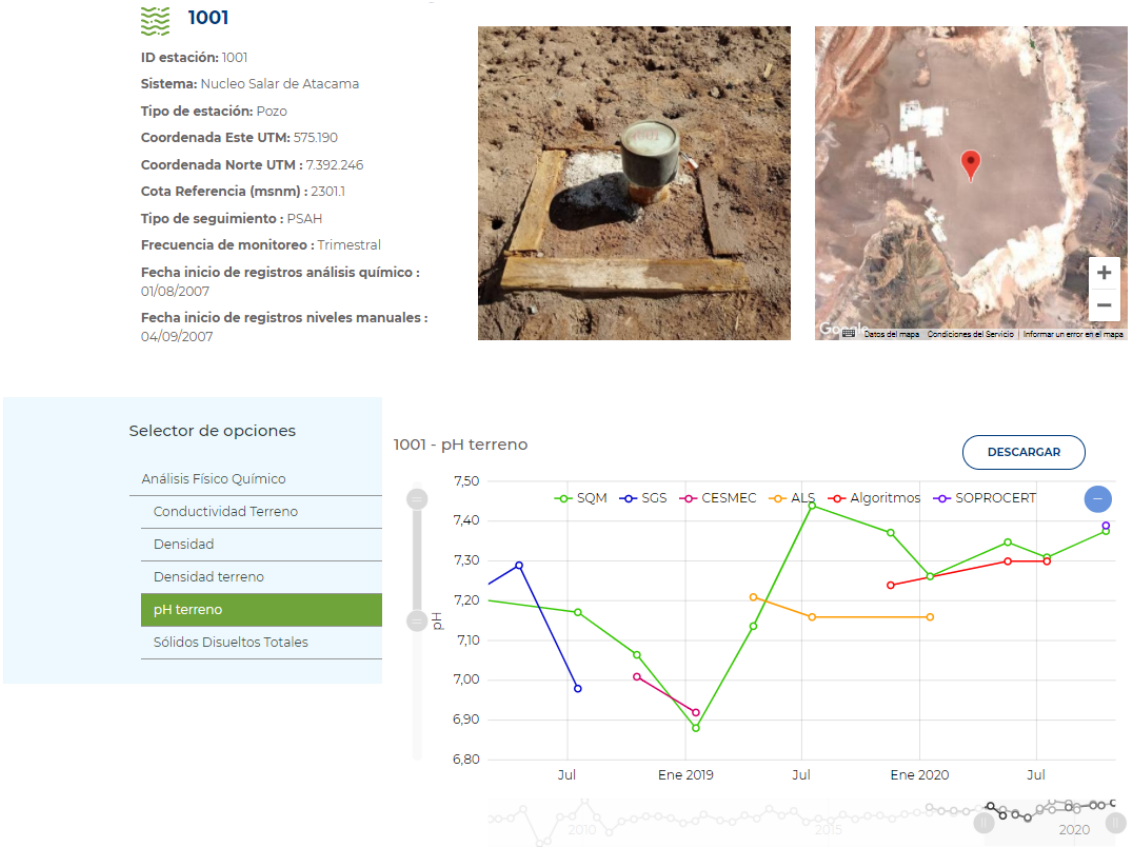


Figura B.2: Plataforma monitoreo SQM

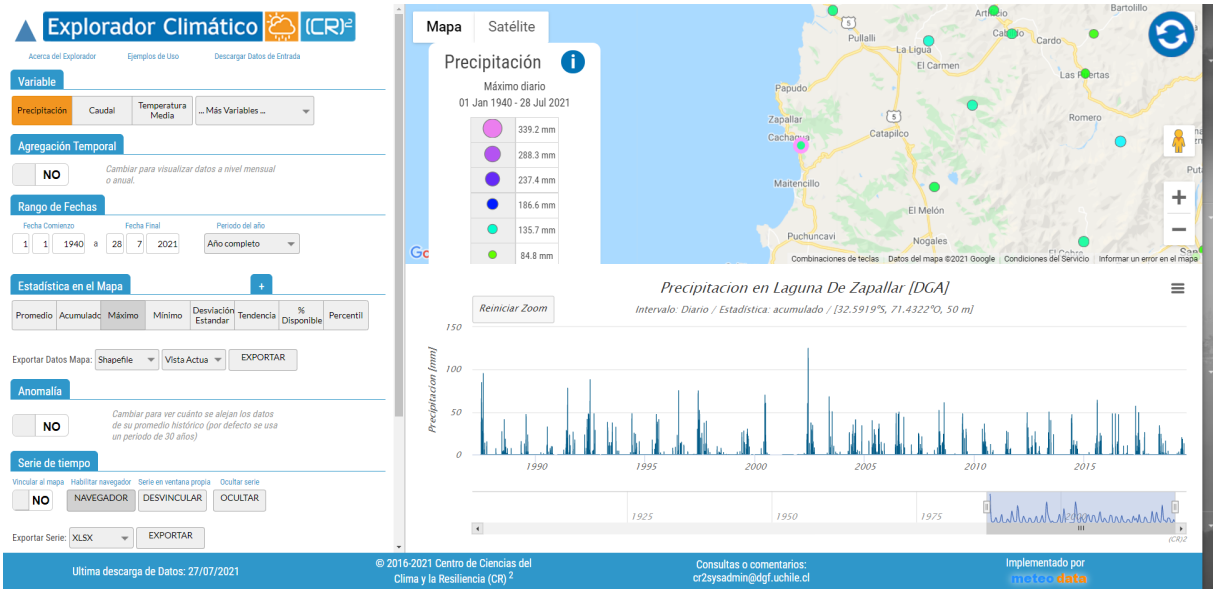


Figura B.3: Plataforma monitoreo CR2