DESARROLLO DE UN SISTEMA PREDICTIVO DEL PH PARA EL MANTENIMIENTO AUTOMATIZADO DE PISCINAS MEDIANTE TÉCNICAS DE MACHINE LEARNING Y BIG DATA

Máster de Programación avanzada en Python para Big Data, hacking y Machine Learning

Proyecto realizado por:

Miriam López Hernández

2024

Índice:

Introducción 3
Objetivos: 4
Marco Teórico: 5
1. Piscinas Inteligentes 5
2. Sistema ChloroMatic Lifeguard de Davey 6
3. Big Data 10
4. Machine Learning en el Mantenimiento Predictivo 12
Metodología 16
Fase 1: Recopilación y Preparación de Datos 16
Fase 2: Análisis Exploratorio y Visualización de Datos 17
Fase 3: Desarrollo de Modelos de ML 18
Fase 4: Integración y Pruebas del Sistema 18
Implementación y Análisis de resultados 20
Fase 2: Análisis Exploratorio y Visualización de Datos 20
Fase 3: Desarrollo de Modelos de ML 37
Fase 4: Integración y Pruebas del Sistema 48
Conclusiones 54
Posibles Mejoras del Sistema: 56
Referencias: 57

Introducción

El mantenimiento adecuado de piscinas es crucial tanto para garantizar la seguridad y el confort de los usuarios, como para prolongar la vida útil de las instalaciones. En la era digital, la integración de tecnologías avanzadas como el Internet de las Cosas, IoT y el análisis de datos masivos con el Big Data, nos ofrece nuevas oportunidades para optimizar estos procesos.

Este proyecto se centra en el desarrollo de un sistema basado en Machine Learning para la predicción de parámetros y el mantenimiento automatizado de piscinas, utilizando datos recopilados de sistemas de monitoreo Davey Lifeguard instalados en 10 piscinas diferentes.

El sistema Davey Lifeguard, que incluye un Clorificador de agua por salinidad, recopila datos continuos de parámetros críticos como pH, ORP, salinidad y temperatura. Estos datos, que fueron recolectados durante un período de seis meses, proporcionan una base sólida para el análisis predictivo y la toma de decisiones automatizada en el sistema.

La implementación de técnicas como el ML nos permitirán anticipar cambios en los parámetros de la piscina, activar los aparatos al tiempo más adecuado y optimizar la dosificación de productos químicos, todo ello podrá ser el objetivo para logra un uso futuro más eficiente de recursos y una mejora en la calidad del agua.

Objetivos:

- ***Objetivo General:*** Desarrollar un sistema predictivo usando algunas de las técnicas de Machine Learning y Big Data aprendidos en el máster para intentar mejorar u optimizar el uso automatizado del sistema Davey ChloroMatic Lifeguard.

- ***Objetivos Específicos:***

  a) Recopilar, preprocesar los datos recopilados de 4 parámetros (pH, ORP, salinidad y temperatura) en diferentes piscinas.

  b) Visualizar y analizar los datos con librerías especializadas de Python y el Big Data para lograr un mejor entendimiento de ellos.

  c) Desarrollar un modelo de clasificación para predecir la necesidad de activación del Clorificador (ChloroMatic Nipper) basado en los niveles de ORP.

  d) Implementar un modelo de series temporales para predecir cambios futuros en el pH del agua de la piscina.

  e) Integrar los modelos desarrollados en un sistema de predicción en cascada que optimice el mantenimiento automatizado de las piscinas.

  f) Evaluar la eficacia y precisión de los modelos predictivos desarrollados utilizando métricas de rendimiento adecuadas y realizar un ajuste de los hiperparámetros.

  g) Proponer recomendaciones para la implementación práctica del sistema predictivo en entornos de mantenimiento de piscinas reales.

Marco Teórico:

1. Piscinas Inteligentes

1.1 Importancia del mantenimiento

El mantenimiento adecuado de una piscina es fundamental para garantizar la seguridad y el confort de los usuarios, así como para prolongar la vida útil de la instalación. Un mantenimiento efectivo implica:

- Prevención de problemas de salud: Un agua bien mantenida reduce el riesgo de infecciones y enfermedades transmitidas por el agua.
- Preservación de la infraestructura: El equilibrio químico adecuado previene daños en los equipos y superficies de la piscina.
- Optimización de costes: Un mantenimiento preventivo reduce la necesidad de reparaciones costosas y el desperdicio de productos químicos.
- Mejora de la experiencia del usuario.

1.2 Parámetros clave en el cuidado del agua

Los principales parámetros que se deben controlar en una piscina, según las directrices de la Organización Mundial de la Salud [1] son:

a) pH:

El pH es una medida de la acidez o alcalinidad del agua en una escala de 0 a 14. Un pH de 7 se considera neutro, por debajo de 7 es ácido, y por encima de 7 es alcalino. En el contexto de las piscinas, el pH afecta directamente la eficacia de los desinfectantes, la comodidad de los nadadores y la durabilidad de los equipos. Un pH equilibrado es crucial para la eficacia del cloro, previene la irritación de ojos y piel, y evita la corrosión o incrustaciones en los equipos de la piscina.

- Rango óptimo: 7.2 - 7.6

b) ORP:

El ORP, Potencial de Oxidación-Reducción, es una medida en milivoltios (mV) que indica la capacidad del agua para oxidar contaminantes. Refleja la actividad del desinfectante en el agua, no su concentración. El ORP es un indicador más preciso de la efectividad de la desinfección que la mera medición de los niveles de cloro, un valor alto indica una mayor capacidad de limpieza. Proporciona una medida en tiempo real de la capacidad del agua para eliminar contaminantes.

- Rango buscado: 600 - 750 mV

c) Salinidad:

La salinidad se refiere a la concentración de sal disuelta en el agua, generalmente medida en partes por millón (ppm). En piscinas con sistemas de cloración salina, la sal se utiliza para producir cloro a

través de un proceso de electrólisis. Mantener el nivel adecuado de salinidad es esencial para el funcionamiento eficiente de los sistemas de cloración salina, asegurando una producción constante de cloro y previniendo daños en el equipo.

- Rango aceptado por el sistema: 1000 - 6000 ppm

d) Temperatura:

La temperatura del agua es un factor crítico que afecta tanto el confort de los nadadores como la química del agua. La temperatura influye en la tasa de reacción de los productos químicos, la proliferación de microorganismos y el confort de los usuarios. Temperaturas más altas pueden acelerar el crecimiento de bacterias y aumentar la demanda de cloro.

- Rango: Varía según el uso, generalmente entre 24°C y 28°C para piscinas recreativas con reguladores de temperatura, pero no todas las piscinas de este proyecto cuentan con un sistema de control de temperatura.

El monitoreo constante y el ajuste de estos parámetros son esenciales para mantener un agua de piscina segura y agradable. Los sistemas automatizados, como el Davey Lifeguard, facilitan este proceso al proporcionar lecturas continuas y permitir ajustes precisos en tiempo real [2]

2. Sistema ChloroMatic Lifeguard de Davey

Ilustración 1. Sistema ChloroMatic Lifeguard Davey. https://daveywater.com/au/

2.1 Estructura y Funcionamiento Químico

Sistema General:

El *ChloroMatic Lifeguard* de Davey es un sistema completo para la desinfección y el control químico del agua en piscinas, que incluye varios componentes clave:

- **Clorador Salino**: El ChloroMatic Nipper usa un sistema que utiliza una célula electrolítica para la generación de cloro a través de la electrólisis para desinfectar el agua de la piscina. Nuestro

sistema es capaz de trabajar con distintos tipos de sal (natural, mineral o de bajo contenido) de hasta 140,000 litros.

- **Sensores de pH, ORP, salinidad y temperatura:** Estos sensores monitorean constantemente los niveles, especialmente de pH y ORP en el agua, para asegurar un buen estado de la piscina y poder mediante los datos aplicar las correcciones necesarias.
- **Controlador**: Regula el funcionamiento del clorador basado en las lecturas de los sensores, ajustando automáticamente la producción de cloro y la dosificación de ácido para mantener el equilibrio químico del agua.
- **Bomba peristáltica**: para la adición controlada de ácido que dosificara el pH.
- **Módulo WiFi**: monitoreo remoto y control a través de aplicación móvil.

2. Funcionamiento del Clorificador y Proceso Químico

- **Electrólisis Salina**: El agua de la piscina, que contiene cloruro de sodio (NaCl), pasa a través de la célula electrolítica del sistema. Al aplicar una corriente eléctrica, el NaCl se descompone en sodio (Na+) y cloro (Cl-).

$$2NaCl + 2H_2O \rightarrow Cl_2(g) + H_2(g) + 2NaOH(aq)$$

- **Producción de Cloro**: El cloro (Cl2) se disuelve en el agua, donde reacciona para formar ácido hipocloroso (HOCl), el agente desinfectante principal.

$$Cl_2(g) + H_2O \rightarrow HOCl + HCl$$

- **Ajuste del pH**: Durante este proceso, el pH del agua puede elevarse debido a la producción de hidróxido de sodio (NaOH). Para contrarrestar este efecto, el sistema inyecta ácido, generalmente ácido clorhídrico (HCl), para mantener el pH dentro del rango deseado (7.2 - 7.6).

$$NaOH + HCl \rightarrow NaCl + H_2O$$

3. Interacción y funcionamiento de los Sensores

- **Sensor ORP**: Mide la capacidad del agua para desinfectar. Si el ORP cae por debajo del umbral configurado el sistema aumenta la producción de cloro.
- **Sensor de pH**: Si el pH se eleva por encima del rango ideal, el sistema inyecta ácido para equilibrarlo.

4. Consideraciones Técnicas

El *ChloroMatic Lifeguard* es un sistema eficiente que automatiza la desinfección de la piscina, asegurando una calidad de agua óptima con intervención mínima. La integración de sensores garantiza que los niveles de cloro y pH se mantengan en un equilibrio constante, proporcionando una experiencia segura y cómoda para los nadadores. [3]

5. Funcionamiento del sistema Lifeguard:

1. Los sensores miden constantemente los parámetros del agua.
2. El controlador analiza estas lecturas y las compara con los rangos ideales preestablecidos.
3. Si se detectan desviaciones, el sistema activa automáticamente los mecanismos de corrección:
    - Ajusta la producción de cloro a través del sistema de cloración salina.
    - Activa la bomba de dosificación de ácido para corregir el pH si es necesario.
4. Todos los datos se registran y pueden ser accedidos remotamente a través de la aplicación móvil.

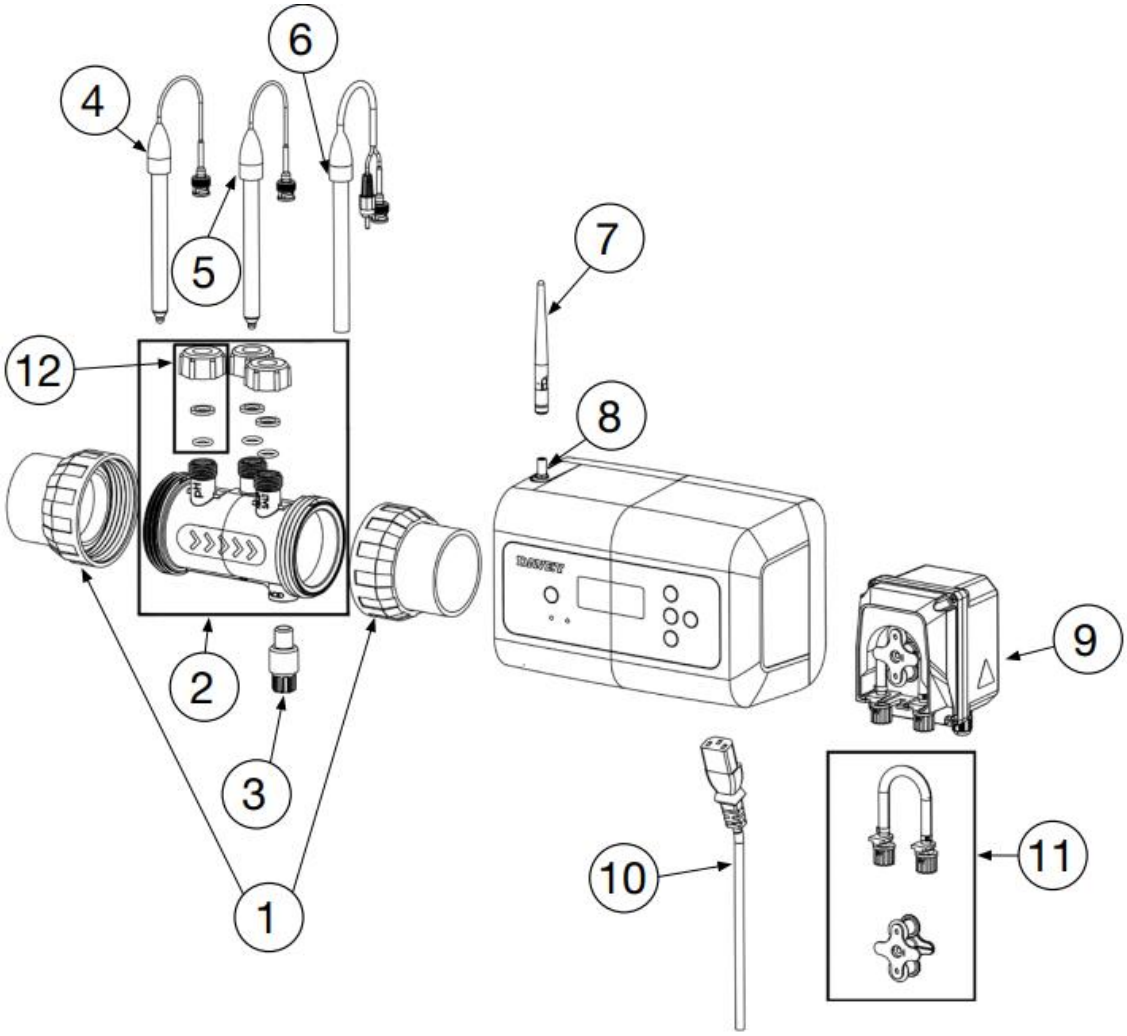


Ilustración 2. Despiece del sistema Davey Lifeguard

ITEM	DESCRIPCIÓN
1	Conjunto de unión de barril EU63mm con O-ring
2	Carcasa de sonda e inyector con kit de sellado de sonda
3	Punto de inyección de ácido
4	Sonda de pH Lifeguard
5	Sonda ORP Lifeguard
6	Sensor de temperatura y sonda TDS
7	Antena estándar Lifeguard
8	Controlador Lifeguard con antena incluida
9	Bomba dosificadora de ácido – 9 conecta a 3
10	Cable de alimentación
11	Kit de servicio de tubo de bomba de ácido con rodillo y tubo
12	Kit de sellado de sonda

Ilustración 3. Aplicación del móvil.

1. **Acceso en tiempo real a las lecturas de los 4 parámetros.**
2. **Histórico de graficas de los últimos días**
3. **Modos disponibles, ajuste de los ciclos de funcionamiento.**

3. Big Data

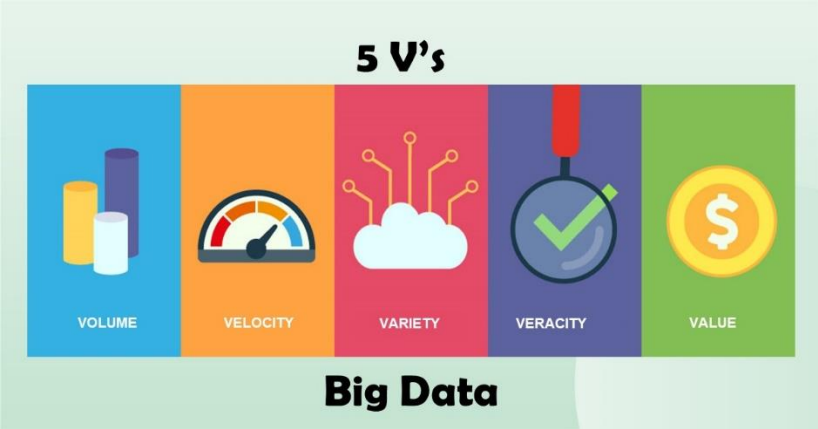


Ilustración 4. 5V's del Big Data. https://www.shiksha.com/online-courses/articles/five-vs-of-big-data/

Big Data se refiere a conjuntos de datos extremadamente grandes y complejos que superan la capacidad de las herramientas de procesamiento de datos tradicionales. Big Data se caracteriza principalmente por las "cinco V" [4] [5]:

- ✓ **Volumen**: Cantidades masivas de datos, que pueden variar desde terabytes hasta petabytes o más. El 90% de los datos en el mundo se han creado en los últimos dos años.
- ✓ **Velocidad:** La rapidez con la que se generan, recopilan y procesan los datos, a menudo en tiempo real o casi real. Por ejemplo, Facebook procesa 500+ terabytes de datos diariamente.
- ✓ **Variedad:** Diversos tipos de datos, incluyendo estructurados (como bases de datos), semiestructurados (como XML) y no estructurados (como texto, imágenes y videos). El 80% de los datos empresariales son no estructurados.
- ✓ **Veracidad:** La confiabilidad y precisión de los datos. Esto es crucial ya que el costo anual de datos de mala calidad en los EE. UU. es de $3.1 trillones.
- ✓ **Valor:** La capacidad de convertir los datos en información útil para la toma de decisiones. Las empresas que adoptan Big Data ven un aumento del >10% en el crecimiento de los ingresos y una reducción del 15% en los costos y 20% en mantenimiento de equipos.

3.1 Funcionamiento, Estrategias y Análisis del Big Data

El proceso de Big Data típicamente incluye la recopilación, almacenamiento, procesamiento y análisis de datos y sus estrategias clave se podrían resumir en [4]:

- **Análisis predictivo**: Esta estrategia utiliza datos históricos, algoritmos estadísticos y técnicas de machine learning para identificar la probabilidad de resultados futuros. En el contexto de mantenimiento de piscinas, puede predecir cuándo es probable que los niveles de pH o cloro se desvíen de los rangos óptimos, permitiendo acciones preventivas.

- **Análisis en tiempo real**: Implica el procesamiento y análisis de datos a medida que se generan. Esta estrategia es crucial para sistemas de monitoreo de piscinas, ya que permite ajustes inmediatos en respuesta a cambios en la química del agua o condiciones ambientales.
- **Machine Learning**: La una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de la experiencia. En el mantenimiento de piscinas, los algoritmos de ML pueden mejorar continuamente la precisión de las predicciones sobre la calidad del agua basándose en patrones históricos y condiciones actuales.
- **Visualización de datos:** Esta estrategia implica la representación gráfica de datos complejos. Nos puede proporcionar interfaces intuitivas que muestren tendencias de calidad del agua, alertas de mantenimiento y predicciones de forma fácilmente comprensible para los operadores.
- **Análisis de datos no estructurados:** Esta estrategia se centra en extraer información valiosa de datos no estructurados como texto, imágenes o videos.

Estas estrategias, cuando se aplican al mantenimiento de piscinas, permiten un enfoque más proactivo y eficiente en la gestión de la calidad del agua y el mantenimiento general de las instalaciones. Y su buen análisis sería importante en nuestro proyecto para: [6]

a. *Reducir de costos*: el análisis permite optimizar el uso de productos químicos y energía, resultando en ahorros operativos sustanciales.
b. *Toma de decisiones más rápida y mejor*: el análisis en tiempo real de datos de sensores IoT en las piscinas permite ajustes inmediatos en parámetros como la dosificación de cloro o la temperatura del agua, basados en condiciones actuales y tendencias previstas.
c. *Desarrollo de nuevos servicios*: permite a los operadores de piscinas ofrecer servicios personalizados, como ajustes automáticos de temperatura o calidad del agua basados en patrones de uso identificados.

Ilustración 5. Implicaciones del analisis. https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

3.2 La importancia del Big Data en el mantenimiento predictivo de piscinas

El mantenimiento predictivo se basa en técnicas de análisis de datos para detectar patrones y anticipar cuándo es probable que se produzcan fallos o cuándo se necesiten ajustes en los sistemas. En el caso de las piscinas un buen uso del dato nos permitiría la:

- ✓ **Predicción de cambios en la química del agua**: al analizar datos históricos y factores que influyen, es posible anticipar fluctuaciones en parámetros como el pH y el ORP.
- ✓ **Optimización de la dosificación de químicos**: el análisis de grandes volúmenes de datos permite ajustar con precisión la cantidad de productos químicos necesarios, minimizando el desperdicio y mejorando la eficiencia.
- ✓ **Detección temprana de anomalías**: los algoritmos pueden detectar pequeñas desviaciones de los patrones normales, lo que permite intervenir antes de que surjan problemas más serios.
- ✓ **Mantenimiento personalizado**: al analizar datos específicos de cada piscina, se pueden diseñar planes de mantenimiento que se adapten a las necesidades particulares de cada instalación.
- ✓ **Eficiencia energética:** el análisis de datos sobre el uso y las condiciones ambientales puede ayudar a optimizar el funcionamiento de bombas y sistemas de calefacción.

