UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

FACULTAD DE INGENIERÍA

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

DETECCIÓN DE ANOMALÍAS EN SERIES DE TIEMPO DE DATOS DEL AGUA UTILIZANDO UN ENFOQUE DE APRENDIZAJE PROFUNDO

POR:

SERGIO ALBERTO VALDÉS RABELO

TESIS PRESENTADA COMO REQUISITO PARA OBTENER EL GRADO DE MAESTRO EN INGENIERÍA EN COMPUTACIÓN

CHIHUAHUA, CHIH., MÉXICO **AGOSTO DE 2020**

Detección de Anomalías en series de tiempo de datos del agua utilizando un enfoque de Aprendizaje Profundo. Tesis presentada por Sergio Alberto Valdés Rabelo como requisito parcial para obtener el grado de Maestro en Ingeniería en Computación, ha sido aprobada y aceptada por:

M.I. Javier González Cantú
Director de la Facultad de Ingeniería

Dr. Alejandro Villalobos Aragón
Secretario de Investigación y Posgrado

M.S.I. Karina Rocío Requena Yáñez
Coordinador(a) Académico

Dr. Luis Carlos González Gurrola
Director(a) de Tesis

Septiembre 2020
Fecha

Comité:

Dr. Luis Carlos González Gurrola
Dr. Miguel Angel Medina Perez
Dr. Alain Manzo Martínez
MI. Jesús Roberto López Santillán

Sergio Alberto Valdés Rabelo
Villa de Antares #9750. Villas del Sol I Etapa, CP: 31124, Chihuahua, Chiahuahua, México.

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

Ing. Sergio Alberto Valdés Rabelo
Presente

En atención a su solicitud relativa al trabajo de tesis para obtener el grado de Maestría en Ingeniería en Computación, nos es grato transcribirle el tema aprobado por esta Dirección, propuesto y dirigido por el director Dr. Luis Carlos González Gurrola para que lo desarrolle como tesis, con el título: **"DETECCIÓN DE ANOMALÍAS EN SERIES DE TIEMPO DE DATOS DEL AGUA UTILIZANDO UN ENFOQUE DE APRENDIZAJE PROFUNDO".**

Índice de Contenido

FACULTAD DE INGENIERÍA
Circuito No.1, Campus Universitario 2
Chihuahua, Chih., México. C.P. 31125
Tel. (614) 442-95-00
www.fing.uach.mx

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

Solicitamos a Usted tomar nota de que el título del trabajo se imprima en lugar visible de los ejemplares de las tesis.

A T E N T A M E N T E

"Naturam subiecit aliis"

EL DIRECTOR

M.I. JAVIER GONZÁLEZ CANTÚ

FACULTAD DE INGENIERÍA U.A.CH. DIRECCIÓN

EL SECRETARIO DE INVESTIGACIÓN Y POSGRADO

DR. ALEJANDRO VILLALOBOS ARAGÓN

FACULTAD DE INGENIERÍA
Circuito No.1, Campus Universitario 2
Chihuahua, Chih., México. C.P. 31125
Tel. (614) 442-95-00
www.fing.uach.mx

DEDICATORIA

A mi familia que son el parte del apoyo fundamental que me permitieron cumplir este sueño, a mi esposa Yaima Roquero Figueroa como pilar imprescindible para todas las metas que me trazo pues sin ella no lo hubiese logrado, a mi hijo Sergio Manuel Valdés Roquero la luz de la alegría y la ocurrencia en nuestra casa.

AGRADECIMIENTOS

Un agradecimiento muy especial para M.S.I. Karina Rocío Requena Yáñez, coordinadora de la maestría, por brindarme la oportunidad de haber cursado este posgrado. Gracias por su apoyo incondicional durante toda esta trayectoria.

Un sincero agradecimiento a mi tutor el Dr. Luis Carlos Gurrola González, quien se hizo responsable de mi plan de estudios y me supo guiar durante el trabajo de tesis. Gracias por enseñarme y brindarme sus conocimientos de manera incondicional.

Un sincero agradecimiento a mi cotutor el Dr. Miguel Ángel Medina Pérez, por su apoyo y guía durante el trabajo de tesis. Gracias por brindarme sus conocimientos de manera incondicional.

A todos los profesores que con sus clases apoyaron mi crecimiento como profesional durante este postgrado.

Al Consejo de la Ciencia y Tecnología (CONACYT) de México por su apoyo económico, sin este no creo haber podido solventar mi estancia en este maravilloso país.

A la Secretaría de Investigación y Posgrado de la Facultad de Ingeniería de la Universidad Autónoma de Chihuahua, por aceptarme siendo extranjero.

RESUMEN

Los suministros naturales de agua potable se consumen a un ritmo mucho más rápido de lo que se están recargando naturalmente [1]. Haciendo una búsqueda local, La Junta Municipal de Aguas y Saneamientos (JMAS) de Chihuahua nos provee acceso a dos conjuntos de datos: uno relacionado con el proceso de distribución de agua y otro con quejas de fugas y faltas de agua reportado por usuarios al Centro de Información y Servicios (*CIS)*. El objetivo de esta investigación es detectar y predecir comportamientos anómalos de las instalaciones de la red hidráulica de la JMAS de Chihuahua aplicando algoritmos de aprendizaje máquina (*ML por sus siglas en inglés)* haciendo uso de los datos proporcionados. Inicialmente se realizó un proceso de obtención, preprocesamiento y visualización de los datos con el apoyo de una herramienta web. Posteriormente, fue necesario el etiquetado de datos por un usuario experto de la JMAS, para esto se desarrolló un módulo de la herramienta web que permite etiquetar las anomalías en las instalaciones de la JMAS. Luego se realizaron un conjunto de experimentos con las técnicas de *ML* para el proceso de detección y predicción de anomalías. Posteriormente, se realiza un análisis de los datos del sistema *CIS* con el objetivo de relacionar esta información para que pueda ser utilizada en el proceso de detección y predicción de anomalías. Aunque se logra extraer información relevante de estos datos no se logra relacionar esta información con el proceso de detección de anomalías debido a que hay información específica que no está presente en los metadatos. Dentro de los resultados principales de esta investigación está la presentación de un conjunto de datos vírgenes que son proporcionados a la comunidad de *ML* para dar continuidad a futuras investigaciones, pues se crearon dos *datasets* uno con todos los datos de cada instalación de la red hidráulica de Chihuahua y el segundo con los datos de dos instalaciones para las cuales se etiquetaron anomalías durante el período de un año. Los experimentos realizados evidencian que, se puede predecir y detectar las anomalías en tiempo real en las instalaciones de la JMAS, mediante el uso de una *LSTM* como regresor y un clasificador del tipo DecisionTree usando ventanas de tiempo deslizantes. En el caso de la instalación Pozo Aeropuerto 3 se obtuvo un 98% de accuracy, 95% de precisión, 97% de recall y 95% de f1 score. Para la instalación Tanque Loma Larga se obtuvo 95% de accuracy, 81% de precisión, 70% de recall y 73% de f1 score. Las enfoques y técnicas de mejores resultados, fueron integrados en una herramienta web que es presentada como prototipo funcional para

predicción y detección de anomalías en tiempo real. Esta herramienta cuenta con los módulos de visualización de los datos, etiquetado de datos, entrenar modelos de *ML* para instalaciones con datos etiquetados y un módulo que le permite predecir comportamientos anómalos de las instalaciones de la red hidráulica una vez que se tienen los modelos de *ML* entrenados.

Palabras clave: detección de anomalías, series de tiempo, aprendizaje profundo, aprendizaje de máquina, datos del agua.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1: INTRODUCCIÓN

El agua es el bien más preciado que como humanidad tenemos, y administrarlo de la mejor manera nos permitirá ampliar su cobertura, calidad y disminuir su desperdicio. Además, es fundamental para el sostén y reproducción de la vida en el planeta, ya que constituye un componente indispensable para el desarrollo de los procesos biológicos que la hacen posible.

Aunque el agua es un recurso natural renovable, debido a la interrupción del ciclo hidrológico, el hecho más presente es que los suministros naturales de agua potable se consumen a un ritmo mucho más rápido de lo que se está llenando y recargando naturalmente. Se está haciendo cada vez más evidente en aguas superficiales y subterráneas, donde el consumo de agua excede significativamente la recarga natural [1].

La sociedad recurre al agua para generar y mantener el crecimiento económico y la prosperidad a través de actividades como la agricultura, la pesca comercial, la producción de energía, la industria, el transporte y el turismo. Es un elemento importante para nuestro propio bienestar donde exigimos un agua potable y limpia para la higiene y el saneamiento.