4. Machine Learning en el Mantenimiento Predictivo

El Machine Learning (ML) es una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente. ML se basa en algoritmos que pueden aprender de los datos, identificar patrones y tomar decisiones con mínima intervención humana. Esta tecnología se aplica en diversos campos, desde reconocimiento de imágenes hasta diagnósticos médicos, pasando por sistemas de recomendación y análisis financiero.

En esencia, el ML utiliza datos históricos para entrenar modelos que pueden hacer predicciones o tomar decisiones sobre nuevos datos. El proceso típico incluye la recopilación de datos, su preprocesamiento, la selección y entrenamiento de modelos, y finaliza con la evaluación y ajuste de hiperparametros para conseguir el modelo óptimo. [7]

4.1 Modelos de ML

Entre los modelos de ML más comúnmente utilizados, los cuales analizaremos y veremos cual se adapta mejor a nuestro caso de estudio son: [8] [9]

- **Regresión Lineal:** Este modelo simple pero efectivo es ideal para predecir valores continuos. Asume una relación lineal entre las variables de entrada y la variable objetivo. Su principal ventaja radica en su interpretabilidad y eficiencia computacional. La regresión lineal se expresa matemáticamente como:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Donde y es la variable objetivo, x son las variables de entrada, β son los coeficientes que debe de aprender, y ε es el error.

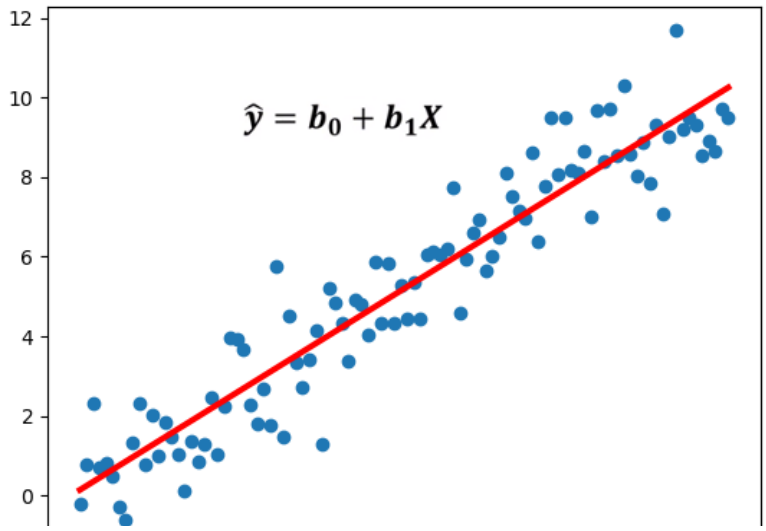


Ilustración 6. Modelo de regresion lineal. https://economipedia.com/definiciones/modelo-de-regresion.html

- **Random Forest:** Como ensemble de árboles de decisión, el Random Forest es excelente para manejar datos no lineales y es robusto frente al overfitting. Este modelo combina múltiples árboles de decisión, donde cada árbol se entrena con una muestra aleatoria de los datos y un subconjunto aleatorio de características.

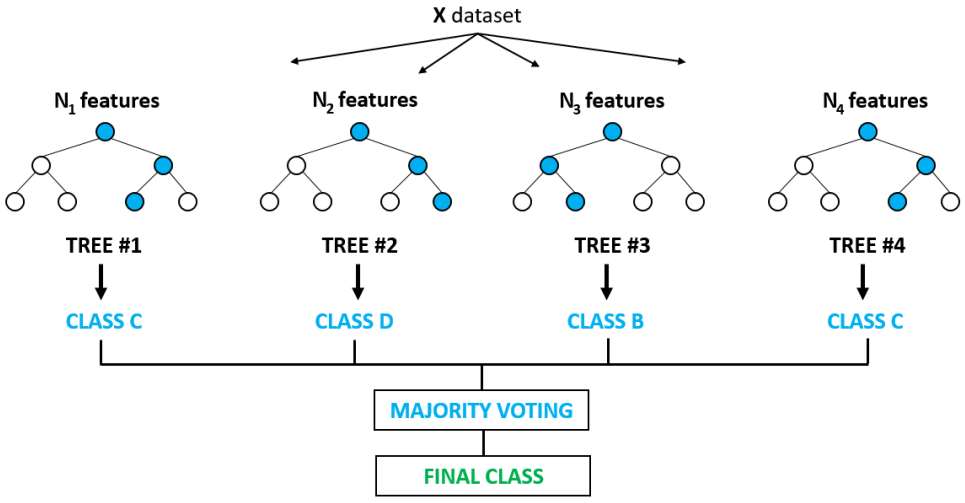


Ilustración 7.Modelo Random Forest. https://rpubs.com/jigbadouin/randomforest1

- **Gradient Boosting:** Este método construye modelos de forma secuencial, mejorando en cada iteración. Ofrece un alto rendimiento predictivo y maneja bien datos heterogéneos. Algoritmos como XGBoost o LightGBM son implementaciones populares de Gradient Boosting.

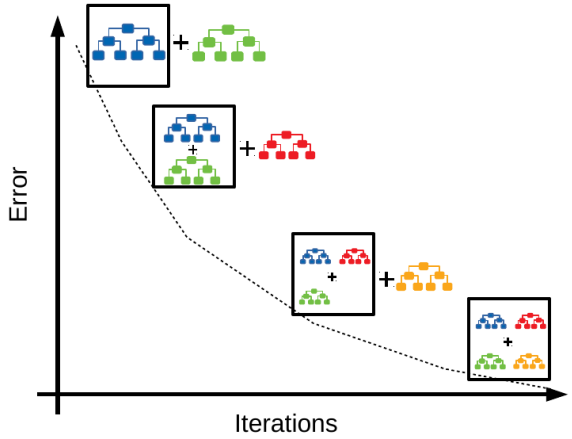


Ilustración 8.Algoritmo Gradient Boostig. https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

- **Redes Neuronales:** Con su capacidad para modelar relaciones complejas y no lineales, las redes neuronales son apropiadas para una amplia gama de tareas. Consisten en capas de "neuronas" interconectadas que transforman los datos de entrada a través de funciones de activación no lineales.

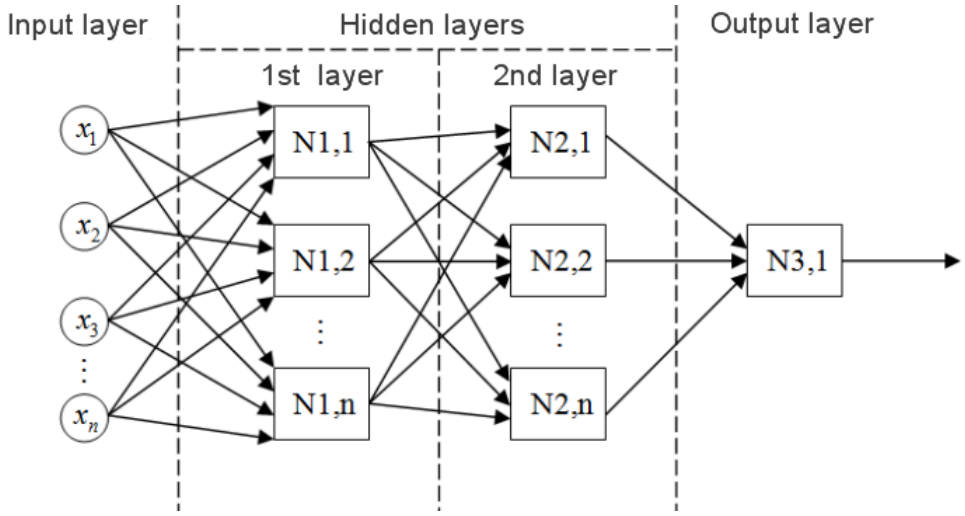


Ilustración 9. Modelo red neuronal multicapa. https://www.mql5.com/es/articles/497

4.2 Ajuste de Hiperparámetros

El ajuste de hiperparámetros es crucial para optimizar el rendimiento de los modelos de ML. Este proceso implica encontrar la mejor configuración de parámetros que no son aprendidos directamente del proceso de entrenamiento. Las técnicas más comunes incluyen: [10] [11]

a. Grid Search: Esta técnica exhaustiva prueba todas las combinaciones posibles de un conjunto predefinido de hiperparámetros.
b. Random Search: Muestrea aleatoriamente combinaciones de hiperparámetros. Es más eficiente que Grid Search, especialmente cuando no todos los hiperparámetros son igualmente importantes.
c. Optimización Bayesiana: Utiliza un enfoque probabilístico para seleccionar los hiperparámetros más prometedores basándose en resultados previos.

La validación cruzada es fundamental para evaluar el rendimiento de los modelos durante el ajuste de hiperparámetros. Técnicas como K-Fold Cross-Validation dividen los datos en K subconjuntos, entrenando el modelo K veces con diferentes combinaciones de datos de entrenamiento y validación.

4.3 Posibles casos de ML para tratamiento de aguas o piscinas

En el contexto específico del tratamiento de aguas y mantenimiento de piscinas, el ML puede aplicarse de diversas formas:

I. **Predicción de parámetros químicos:** Modelos de regresión pueden predecir futuros niveles de pH, cloro, o alcalinidad basándose en datos históricos y condiciones actuales.

II. **Clasificación de estados del agua**: Algoritmos de clasificación pueden categorizar el estado del agua (por ejemplo, "óptimo", "necesita acción", "crítico") basándose en múltiples parámetros.

III. **Optimización de dosificación**: Técnicas de aprendizaje por refuerzo podrían optimizar la dosificación de productos químicos, minimizando el uso de recursos mientras se mantiene la calidad del agua.

IV. **Detección de anomalías:** Modelos no supervisados podrían identificar patrones inusuales en los datos de la piscina, alertando sobre posibles problemas antes de que se vuelvan críticos.

V. **Predicción de mantenimiento**: Algoritmos de series temporales podrían predecir cuándo será necesario realizar tareas de mantenimiento, como la limpieza de filtros.

Metodología

Para abordar el objetivo de optimizar el funcionamiento del sistema Davey ChloroMatic Lifeguard, hemos seguido una metodología estructurada en fases, inspirada en el marco de trabajo Scrum, adaptada a las necesidades específicas de nuestro proyecto de análisis de datos y ML.

Fase 1: Recopilación y Preparación de Datos

1.1 Recolección de Datos:

Junto con la activa colaboración de mi actual empresa en Australia, Davey Water Products, pudimos recopilar los datos de 10 piscinas reales de diferentes tamaños todas equipadas con el sistema Davey ChloroMatic Lifeguard durante un período de seis meses(en la mayoría de casos), sin información sensible de clientes. Los datos incluyen mediciones de cuatro parámetros clave: pH, ORP, salinidad y temperatura. Estos datos fueron registrados por los sensores del sistema a intervalos regulares durante los ciclos de funcionamiento.

Proceso de recolección:

El proceso se ha basado en el funcionamiento cíclico del sistema, que opera de la siguiente manera:

- Ciclos programados: El sistema Lifeguard está configurado para funcionar en ciclos específicos, generalmente de 2 a 8 horas de duración, dependiendo del tamaño de la piscina y sus necesidades de filtración.
- Frecuencia de ciclos: Típicamente, se programan 1 o 2 ciclos diarios, preferentemente durante la noche o en horas de bajo uso de la piscina.
- Captura de datos: Durante cada ciclo, los sensores del sistema registran continuamente los cuatro parámetros clave. Además de los parámetros principales, el sistema también registra información sobre los tiempos de activación del clorificador.
- Intervalo de muestreo: Los datos se registran cada 15 minutos durante los ciclos activos, proporcionando una imagen detallada de las fluctuaciones en los parámetros del agua.
- Almacenamiento de datos: La información capturada se almacena localmente en el controlador del sistema y se transmite a través de la conexión WiFi a una base de datos centralizada.
- Adaptabilidad del sistema: El Lifeguard ajusta automáticamente sus operaciones basándose en las lecturas de los sensores, activando la dosificación de productos químicos o modificando los tiempos de filtración según sea necesario.

Este enfoque de recolección de datos nos ha permitido obtener un conjunto de información completo y detallado sobre el funcionamiento de las piscinas a lo largo del tiempo, capturando tanto las variaciones diarias como las tendencias a largo plazo en la calidad del agua.

Entendimiento de nuestro sistema:

El sistema ChloroMatic Nipper está diseñado para operar en ciclos específicos, optimizando así la eficiencia energética y el mantenimiento del agua. Según la documentación de DWP, estos ciclos de funcionamiento son programables y típicamente se configuran para durar entre 2 y 8 horas, dependiendo del tamaño de la piscina y sus necesidades de filtración. Generalmente, se programan uno o dos ciclos diarios, preferentemente durante la noche o en horas de bajo uso de la piscina.

Durante cada ciclo activo, el sistema realiza varias funciones críticas:

- ✓ Filtración del agua
- ✓ Circulación para prevenir el estancamiento
- ✓ Producción de cloro mediante electrólisis salina
- ✓ Monitoreo y ajuste del pH mediante dosificación de ácido

El clorador salino, componente clave del sistema, utiliza una célula electrolítica para generar cloro a partir de la sal disuelta en el agua. Este proceso es controlado por el sistema Lifeguard, que ajusta la producción de cloro basándose en las lecturas de ORP (Potencial de Oxidación-Reducción).

1.2 Limpieza y Preprocesamiento de Datos:

Una vez obtenidos los datos brutos, procedimos a su limpieza y preprocesamiento. Este proceso incluyó:

- Eliminación de valores atípicos y erróneos.
- Manejo de valores faltantes mediante técnicas de imputación (no había Nah).
- Normalización de los datos para asegurar la consistencia entre las diferentes escalas de medición.
- Transformación de variables categóricas si fue necesario.

Fase 2: Análisis Exploratorio y Visualización de Datos

2.1 Análisis Estadístico Descriptivo:

Realizamos un análisis estadístico descriptivo de los datos para entender mejor su distribución y características. Esto incluyó el cálculo de medidas de tendencia central, dispersión y la identificación de patrones temporales.

2.2 Visualización de Datos:

Utilizamos diversas técnicas de visualización para representar gráficamente los datos y obtener insights, destacando:

- Gráficos de series temporales para cada parámetro.
- Diagramas de dispersión para analizar correlaciones entre variables.
- Histogramas y diagramas de caja para entender la distribución de los datos.
- Mapas de calor para visualizar la matriz de correlación entre variables.

Fase 3: Desarrollo de Modelos de ML

3.1 Selección de Modelos:

Basándonos en la naturaleza de nuestro problema y las conclusiones obtenidas del análisis exploratorio, decidimos implementar y comparar varios modelos de ML, para la activación del clorificador un simple modelo de clasificación que decide si activa o no, y para la predicción del pH si se decidió investigar estas cuatro opciones:

- **Regresión Lineal:** Para establecer una línea base de predicción.
- **Random Forest:** Para capturar relaciones no lineales entre variables.
- **Gradient Boosting:** Por su alta capacidad predictiva en problemas de series temporales.
- **Redes Neuronales (LSTM):** Para modelar dependencias temporales complejas.

3.2 Preparación de Datos para Modelado:

Se divide nuestro conjunto de datos en conjuntos de entrenamiento, validación y prueba. Además, creamos características adicionales basadas en nuestro conocimiento del dominio y los patrones observados en el análisis exploratorio.

3.3 Entrenamiento y Validación de Modelos:

Entrenamos cada uno de los modelos seleccionados utilizando el conjunto de datos de entrenamiento. Utilizamos validación cruzada para ajustar los hiperparámetros de cada modelo y evitar el sobreajuste.

3.4 Evaluación y Comparación de Modelos:

Evaluamos el rendimiento de cada modelo utilizando métricas apropiadas como RMSE (Root Mean Square Error) para las predicciones de pH y precisión, recall y F1-score para la clasificación de la necesidad de activación del clorificador. Comparamos los resultados de los diferentes modelos para seleccionar el más adecuado para cada tarea.

3.5 Ajuste de Hiperparámetros:

Realizamos un ajuste fino de los hiperparámetros de los modelos seleccionados utilizando técnicas como Grid Search y Random Search dando lugar a una pequeña optimización del modelo.

3.5 Visualizacion

Representamos los resultados para tener una mejor visión de ellos,

Fase 4: Integración y Pruebas del Sistema

4.1 Pruebas:

Realizamos pruebas exhaustivas del sistema integrado utilizando datos no vistos previamente para validar su rendimiento en condiciones reales.

4.2 Interpretación de Resultados:

Analizamos en profundidad los resultados obtenidos, comparándolos con el funcionamiento actual del sistema Davey ChloroMatic Lifeguard.

4.3 Elaboración de Conclusiones y Recomendaciones:

Basándonos en los resultados, elaboramos conclusiones sobre la eficacia del sistema predictivo desarrollado y formulamos recomendaciones para su implementación práctica en entornos de mantenimiento de piscinas reales.

A lo largo de todo el proceso, hemos utilizado herramientas y bibliotecas de Python especializadas en análisis de datos ML, incluyendo pandas para el manejo de datos, scikit-learn para el modelado, matplotlib y seaborn para la visualización, y TensorFlow para la implementación de modelos.

Como las condiciones de cada Device/Piscina son diferentes, es decir, a pesar de tener los datos de 10 devices, tras la fase exploratoria se decide enfocar la mejora del modelo en la piscina con datos de mejor y mayor calidad, todo esto será explicado según vayamos viendo la implementación y las diferentes fases, para poder entender el desarrollo de nuestro proyecto, dado que es un caso real hay muchos factores que tenían que ser analizados particularmente y tomar decisiones en base a los resultados que se iban obteniendo.

Implementación y Análisis de resultados

Fase 2: Análisis Exploratorio y Visualización de Datos

En esta sección, describimos el proceso de implementación de nuestro análisis de datos, siguiendo las fases establecidas en la metodología. Nos enfocamos en la preparación, exploración y visualización de los datos recolectados del sistema.

Carga de Datos y Preprocesamiento

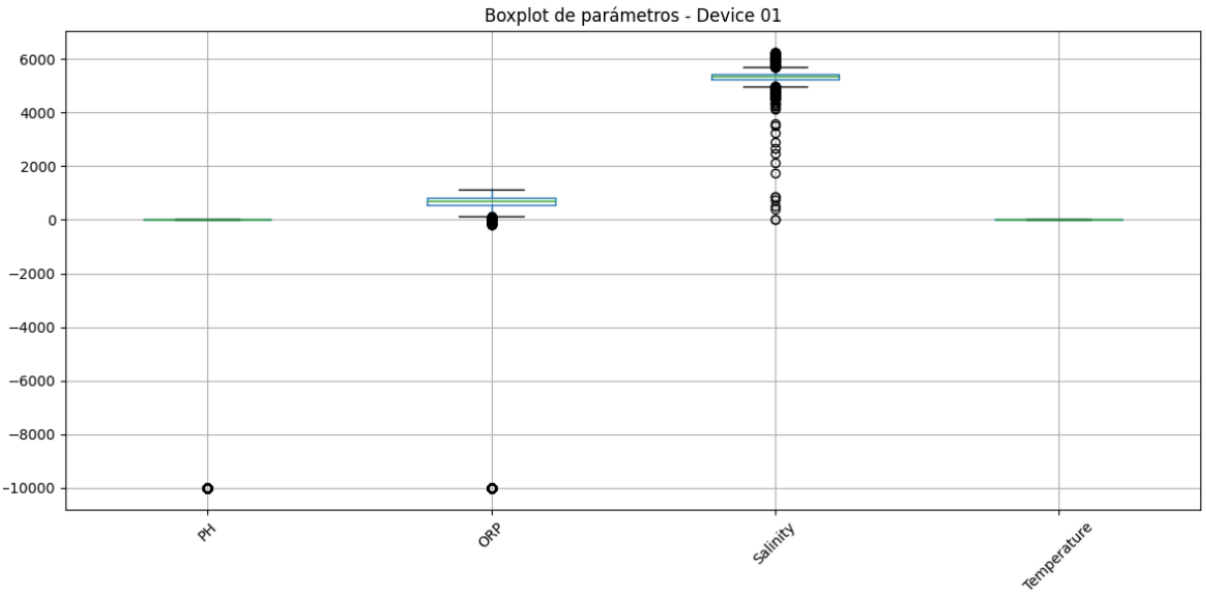
Utilizamos la biblioteca panda para cargar los datos desde archivos CSV. El conjunto de datos principal, 'Data_Pool.csv', contiene información de múltiples dispositivos Lifeguard. Y se realiza una primera vista a los datos que columnas tiene, cuantas filas posee, nulos, tipos de dato y estadísticas descriptivas para cada dispositivo. Además se ajusta la columna de tiempo al formato adecuado.

- *Análisis de resultados*
  - Hay 7 columnas: 'Device', 'Time', 'PH', 'ORP', 'Salinity', 'Temperature', 'Flow status'.
  - 141102 entradas de datos, 0 nulos y su dtypes: datetime64[ns](1), float64(1), int64(4), object(1)
  - En las estadísticas descripticas por cada Device vemos datos que no parecen ser correctos a primera vista como puede ser valores mínimos de -9999 en columnas como el pH y ORP, estos datos corrompen la media, así que se trataron mas adelante.

Análisis Exploratorio de Datos

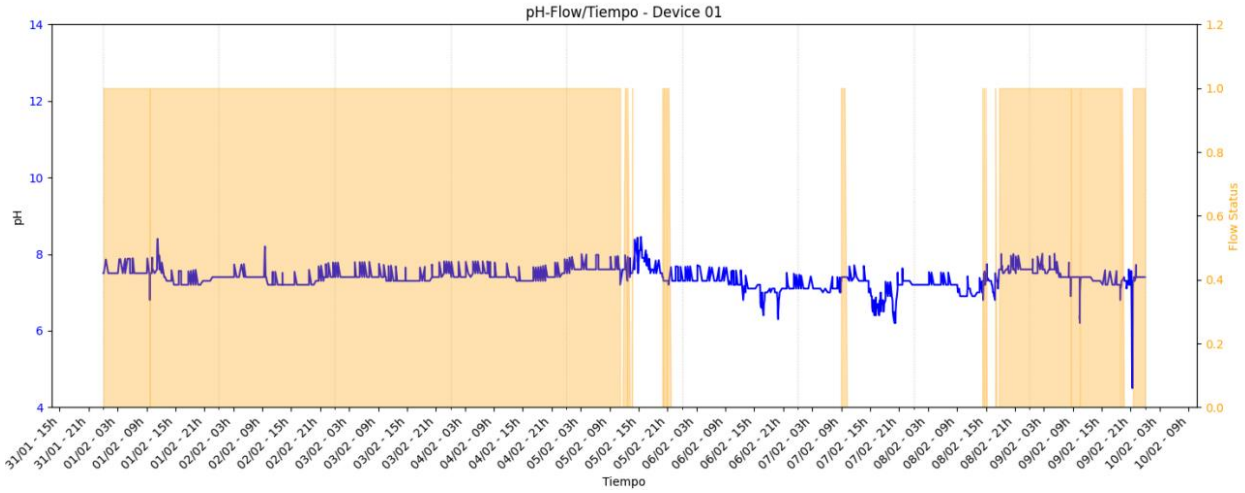
En esta sección creamos un diagrama de cajas para el Device 01 para visualizar outliers. Después se eliminaron valores anómalos (-9999) identificados en las estadísticas descriptivas. Generamos unos gráficos de pH vs. Tiempo con estado de Flow para un período de 10 días en febrero de todos los dispositivos para conocer el funcionamiento del dispositivo y sus ciclos. Se desarrolló una función para calcular días activos y promedio diario y tras analizar los datos se decidió aplicar un filtrado global eliminando todas las filas con Flow = 0.

- *Análisis de resultados*
  - ✓ **Visualización inicial de outliers:** Comenzamos creando un diagrama de cajas (boxplot) para el Device 01. Esta visualización nos permitió identificar rápidamente la presencia y distribución de valores atípicos en los parámetros principales: pH, ORP, salinidad y temperatura y confirmaron la existencia de datos anómalos con valores de -99999.

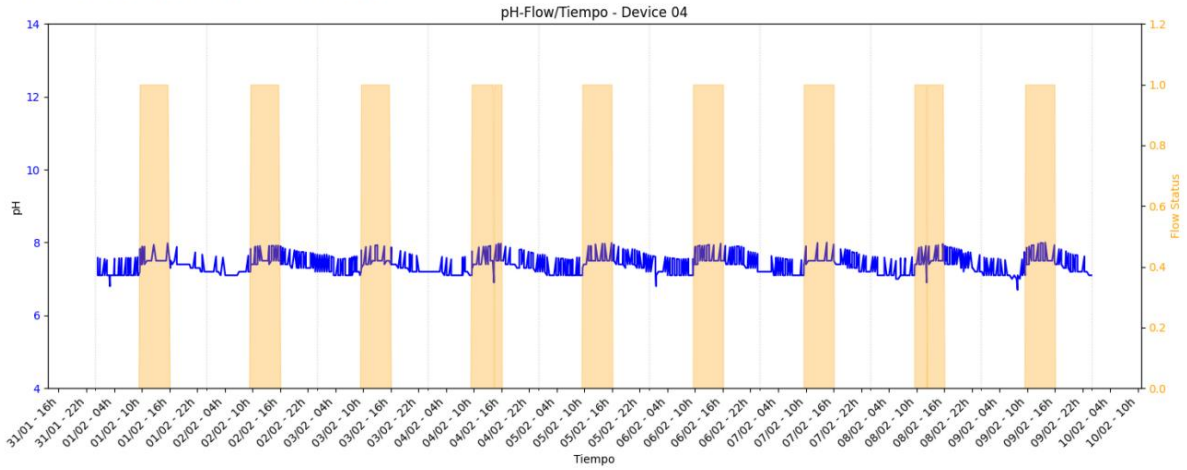


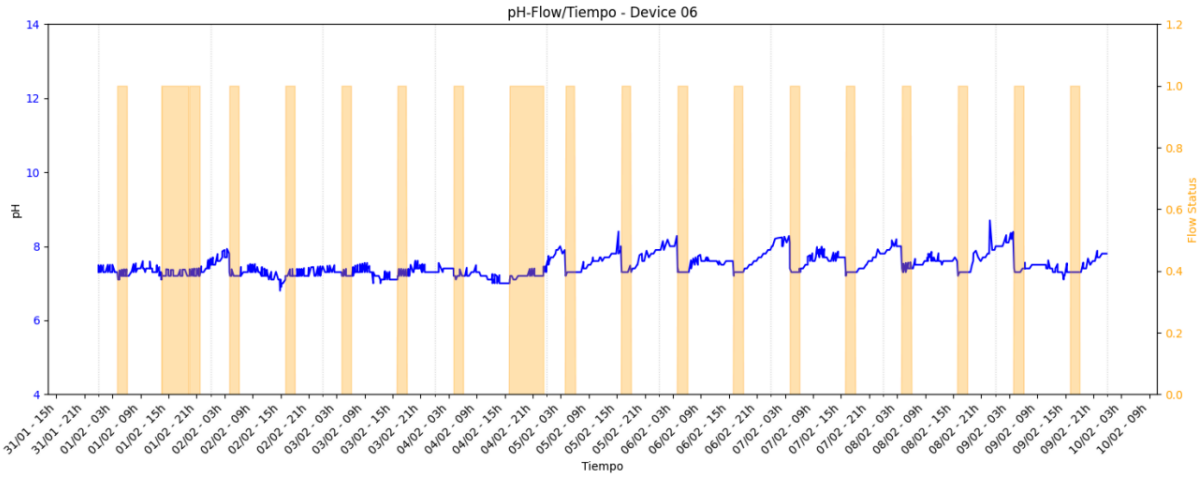
  - ✓ **Limpieza de datos anómalos**: Tras identificar valores claramente erróneos en nuestras estadísticas descriptivas iniciales, eliminamos los valores negativos de nuestro conjunto de datos. Estos valores representaban claramente errores de medición o lecturas fallidas del sistema.

  - ✓ **Análisis de patrones temporales**: Para comprender mejor los ciclos de funcionamiento de los sistemas, generamos gráficos de pH versus tiempo para todos los dispositivos. Incluimos también el estado de Flow en estos gráficos, lo que nos permitió visualizar la relación entre el pH y los períodos de actividad del sistema. Nos enfocamos en un período de 10 días en febrero para obtener una muestra representativa de los patrones de funcionamiento más fácil de ver y precisa para comprender las etapas.



Número de activaciones: 383





En nuestro análisis de los patrones de funcionamiento, observamos una diversidad significativa entre los dispositivos.

El Device 01 presenta un comportamiento distintivo, con la bomba activa durante la mayor parte del tiempo monitorizado. Este patrón contrasta notablemente con la mayoría de los otros dispositivos.

Para la mayoría de los sistemas, como se ejemplifica claramente en el gráfico del Device 04, identificamos un patrón de funcionamiento más estructurado. Estos dispositivos típicamente operan en un ciclo diario bien definido, con un período de actividad de aproximadamente ocho horas. Durante este intervalo, el sistema funciona a plena capacidad, realizando sus tareas de filtración y mantenimiento de la calidad del agua.

Por otra parte, el Device 06 se destaca como otra excepción interesante. A diferencia del patrón común de un ciclo diario extenso, este dispositivo opera con dos ciclos de limpieza más cortos cada día. Esta variación podría indicar una estrategia de mantenimiento diferente, posiblemente adaptada a necesidades específicas de la piscina o preferencias del usuario.

Todas estas variaciones en los patrones de funcionamiento subrayan la importancia de considerar las características individuales de cada instalación en nuestro análisis y en el desarrollo de modelos predictivos. La diversidad observada sugiere que factores como el tamaño de la piscina, el uso personalizado del sistema y las condiciones ambientales específicas pueden influir significativamente en la programación y operación de estos sistemas de mantenimiento.

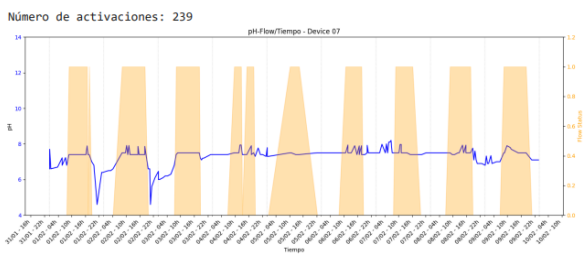
- ✓ **Ciclos de operación**: Desarrollamos una función personalizada para calcular con precisión los días activos y el promedio diario de funcionamiento de cada dispositivo. Esta función nos permitió establecer un criterio consistente, considerando como día activo aquel en el que el sistema funcionó al menos una hora en total.

Resultados del análisis de activación por dispositivo:

	Dispositivo	Total activo (h)	Numero de ciclos	Dias activos	Media diaria (h)
0	Device 01	1607.09	296	148	10.86
1	Device 02	177.74	57	20	8.89
2	Device 02.2	657.32	33	4	164.33
3	Device 03	1175.57	305	165	7.12
4	Device 04	1189.59	337	183	6.50
5	Device 05	1679.09	289	101	16.62
6	Device 06	796.18	545	181	4.40
7	Device 07	1801.55	296	183	9.84
8	Device08	1801.55	296	183	9.84
9	Device09	177.74	57	20	8.89
10	Device10	1289.92	338	143	9.02

En la tabla anterior se observa que la media de funcionamiento diaria esta en torno a las 8 horas, destaca de forma anómala el comportamiento del device 02.2 que además en las estadísticas descripticas se vio que tenia una cantidad de data muy inferior a la del resto puede que sus datos no

se hayan recolectado o suministrado de manera correcta. Además con estos datos y las graficas se observaron que varios de los devices tenían los mismos datos por lo que eran el mismo.



- ✓ **Eliminación de redundancias**: Basándonos en nuestras observaciones previas, eliminamos los dispositivos que mostraban lecturas duplicadas. Específicamente, removimos los Devices 9 y 8, ya que eran duplicados de los Devices 2 y 7 respectivamente.

- ✓ **Filtrado**: El paso final del análisis fue el filtrado global para todo el conjunto, la eliminación de todas las filas donde el valor de Flow era 0, indicativo de períodos en los que la bomba estaba inactiva. Esta decisión se fundamentó en nuestras observaciones detalladas de los patrones de datos y el comportamiento del sistema.

  Al examinar los gráficos, notamos fluctuaciones anómalas en los valores de pH durante los períodos de inactividad de la bomba. Estas variaciones abruptas no reflejan el comportamiento típico de los procesos químicos en una piscina, que generalmente ocurren de manera gradual. La explicación de estas lecturas atípicas radica en la naturaleza del sistema ChloroMatic Lifeguard.

  Cuando la bomba está inactiva, los sensores continúan registrando mediciones, pero estas corresponden al agua estancada dentro de la célula con los sensores, no a la totalidad de la piscina. Esta agua estática puede experimentar cambios químicos localizados debido a varios factores:

- Proximidad a la célula electrolítica: El agua cercana a la célula puede tener concentraciones temporalmente elevadas de cloro.
- Cercanía al sistema de dosificación de ácido: Puede resultar en lecturas de pH no representativas del conjunto de la piscina.
- Precipitación de minerales: En agua estancada, puede ocurrir sedimentación de minerales, alterando las lecturas.
- Variaciones de temperatura: El agua dentro de esta célula puede calentarse o enfriarse más rápidamente que el volumen total de la piscina.

  Además, nuestro conocimiento del funcionamiento del sistema nos indica que las mediciones más precisas y representativas se obtienen cuando hay un flujo constante de agua a través del

sistema. Este flujo asegura una mezcla homogénea y una distribución uniforme de los químicos en toda la piscina.

Por estas razones, determinamos que los datos recopilados durante los períodos de inactividad de la bomba no representan de manera fiable las condiciones reales de la piscina. Su inclusión podría introducir sesgos significativos en nuestros análisis posteriores y en el desarrollo de modelos predictivos. Al eliminar estos datos, aseguramos que nuestro conjunto final se componga exclusivamente de mediciones tomadas bajo condiciones operativas óptimas, proporcionando así una base más sólida y representativa para nuestros estudios subsiguientes.

Estos hallazgos nos proporcionaron información crucial sobre el funcionamiento real de los sistemas Lifeguard en diferentes piscinas, revelando patrones de uso, posibles problemas de registro de datos y la importancia de considerar solo los períodos de funcionamiento activo. El conocimiento de esta información fue fundamental para el desarrollo de modelos predictivos precisos y relevantes en las siguientes fases de nuestro proyecto.

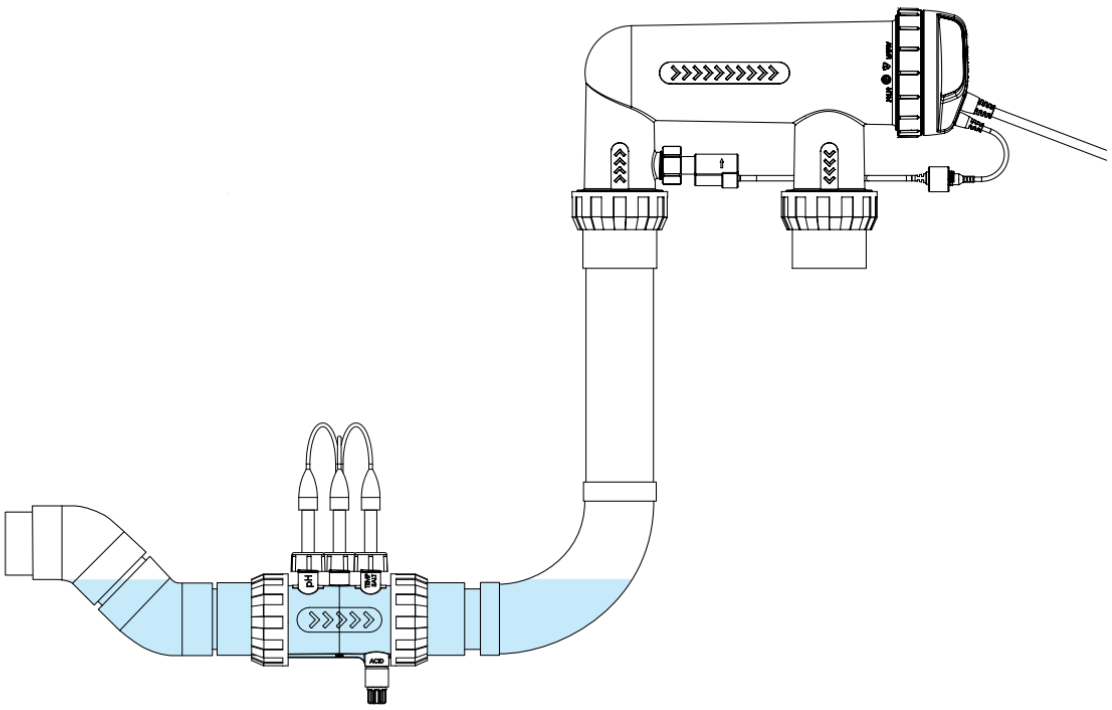


Ilustración 10. Probe and injection point housing & Nipper Cell

Detección de outliers:

El siguiente paso era enfocarse en la detección y tratamiento de outliers, así como en la identificación de valores críticos para los parámetros clave de la piscina.

1. Método de detección de outliers: Implementamos el método del Rango Intercuartílico (IQR) para identificar y eliminar outliers en nuestro conjunto de datos. Este método es robusto y menos sensible a valores extremos que otros enfoques basados en la media y la desviación estándar.
2. Definición de valores críticos: Establecimos rangos críticos para pH, ORP y salinidad basándonos en estándares de la industria y las especificaciones del sistema ChloroMatic Lifeguard. Estos rangos nos ayudaron a identificar valores que, aunque no fueran necesariamente outliers estadísticos, podrían indicar condiciones problemáticas en la piscina. No se llego a poner ningún filtro finalmente porque muchos de los devices deben de tener sensores mal calibrados y se vieron representaciones desfasadas por lo que para poder sacar conclusiones y no manipular los datos para los futuros modelos no se establecieron limites muy restrictivos, reconociendo que cada piscina podría tener características únicas que afecten sus rangos normales de operación.

Parámetro	Valores Estándar	Mínimo Peligroso	Máximo Peligroso
pH	7.2 - 7.6	< 6.8	> 8.0
ORP (mV)	625 - 750	< 600	> 900
Salinidad (ppm)	2700 - 3400	< 2000	> 6000

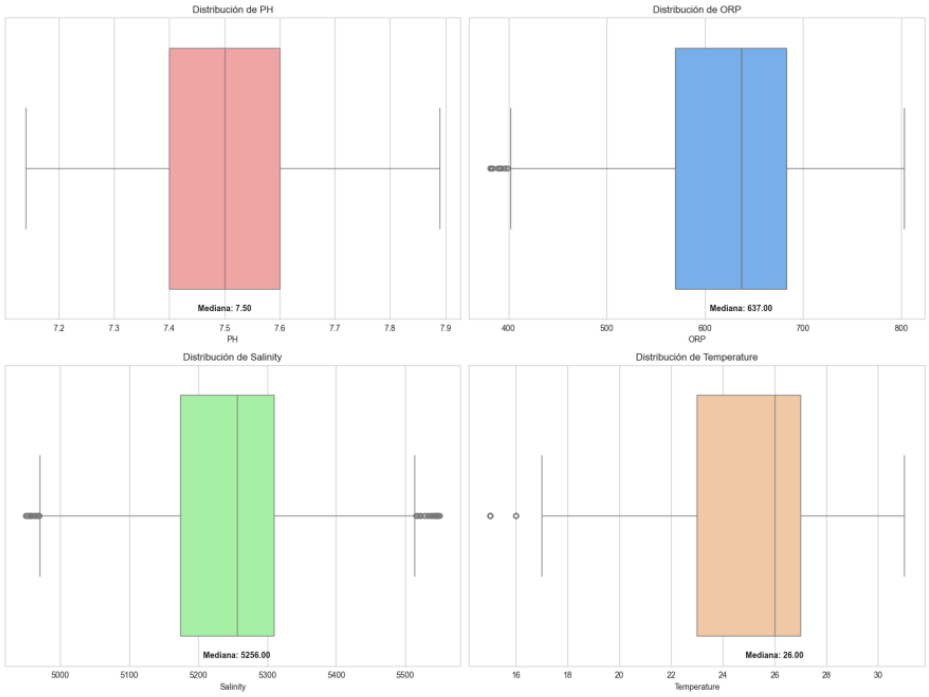
Análisis de Resultados:

- **Eficacia del método IQR**: El método IQR demostró ser eficaz para identificar valores atípicos mientras respetaba la variabilidad natural de los parámetros de la piscina. Esta aproximación nos permitió mantener datos válidos aunque inusuales, preservando la integridad de nuestro conjunto de datos.
- **Patrones de outliers:** Identificamos patrones distintivos en la distribución de outliers entre los diferentes dispositivos. Esta variabilidad podría indicar problemas específicos en ciertas piscinas, diferencias en el mantenimiento, o posibles fallos en los sensores correspondientes.
- **Evaluación de parámetros críticos:**
    - pH: Establecimos un rango crítico entre 6.8 y 8.0, marcando valores fuera de este rango para revisión.

- ORP: Definimos valores críticos por debajo de 600 mV y por encima de 900 mV.
- Salinidad: Consideramos críticos los valores fuera del rango 2000-6000 ppm, basándonos en las especificaciones del sistema. Niveles fuera de este rango podrían afectar la eficacia del sistema de cloración.

Estos rangos nos ayudaron a identificar lecturas potencialmente problemáticas, ya sea por condiciones reales de la piscina o por posibles fallos en la calibración de los sensores.

- **Impacto en la calidad de datos:** Aunque la aplicación de estos criterios redujo nuestro conjunto de datos, mejoró significativamente su calidad y representatividad. Este proceso es crucial para establecer una base de datos más confiable para nuestros análisis posteriores.



- **Variabilidad entre dispositivos**: Observamos diferencias significativas en la frecuencia y naturaleza de los outliers y valores críticos entre los distintos dispositivos. Esta variabilidad refuerza la importancia de considerar cada piscina como un caso individual en nuestros análisis futuros, reconociendo que factores como el mantenimiento, el uso y las condiciones ambientales pueden influir significativamente en los parámetros medidos.

- **Consideraciones sobre calibración**: La presencia de lecturas consistentemente fuera de rango en algunos dispositivos sugiere posibles problemas de calibración. Esto destaca la importancia del mantenimiento regular y la calibración de los sensores para garantizar la precisión de los datos recopilados. Se verán mas detalladamente en las representaciones graficas mostradas en el siguiente paso.