Debido a su importancia, es vital una adecuada gestión del recurso hídrico para lograr una distribución satisfactoria. El agua es suministrada a nuestros hogares por una Red de Distribución de Agua que es propiedad y está mantenida tanto por empresas privadas como por organismos o instancias gubernamentales.

Para la gestión de estas Redes de Distribución se utilizan sistemas de monitoreo que utilizan diferentes tipos de sensores a lo largo de esta red con el objetivo de captar en tiempo real el estado de la misma y apoyar en la toma de decisiones en los Centros de Monitoreo. Estos sistemas de monitoreo son capaces de guardar toda la información que se obtienen de estos sensores con el objetivo de mantener un histórico para poder ser analizados e inferir conocimiento a ser utilizado en procesos futuros. Sin embargo, la mayoría de las compañías de agua todavía carecen de capacidades avanzadas de predicción y reporte en tiempo real para monitorear los factores cambiantes [2].

Lamentablemente contar con una Red de distribución del Agua y con estos sistemas de monitoreo no es suficiente, pues alrededor de un tercio de las empresas de servicios de agua en todo el mundo reportan una pérdida del 40% de agua limpia debido a fugas en este

proceso de distribución [3]. Además, debido a la limitada disponibilidad de vías de agua, el uso intensivo de agua conduce a un deterioro de su calidad en la medida en que se vuelve prácticamente inutilizable para otros fines [1].

Los avances tecnológicos han permitido la producción de grandes volúmenes de datos asociados con el funcionamiento de estos sectores. Y se está comenzando a ver que las técnicas de Estadísticas y *Machine Learning(ML)* pueden ayudar a dilucidar patrones en estos datos, desde la disponibilidad de agua, el transporte, el uso, el suministro de combustible y la demanda de los clientes [4].

El crecimiento ubicuo de los datos proporciona valor agregado en casi todos los dominios de la sociedad y el dominio del agua no es una excepción. Desde prevenir desastres provocados por el hombre, como desbordamientos de ríos, hasta inundaciones naturales, además aumentar la conciencia pública sobre la conservación del agua y minimizar los impactos de la sequía en regiones áridas. Finalmente, simplemente ahorrar costos al mejorar la confiabilidad de las obras públicas de agua de una ciudad tiene un valor sumamente importante. Todos estos objetivos son posibles con el uso efectivo de las tecnologías de Big Data[1] [5].

En la era de Big Data, las series de tiempo se generan en cantidades masivas. Hoy en día, los sensores y los dispositivos de Internet de las cosas (IoT) son omnipresentes y producen datos continuamente. Si bien los datos recopilados por estos dispositivos son valiosos y pueden proporcionar información significativa, existe una creciente necesidad de desarrollar algoritmos que puedan procesar estos datos de manera eficiente [6].

El progreso realizado en *ML* y en particular *Aprendizaje Profundo (DL por sus siglas en inglés)*, permite construir modelos que aprenden directamente de los datos sin un procesamiento previo extenso y un conocimiento significativo del dominio por parte de expertos en el campo de la aplicación [6].

---

[1] Big data es un término que describe el gran volumen de datos – estructurados y no estructurados – es un campo que trata formas de analizar, extraer información de forma sistemática o tratar con conjuntos de datos que son demasiado grandes o complejos para ser tratados por el software de aplicación de procesamiento de datos tradicional.

La detección de anomalías se refiere al problema de encontrar patrones en datos que no se ajustan al comportamiento esperado. Estos patrones no conformes a menudo se denominan anomalías, valores atípicos, observaciones discordantes, excepciones, aberraciones, sorpresas, peculiaridades o contaminantes en diferentes dominios de aplicación [7].

Detectar cuando un proceso variable se desvía de su patrón normal puede resultar un indicador valioso, si esta detección se puede hacer con una frecuencia razonablemente baja de falsos positivos. Una técnica útil para encontrar estas desviaciones es aplicar la ***Detección de Anomalías (AD por sus siglas en inglés)***. La detección de anomalías es el concepto de encontrar patrones desviados en los datos y se aplica en una gran variedad de dominios, por ejemplo, detectar fraudes con tarjetas de crédito, intrusiones en la red y mal funcionamiento del sensor. Independientemente del dominio de la aplicación, la capacidad de encontrar anomalías ha demostrado ser fructífera ya que las anomalías contienen información valiosa [8].

La definición de una anomalía puede variar con el dominio de la aplicación y, a veces, incluso dentro del dominio de la aplicación. Esto dificulta la implementación de técnicas de detección de anomalías. Otro problema es que las anomalías en general son eventos raros. Esto significa que crear un conjunto de datos con etiquetas de comportamiento normal y anormal que sea lo suficientemente grande como para capturar tanto las anomalías como el comportamiento normal no solo es difícil, sino que también consume mucho tiempo [8].

Un enfoque más razonable es extraer la noción de normalidad de los valores históricos, lo que resulta en un algoritmo de detección de anomalías usando un enfoque **Semi-Supervisado**. Además, para que la detección de anomalías funcione como una ayuda de monitoreo, debe ser capaz de detectar anomalías de series temporales de transmisión, es decir, ejecutarse en un entorno en línea [8].

Debido a la importancia que tiene mejorar todo el proceso de Distribución del Agua y a la cantidad y tipo de información que se tiene actualmente de sistemas de monitoreo. Se han realizado varios estudios que emplean algoritmos de *ML* con el objetivo de mejorar el proceso de distribución del agua, así como para afrontar los principales problemas que presenta esta área.

1.1. ANTECEDENTES

A continuación, se presenta un conjunto de trabajos relacionados con la detección de anomalías en series de tiempo.

La detección de anomalías es un viejo problema que se ha abordado utilizando diferentes enfoques a lo largo del tiempo [7], y ha ganado mucha atención de investigación en dominios de aplicaciones que involucran grandes conjuntos de datos adquiridos de sistemas críticos.

En 2014 se publica el trabajo citado en [9] donde se proporciona una visión general actualizada y estructurada de estudios y enfoques para la detección de anomalías que han aparecido en la literatura de *Machine Learning* y procesamiento de señales realizados en la década anterior a esa fecha. Además, se aporta una visión general en profundidad sobre los enfoques de detección de anomalías que se habían propuesto hasta ese momento y se provee una clasificación de las técnicas de detección de anomalías en las siguientes cinco categorías generales:

I. **Técnicas probabilísticas:** utiliza métodos probabilísticos que a menudo implican una estimación de densidad de la clase "normal". Estos métodos suponen que las áreas de baja densidad en el conjunto de entrenamiento indican que estas áreas tienen una baja probabilidad de contener objetos "normales".
II. **Técnicas basadas en la distancia:** incluye los conceptos de análisis de agrupación y vecino más cercano que también se han utilizado en problemas de clasificación. La suposición aquí es que los datos "normales" están estrechamente agrupados, mientras que los datos novedosos se producen lejos de sus vecinos más cercanos.
III. **Técnicas basadas en la reconstrucción:** implica entrenar un modelo de regresión utilizando el conjunto de entrenamiento. Cuando los datos anormales se mapean utilizando el modelo entrenado, el error de reconstrucción entre el objetivo de regresión y el valor real observado da lugar a un alto puntaje de novedad. Las redes neuronales, por ejemplo, pueden usarse de esta manera y pueden ofrecer cualquiera de las mismas ventajas para la detección de novedades que para los problemas de clasificación regulares.

IV. **Técnicas basadas en el dominio:** utiliza métodos de dominio para caracterizar los datos de entrenamiento. Estos métodos suelen tratar de describir un dominio que contiene datos "normales" definiendo un límite alrededor de la clase "normal" de modo que siga la distribución de los datos, pero no proporciona explícitamente una distribución en regiones de alta densidad.

V. **Técnicas informativas:** calcula el contenido de la información en los datos del entrenamiento utilizando medidas teóricas de la información, como la entropía o la complejidad de *Kolmogorov*. La hipótesis principal aquí es que los datos novedosos alteran significativamente el contenido de la información en un conjunto de datos.

Es publicado en el 2006 el trabajo presentado en [10], en el cual se utiliza un método de búsqueda de proyección para desarrollar un procedimiento para detectar valores atípicos en una serie de tiempo multivariada. Además, se muestra que la prueba de valores atípicos en algunas direcciones de proyección puede ser más poderosa que probar la serie multivariada directamente. Las direcciones óptimas para detectar anomalías se encuentran mediante la optimización numérica del coeficiente de *kurtosis* de la serie proyectada.