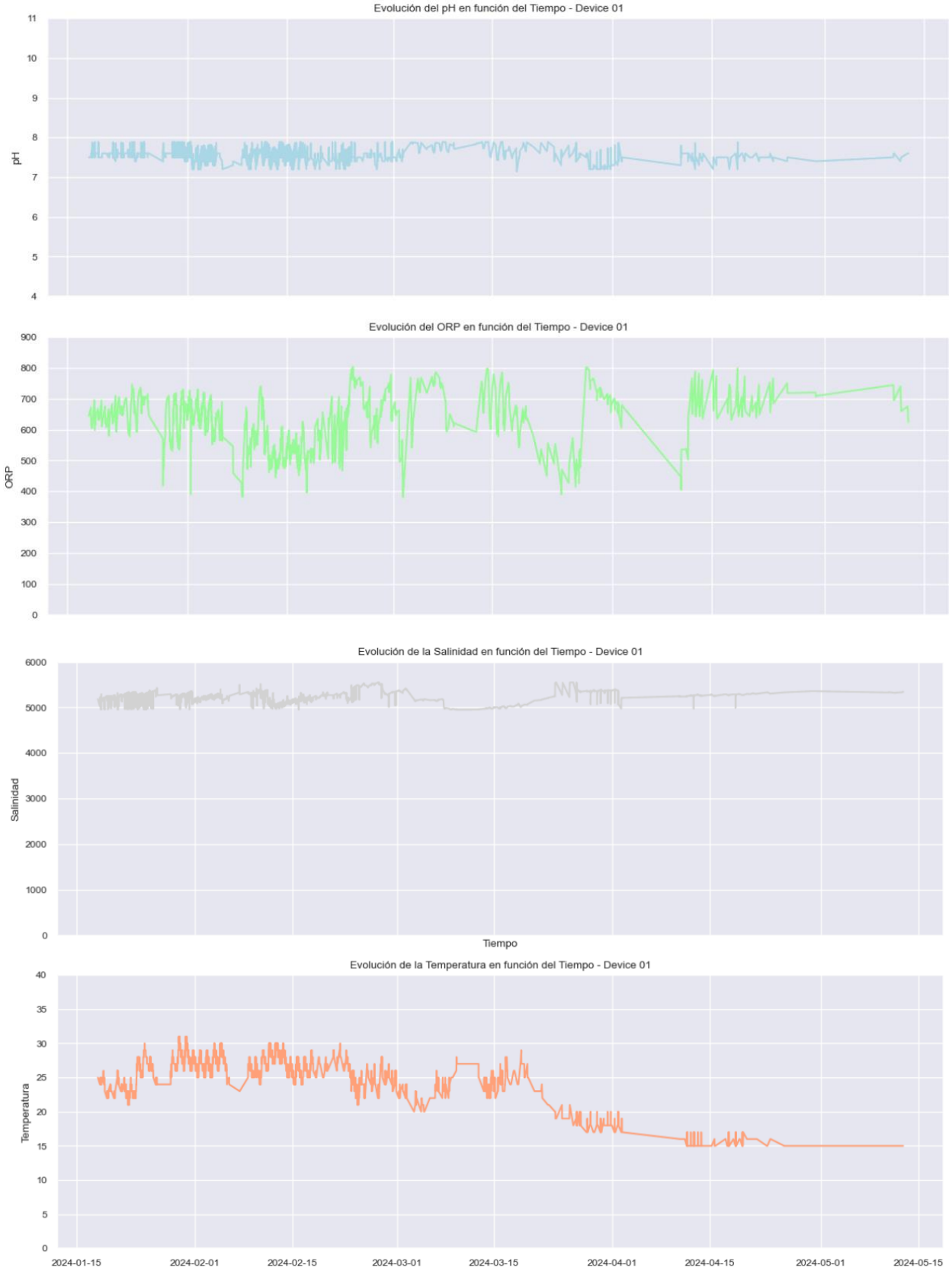
Visualización de datos:

Tras la limpieza y eliminación de Outliers se sacaron las estadísticas descriptivas por cada dispositivo y representaciones de parámetro vs tiempo por cada una de ellas. Esta tabla siguiente contiene un resumen de los datos:

Dev.	pH			ORP			Salinidad			Temperatura		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
1	7.14	7.52	7.89	381	625	803	4950	5240	5550	15	25.01	31
2	8.9	9.81	10	178	283	436	4478	5254	5783	26	30.24	34
3	7.15	7.35	7.53	454	563	687	4470	5623	6589	14	22.93	35
4	7	7.62	8.33	593	660	727	2905	3846	4991	22	25.7	29
5	6.9	7.44	8	586	677	754	3796	4017	4238	11	24.67	34
6	7.2	7.35	7.55	468	640	803	3342	3689	4228	1	23.25	42
7	7.2	7.38	7.51	567	614	662	3655	4339	4908	2	17.92	30
10	7.3	7.65	9.3	596	648	716	4006	4280	4527	16	25.96	30

Análisis por dispositivo:

- Device 01:
  - Amplio rango de pH (7.14-7.89) y ORP (381-803 mV).
  - Muestra fluctuaciones significativas, sugiriendo un sistema muy activo o posiblemente necesitado de ajuste.

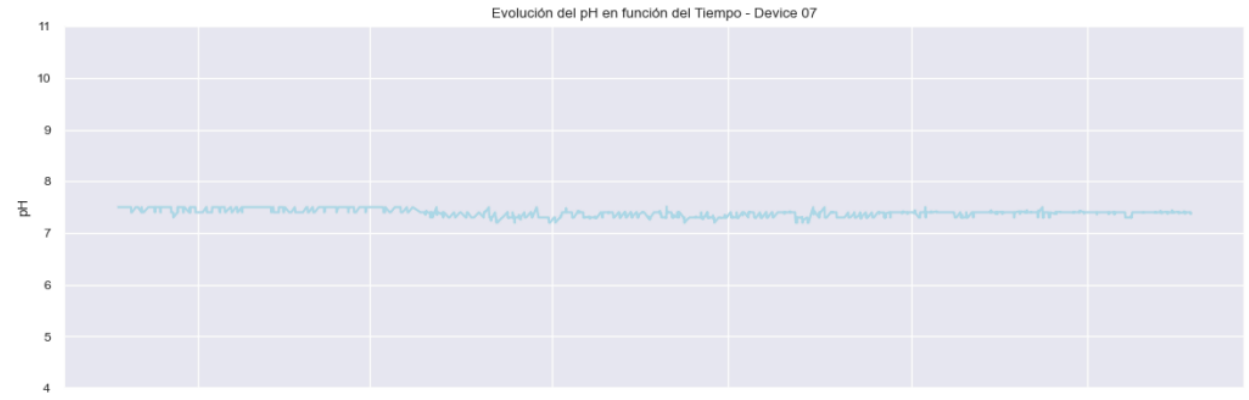

Evolución del pH en función del Tiempo - Device 01
pH
Evolución del ORP en función del Tiempo - Device 01
ORP
Evolución de la Salinidad en función del Tiempo - Device 01
Salinidad
Tiempo
Evolución de la Temperatura en función del Tiempo - Device 01
Temperatura
2024-01-15
2024-02-01
2024-02-15
2024-03-01
2024-03-15
2024-04-01
2024-04-15
2024-05-01
2024-05-15

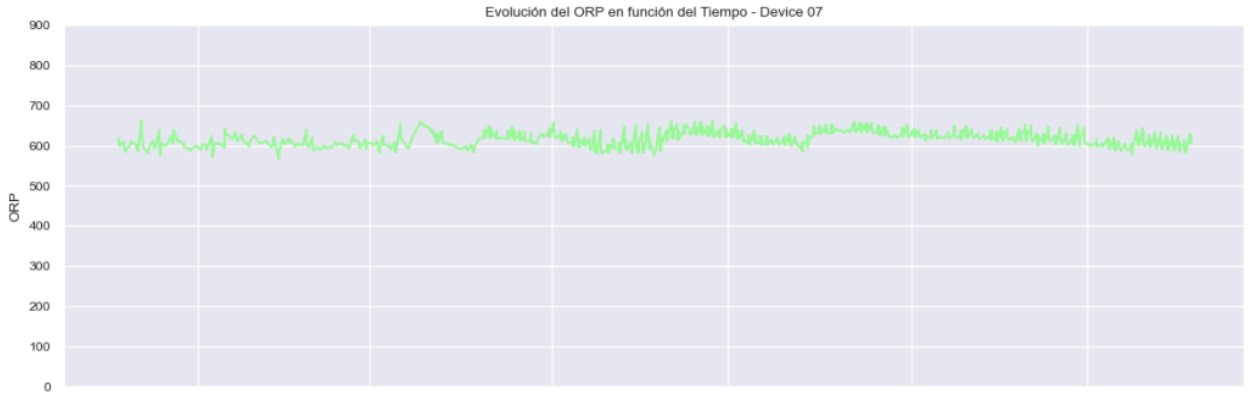
- Device 02 y Device 09:
  - Valores de pH inusualmente altos (8.9-10.0) y ORP bajos (178-436 mV).
  - Posible problema de calibración o un sistema de tratamiento de agua atípico.
- Device 03:
  - Rango de pH estable (7.15-7.53) pero ORP relativamente bajo (454-687 mV).
  - Podría indicar un buen control de pH pero insuficiente cloración.
- Device 04:
  - Amplio rango de pH (7.0-8.33) pero ORP estable y adecuado (593-727 mV).
  - Sugiere un sistema que mantiene bien la desinfección, pero podría mejorar en el control de pH.

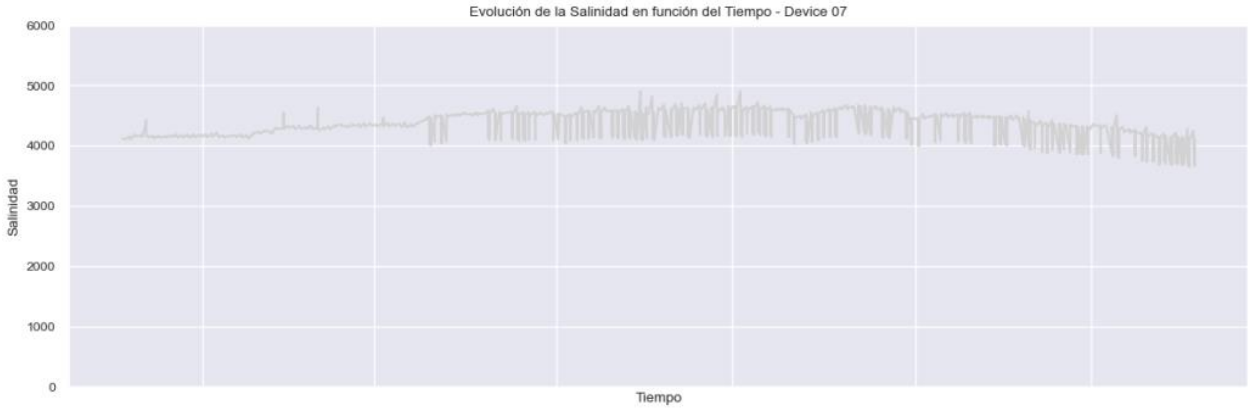
```
Datos limpios para Device 04 guardados en datos_limpios_Device 04.csv

Estadísticas Device 04:
                PH          ORP     Salinity  Temperature
count  4294.000000  4294.000000  4294.000000  4294.000000
mean      7.619506   659.754541  3845.663484    25.695622
std       0.242965    23.637339   463.643672     1.471935
min       7.000000   593.000000  2905.000000    22.000000
25%       7.500000   643.000000  3480.000000    25.000000
50%       7.500000   660.000000  3822.000000    26.000000
75%       7.810000   677.000000  4218.000000    27.000000
max       8.330000   727.000000  4991.000000    29.000000
Porcentaje de datos conservados para Device 04: 89.18%
Imagen guardada como ParametrosVSTiempo_Device 04.png
```

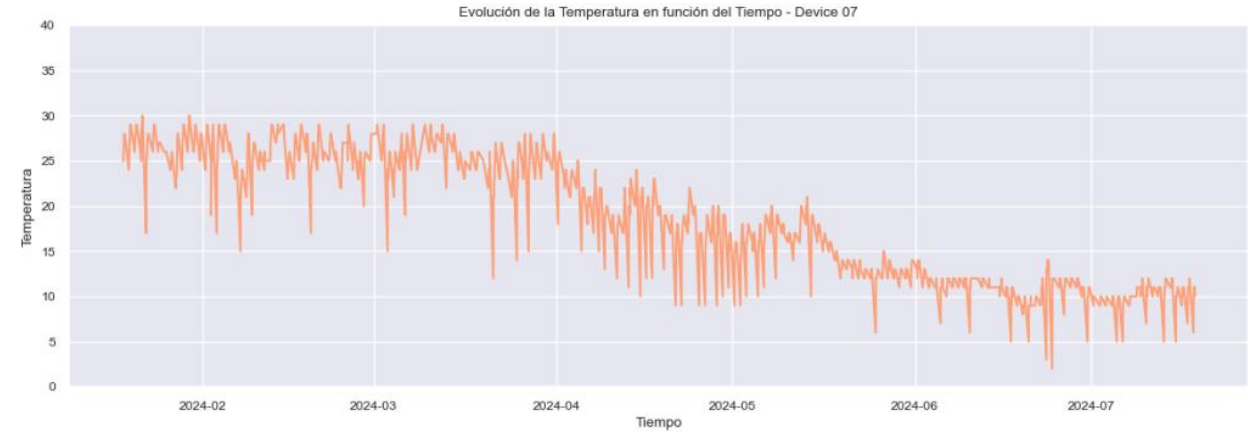
Ilustración 11. Ejemplo de salida de Estadísticas descriptivas de cada dispositivo

- Device 05:
  - Rango de pH amplio (6.9-8.0) y ORP adecuado (586-754 mV).
  - Indica un sistema que funciona bien en general pero con cierta variabilidad en pH.
- Device 06:
  - pH estable (7.2-7.55) y amplio rango de ORP (468-803 mV).
  - Sugiere buen control de pH pero posible variabilidad en la eficacia de desinfección.
- Device 07:
  - Rango de pH muy estable (7.2-7.51) y ORP moderado (567-662 mV).
  - Indica un sistema bien controlado y estable.


Evolución del pH en función del Tiempo - Device 07
pH
11
10
9
8
7
6
5
4


Evolución del ORP en función del Tiempo - Device 07
ORP
900
800
700
600
500
400
300
200
100
0


Evolución de la Salinidad en función del Tiempo - Device 07
Salinidad
6000
5000
4000
3000
2000
1000
0
Tiempo


Evolución de la Temperatura en función del Tiempo - Device 07
Temperatura
40
35
30
25
20
15
10
5
0
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
Tiempo

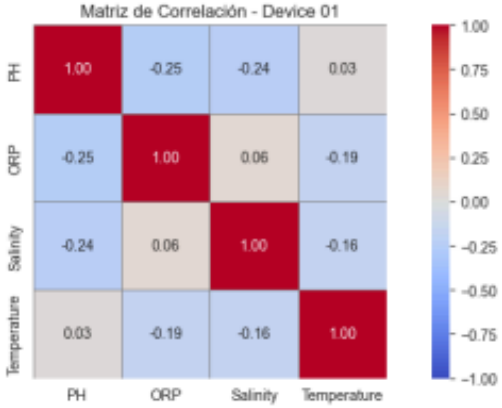
- Device 10:
    - Amplio rango de pH (7.3-9.3) y ORP moderado (596-716 mV).
    - Sugiere desafíos en el control de pH pero desinfección relativamente estable.

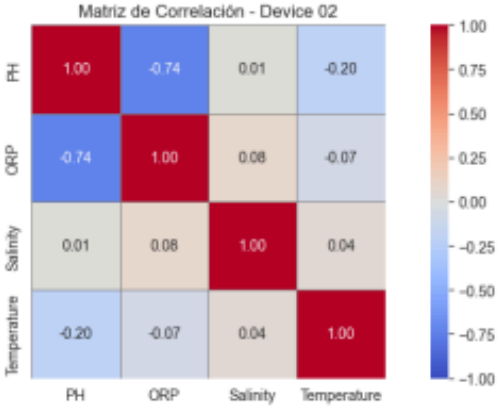
Análisis del conjunto, Patrones y Observaciones:

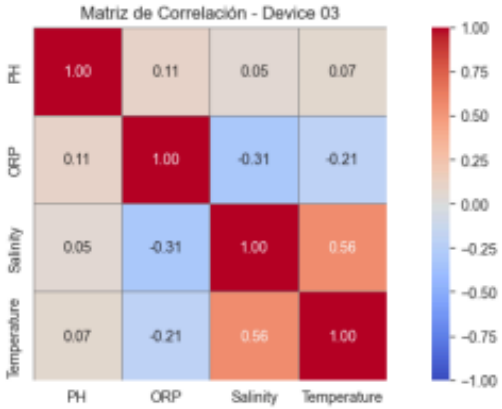
1. Efecto estacional:
    - En Australia, enero es verano y julio invierno. Se observa una tendencia general de **disminución en las temperaturas** del agua hacia julio.
    - Varios dispositivos muestran una reducción en la variabilidad de los parámetros durante los meses más fríos, consistente con un **menor uso de la piscina** y posible reducción en el funcionamiento del sistema.
2. Comportamiento del ORP anómalo:
    - Valores de ORP entre 300-400 mV son considerados bajos y generalmente indican una desinfección insuficiente.
    - El rango 500-700 mV es más típico, con un objetivo ideal entre 630-650 mV.
    - Devices 02 y 09 muestran ORP consistentemente bajo, indicando problemas graves de desinfección o fallos de calibración.
    - La mayoría de los otros dispositivos mantienen ORP en rangos aceptables, con fluctuaciones ocasionales. Esto se justificara próximamente.
3. Patrones de uso y mantenimiento:
    - Se observan patrones cíclicos en varios dispositivos, probablemente relacionados con los ciclos de filtración y cloración.
    - Algunos dispositivos (como el 01 y 04) muestran cambios más frecuentes, sugiriendo un uso más intensivo o sistemas más reactivos.
    - Otros (como el 07 y 08) muestran patrones más estables, indicando posiblemente un uso menos intensivo o sistemas mejor equilibrados.
4. Posibles fallos de calibración:
    - Device 02: Claro problemas de calibración en el sensor pH y ORP.
5. Falta de correlación:

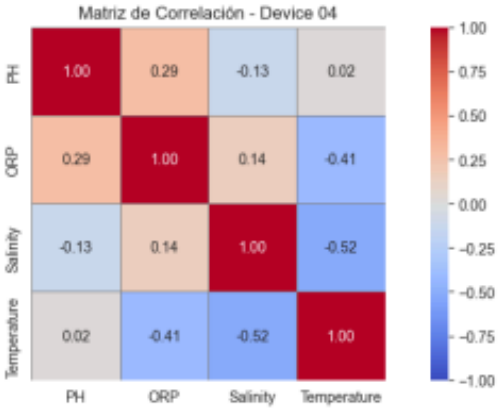
Problema de registro de datos: El desplazamiento temporal de las lecturas debido al problema de registro donde se distorsiona las relaciones reales entre los parámetros. Además de las condiciones únicas de cada piscinas donde opera en una piscina con características específicas (tamaño, uso, entorno) y el funcionamiento del propio sistema que corrompe el proceso natural de la química porque al subir el ORP sube el pH pero este al alcanzar cierta basicidad activa la bomba de acidez, rompiendo así el patrón de comportamiento natural que tendrían los parámetros.

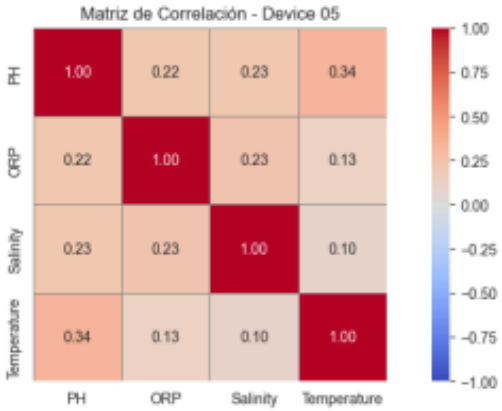
Matrices de Correlación generales:

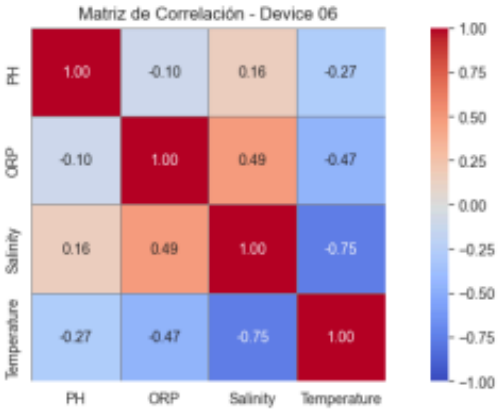


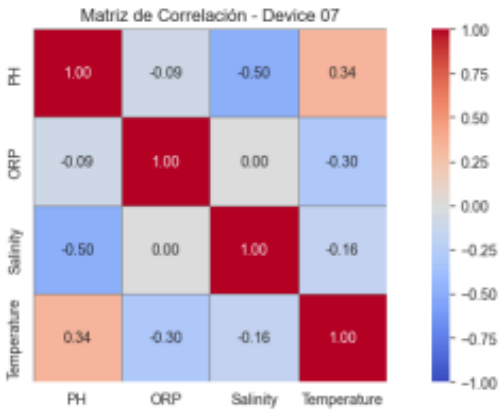


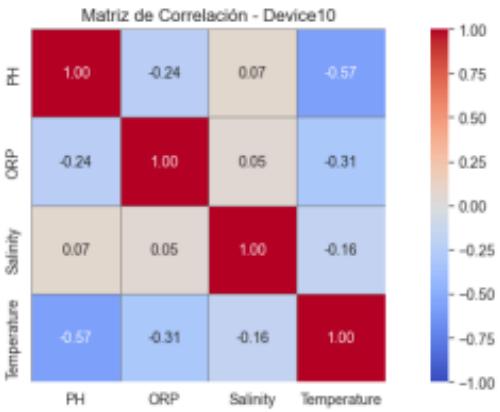






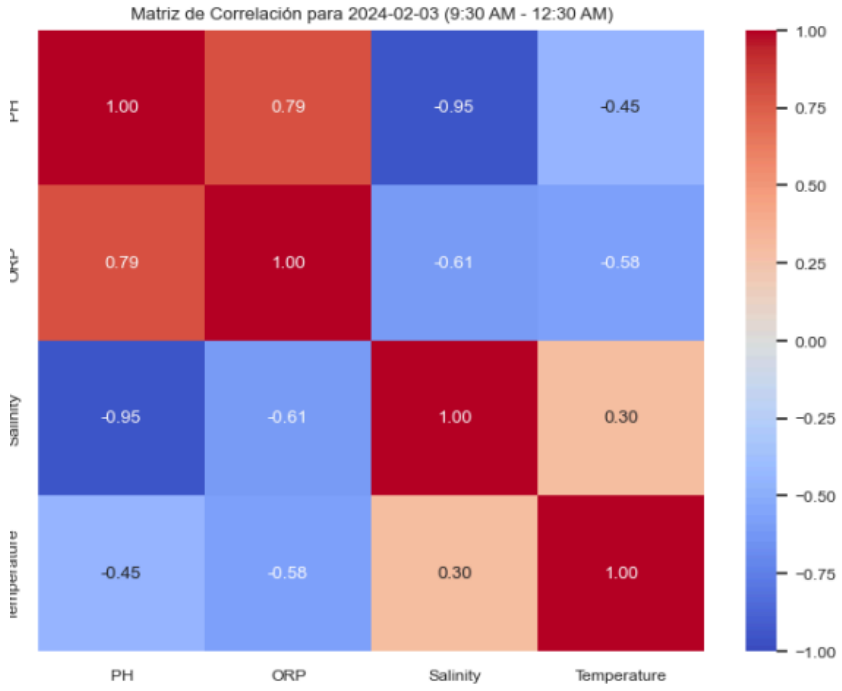






Matriz temporal acotada.

En un dispositivo bien mantenido como el 07 y acotando los tiempos vemos mejor los patrones de correlación son más evidentes y consistentes cuando se analiza un período de tiempo específico. Esto demuestra la importancia de considerar ventanas de tiempo más cortas para observar las relaciones reales entre parámetros, evitando las distorsiones causadas por el problema de registro de datos a largo plazo.



- ✓ Fuerte correlación negativa entre pH y Salinidad (-0.95): Esto sugiere que cuando el pH aumenta, la salinidad tiende a disminuir y viceversa. Esta relación podría reflejar el funcionamiento del sistema de cloración salina, donde la producción de cloro afecta tanto al pH como a la concentración de sal.
- ✓ Correlación positiva entre pH y ORP (0.79): Contrariamente a la relación inversa teórica, aquí vemos una correlación positiva. Esto podría indicar que durante este período específico, otros factores (como la adición de cloro) están influyendo más en el ORP que el pH.
- ✓ Correlación negativa moderada entre ORP y Salinidad (-0.61): Esta relación sugiere que niveles más altos de salinidad están asociados con niveles más bajos de ORP. Podría indicar que el sistema está ajustando la producción de cloro en respuesta a los cambios en la salinidad.
- ✓ Correlaciones moderadas con la Temperatura: La temperatura muestra correlaciones negativas moderadas con pH (-0.45) y ORP (-0.58), y una correlación positiva débil con la Salinidad (0.30). Esto refleja cómo los cambios de temperatura afectan a los procesos químicos en la piscina.
- ✓ Importancia del contexto temporal: Esta matriz esta acotada en un tiempo en el que el dispositivo 07 empezaba su ciclo de tratamiento.

Problema de registro de datos y su impacto en el análisis:

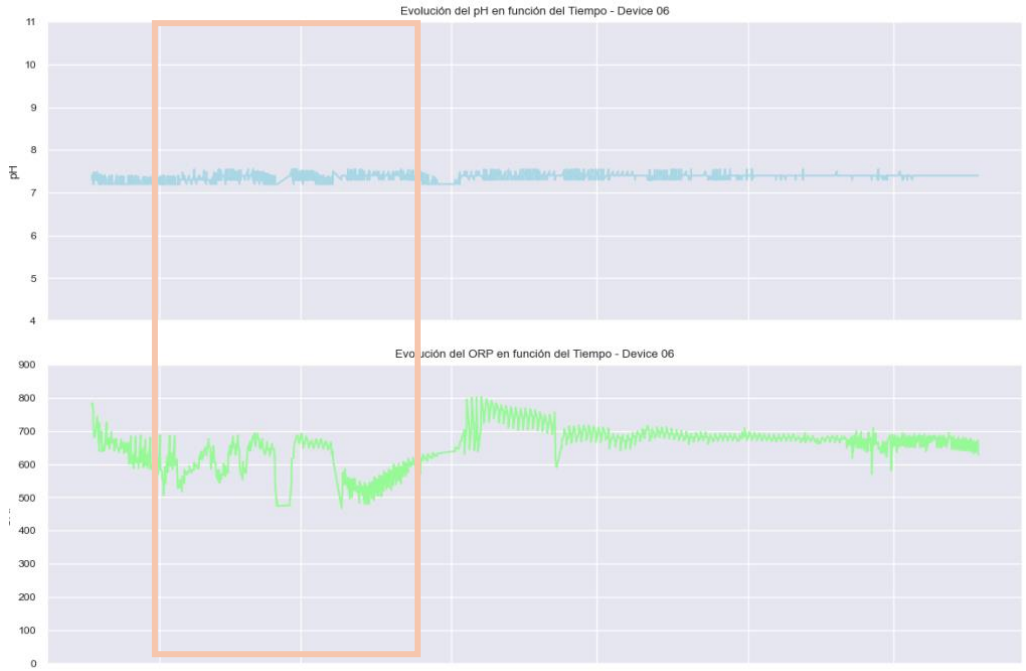
Un aspecto crítico que ha afectado significativamente a la interpretación de nuestros datos es un problema inherente al sistema de registro del Lifeguard. Durante el tratamiento de datos y su visualización se ha identificado que cuando el sensor no logra tomar una lectura o se salta una medición, en lugar de registrar un valor nulo o cero, el sistema simplemente desplaza los datos subsiguientes. Este comportamiento ha resultado en un desajuste acumulativo a lo largo del tiempo entre las lecturas registradas y los momentos reales en que estas ocurrieron.

Estos problemas surgen cuando nos enfrentamos a **escenarios reales** como este, donde la empresa que ha proporcionado los datos y en la que actualmente trabajo, almacenaba este tipo de información, pero nunca le había dado un uso real para haber detectado este problema y que se ha localizado en una de las actualizaciones del software que se hizo en el sistema.

El problema no esta en el funcionamiento sino en el almacenamiento porque estos dispositivos tienen alarmas y si las piscinas estuvieran mal mantenidas la aplicación avisaría al dueño de su peligrosidad y mantenimiento en los químicos, las lecturas en el momento son correctas y siguen la variabilidad de los setpoints que se les da.

Implicaciones del problema de registro:

- **Desfase temporal**: Con el paso del tiempo, las lecturas registradas para un momento dado podrían corresponder en realidad a un período completamente diferente. Esto explica por qué observamos aparentes inconsistencias en la relación entre pH y ORP como podemos ver en la siguiente imagen, ej Device 06 o la grafica anterior Device 01.



- **Datos de inactividad mezclados**: Como resultado de este desplazamiento, es posible que estemos analizando datos que incluyen períodos donde la bomba ni siquiera estaba en funcionamiento, lo cual explica algunas de las lecturas aparentemente incorrectas o improbables

- **Dificultad en la Identificación de Patrones**: La falta de sincronización entre los datos de diferentes sensores hizo que fuera más difícil identificar patrones y correlaciones precisas entre las variables. Esto es crítico en un sistema predictivo, donde las relaciones entre variables son la base para construir modelos precisos.

- **Posibles Sesgos en el Modelo**: Al trabajar con datos desalineados, existe el riesgo de introducir sesgos en los modelos predictivos. Por ejemplo, si los datos de pH se correlacionan incorrectamente con datos de temperatura que no corresponden al mismo momento en el tiempo, el modelo podría aprender relaciones espurias, lo que degradaría su rendimiento en escenarios reales.

Justificación del comportamiento observado:

I. Antigüedad: Los dispositivos se ven afectados más o menos por este desfase en función de su vida útil algunos de ellos llevan menos activos que otros, se ha visto que el device 01 lleva colectando data más de 5 años cuando otros como el 7 llevan menos pueden no haber notado o se puede haber corregido el problema en las últimas actualizaciones o simplemente tenía mejor conexión a internet.

II. Falta de patrones claros: La dificultad para identificar patrones claros y consistentes en algunos dispositivos se debe probablemente a que estamos analizando datos que no están correctamente alineados en el tiempo (grafica previa)

III. Comportamiento improbable del sistema: El hecho de que los datos sugieran que algunos sistemas funcionaban continuamente sin pausas o sin alertar a los usuarios sobre condiciones críticas se explica por este **problema en el registro** y recolección de la data, **no de funcionamiento.**

IV. Enfoque en tendencias generales: En lugar de centrarnos en correlaciones específicas o eventos puntuales, pensamos que era mejor examinar tendencias muy generales y rangos de valores para cada parámetro. Por eso por ejemplo a la hora de representar la matriz de correlación al no encontrar patrones obvios o repetitivos en todos los conjuntos se observó el Device 07 uno de los más estables, al inicio de cada ciclo durante un periodo de tiempo, confirmando que de alguna manera si existe alguna correlación como se vio en el apartado previo.

Fase 3: Desarrollo de Modelos de ML

En el proyecto nos centramos en desarrollar un sistema predictivo para optimizar el funcionamiento del sistema Lifeguard en piscinas. Después de un análisis exploratorio de datos de 10 piscinas diferentes, se decidió enfocar el estudio en dos dispositivos específicos:

- Device 01: Elegido por su gran cantidad de datos y alta variabilidad, lo que proporciona un conjunto de datos rico para el entrenamiento de modelos más complejos.
- Device 07: Seleccionado por tener un conjunto de datos más pequeño, pero de mejor calidad, con patrones más consistentes y representativos del funcionamiento típico de una piscina.

Esta decisión de entrenar conjuntos por separado también se basó en la observación de que cada piscina tiene condiciones únicas influenciadas por factores como el clima, la localización, el tamaño de la piscina y los puntos de ajuste específicos. Por lo tanto, la opción de crear modelos personalizados tiene más sentido que un modelo general que no se adapte a las características propias de cada sistema por ello estas fueron las consideraciones importantes:

A. **Variabilidad entre piscinas:** Cada piscina tiene condiciones únicas basadas en su localización, tamaño, uso y entorno. Esto se refleja en la variabilidad de los rangos de pH, ORP y otros parámetros entre los diferentes dispositivos, como se muestra en la tabla de resumen.
B. **Optimización específica**: Al entrenar un modelo para cada piscina individualmente, podemos optimizar el rendimiento para las condiciones específicas de esa piscina, lo que potencialmente conduce a predicciones más precisas y relevantes.
C. **Control de variables**: Comenzar con un solo dispositivo nos permite tener un mejor control sobre las variables y entender mejor cómo el modelo responde a un conjunto específico de condiciones antes de introducir la complejidad adicional de múltiples piscinas.
D. **Escalabilidad del enfoque**: Una vez que hayamos desarrollado y validado nuestro enfoque con un dispositivo, podremos escalar el proceso a los demás dispositivos, ya sea entrenando modelos individuales para cada uno o explorando técnicas de transferencia de aprendizaje.
E. **Comparación de rendimiento**: Este enfoque nos permitirá comparar el rendimiento del modelo entre diferentes piscinas, lo que puede proporcionar insights valiosos sobre qué factores influyen más en la precisión de las predicciones.
F. **Personalización del mantenimiento**: Al modelar cada piscina por separado, podemos ofrecer recomendaciones de mantenimiento más personalizadas y precisas para cada instalación.

Implementación de los modelos:

El proceso de modelado se dividió en dos partes principales:

1. **Modelo de Clasificación para la Activación del Clorificador:**
   - Objetivo: Predecir cuándo se necesita activar el clorificador basándose en los niveles de ORP.

- Método: Se utilizó un modelo de Random Forest Classifier. Tenia unas condiciones de uso en las que el clorificador se activa cuando el ORP esta 50mV por debajo del umbral establecido y deja de generar cuando se hacerca 20mV porque la química resultante continua un tiempo y no se quiere que suba demasiado.

- Importancia: Este modelo ayuda a optimizar el uso del clorificador, mejorando la eficiencia energética y la calidad del agua. El uso de Random Forest permite capturar relaciones no lineales entre las variables de entrada y la decisión de activación. Obviamente el resto de características por el tipo de modelo no tenían peso sobre esta decisión en el modelo.

2. <u>Modelo de Predicción del pH:</u>

- **Objetivo**: Predecir cambios futuros en el pH del agua de la piscina.
- **Métodos comparados**: Regresión Lineal, Random Forest, Gradient Boosting, y en el caso del Device 01, también Redes Neuronales.
- **Características clave**: Se incluyeron variables de rezago (lag) para pH y ORP. Aquí realmente no hicieron falta demasiadas instrucciones de uso porque se quería ver si sabiendo cuando se activa el clorificador y el aumento de las variables se podía predecir en que momento seria adecuada la activación de la bomba de acidez – 'Predict_Cl'
- **Importancia del rezago**: Las variables de rezago capturan la dependencia temporal de los valores de pH, permitiendo al modelo aprender de los patrones históricos recientes. Y ajuste de hiperparametros en el device 01 del modelo con mejores resultados y en el device 07 se quisieron analizar los tres mejores modelos.

```
Forma de los datos de entrenamiento: (3466, 10)
Forma de los datos de prueba: (867, 10)
Características utilizadas: ['ORP', 'Salinity', 'Temperature', 'PH_lag_1', 'PH_lag_
2', 'PH_lag_3', 'ORP_lag_1', 'ORP_lag_2', 'ORP_lag_3', 'Predict_Cl']
```

- **Elección del modelo**: Para ambos dispositivos, se optó por comparar varios modelos para capturar las posibles relaciones no lineales en los datos. La elección final se basó en el rendimiento medido por el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2). Uso del MSE fue métrica principal porque:

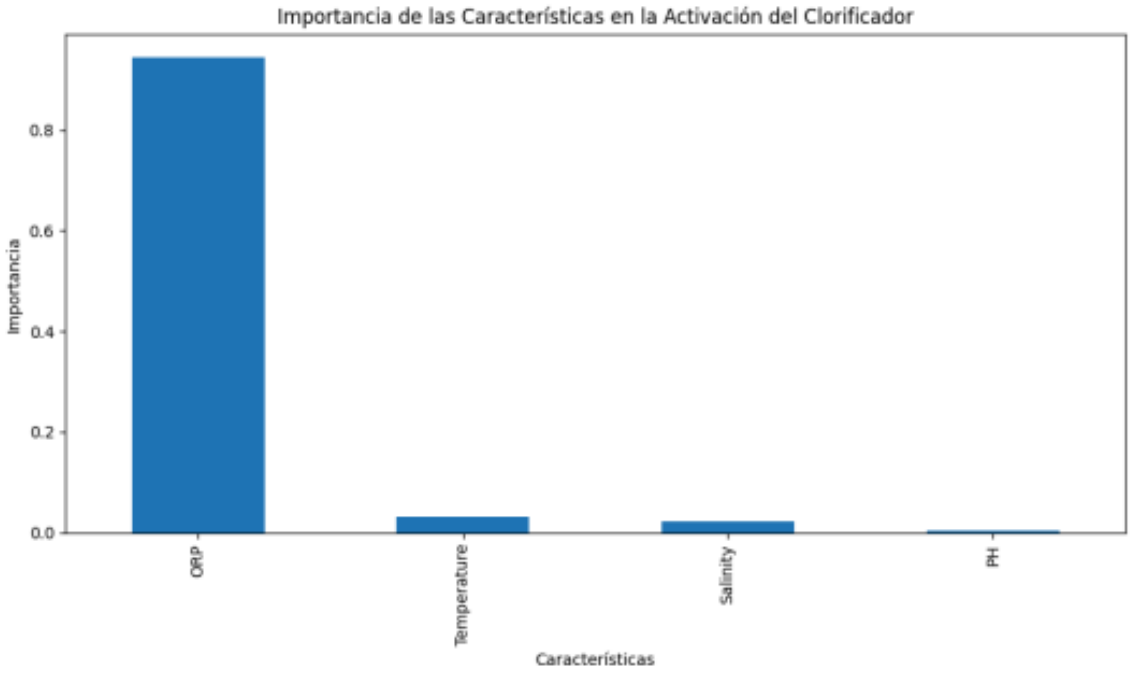
  - Penaliza errores grandes más que errores pequeños, lo cual es crucial en el control de pH donde desviaciones significativas pueden ser problemáticas.
  - Es sensible a outliers, lo que ayuda a identificar modelos que puedan estar fallando en casos extremos.
  - Está en las mismas unidades que la variable objetivo (pH), facilitando la interpretación.

Análisis de resultados:

Predicción de la Activación del Clorificador:

La decisión de activar el clorificador es fundamentalmente un problema de clasificación binaria (activar o no activar). Este modelo tuvo un acierto del 100% debido a que era igual de eficiente que lo sería un simple if-else. Pero se decidió usar los conocimientos para probarlo y luego ver de hecho si la combinación de esto afectaba a la predicción del pH añadiendo su resultado como un input de entrada al siguiente sistema. Dio resultados iguales en ambos dispositivos dada la simplicidad y la importancia de los otros parámetros no era relevante.

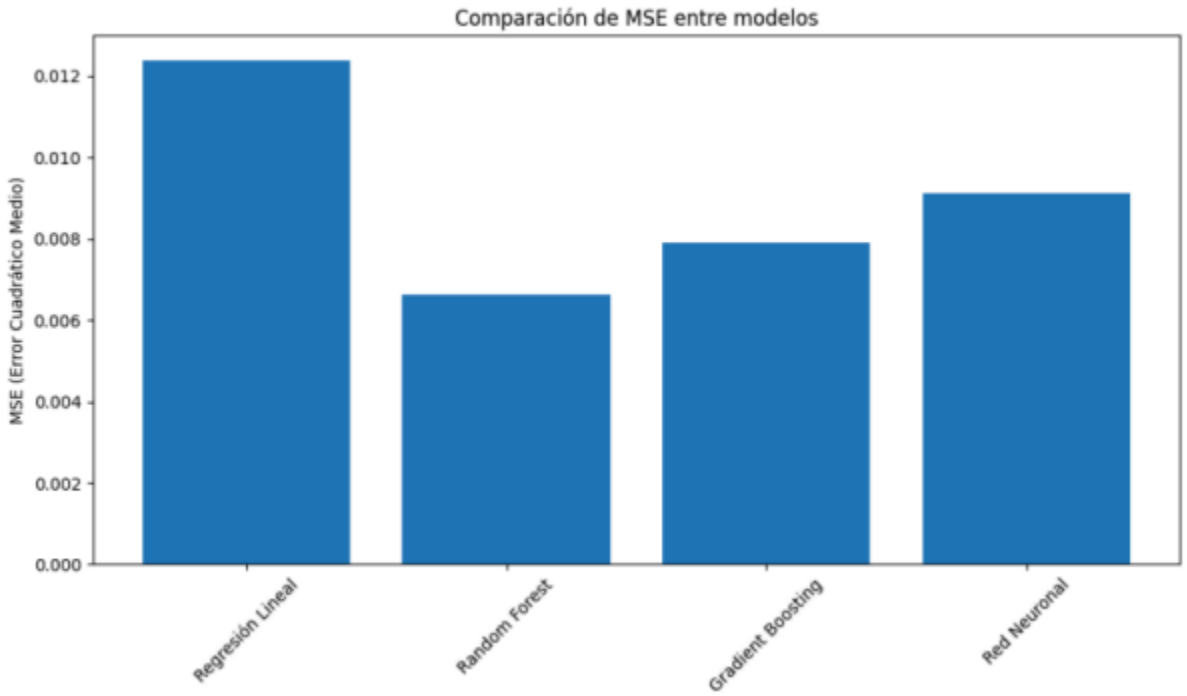
```
Informe de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       513
           1       1.00      1.00      1.00       357

    accuracy                           1.00       870
   macro avg       1.00      1.00      1.00       870
weighted avg       1.00      1.00      1.00       870


Matriz de confusión:
[[513   0]
 [  0 357]]
```



Predicción del pH - Modelos analizados:

Device 01:

Modelo	MSE	R2	Conclusiones
Regresión Lineal	0.0124	0.4177	A pesar de su simplicidad y velocidad, este modelo no es ideal para este tipo de problema donde las interacciones entre las variables no son lineales.
Random Forest	0.0066	0.6891	Captura bien las relaciones no lineales, mejor desempeño global.
Gradient Boosting	0.0079	0.6283	Similar a Random Forest pero con un desempeño ligeramente inferior.
Red Neuronal	0.0091	0.5717	Buen rendimiento pero sobreajustado y menor que Random Forest.