En 2007 se publica el artículo descrito en [11], el cual expresa que en el análisis de series de tiempo, se supone que los datos observados se pueden modelar como derivados de una serie de regímenes de dinámica, por ejemplo un *Switching Kalman Filter (SKF)*. En el cual se aplica un método a los datos de monitoreo fisiológico de bebés prematuros que reciben cuidados intensivos, y demuestran que el modelo es efectivo para detectar secuencias anormales de observaciones que no están modeladas por los regímenes conocidos.

Luego en 2009 se presenta el trabajo citado en [12], aquí se evalúan algoritmos de aprendizaje competitivo en la tarea de identificar patrones anómalos en datos de series de tiempo. La metodología empleada en este trabajo consiste en calcular los umbrales de decisión a partir de la distribución de los errores de cuantificación producidos por los datos normales de entrenamiento, estos umbrales se utilizan para clasificar las muestras de datos entrantes como normales o anormales. Para este propósito, se llevan a cabo comparaciones de rendimiento entre cinco redes neuronales competitivas (*Self-Organizing Map SOM*, Kangas Model, *Temporal Kohonen Map*, *Recurrent SOM* y Fuzzy ART) en datos de series de tiempo simuladas y del mundo real. Los resultados obtenidos en este trabajo indican que

las variantes temporales del *SOM* son más adecuadas para manejar datos de series temporales que las redes neuronales competitivas estáticas.

En 2011 se publica la investigación descrita en [13], en la cual se presenta el uso de *Support Vector Machine (SVM)* para la detección de anomalías a partir de datos de series de tiempo de presión y flujo de agua de *Water Distributions Systems (WDS)*. Las conclusiones claves de este trabajo y las instrucciones para el trabajo futuro son las siguientes:

- El enfoque basado en *SVM* se puede utilizar para lograr la detección de anomalías a partir de datos de series de tiempo *WDS*. La anomalía puede incluir una variedad de eventos como explosiones de tuberías, descarga de hidrantes y fallas del sensor.
- El sistema se aplicó a un conjunto de datos históricos durante un período de prueba de seis meses y ha demostrado la capacidad de detectar anomalías en los patrones de entrada y presión. Alrededor del 78% de las detecciones pudieron correlacionarse con información operativa e interpretación manual de datos.
- Se demostró que la metodología propuesta usando *Support Vector Regresion (SVR)* puede proporcionar una generación de alertas más rápida que un sistema *ANN / FIS* desarrollado previamente.
- La técnica *SVR* muestra el potencial para la operación en línea con un esquema adecuado para el manejo de la calidad de los datos, la selección de datos de capacitación y el programa de reciclaje [13].

En 2016 se presenta el artículo citado en [14], en el cual se prueban tres enfoques diferentes de *Machine Learning* para determinar su viabilidad para detectar anomalías. En el primer enfoque se emplean modelos de *Support Vector Regression* (SVR) por cada día de la semana los cuales son entrenados para predecir las señales de medición y se comparan con modelos sencillos utilizando estimaciones medias y medias, respectivamente. Utilizando *SVR* o los datos almacenados como reconocedores de patrones en tiempo real en todas las señales disponibles, se pueden detectar fugas de agua. El segundo enfoque utiliza proyecciones ortogonales adaptativas e informa un evento cuando el número de variables ocultas requeridas para describir los datos de transmisión a un grado definido por el usuario (umbral de nivel de energía) aumenta. Como tercer enfoque, se aplican técnicas de *clustering (unsupervised learning)* para detectar anomalías y patrones subyacentes de los flujos de datos

sin procesar. Los resultados preliminares de este trabajo indican que el conjunto de datos que se posee es demasiado limitado en la cantidad de eventos y patrones para aprovechar el potencial de las técnicas empleadas.

También en 2016 se publica la tesis introducida en [8] la cual proporciona un algoritmo de detección de anomalías como ayuda de monitoreo aplicada a datos de series de tiempo de la industria de la pulpa y el papel, desarrollada para la empresa Eurocon MOPSies AB. El algoritmo propuesto está diseñado para ser generalmente aplicable a series de tiempo específicas al proporcionar métodos para adaptar los parámetros a los datos de entrada. El algoritmo de detección de anomalías se ejecuta en un entorno no supervisado utilizando un enfoque estadístico para la detección. Como ayuda para la evaluación, las series temporales se inspeccionaron visualmente para etiquetar manualmente los patrones desviados. Se muestra que el algoritmo de detección de anomalías puede encontrar estos patrones desviados. Sin embargo, no se pudo determinar si estos patrones son anomalías con respecto al proceso subyacente ya que no había datos de prueba etiquetados disponibles [8].

En 2017 es publicada la tesis citada en [15], en la cual se aplica *Deep Learning (DL)* para la detección de anomalías en datos de sensores multivariados de máquinas lavadoras-secadoras. En este trabajo se seleccionó un modelo de *autoencoder* basado en *Long short-term memory (LSTM)* para la detección de anomalías. El modelo propuesto aprende a reconstruir secuencias normales con alta precisión, además se puede calcular una puntuación de anomalía en función del error de reconstrucción del modelo. Los patrones anómalos mostraron un puntaje de anomalía significativamente mayor. Por lo tanto, se concluye que el modelo de *autoencoder* se puede usar para detectar patrones no vistos previamente en los datos del sensor multivariado. Los experimentos en modelos de autoencoder también encontraron que el procesamiento posterior de la puntuación de anomalía es útil para filtrar los picos relacionados con la pérdida de información a través de la aniquilación de las secuencias de entrada [15].

También en el 2017 se presenta la investigación mostrada en [16] en la cual se presenta un modelo basado en algoritmos de *Deep Learning* de *LSTM* y *GRU* para facilitar la detección de anomalías en los imanes superconductores del Gran Colisionador de Hadrones. Utilizamos datos de alta resolución disponibles en la base de datos Post Mortem para entrenar un conjunto de modelos y elegimos el mejor conjunto posible de sus hiper-

parámetros. El método presentado no requiere una tediosa configuración manual del umbral ni la atención del operador en la etapa de configuración del sistema, en cambio, se propone el enfoque automático, que logra de acuerdo con los resultados presentados en experimentos una precisión del 99%. Esto se alcanza con la siguiente arquitectura de la red: *LSTM* de capa única, 128 celdas, 20 épocas de entrenamiento, look_back = 16, look_ahead = 128, grid = 100 y optimizador Adam [16].

Es publicado también en el 2017 el artículo presentado en [17], en el cual se propone un enfoque de segmentación de series temporales basado en *Redes Neuronales Convolucionales (CNN)* para la detección de anomalías. Además, se propone un *transfer learning framework* que capacita previamente a un modelo en un conjunto de datos de series temporales sintéticas univariadas a gran escala y luego ajusta sus pesos a pequeña escala, univariadas o multivariadas con clases de anomalías nunca antes vistas. El enfoque propuesto se probó en múltiples conjuntos de datos sintéticos y reales con éxito.

Nuevamente en 2017, es publicada la investigación mostrada en [18], en la que se propone y evalúa la aplicación de un algoritmo de *Machine Learning* usando un enfoque no supervisado para la detección de anomalías en un *Sistema Ciber-Físico (CPS por sus siglas en inglés)*. En este trabajo se comparan dos métodos: una *Red Neuronal Profunda (DNN por sus siglas en inglés)* adaptadas a datos de series temporales generadas por un *CPS* y *One-Class Support Vector Machines (OC-SVM)*. Para ambos métodos, primero se entrenan detectores utilizando un registro generado por una *Planta de Tratamiento de Agua (SWaT por sus siglas en inglés)* que funciona en condiciones normales, luego, se evalúa el rendimiento de ambos métodos usando un registro generado por *SWaT* que opera bajo 36 escenarios de ataque diferentes. En los resultados se encuentra que la red neuronal *DNN* genera menos falsos positivos que el *SVM* de una clase, mientras que el *SVM* detecta mayor cantidad de anomalías. En general, la red neuronal *DNN* tiene una medida F1-score ligeramente mejor que el *SVM*.

Es presentado en 2017, el artículo citado en [19] donde se presenta un nuevo algoritmo de procesamiento de señales inspirado en el paradigma de *Deep Learning* que combina *wavelets*, redes neuronales y la transformación de Hilbert. El algoritmo funciona de manera robusta y es transferible. La estructura de red neuronal propuesta facilita el aprendizaje de la interdependencia de patrones a corto y largo plazo; una tarea usualmente

difícil de lograr usando algoritmos estándar de entrenamiento de redes neuronales. El documento proporciona pautas para seleccionar el tamaño del búfer de la red neuronal, el algoritmo de entrenamiento y las características de detección de anomalías. El algoritmo aprende el comportamiento normal del sistema y no requiere la existencia de datos anómalos para evaluar su significado estadístico. Se muestra que el método puede detectar automáticamente anomalías en la Señal Eléctrica Sísmica que podrían usarse para predecir la actividad sísmica. Además, el método se puede utilizar en combinación con el *crowdsourcing* de datos de teléfonos inteligentes para localizar defectos en la carretera, como baches, para intervención y reparación.

En 2018 se publica la investigación presentada en [20], la cual propone el uso de un algoritmo multiobjetivo para resolver el problema de la detección de anomalías en línea, para la calidad del agua potable. Dicho problema consiste en un conjunto de datos desequilibrados donde los eventos anómalos, la clase minoritaria, deben detectarse correctamente en función de una serie de tiempo que denota datos de calidad del agua y datos operativos. Los sistemas propuestos se prueban con validación de espera durante la optimización y se espera que generalicen bien las predicciones para futuros datos de prueba.

En 2019 es publicada la tesis mostrada en [6], en la cual se propone un *framework*, no supervisado y escalable para la detección de anomalías en datos de series de tiempo. El enfoque propuesto se basa en un *Autoencoder Variacional (VA por sus siglas en inglés)*, un modelo generativo profundo que combina inferencia variacional con aprendizaje profundo. Además, la arquitectura integra redes neuronales recurrentes para capturar la naturaleza secuencial de los datos de series temporales y sus dependencias temporales. Además, se introduce un mecanismo de atención para mejorar el rendimiento del proceso de *encoding-decoding*. Los resultados sobre la generación de energía solar y los datos de series de tiempo del electrocardiograma muestran la capacidad del modelo propuesto para detectar patrones anómalos en series de tiempo de diferentes campos de aplicación, al tiempo que proporcionan representaciones de datos estructurados y expresivos [6].

En 2020 se publica el artículo citado en [21] en el cual se introduce un algoritmo que procesa series de tiempo de longitud variable, muestreadas de forma irregular con valores faltantes en el proceso de detección de anomalías. En este trabajo se propone un enfoque que utiliza una *Long Short Term Memory (LSTM)* para extraer características temporales y

encontrar los vectores de características más relevantes para la detección de anomalías. Después de obtener las características más relevantes usando la *LSTM*, se etiquetan las series de tiempo usando un modelo *Support Vector Data Descriptor (SVDD)*. Los resultados presentados mediante experimentos en conjuntos de datos de la vida real, comunican que el algoritmo propuesto supera el enfoque estándar gracias a la combinación de la *LSTM* con el modelo *SVDD* y su optimización conjunta.

Luego de hacer un análisis de lo anteriormente expuesto se evidencia que las técnicas de *Machine Learning* han sido ampliamente utilizadas en el área de la detección de anomalías en series de tiempo tanto univariadas como multivariadas y que seguirán siendo empleadas debido a los resultados positivos obtenidos. Se evidencia además el papel fundamental que juegan los datos o registros asociados a este proceso pues permiten la realización de análisis para la toma de decisiones en tiempo real, así como predicciones acerca de este proceso. Esto nos muestra la necesidad de emplear técnicas de *Data Science* y *Machine Learning* para lograr obtener resultados satisfactorios en la investigación.

Aunque el problema de la pérdida de agua en los Sistemas de Distribución de Agua es global, las soluciones deben adaptarse a las circunstancias locales debido a la variación en las causas de las pérdidas de agua y los mecanismos disponibles para gestionarlas [22].

El abasto de agua potable para casi un millón de habitantes de la capital del estado de Chihuahua está en riesgo por la sobreexplotación de los acuíferos Sauz-Encinillas, Tabalaopa-Aldama y Chihuahua-Sacramento. Además, la falta de mantenimiento de las instalaciones de la red hidráulica de Chihuahua (tubería, pozos, válvulas, tanques de almacenamiento, etc). podría derivar en una situación “de emergencia” por el aumento de la población (200 mil personas en 10 años) y la falta de lluvias [23]. Todo esto influye directamente en la ocurrencia de fallas en el suministro de agua potable a la población y se evidencia en las figuras 1 y 2 que muestran reportes de falta y fugas de agua en los reportes realizados por los usuarios al Centro de Información y Servicios (CIS) de la Junta Municipal de Agua y Saneamiento de Chihuahua (JMAS) de Chihuahua.

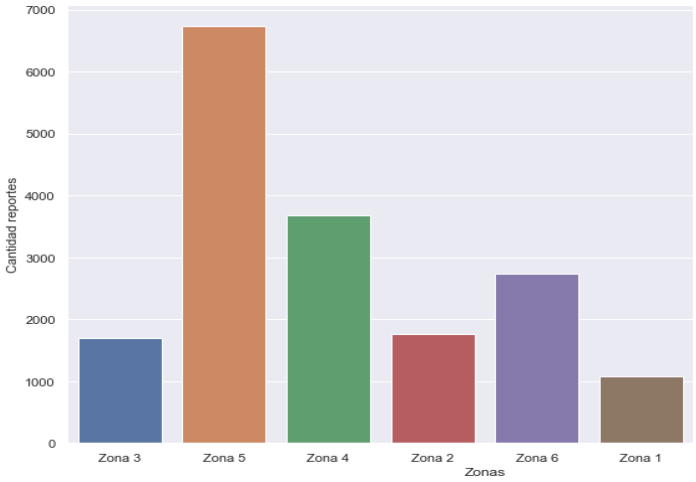

Figura 1. Reportes de falta de agua en el período Jun 2018 - Jun 2019 (17,675).

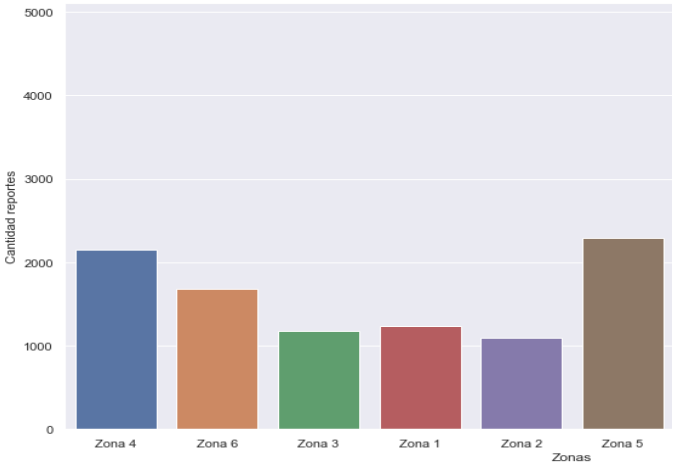

Figura 2. Reportes de fuga de agua en el período Jun 2018 - Jun 2019 (18,628).

La distribución del agua en el municipio de Chihuahua está soportada por 2 grandes redes de conducción: la de Suministro y la Hidráulica. La JMAS ha hecho grandes esfuerzos por desarrollar y consolidar sistemas de censado, telemetría, bases de datos y visualizaciones para administrar de la mejor forma el vital líquido. Por su misma naturaleza, estos sistemas generan cientos y miles de datos de forma diaria, lo cual implica un reto para su interpretación a mediano y largo plazo.

La ciencia de datos o Data Science es una nueva disciplina que busca ordenar, relacionar, interpretar y ofrecer un entendimiento de grandes volúmenes de datos, como los que generan los sistemas de monitoreo en la red hidráulica.

Actualmente se cuenta con un gran volumen de información provenientes de los sistemas de censado y telemetría de la JMAS. Aquí es donde se presenta un área de oportunidad para este trabajo; la cual es realizar un análisis de la información con el objetivo de inferir conocimientos a ser utilizados en procesos futuros. Con el análisis y procesamiento de la información utilizando técnicas de *Data Science* y *ML*, se podría predecir estados anómalos en las diferentes instalaciones de la Red Hidráulica. Estos estados pueden ocasionar la falta y fugas de agua en las diferentes zonas perjudicando directamente el proceso de distribución del agua. Entre otros posibles beneficios se encuentran: pronosticar los niveles de presión del agua necesarios para abastecer una zona residencial, predecir el tamaño de las fugas y su localización, determinar las configuraciones de la Red Hidráulica que minimizan la posibilidad de ocurrencia de falta de agua en zonas residenciales.

Considerando la situación problémica anterior se plantean las siguientes preguntas de investigación:

1.2. PREGUNTAS DE INVESTIGACIÓN

¿Cómo presentar los datos obtenidos de forma que puedan ser utilizados por la comunidad de Machine Learning?