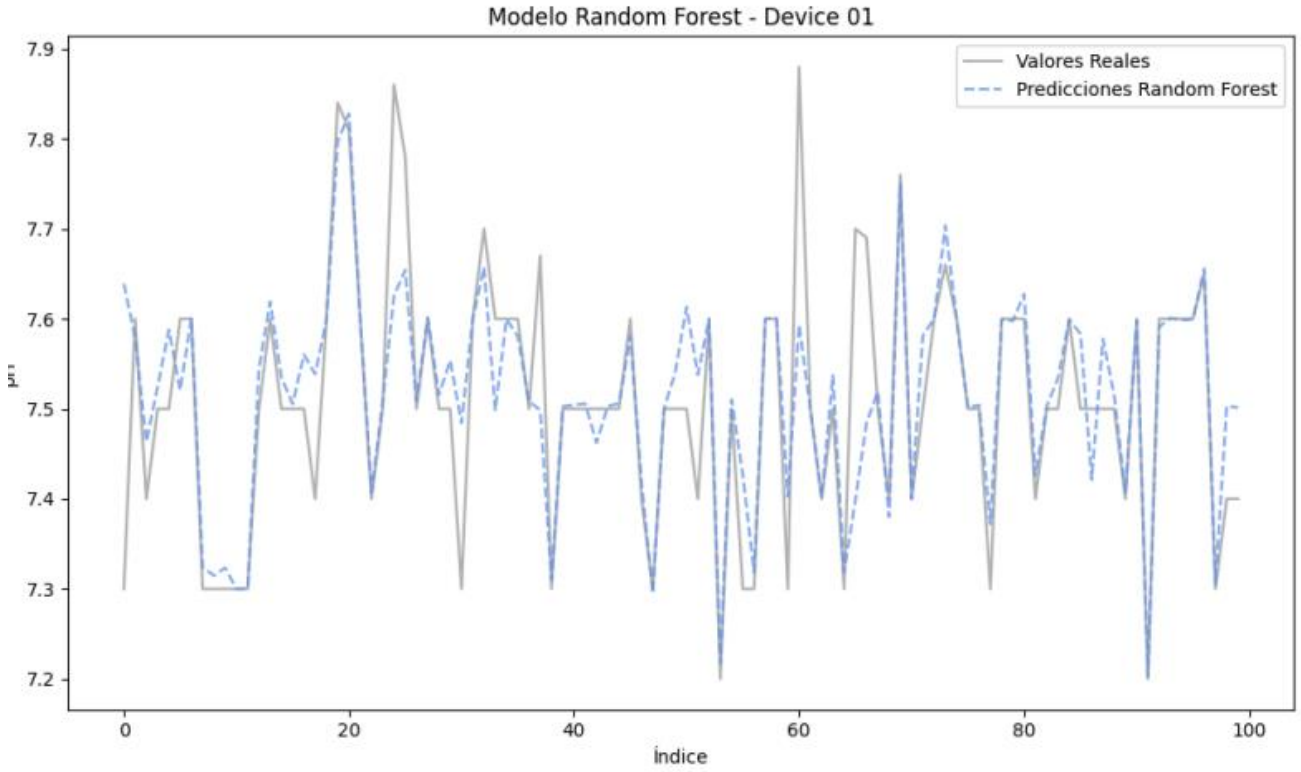
Análisis de las gráficas y resultados:

La mejor opción para entrenar nuestro modelo fue Random Forest Funcionó porque:

1. **Captura de Relaciones No Lineales**: A diferencia de la regresión lineal, Random Forest es un modelo basado en árboles de decisión que puede capturar relaciones no lineales complejas en los datos, lo cual es crucial en sistemas como el control del pH de piscinas donde la relación entre variables como ORP, salinidad y temperatura con el pH no es lineal.

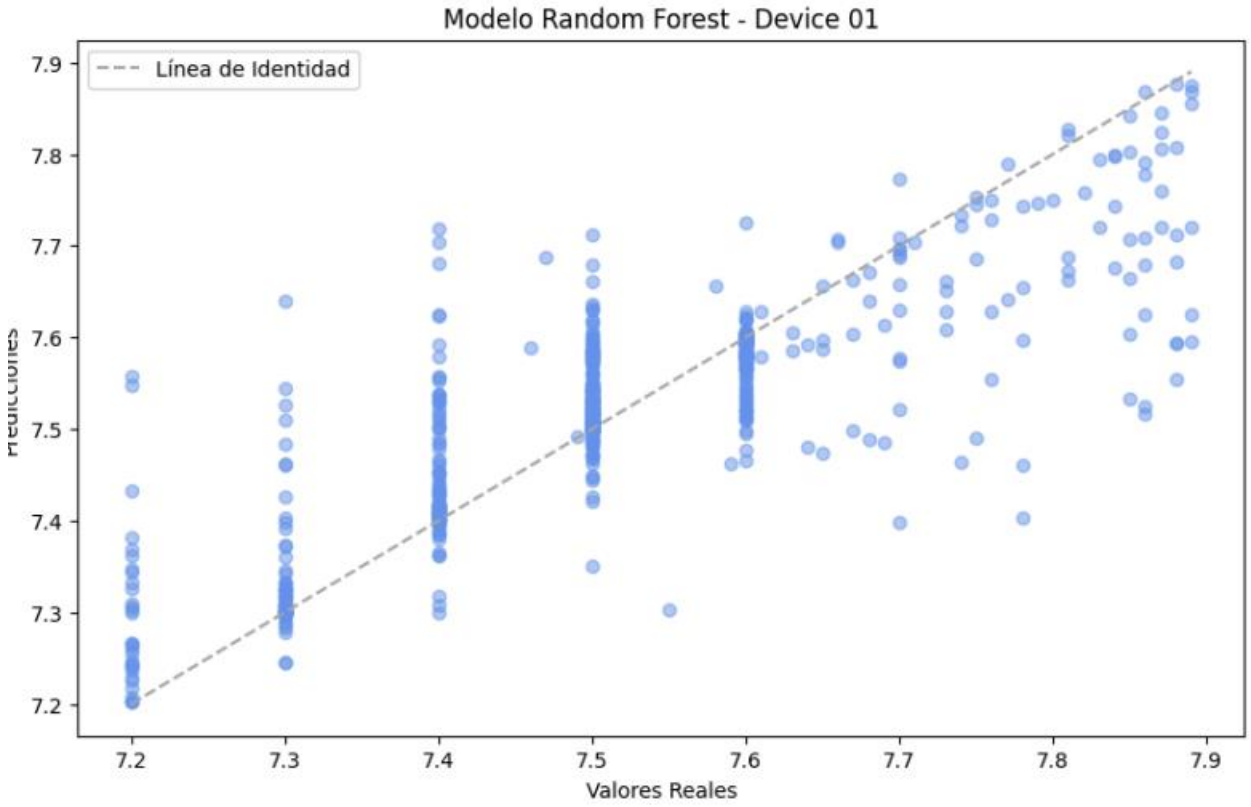
2. **Robustez y Estabilidad**: Random Forest promedia los resultados de múltiples árboles de decisión, lo que reduce el riesgo de sobreajuste (overfitting) y mejora la robustez del modelo ante la variabilidad de los datos.
3. **Importancia de Variables**: El análisis de importancia de características mostró que las variables como el ORP y los valores anteriores de pH eran cruciales para la predicción, algo que Random Forest pudo capturar efectivamente.

   Después el ajuste de hiperparametros las predicciones del sistema quedaron así:



Comparación de la grafica temporal entre Valores Reales y Predicciones

En esta gráfica con los 100 primeros valores predichos mustra en la línea continua gris que representa los valores reales de pH a lo largo de un conjunto de datos de prueba. La línea discontinua azul representa las predicciones del modelo Random Forest.



Grafica de dispersión

La línea de identidad nos muestra donde tendrían que haber sido los valores reales y los marcadores son los predichos por nuestro modelo.

Análisis de los resultados:

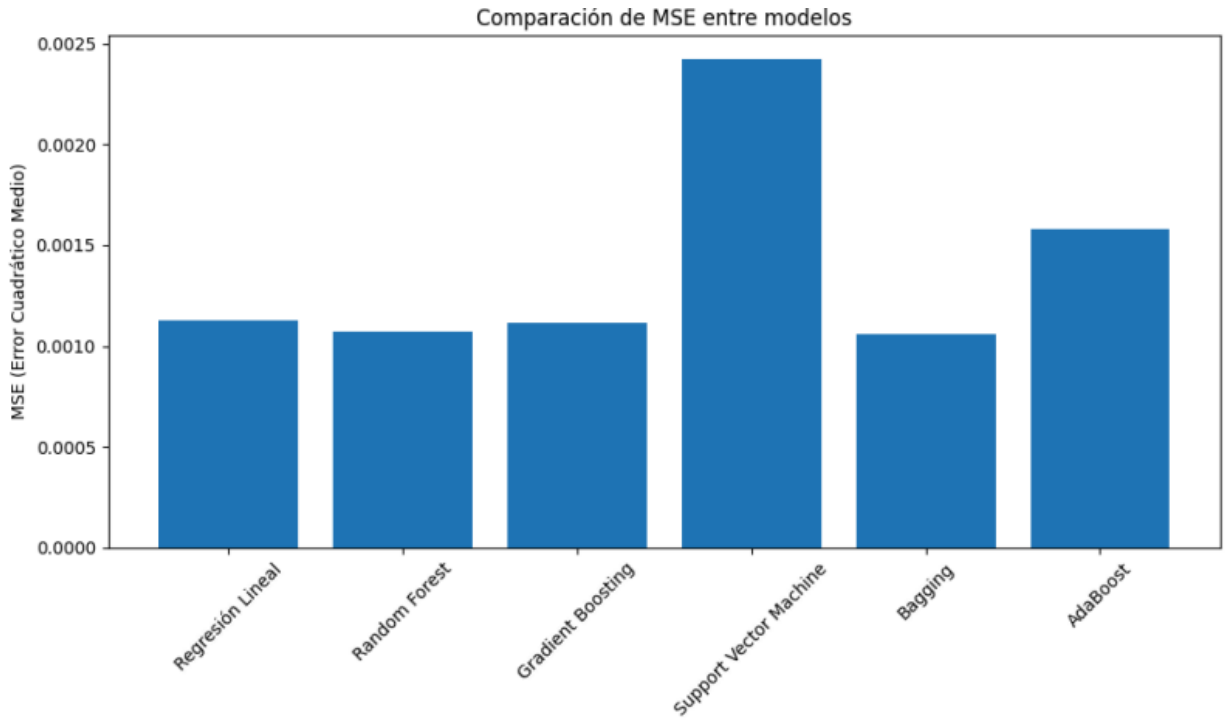
- **Tendencia General**: Hemos observado que las predicciones del modelo siguen de manera general la misma tendencia que los valores reales. Esto indica que el modelo es capaz de captar la estructura subyacente en los datos.
- **Discrepancias Notables**: Vemos que hay momentos en que las predicciones del modelo no coinciden exactamente con los valores reales, especialmente en los puntos más extremos, esto se ve fácilmente en la segunda grafica. Por ejemplo, cuando el valor de pH real presenta picos o caídas abruptas, el modelo tiene dificultades para ajustarse con precisión, lo que puede indicar que el modelo no captura bien los cambios bruscos en el pH.
- **Retraso en la Respuesta**: El modelo parece tener un cierto retraso en la respuesta en algunas secciones, es decir, cuando hay un cambio abrupto en los valores reales, el modelo tarda un poco en reaccionar a dicho cambio.

La dispersión observada sugiere que el modelo tiene problemas para predecir con precisión en los valores extremos del pH. Esto podría ser debido a la falta de suficientes datos extremos durante el entrenamiento o a que el modelo está sobreajustado para predecir en el rango medio, donde probablemente haya más datos disponibles. Aunque el modelo es fuerte en la mayoría de los casos,

podría beneficiarse de una mayor diversidad en los datos de entrenamiento o de una mayor optimización de sus hiperparámetros.

Device 07:

Modelo	MSE	Observaciones
Regresión Lineal	0.0015	Modelo simple, no captura bien las relaciones no lineales complejas del sistema.
Random Forest	0.0013	Captura bien las relaciones no lineales, mostrando mejor rendimiento general.
Gradient Boosting	0.0013	Similar a Random Forest, pero con un comportamiento ligeramente menos estable.
Support Vector Machine	0.0025	El peor desempeño debido a su ineficiencia para capturar las complejidades de las relaciones.
Bagging	0.0014	Mejor que la regresión lineal, pero menos efectivo que Random Forest y Gradient Boosting.
AdaBoost	0.0020	Mayor error, sufre en la predicción de casos extremos.



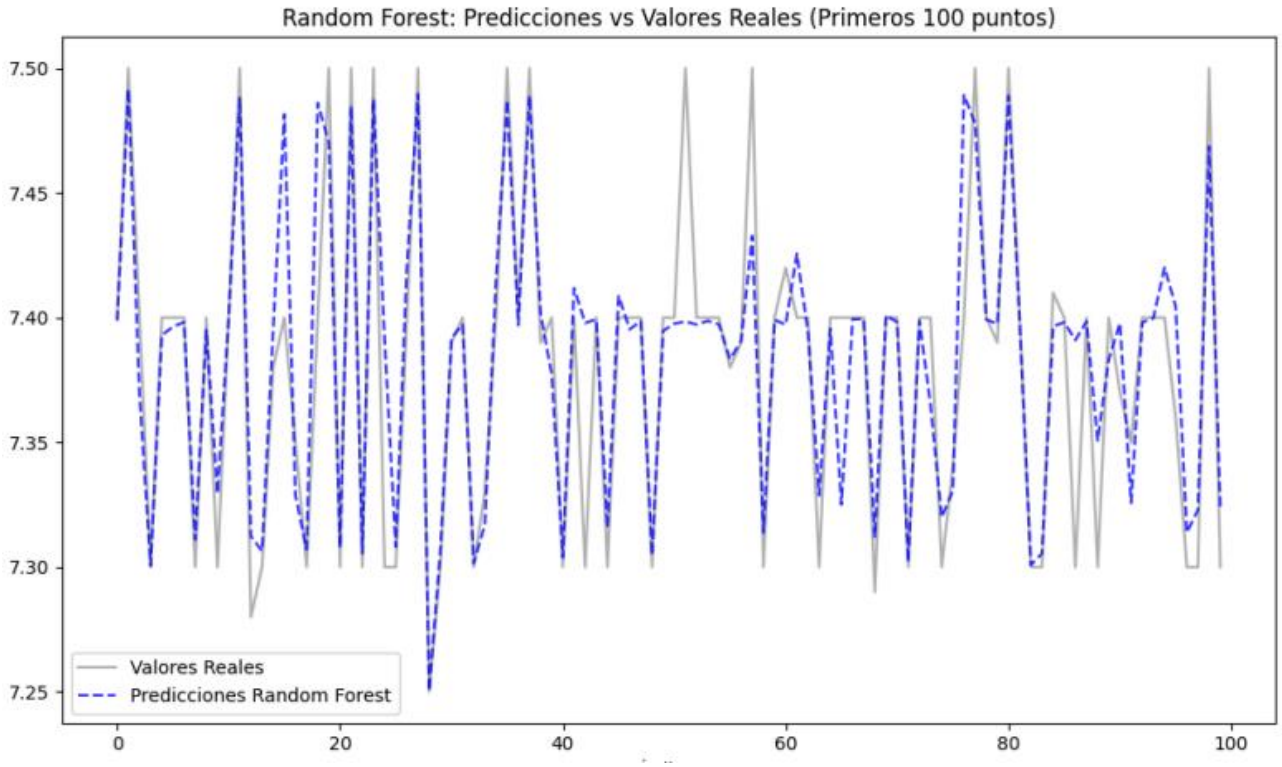
Análisis de las gráficas y resultados:

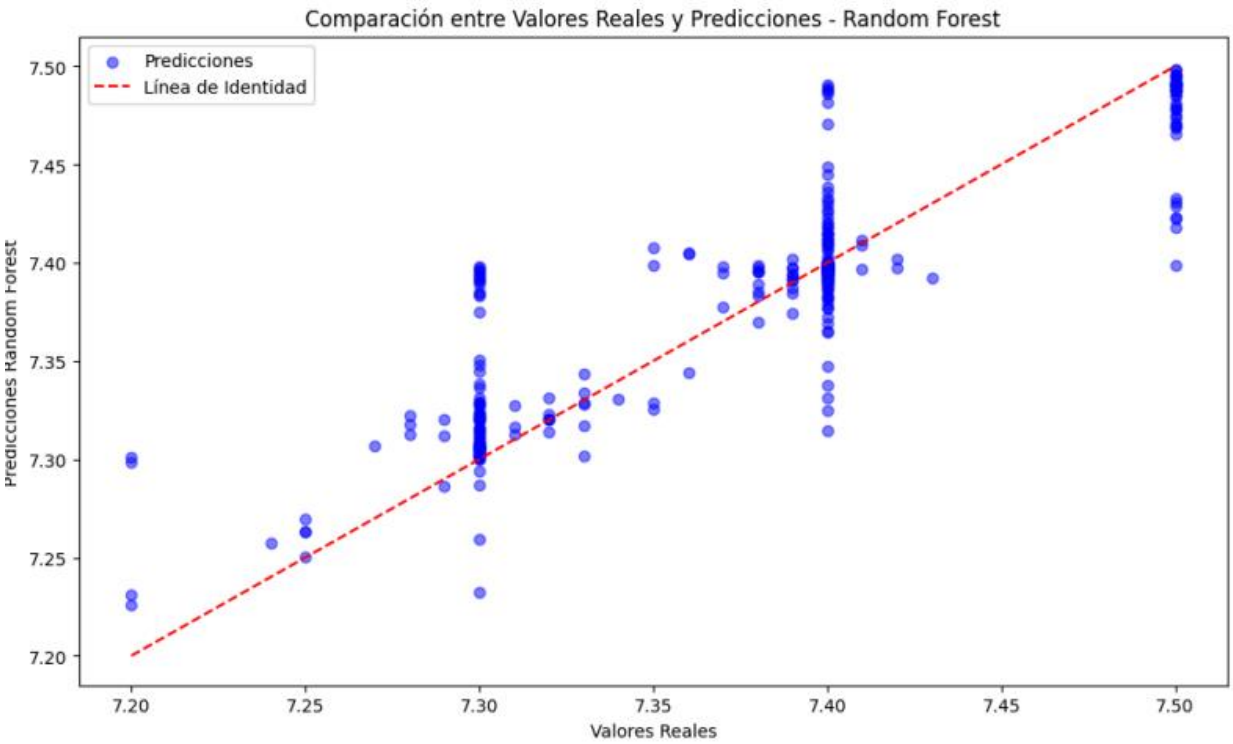
Los peores modelos fueron:

- **Regresión Lineal,** explicado en el anterior caso.
- **Support Vector Machine**: Sin duda vemos que este modelo tiene el peor rendimiento de todos los probados, con el MSE más alto. Es menos efectivo en la captura de relaciones no lineales complejas que son cruciales en este problema. Aunque es poderoso en ciertos contextos, no es adecuado para este tipo de problemas donde las relaciones no lineales predominan y el ruido es significativo.
- **AdaBoost**: Aunque sabemos que puede mejorar las predicciones al enfocarse en ejemplos difíciles, en este caso, su rendimiento es inferior debido a su tendencia a ser influenciado por ruido y valores atípicos, lo que resulta en un mayor MSE.