¿Como permitir que un usuario experto pueda identificar y registrar comportamientos que pueden llevar a la falta o fugas de agua en las instalaciones de la Red Hidráulica?

¿Los datos actuales son suficientes para predecir y detectar anomalías en tiempo real?

¿Qué información aporta los datos recopilados por el sistema CIS?, ¿esta información puede ser utilizada en el proceso de predecir y detectar anomalías en tiempo real?

¿Cómo permitir que los profesionales del Centro de Control de Aguas y Filtros de la JMAS puedan tomar acciones preventivas relacionadas a la falta y fugas de agua en las instalaciones de la Red Hidráulica?

1.3. OBJETIVOS

El **objetivo general** de la presente investigación es aplicar algoritmos de Detección de Anomalías a colecciones de datos provenientes de la red de hidráulica de la JMAS, con el

objetivo detectar y predecir comportamientos anómalos de las instalaciones de la red que pueden afectar de manera negativa al suministro de agua en el municipio de Chihuahua.

A partir de este objetivo general se derivan los siguientes Objetivos Específicos:

1. Diseñar e implementar un mecanismo que permita etiquetar datos de eventos anómalos.
2. Analizar datos provenientes de sensores en instalaciones de la Red Hidráulica por medio de método de Machine Learning.
3. Explorar los datos del sistema CIS con el objetivo de verificar que información puede aportar al proceso de detección y predicción de anomalías en tiempo real.
4. Integrar las técnicas de Aprendizaje Máquina de mejores resultados, en una herramienta Web, que permita el apoyo en la toma de decisiones a los profesionales de las JMAS, para brindar un mejor servicio de la administración del agua.

1.4. JUSTIFICACIÓN

El agua proporciona sustento, apoya a la industria y riega los campos. Pero las administraciones de las ciudades están luchando para satisfacer la creciente demanda de las poblaciones en crecimiento mientras se enfrentan a problemas como la calidad del agua, las inundaciones, la sequía y el envejecimiento de la infraestructura. Según las Organizaciones de las Naciones Unidas, alrededor de dos tercios de la población mundial, 4,600 millones de personas, enfrentarán condiciones de estrés hídrico en la próxima década [24].

El aumento en los costos de bombeo, tratamiento y distribución del agua están impulsando a las empresas de agua para combatir los principales problemas que estas enfrentan. La pérdida de agua es uno de los problemas más críticos, debido a esto su interés por desarrollar métodos para detectar, localizar y reparar fugas. Los métodos tradicionales de detección de fugas en tuberías requieren una inspección periódica con participación humana, lo que lo hace lento e ineficiente para la detección de fugas de manera oportuna [3].

Las ciudades inteligentes utilizan las Tecnologías de la Información y las Comunicaciones (TIC) para lograr un suministro de agua sostenible, eficiente y limpio. La mayoría de las personas se refieren a un sistema de agua habilitado para las TIC como un "Sistema de Agua Inteligente" o una "Red de Agua Inteligente". La necesidad de estos sistemas es impulsada por cuatro realidades urgentes [24]:

- El agua es escasa.
- El agua está en riesgo.
- El agua está subvaluada.
- La infraestructura del agua es cara.

Los organismos que tienen la tarea de la gestión y distribución de este preciado líquido deben hacer uso de las tecnologías existentes con el objetivo de preservar y mejorar la gestión y el suministro de agua, mientras mantienen el costo del agua lo más bajo posible. Pues ya existen regiones donde periódicamente escasea el agua y otras donde el agua es prohibitivamente cara.

Con este trabajo se busca trabajar en conjunto con la Junta Municipal de Agua y Saneamiento de Chihuahua (JMAS) con el objetivo de aplicar las TIC utilizando un enfoque semi-supervisado de Detección de Anomalías en las Instalaciones de la Red Hidráulica que pueden conllevar a afectar los procesos de administración y distribución del agua. Teniendo como principal interés brindar herramientas que permitan apoyo tecnológico para poder detectar y predecir estos eventos anómalos.

Los beneficios que este proyecto generará serán el desarrollo de herramientas para la JMAS que beneficien, den soporte y precisión a la toma de decisiones con respecto a los procesos de administración y distribución del agua, posicionando a la JMAS como un organismo de vanguardia entre sus pares. El desarrollo de estas herramientas será utilizando tecnologías libres con el objetivo de evitar el gasto en tecnologías propietarias que incrementarían los gastos relacionados a este proceso.

CAPÍTULO 2: MARCO TEÓRICO

Con el objetivo de facilitar la comprensión del alcance de la presente investigación, en este capítulo se exponen conceptos asociados al dominio del problema, para ello se analizan las características y aplicaciones de los mismos, así como las tecnologías más utilizadas.

El agua es suministrada a nuestro hogar por una ***Red de Distribución de Agua (WDN, por sus siglas en inglés)***, que es propiedad y es mantenida por las compañías de servicios de agua. Estas compañías enfrentan el aumento de los costos de instalación de nuevas tuberías para atender a la creciente población, así como el mantenimiento y la sustitución del sistema de envejecimiento.

La *WDN* es una infraestructura hidráulica que consta de varios componentes, como tuberías, bombas, válvulas, depósitos, tanques de almacenamiento, medidores y otros componentes que conectan las plantas de tratamiento de agua con los grifos de los consumidores. Estas redes de distribución están diseñadas para satisfacer las demandas máximas. El propósito del sistema de tuberías conectadas en una red es suministrar agua a una presión y flujo adecuados. El agua en la red de suministro se mantiene a presión positiva para garantizar que el agua llegue a todas las partes de la red. El agua normalmente está presurizada por bombas que bombean agua a los tanques de almacenamiento, que se construyen en el punto más alto de la red. Una sola red puede tener varios de estos reservorios [3].

Con la disponibilidad de sensores de vanguardia y el crecimiento de los servicios habilitados de *Internet de las Cosas (IoT por sus siglas en inglés)*, es posible construir un sistema autónomo que recopile continuamente datos de campo de forma remota por los nodos de sensores implementados en la *WDN*. Estos nodos sirven como un componente básico de una red que se conecta con una unidad de monitoreo central de forma inalámbrica. Dichas redes se denominan ***Redes de Sensores Inalámbricos (WSN por sus siglas en inglés)*** como se describe en la siguiente sección.

La *WSN* es una red de una gran cantidad de nodos de sensores que detectan y monitorean cooperativamente los parámetros que involucran condiciones físicas o ambientales. Es una infraestructura que es una combinación única de detección, computación

y segmento de comunicación que proporciona a un administrador la capacidad de instrumentar, observar y reaccionar ante eventos y fenómenos en un entorno determinado [3]. Estos sensores son distribuidos por toda la *WDN* con el objetivo de monitoreo, continuo y en tiempo real de la red hidráulica para dar soporte a la toma de decisiones relacionadas en el proceso de distribución del agua.

Colocación de nodos sensores: Dada cualquier red de distribución de agua, se necesita un mecanismo para averiguar automáticamente cuántos nodos de sensores se necesitan y dónde colocarlos en la red. La colocación del nodo sensor tendría un impacto directo en la eficacia de la localización de la fuga en la red [3].

Además, de encontrar el número correcto de sensores y su ubicación en la *WDN*, también es conveniente definir una zona cerca de cada ubicación del nodo del sensor conocida como zona de fuga. Cada nodo sensor representará una zona que consta de varios nodos de unión cercanos. Esto ayuda a identificar en qué zona ha ocurrido la fuga en los datos actuales [3].

Debe haber una distribución equitativa del tamaño de la zona de fuga. El tamaño no debe ser demasiado plano ni demasiado grande. Una distribución sesgada del tamaño de la zona de fuga tendrá un efecto negativo en la resolución de la localización de la fuga. Por un lado, los sensores que se representan solo en la red darían como resultado un uso redundante de los recursos; por otro lado, el nodo sensor que representa una mayor cantidad de nodos daría lugar a un rendimiento deficiente de la localización de fugas [3].

2.1. DETECCIÓN DE ANOMALÍAS (AD POR SUS SIGLAS EN INGLÉS)

Las anomalías se definen como patrones en los datos que no se ajustan al comportamiento normal o esperado. El problema de encontrar tales patrones se conoce como detección de anomalías. La detección de anomalías se puede aplicar a cualquier tipo de datos, binarios, discretos o continuos, univariadas o multivariadas. Se debe proporcionar un conjunto de datos iniciales, que se denominará conjunto de entrenamiento. El conjunto de datos que se probará si contiene puntos anómalos y se denomina conjunto de detección. Los conjuntos de entrenamiento y detección pueden ser los mismos, además pueden cambiar con el tiempo, por ejemplo, en el caso de la transmisión de datos [8].