Modelos que se quisieron comparar con un mejor ajuste de los hiperparametros:

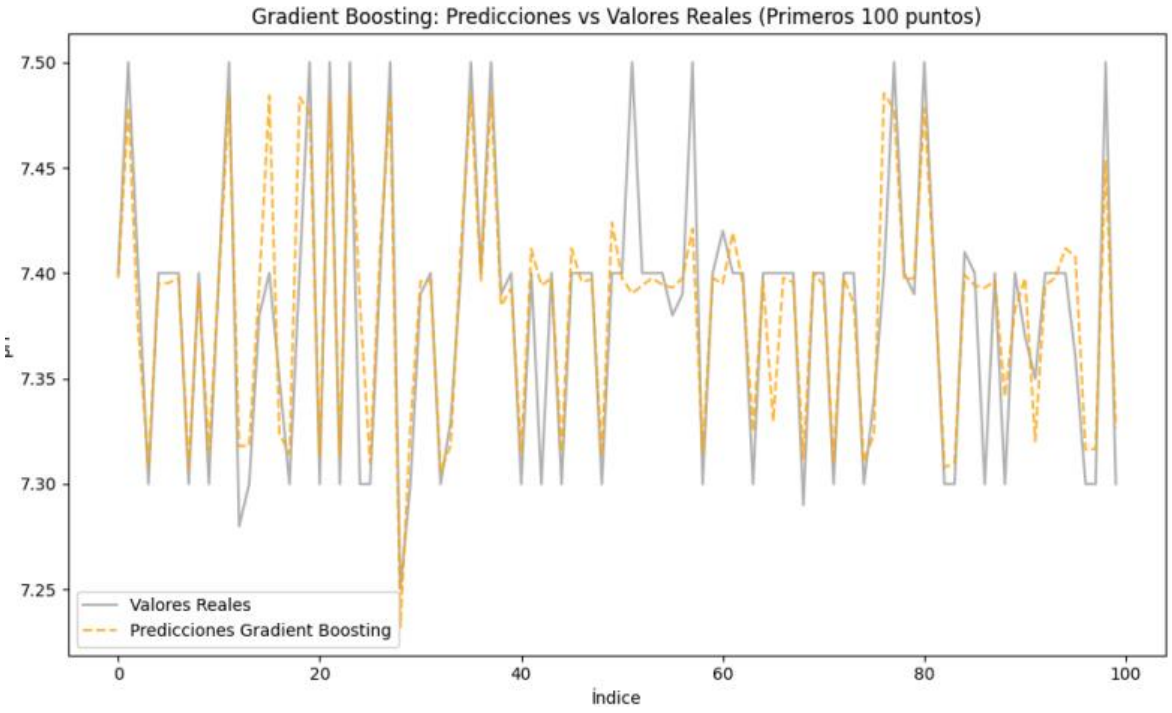
1. **Random Forest**: Como en el caso anterior muestra un rendimiento superior con un MSE más bajo, lo que indica que captura mejor las complejidades del sistema. Su capacidad para manejar datos ruidosos y no lineales le permite sobresalir en comparación con otros modelos. Este modelo es robusto y flexible, funcionando bien incluso en situaciones donde los datos pueden ser variados o ruidosos. Es el modelo más recomendable para problemas similares en este contexto.

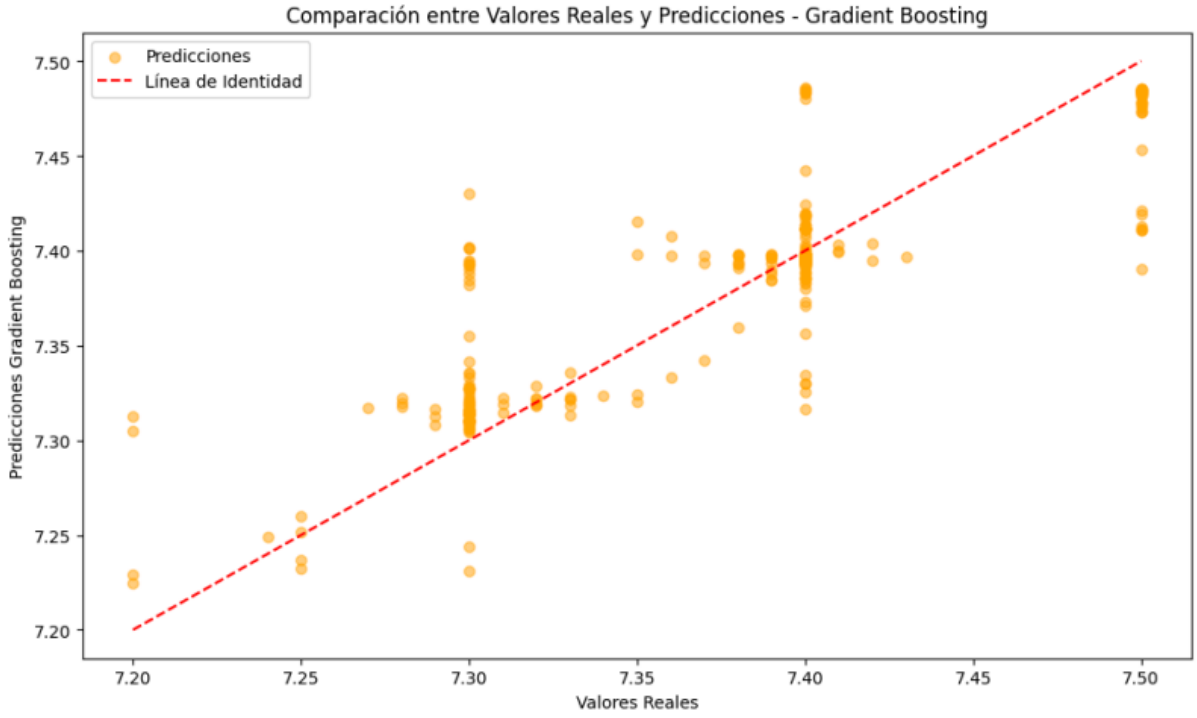




Con esta representación gráfica vemos que el favorito es Random Forest, el cual demuestra ser un modelo robusto con una buena capacidad para seguir la tendencia general de los datos y predecir con precisión en la mayoría de los casos. Las ligeras discrepancias en los extremos son mínimas en comparación con su rendimiento general.

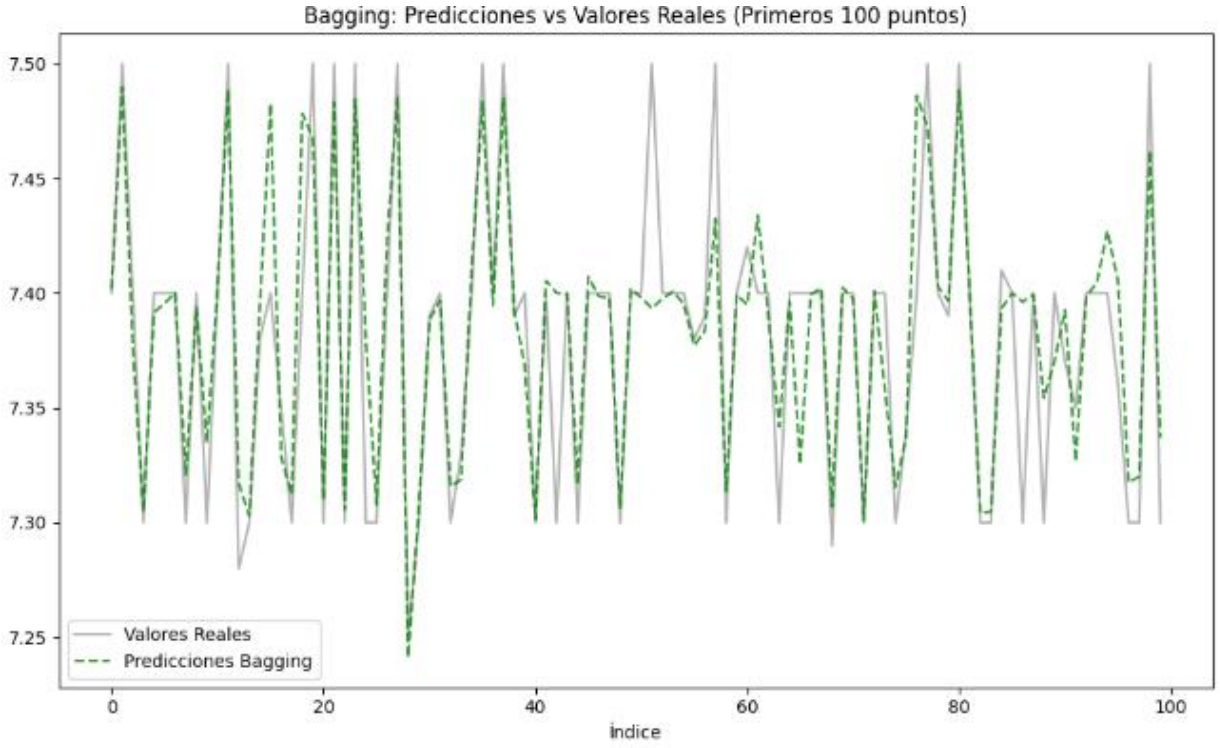
2. **Gradient Boosting**: sabemos que a pesar de que la muestra un MSE similar al de Random Forest, Gradient Boosting es menos estable y más susceptible a sobreajustes si no se configura cuidadosamente. Su proceso iterativo puede capturar más complejidades, pero también puede exagerar patrones en los datos. Es útil en situaciones donde se requiere un ajuste fino y los datos tienen muchas interacciones sutiles, pero requiere más atención en su configuración para evitar sobreajuste.

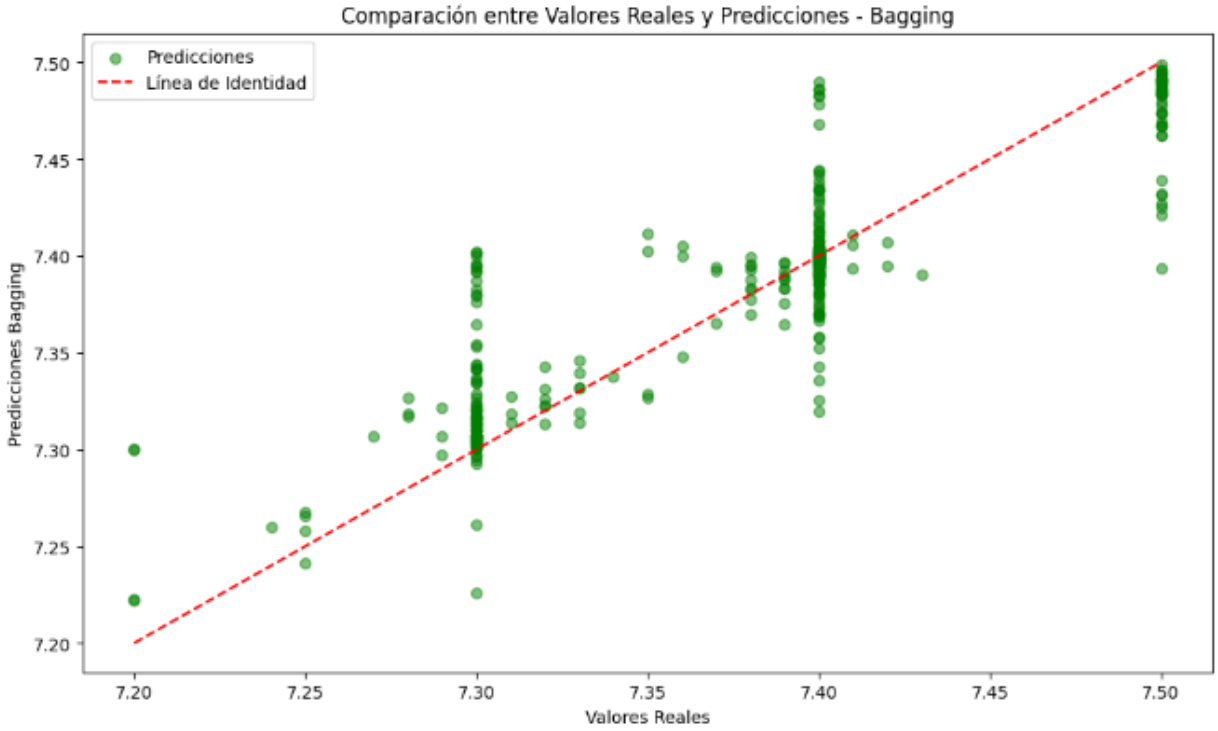




Observamos en las gráficas correspondientes que Gradient Boosting es competente para capturar la tendencia general, pero su tendencia a suavizar los resultados y la mayor dispersión en los valores extremos indica que es menos preciso que Random Forest, especialmente en situaciones fuera del rango medio de pH.

3. **Bagging**: mejora sobre la regresión lineal y SVM, pero no alcanza el nivel de precisión de Random Forest o Gradient Boosting. Su enfoque en la reducción de la varianza a través de múltiples modelos no es suficiente para superar los modelos basados en árboles más sofisticados.





Las gráficas anteriores de Bagging muestran un buen ajuste a los valores reales, pero con más variabilidad en comparación con Random Forest. Las predicciones son precisas en general, pero presentan más desviaciones en algunos puntos en comparación con los otros modelos

Conclusión de los modelos:

Random Forest combina la robustez de múltiples árboles de decisión, lo que le permite manejar bien el ruido y las variaciones en los datos. Esto es especialmente útil en un sistema complejo como la predicción de pH, donde las relaciones no lineales y el ruido son comunes. A diferencia de otros modelos, Random Forest generaliza bien sin sobreajustarse a los datos de entrenamiento, lo que se refleja en su MSE más bajo y su estabilidad en predicciones en todo el rango de pH. La importancia de las variables esta mejor distribuida, lo que le permite captar la contribución de múltiples factores en el comportamiento del pH, a diferencia de otros modelos que podrían depender demasiado de unas pocas variables clave.

Fase 4: Integración y Pruebas del Sistema

En esta fase, se decidió evaluar la capacidad de los modelos entrenados para ser aplicados en dispositivos distintos a aquellos con los que fueron entrenados originalmente. A pesar de que se reconoce que cada piscina tiene sus propias condiciones y factores ambientales que influyen significativamente en el funcionamiento del sistema, se exploró la viabilidad de usar un modelo entrenado en un dispositivo Device 07 y aplicarlo a otro Device 03. Posteriormente, se combinó el modelo entrenado en Device 01 con el de Device 07 para evaluar si un modelo integrado podría ser eficaz en el mantenimiento general de cualquier piscina, al menos en sus primeros ciclos de uso, con la idea de adaptarlo en el futuro según los datos recogidos durante su funcionamiento.

4.1 Pruebas:

Implementación y Pasos Seguidos

Carga y Preparación de Datos del modelo previamente entrenado y guardado del Device 07 utilizando la técnica Random Forest. Se cargaron los datos del Device 03 y se prepararon de manera similar a como se realizó con el Device 07, asegurándose de mantener la coherencia en las características utilizadas (ORP, Salinidad, Temperatura, valores rezagados de pH y ORP, etc.). Se repitió con el Device 06 y un conjunto de modelos.

Dev.	pH		
	Min	Avg	Max
1	7.14	7.52	7.89
7	7.2	7.38	7.51
3	7.15	7.35	7.53
6	7.2	7.35	7.55

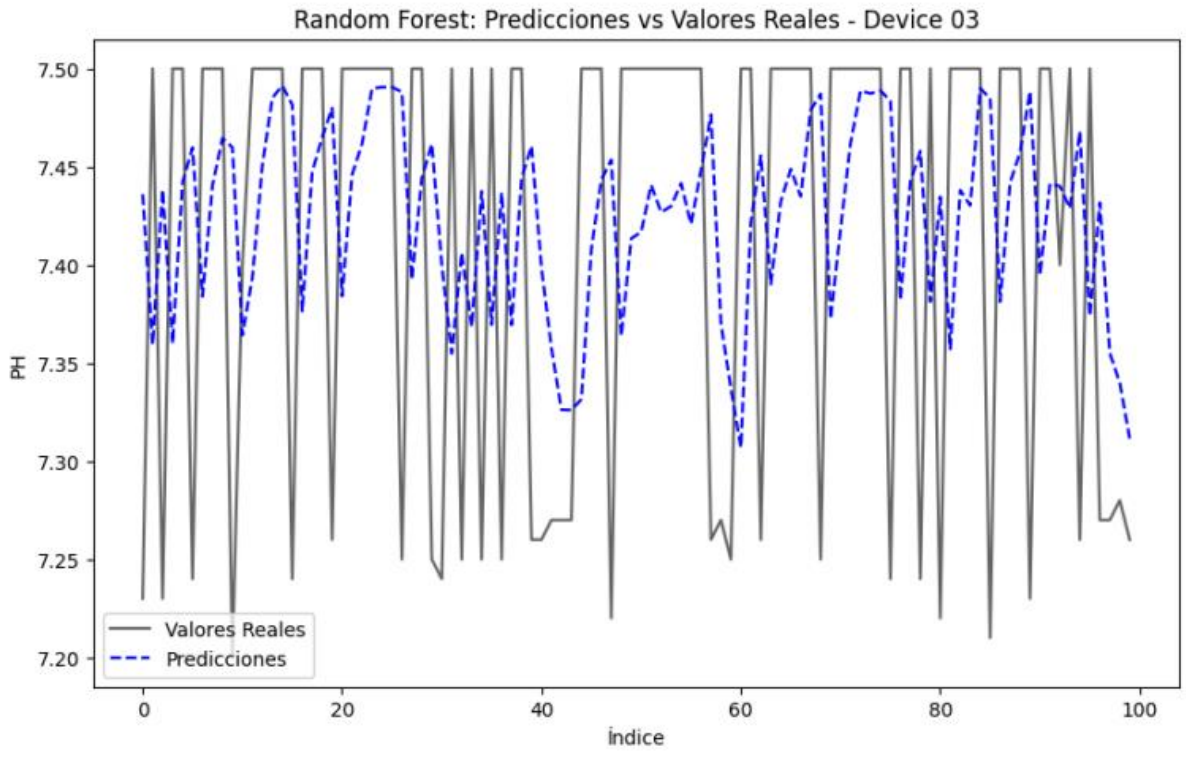
Evaluación de Resultados:

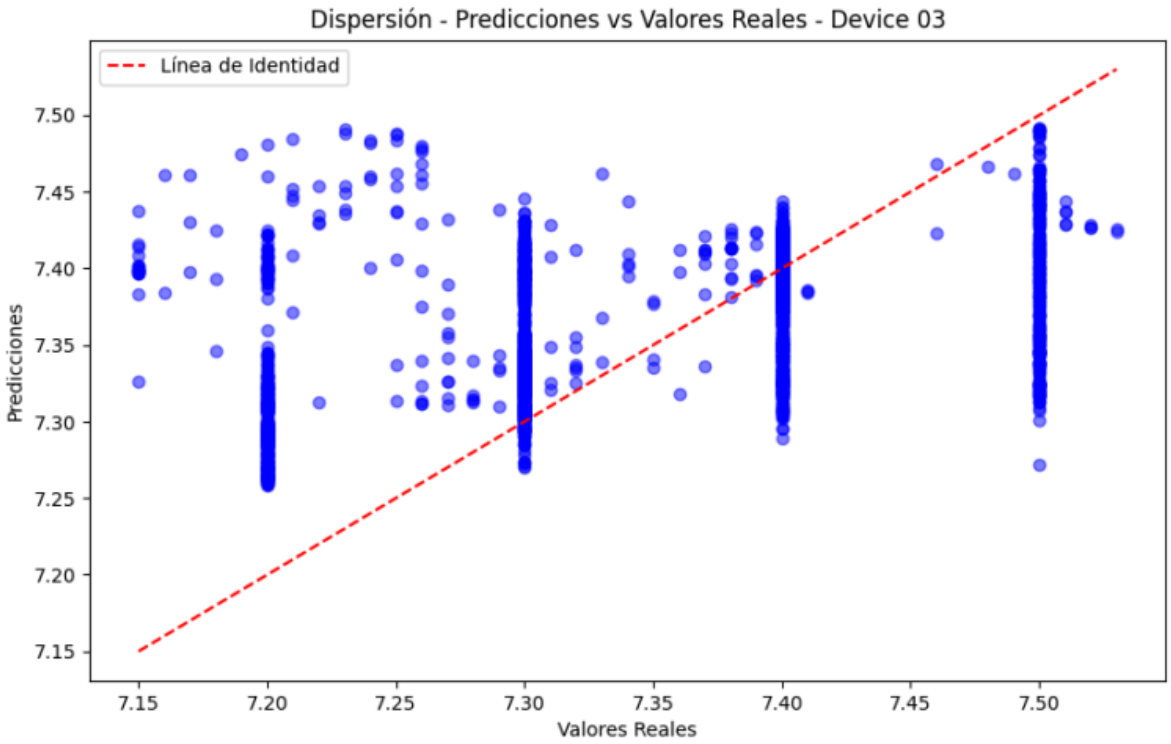
Modelo Entrenado 07 – Device 03

Métricas Evaluadas

1. **MSE: 0.0045**. El Error cuadrático medio obtenido es relativamente alto si consideramos el rango estrecho en el que el pH típicamente varía en piscinas bien mantenidas (7.2 - 7.8). Un MSE como este nos sugiere que el modelo comete errores significativos en sus predicciones, lo cual puede ser problemático en aplicaciones donde incluso pequeños errores pueden tener un impacto considerable.
2. **MAE: 0.0428**. Aunque pensemos que este Error Absoluto puede parecer pequeño, es significativo en el contexto de control de pH, donde la precisión es crítica para mantener la calidad del agua.
3. **R^2: 0.3155**. Vemos que el modelo solo puede explicar alrededor del 31.55% de la variabilidad en los valores reales de pH, lo cual es insuficiente. Un valor del Coeficiente de Determinación bajo como este sugiere que el modelo falla en capturar las relaciones complejas entre las variables predictoras y la variable objetivo - pH, especialmente cuando se transfiere a un nuevo dispositivo.

Análisis de las Gráficas





Las gráficas de dispersión y comparación temporal ofrecen información complementaria sobre el rendimiento del modelo:

- **Dispersión**: La dispersión significativa de los puntos alrededor de la línea de identidad en la gráfica de dispersión indica que el modelo tiene dificultades para predecir con precisión los valores extremos de pH. Existe un sesgo donde el modelo tiende a sobreestimar los valores más bajos y subestimar los valores más altos, lo que sugiere un sobreajuste a los valores medianos.
- **Comparación Temporal**: La gráfica temporal muestra que, aunque las predicciones siguen la tendencia general, no capturan adecuadamente los cambios abruptos o picos en el pH. Esto sugiere que el modelo no es lo suficientemente dinámico para ajustarse a las fluctuaciones rápidas que pueden ocurrir en el sistema.

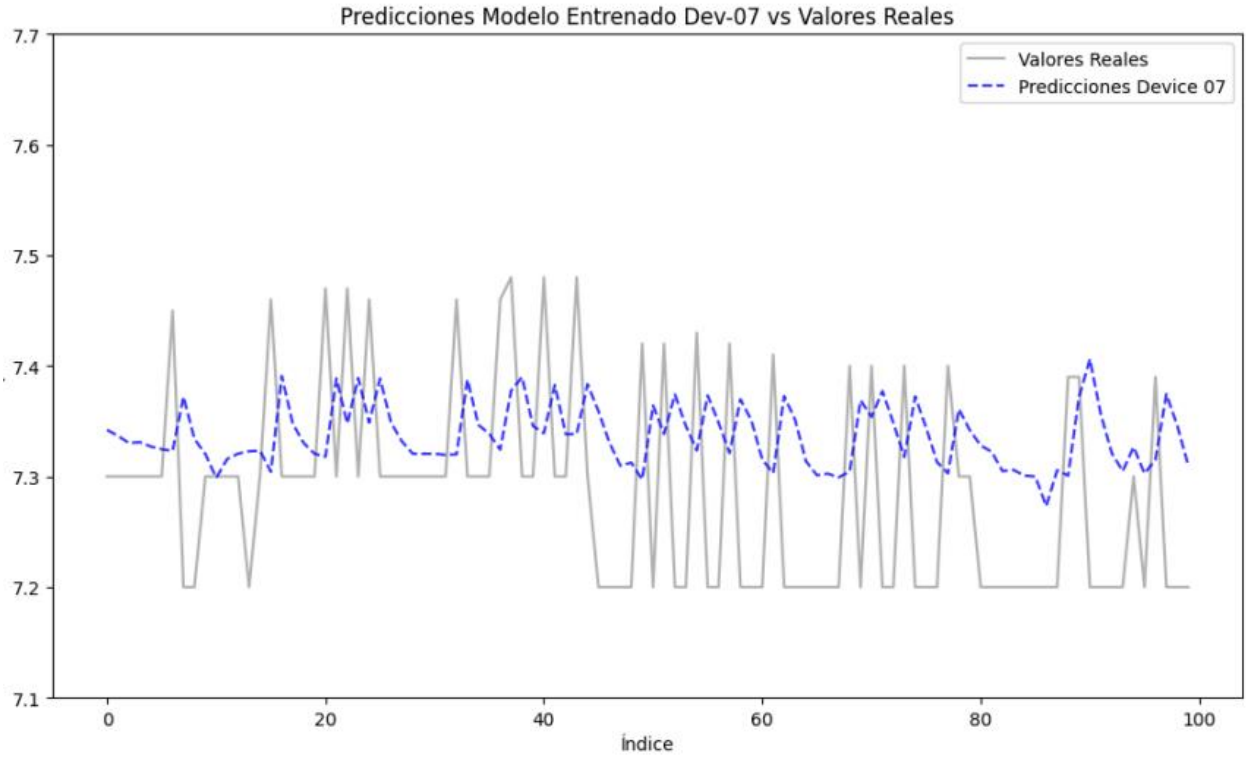
Conclusión.

El análisis de las métricas y las gráficas nos lleva a la conclusión de que el modelo de Random Forest entrenado en el Device 07 no generaliza bien cuando se aplica al Device 03. A pesar de que el modelo es capaz de seguir la tendencia general del pH, la falta de precisión en los valores extremos y la baja capacidad explicativa reflejada en el R^2 nos sugieren que el modelo está insuficientemente adaptado a las particularidades de cada dispositivo.

Combinación de Modelos Entrenados 01 & 07 – Device 06

Modelo 07

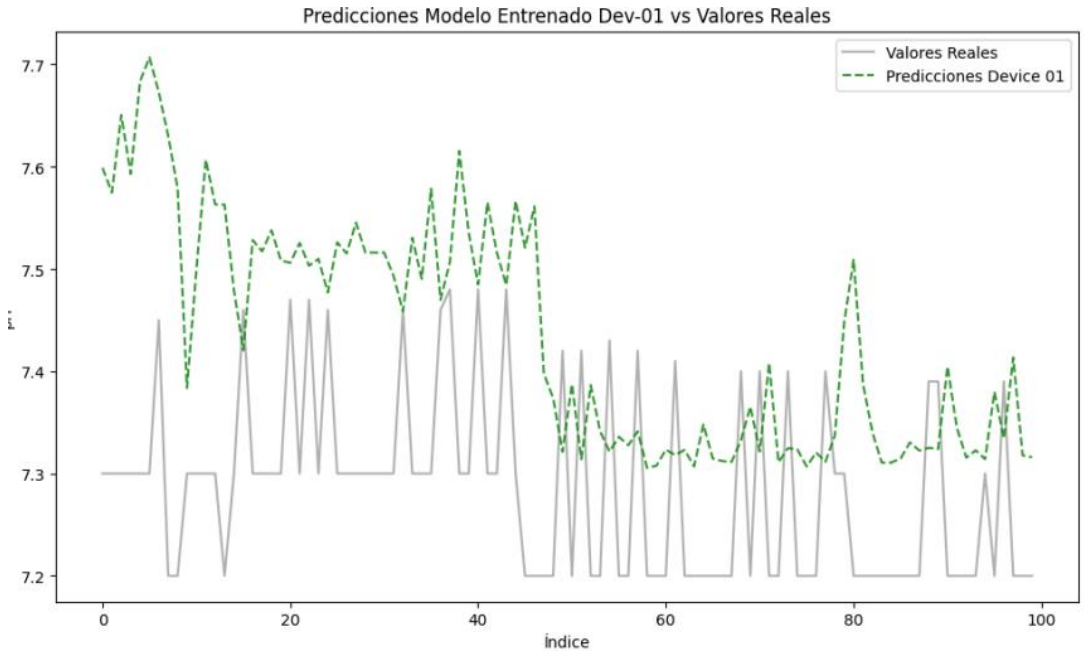
- **MSE: 0.0076, R^2: 0.1196**
    - El modelo entrenado en el Device 07 tiene un rendimiento aceptable, pero lejos de ser ideal. El R^2 bajo indica que no captura bien la variabilidad del pH en el Device 06 seguramente por el modelo 07 sus valores del pH siempre estaban acotados en valores de entre 7.2-7.4 y no fue entrenado para valores más allá.
    - La gráfica temporal muestra que el modelo sigue algunas tendencias generales, pero con notables discrepancias en los picos. La gráfica de dispersión muestra cierta alineación con la línea de identidad, aunque con variabilidad, indicando errores de predicción.



Modelo 01

- **MSE: 0.0381, R^2: -3.4098**
    - Este modelo muestra un rendimiento muy pobre cuando se aplica al Device 06, con un R^2 negativo que indica que es peor que un promedio simple.
    - La gráfica temporal revela que el modelo no sigue adecuadamente los valores reales, con desviaciones significativas. La gráfica de dispersión muestra una gran dispersión de puntos lejos de la línea de identidad, lo que subraya los errores graves de

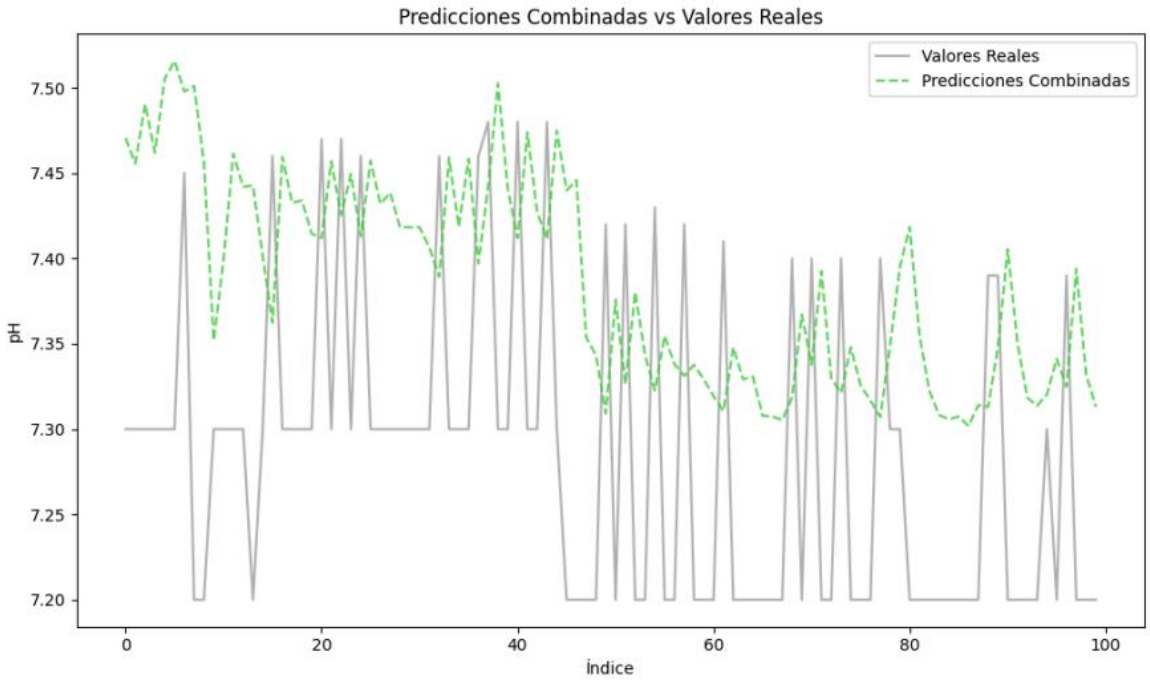
predicción. A primera vista encontramos un desfase en la predicción de valores, posiblemente por la diferencia de pH con el que el modelo fue entrenado.



Modelo Combinado

- **MSE: 0.0156, R^2: -0.8053**
    - **Desempeño**: La combinación de los dos modelos no mejora el rendimiento, sino que lo empeora. El R^2 negativo y el MSE elevado indican un fracaso en la predicción.
    - **Gráficas**: La gráfica de comparación temporal muestra que las predicciones del modelo combinado no siguen de cerca los valores reales de pH. Existe una notable falta de coincidencia entre las líneas de predicciones y los valores reales, lo que indica que el modelo combinado tiene dificultades para capturar con precisión las fluctuaciones del pH. Las predicciones parecen estar suavizadas, probablemente debido al promediado de las predicciones de los dos modelos individuales (Device 01 y Device 07). Este proceso de suavización ha llevado a un rendimiento general menos dinámico y menos preciso, reflejando una incapacidad del modelo combinado para adaptarse a las variaciones rápidas y específicas del pH en el Device 06.

      La dispersión significativa de los puntos en la gráfica, lejos de la línea de identidad (ideal), subraya la falta de ajuste del modelo. La gran dispersión y la falta de alineación con los valores reales sugieren que el modelo no solo falla en mejorar las predicciones, sino que en realidad puede estar empeorando el rendimiento, como lo demuestra el R^2 negativo. Esto indica que el modelo combinado no es efectivo para generalizar a las condiciones del nuevo dispositivo, fallando en su objetivo de proporcionar predicciones precisas.



Conclusión de las Pruebas de Integracion:

- **Robustez**: Aunque Random Forest es robusto y puede manejar datos ruidosos y relaciones no lineales, su aplicación directa a dispositivos diferentes sin reentrenamiento o ajustes específicos resulta en una pérdida significativa de precisión.
- **Necesidad de Personalización**: Los resultados subrayan la importancia de personalizar o ajustar los modelos para cada dispositivo específico. Un enfoque personalizado o un proceso iterativo que permita adaptar el modelo a nuevos datos en tiempo real sería más efectivo.
- **Individualización de Modelos**: Los resultados refuerzan la idea de que cada dispositivo o piscina puede requerir su propio modelo entrenado y optimizado para sus condiciones específicas. Un enfoque de "talla única" no parece ser adecuado para este tipo de problemas.
- **Combinación de Modelos**: Vemos que solo deberíamos considerarla si los modelos individuales hubieran tenido un rendimiento compatible y complementario. En este caso, los modelos entrenados en condiciones muy distintas no se beneficiaron de la combinación, sino que resultaron en un desempeño peor.

En resumen, si bien el modelo muestra cierta capacidad de predicción, su falta de precisión en un nuevo dispositivo indica que es crucial adaptar o reentrenar los modelos para cada dispositivo, especialmente si se espera mantener la precisión en aplicaciones críticas como el control del pH en piscinas.

Conclusiones

Fase 1: Recopilación de Datos

Durante la primera fase, nos enfocamos en la recopilación de datos provenientes de 10 dispositivos diferentes instalados en piscinas, obteniendo información clave sobre parámetros como el pH, ORP, salinidad y temperatura. Si bien logramos una base de datos amplia y representativa, observamos variaciones significativas entre dispositivos. Estas diferencias se deben a las diversas condiciones ambientales y operativas de cada piscina, que influyen directamente en la calidad y consistencia de los datos recolectados.

Fase 2: Limpieza y Preprocesamiento de Datos

La limpieza y el preprocesamiento de datos representaron una etapa crucial y desafiante del proyecto. Durante esta fase, nos enfrentamos a varios problemas significativos:

1. **Valores Atípicos y Erróneos**: Implementamos un proceso de detección y eliminación de outliers utilizando el método del Rango Intercuartílico (IQR). Esto nos permitió identificar y excluir valores extremos que podrían distorsionar nuestros modelos predictivos. Sin embargo, decidimos no establecer límites excesivamente estrictos para evitar la exclusión de datos válidos, considerando la singularidad de cada piscina.
2. **Desfase Temporal en el Registro de Datos**: Se detectó un desfase temporal en los registros, especialmente en los dispositivos más antiguos, lo que complicó la alineación de los datos y dificultó la identificación de patrones claros. Optamos por centrarnos en tendencias generales, lo que nos permitió mitigar el impacto de este problema sin comprometer la calidad del análisis.
3. **Exclusión de Datos de Inactividad**: Observamos que los datos recolectados durante los períodos de inactividad de las bombas no eran representativos debido a que el mismo agua permanencia en la célula y cambiaba las condiciones operativas reales de la piscina, ya que durante estos períodos los ciclos de funcionamiento de la piscina, incluyendo la circulación del agua por nuestro sistema, estaban detenidos o funcionando a una capacidad mínima. Estos datos, de haber sido incluidos, podrían haber introducido sesgos significativos en los modelos predictivos, especialmente en el contexto de la predicción del pH, que está directamente influenciado por el flujo constante y la calidad del agua, y que además no eran reales.

Esta etapa de limpieza y preprocesamiento fue esencial para garantizar que los datos utilizados en los modelos fueran de alta calidad, lo que a su vez mejoró la fiabilidad y precisión de las predicciones.

Fase 3: Desarrollo de Modelos de ML

En esta fase, optamos por desarrollar modelos personalizados para cada piscina debido a las condiciones particulares observadas en cada dispositivo y a la calidad variable de los datos recolectados. Cada piscina presenta un conjunto único de características ambientales y operativas, lo que significa que un enfoque generalizado no sería adecuado para capturar las complejidades de cada sistema. Esto nos llevó a elegir modelos que pudieran adaptarse a estas diferencias específicas y optimizar su rendimiento en función de las particularidades.

Evaluamos y comparamos varios modelos de ML, comenzando con una **Regresión Lineal** como línea base. Sin embargo, rápidamente nos dimos cuenta de que este enfoque era insuficiente para capturar las complejidades no lineales del sistema, presentando un rendimiento modesto en términos de MSE y R^2.

El modelo de **Random Forest** fue elegido como el más prometedor debido a su capacidad para manejar relaciones no lineales complejas y su robustez frente a datos ruidosos. Este modelo demostró ser eficaz en capturar la importancia de variables críticas como el ORP y los valores históricos de pH. Para optimizar su rendimiento, llevamos a cabo un ajuste de hiperparámetros utilizando técnicas como Grid Search y Random Search. Este proceso de ajuste fue clave para mejorar la precisión del modelo, reduciendo el MSE y mejorando el coeficiente de determinación.

También se concluye que otros modelos como **Gradient Boosting** y **AdaBoost**. Aunque conceptualmente similares presentaron un rendimiento ligeramente inferior. AdaBoost, en particular, mostró una sensibilidad al ruido que afectó su precisión en la predicción de casos extremos. Por otro lado, **Redes Neuronales** y **Support Vector Machine (SVM)** fueron considerados, pero mientras que las redes neuronales mostraron potencial, también evidenciaron un sobreajuste. SVM, sin embargo, no logró capturar las complejidades no lineales necesarias para este tipo de predicción, por esto ambos fueron descartados tras obtener los peores resultados.

Concluimos que la personalización de los modelos era fundamental para maximizar su rendimiento, ya que las diferencias significativas entre las piscinas requerían soluciones adaptadas a las condiciones específicas de cada dispositivo, pero que podía ser buena practica demostrar probando los modelos entrenados en los datos de otros dispositivos.

Fase 4: Integración y Pruebas del Sistema

Durante la evaluación pudimos observar que los modelos entrenados en un dispositivo **no generalizaban bien** cuando se aplicaban a otros, a menos que se realizaran ajustes significativos. Esto reforzó a un más la idea de que condiciones operativas y ambientales de cada piscina son muy determinantes. Aunque Random Forest demostraba ser robusto, su aplicación directa sin un reentrenamiento o ajuste específico resultó en una pérdida notable de precisión.

Además, también intentamos combinar modelos entrenados en diferentes dispositivos, pero esta estrategia no mejoró el rendimiento, sino que empeoró. La combinación de modelos solo debería considerarse si su rendimiento es compatible y complementario.

Posibles Mejoras del Sistema:

1. **Mejora en la Recolección de Datos**: Se debería poner especial énfasis en corregir la manera de registrar los datos donde por ejemplo el registro sea por tiempo y que capture valores nulos en los sensores cuando un dato no se lee correctamente. Esto permitiría un tratamiento más efectivo de los datos faltantes mediante técnicas de imputación como la media o la mediana, asegurando que el análisis no se vea comprometido por la falta de datos.
2. **Ajuste Dinámico de Modelos**: Posible desarrollo de un sistema que permita el ajuste continuo de los modelos basado en los datos nuevos recolectados durante la operación. Esto mejoraría la capacidad del sistema para adaptarse a las condiciones cambiantes de las piscinas, reduciendo el riesgo de sobreajuste y aumentando la precisión de las predicciones. Si el modelo sigue aprendiendo su precisión y adaptabilidad a condiciones extremas mejoraría con consistencia.
3. **Expansión del Conjunto de Características**: La inclusión de nuevas variables en el análisis, como el clima, el uso activo y otros factores ambientales que podrían influir en el pH, podría proporcionar un modelo más completo y preciso.
4. **Sistema Modular y Escalable**: Es posible el desarrollo de un sistema predictivo modular que pueda integrarse fácilmente en diferentes dispositivos y adaptarse a las necesidades específicas de cada piscina sin requerir un rediseño completo del modelo, sino que necesitaría un continuo aprendizaje pero comenzaría con una base sólida.
5. **Pruebas de Largo Plazo con Datos Nuevos**: Se va a continuar recolectando datos y realizar pruebas con datos completamente nuevos de temporadas futuras de los mismos devices entrenados 01 y 07. Esto permitirá evaluar la durabilidad y adaptabilidad del sistema predictivo a lo largo del tiempo, ayudando a identificar posibles casos de sobreaprendizaje y permitiendo ajustes antes de implementar el sistema a gran escala. Además colectar datos o usar mas periodos de tiempo harán que el modelo sea capaz de ser más preciso.

Referencias:

[1] OMS, 2006. [En línea]. Available: https://apps.who.int/iris/handle/10665/43336.

[2] DWP, «https://daveywater.com/au,» 2024. [En línea].

[3] chloromatic-lifeguard, 2024. [En línea]. Available: https://daveywater.com/au/product/chloromatic-lifeguard/.

[4] G. Cloud, «What is big data?,» [En línea]. Available: https://cloud.google.com/learn/what-is-big-data.

[5] Shiksha, «Five V's of Big Data,» 2023. [En línea]. Available: https://www.shiksha.com/online-courses/articles/five-vs-of-big-data/.

[6] SAS, «History and evolution of big data analytics,» [En línea]. Available: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

[7] EIP, *Manual "Fundamentos de IA y Machine Learning", Lección 1,* 2023.

[8] EIP, *Manual "Regresión Lineal, Regresión Logística y kNN"*.

[9] EIP, *Manual "Redes Neuronales Artificiales"*.

[10] EIP, *Manual "Ajuste de hiperparámetros y flujos de trabajo en Machine Learning",* 2023.

[11] EIP, *Manual "Técnicas y métricas de evaluación de modelos de Machine Learning",* 2023.