TIPOS DE ANOMALÍAS

Existen principalmente tres tipos de anomalías que se estudian en la literatura, a saber, anomalías puntuales, anomalías secuenciales y anomalías contextuales. Una breve descripción de estos se da a continuación [8].

- **Anomalías puntuales:** Si un único punto se desvía del patrón normal considerado, se denomina anomalía puntual. Esta es la forma más simple de una anomalía y es la forma más investigada. Un ejemplo de anomalía puntual es si un valor del proceso de repente es muy bajo o alto.

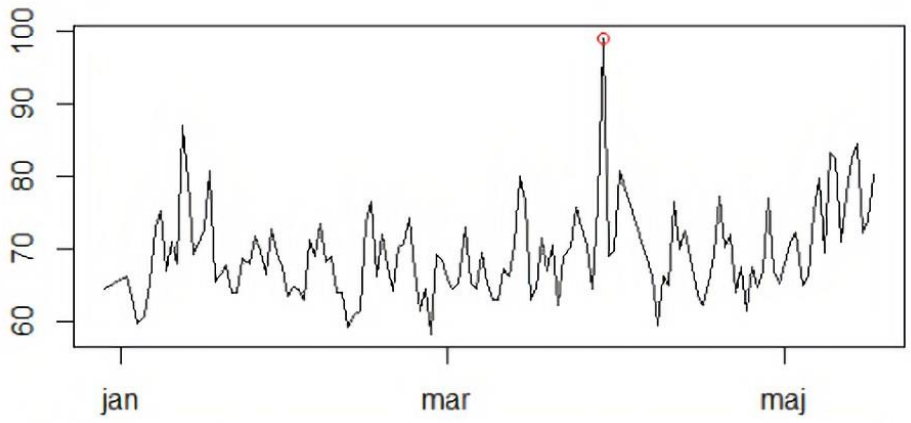

Figura 3. Anomalía puntual [8].

- **Anomalías secuenciales**: Si una secuencia o colección de puntos es anómala con respecto al resto de los datos, pero no los puntos en sí, se denomina anomalía secuencial o colectiva. Dado que esta tesis trata las anomalías en series de tiempo, nos referiremos a este tipo de anomalía como una anomalía secuencial.

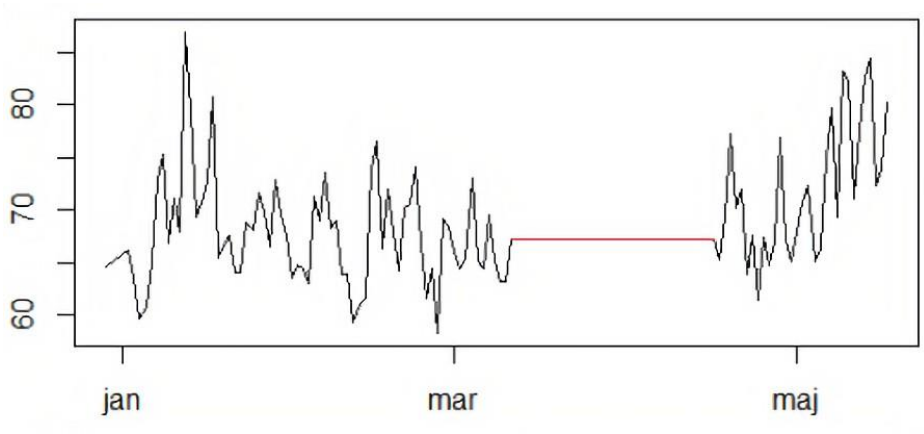

Figura 4. Anomalía secuencial [8].

- **Anomalías contextuales:** Si un punto o una secuencia de puntos se consideran como una anomalía con respecto a su vecindad local, pero no de otra manera, se conoce como una anomalía contextual.

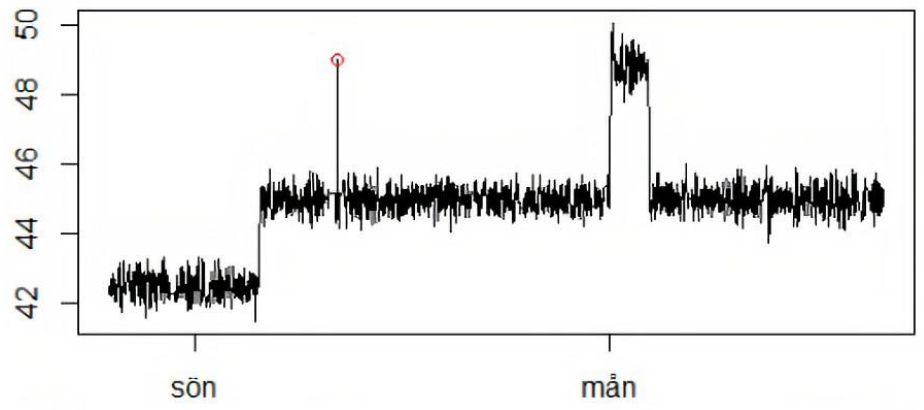

Figura 5. Anomalía contextual [8].

Aprendizaje Máquina (ML por sus siglas en inglés) es una clase de métodos en el análisis de datos que aprende patrones y puntos de vista ocultos en los datos sin ser programado explícitamente para ellos. Gracias a los dispositivos informáticos mejores y más potentes, hay una tendencia creciente a aplicar el *ML* en varios casos recientemente, como la detección de fallas en el equipo, el reconocimiento de patrones e imágenes, el filtrado de correo no deseado y la detección de fraudes [3].

Algoritmos *ML* realizan análisis predictivo. Cuando las variables de salida toman valores continuos, se denominan Regresión, mientras que cuando toma las etiquetas de clase

se denomina Clasificación. La regresión se refiere a la estimación de una respuesta, mientras que la Clasificación se refiere a la identificación de miembros del grupo [3].

ML tiene dos ramas principales: *Aprendizaje Supervisado* y *Aprendizaje No Supervisado*, y muchas ramas secundarias que las unen.

2.2. APRENDIZAJE SUPERVISADO (SL POR SUS SIGLAS EN INGLÉS)

En el aprendizaje supervisado, se tiene acceso a etiquetas *(labels)*, que se usan para mejorar el rendimiento en alguna tarea. El principio detrás del enfoque de *SL* es aprender la función de mapeo f: x → Y que mapea la entrada *x* a la salida *y*. Las variables de entrada *x* se componen de una o más variables o predictores independientes, mientras que la salida consta de una variable independiente o de predicción y. El aprendizaje se realiza mediante la aplicación del algoritmo de *ML* en los "datos de entrenamiento" mediante los cuales obtenemos el modelo aprendido como salida. Las principales técnicas de aprendizaje supervisado que se han utilizado en el nexo entre la energía y el agua comprenden *Análisis de Regresión*, *Redes Neuronales Artificiales (ANN por sus siglas en inglés)*, *Máquinas de Vectores de Soporte (SVM sus siglas en inglés)* y el análisis de series de tiempo [4].

Análisis de regresión

Es una técnica de aprendizaje supervisado que se basa en estimar la relación entre una variable dependiente (y) con una o más variables independientes (x). Existen diferentes formas de técnicas de regresión que se basan en el número de variables independientes, el tipo de variables dependientes y la complejidad de la relación que se modela entre estas variables [4].

Redes neuronale artificiales (ann por sus siglas en inglés)

Es un importante algoritmo de aprendizaje automático supervisado y es uno de los algoritmos potentes debido a su capacidad para aprender cualquier relación funcional entre una variable dependiente y una o más variables independientes. Además, maneja datos no lineales de manera efectiva debido al uso de las funciones de activación. El propósito de la función de activación como *Sigmoid, ReLU y Tanh* es manejar efectivamente la relación no lineal entre la variable de salida y las variables de entrada. Una arquitectura *ANN* típica consiste en dos capas (una capa oculta y una capa de salida) [4].

El uso de *ANN* ha demostrado ser útil para estimar eficientemente los niveles de agua subterránea en comparación con los métodos de simulación hidrológica. Los resultados de este método de *ANN* híbrido mostraron que las relaciones complejas y no lineales entre la precipitación, la temperatura, el flujo de la corriente, los índices climáticos, la demanda de riego y los niveles de agua subterránea podrían representarse y reproducirse con el método [4].

Máquina de vectores de soporte (svm por sus siglas en inglés)

Es la poderosa técnica de aprendizaje supervisado que se usa tanto para la clasificación como para la regresión. El uso de *SVM* con funciones de kernel es mapear datos no lineales del espacio de entrada a un espacio dimensional más alto para que sea linealmente separable. El uso de este truco del núcleo en *SVM* se ha explorado más a fondo, no solo en otros dominios, sino también en el pronóstico de lluvias [4].

SVM se aplica a las estimaciones futuras de la disponibilidad de agua y a la predicción de la calidad del aire y del agua. Las técnicas de regresión lineal de mínimos cuadrados y *SVM* se utilizan para predecir la generación de energía solar en función de los pronósticos meteorológicos [4].

Arboles de decisión (dt por sus siglas en inglés)

Son un método de aprendizaje automático supervisado utilizado para la clasificación y la regresión. Cuanto más profundo es el árbol, más complejas son las reglas de decisión y más se ajustan al modelo. Las ventajas de esta técnica incluyen fácil de entender, interpretar y visualizar, mientras que las desventajas incluyen una alta variación que conduce a un problema de sobreajuste. Los modelos de árbol de decisión pueden producir reglas o declaraciones lógicas que son fáciles de interpretar, pero no funcionan tan bien como las redes neuronales para datos no lineales y tienden a ser susceptibles al ruido [4].

2.3. APRENDIZAJE NO SUPERVISADO (UL POR SUS SIGLAS EN INGLÉS)

En el aprendizaje no supervisado, las etiquetas no están disponibles. Por lo tanto, la tarea no está bien definida y el rendimiento no se puede medir con tanta claridad. Estos enfoques se basan en encontrar una estructura oculta a partir de datos no etiquetados. A diferencia del enfoque de aprendizaje supervisado en el que tenemos datos etiquetados

conocidos de las variables de entrada y salida, en el *UL* aprendemos los patrones ocultos, las asociaciones, las similitudes entre las entradas sin ninguna variable de salida conocida [4].

En lugar de intentar hacer predicciones, los algoritmos de aprendizaje no supervisados intentarán aprender la estructura subyacente de los datos. Existen varias familias de algoritmos las cuales son las siguientes:

Extracción de características (fe por sus siglas en inglés)

Con el aprendizaje no supervisado, podemos aprender nuevas representaciones de las características originales de los datos en un campo conocido como extracción de características. La extracción de características se puede usar para reducir el número de características originales a un subconjunto más pequeño, realizando efectivamente la reducción de dimensionalidad. Pero la *extracción de características* también puede generar nuevas representaciones de características para ayudar a mejorar el rendimiento en problemas de aprendizaje supervisado.

Autoencoders

Para generar nuevas representaciones de características, podemos usar una red neuronal no recurrente de avance para realizar el aprendizaje de representación, donde el número de nodos en la capa de salida coincide con el número de nodos en la capa de entrada. Esta red neuronal se conoce como *autoencoder* y reconstruye efectivamente las características originales, aprendiendo una nueva representación usando las capas ocultas en el medio. Cada capa oculta del *autoencoder* aprende una representación de las características originales, y las capas posteriores se basan en la representación aprendida por las capas anteriores. Capa por capa, el *autoencoder* aprende representaciones cada vez más complicadas de las más simples. La capa de salida es la representación final recién aprendida de las características originales. Esta representación aprendida se puede utilizar como entrada en un modelo de aprendizaje supervisado con el objetivo de mejorar el error de generalización [25].

2.4. APRENDIZAJE PROFUNDO NO SUPERVISADO (DUL POR SUS SIGLAS EN INGLÉS)

Hasta hace muy poco, el entrenamiento de redes neuronales profundas era computablemente intratable. En estas redes neuronales, las capas ocultas aprenden

representaciones internas para ayudar a resolver el problema en cuestión. Las representaciones mejoran con el tiempo en función de cómo la red neuronal utiliza el gradiente de la función de error en cada iteración de entrenamiento para actualizar los pesos de los distintos nodos.

Estas actualizaciones son computacionalmente costosas y pueden ocurrir dos tipos principales de problemas en el proceso. Primero, el gradiente de la función de error puede volverse muy pequeño y, dado que *Backpropagation* se basa en multiplicar estos pequeños pesos juntos, los pesos de la red pueden actualizarse muy lentamente o no hacerlo, evitando la capacitación adecuada de la red. Esto se conoce como *vanishing gradient problema* [25].

Redes neuronales profunas (dnm por sus siglas en inglés)

El término *deep* en *deep neural network* se refiere al número de capas a través de las cuales se extraen las características de los datos. Las arquitecturas profundas superan las limitaciones de los enfoques tradicionales de aprendizaje automático de escalabilidad y generalización a nuevas variaciones dentro de los datos y la necesidad de ingeniería manual de características.

DNN es una red neuronal artificial *ANN* con múltiples capas entre las capas de entrada y salida. *DNN* encuentra la manipulación matemática correcta para convertir la entrada en la salida, ya sea una relación lineal o no lineal. La red se mueve a través de las capas calculando la probabilidad de cada salida. Por ejemplo, una *DNN* que está entrenada para reconocer razas de perros revisará la imagen dada y calculará la probabilidad de que el perro en la imagen sea una raza determinada. El usuario puede revisar los resultados y seleccionar qué probabilidades debe mostrar la red (por encima de cierto umbral, etc.) y devolver la etiqueta propuesta. Cada manipulación matemática como tal se considera una capa, y los *DNN* complejos tienen muchas capas, de ahí el nombre de redes "profundas".

DNN pueden modelar relaciones complejas no lineales. Las arquitecturas de *DNN* generan modelos de composición donde el objeto se expresa como una composición en capas de primitivas. Las capas adicionales permiten la composición de características de capas inferiores, potencialmente modelando datos complejos con menos unidades que una red superficial de rendimiento similar.

Red Neuronal Recurrente de memoria larga a corto plazo (LSTM por sus siglas en inglés)

La estructura interna *LSTM* se basa en un conjunto de celdas conectadas. La estructura de una sola celda se presenta en la Figura 6.

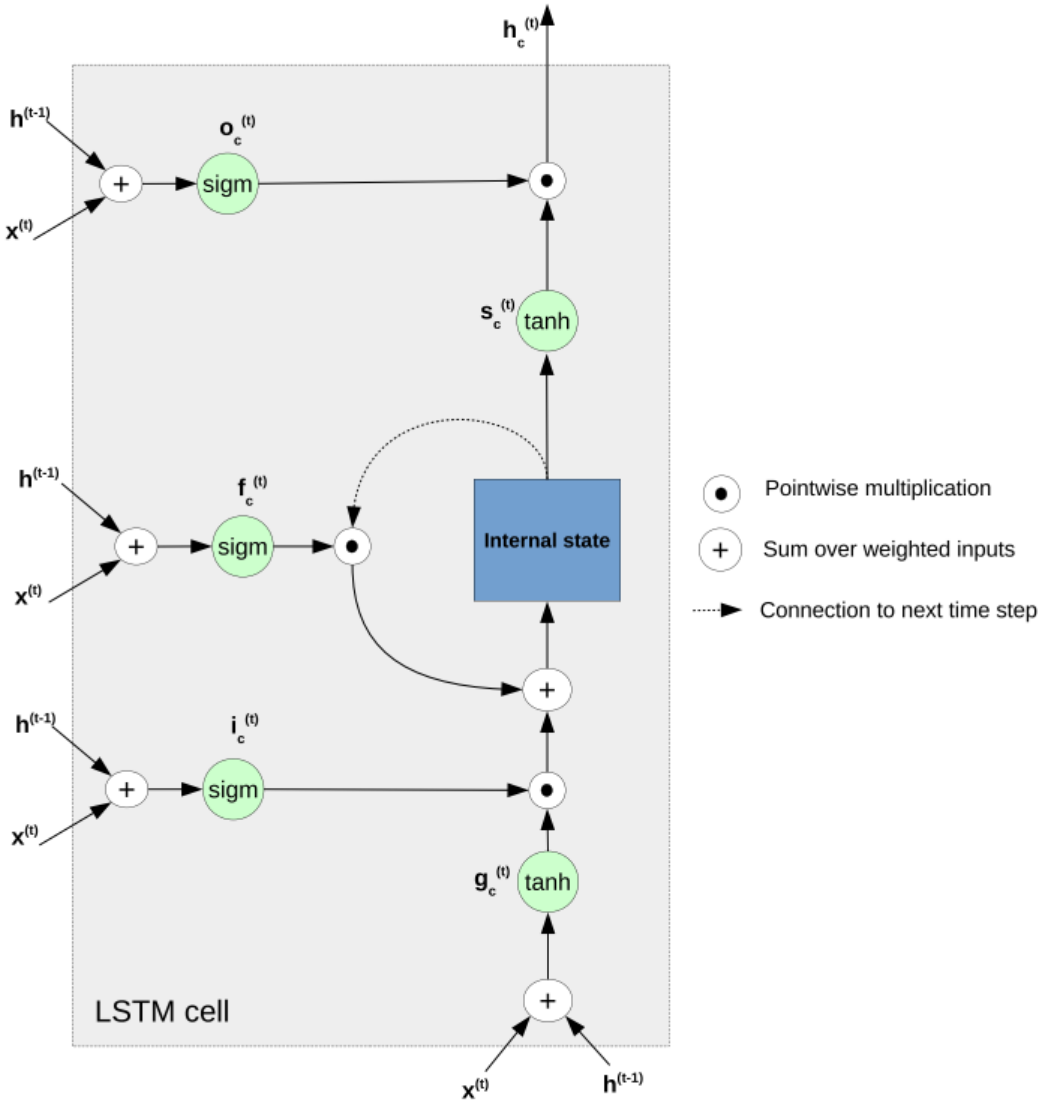

Figura 6. Arquitectura de celda *LSTM* [16].

Contiene una conexión de retroalimentación que almacena el estado temporal de la célula, tres puertas y dos nodos que sirven como interfaz para la propagación de información dentro de la red. Hay tres puertas diferentes en cada celda *LSTM* [16]:

- *input gate* $i_c^{(t)}$ que controla las activaciones de entrada en el elemento de memoria.

- *output gate* $o_c^{(t)}$ controla el flujo de salida celular de las activaciones al resto de la red.
- *forget gate* $f_c^{(t)}$ escala el estado interno de la celda antes de sumarlo con la entrada a través de la conexión auto-recurrente de la celda. Esto permite el olvido gradual en la memoria celular.

La red *LSTM* aprende cuándo permitir una activación en los estados internos de sus celdas y una activación de las salidas. Este es un mecanismo de activación y todas las puertas se consideran componentes separados de la celda *LSTM* con su propia capacidad de aprendizaje. Esto significa que las células se adaptan durante el proceso de capacitación para preservar un flujo de información adecuado en toda la red como unidades separadas. Esto significa, que cuando las puertas están cerradas, el estado interno de la celda no se ve afectado [16].

2.5. APRENDIZAJE SEMI-SUPERVISADO

Aunque el aprendizaje supervisado y el aprendizaje no supervisado son dos ramas principales distintas del aprendizaje automático, los algoritmos de cada rama se pueden mezclar como parte de una tubería de aprendizaje automático. Por lo general, esta combinación de supervisados y no supervisados se usa cuando queremos aprovechar al máximo las pocas etiquetas que tenemos o cuando queremos encontrar patrones nuevos pero desconocidos de datos no etiquetados, además de los patrones conocidos de los datos etiquetados. Este tipo de problemas se resuelven utilizando un híbrido de aprendizaje supervisado y no supervisado conocido como *Aprendizaje Semi-Supervisado* [25].

Con los sistemas de detección de anomalías, podemos tomar un problema no supervisado y eventualmente crear uno semi-supervisado con el enfoque de *cluster and label,* en el cual se crean agrupaciones y se establece una etiqueta a cada grupo. Con el tiempo, podemos ejecutar algoritmos supervisados en los datos etiquetados junto con los algoritmos no supervisados. Para aplicaciones de *ML* exitosas, los sistemas no supervisados y los sistemas supervisados deben usarse en conjunto, complementándose entre sí [25].

El sistema supervisado encuentra los patrones conocidos con un alto nivel de precisión, mientras que el sistema no supervisado descubre nuevos patrones que pueden ser

de interés. Una vez que estos patrones son descubiertos por la *IA* no supervisada, los patrones son etiquetados por humanos, haciendo la transición de más datos de no etiquetados a etiquetados [25].

2.6. MÉTRICAS DE EVALUACIÓN

Se utilizan diferentes métricas para evaluar la efectividad y la eficiencia de los métodos de detección de anomalías en los experimentos realizados en esta investigación. La eficacia de las técnicas de detección de anomalías se puede evaluar de acuerdo con cuántos puntos de datos anómalos se identifican correctamente y también de acuerdo con cuántos datos normales se clasifican incorrectamente como datos anómalos, este último también se conoce como tasa de falsas alarmas [9].

A menudo se utilizan las métricas de rendimiento, se basan en una matriz de confusión que informa el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).

En los artículos revisados en el apartado de trabajos relacionados [6], [15], [26]–[30] se puede apreciar el uso de las siguientes métricas de evaluación:

Para conjuntos de datos desequilibrados cuando la clase minoritaria es una clase importante, a menudo se utilizan métricas de rendimiento mencionadas por Bekkar, Djemaa y Alitouche (2013). Se basan en una matriz de confusión que informa el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).

Accuracy

Medida estadística de qué tan bien una prueba de clasificación binaria identifica o excluye correctamente una condición. Es la proporción de resultados verdaderos (tanto positivos verdaderos como negativos verdaderos) entre el número total de casos examinados.

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FP}\right)$$

Precision

El número de ejemplos positivos correctamente clasificados dividido por el número de ejemplos etiquetados por el sistema como positivos [26].

$$Precision = \left(\frac{TP}{TP + FP}\right)$$

Recall

El número de ejemplos positivos correctamente clasificados dividido por el número de ejemplos positivos en los datos [26].

$$Recall = TPR = \left(\frac{TP}{TP + FN}\right)$$

F1 score

La media armónica entre *precision* y *recall* [26].

$$F1score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

Receiver operating characteristic (roc) curve

Las curvas características de funcionamiento del receptor (ROC) se utilizan generalmente para representar el equilibrio entre la tasa de detección positiva (TP) y la tasa de falsa alarma (FP). Las técnicas de detección de anomalías deben tener como objetivo tener una tasa de detección positiva (TP) alta mientras se mantiene baja la tasa de falsa alarma (FP).

Area under the curve (AUC)

Para comparar clasificadores, es posible que deseemos reducir el rendimiento de ROC a un único valor escalar que represente el rendimiento esperado. Un método común es calcular el área bajo la curva ROC, abreviada AUC. El AUC tiene una propiedad estadística importante: el AUC de un clasificador es equivalente a la probabilidad de que el clasificador clasifique una instancia positiva elegida aleatoriamente por encima de una instancia negativa elegida aleatoriamente [31].

Cuando el algoritmo retorna clases discretas el AUC se puede calcular de la siguiente forma:

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

En cambio, cuando se tienen clases continuas como resultado de los algoritmos se puede calcular el AUC utilizando el algoritmo mostrado en la figura 7:

```
1: L_sorted ← L sorted decreasing by f scores
2: FP ← TP ← 0
3: FP_prev ← TP_prev ← 0
4: A ← 0
5: f_prev ← −∞
6: i ← 1
7: while i ⩽ |L_sorted| do
8:    if f(i) ≠ f_prev then
9:       A ← A + TRAPEZOID_AREA(FP, FP_prev,
                TP, TP_prev)
10:      f_prev ← f(i)
11:      FP_prev ← FP
12:      TP_prev ← TP
13:   end if
14:   if i is a positive example then
15:      TP ← TP + 1
16:   else       /* i is a negative example */
17:      FP ← FP + 1
18:   end if
19:   i ← i + 1
20: end while
21: A ← A + TRAPEZOID_AREA(N, FP_prev, N, TP_prev)
22: A ← A/(P × N)  /* scale from P × N onto the unit
    square */
23: end
 1: function TRAPEZOID_AREA(X1, X2, Y1, Y2)
 2: Base ← |X1 − X2|
 3: Height_avg ← (Y1 + Y2)/2
 4: return Base × Height_avg
 5: end function
```

Figura 7. Pseudocódigo de algoritmo para calcular AUC cuando se tienen clases continuas [31].

Entradas del Algoritmo:

- L: el conjunto de ejemplos de prueba.
- f (i), el clasificador probabilístico estima que el ejemplo *i* es positivo.
- P, N: el número de ejemplos positivos y negativos respectivamente. Los cuales deben ser ambos mayores que cero.

Salida del Algoritmo:

- A: el área bajo la curva ROC.

CAPÍTULO 3: METODOLOGÍA

Este capítulo describe los procedimientos y enfoques utilizados para la detección y predicción de anomalías en datos de series tiempo y su implementación en los algoritmos propuestos empleando un enfoque supervisado.

3.1. PROCEDIMIENTO DE OBTENCIÓN Y VISUALIZACIÓN DE LOS DATOS DE LA RED HIDRÁULICA

3.1.1. DESCRIPCIÓN Y OBTENCIÓN DE LOS DATOS

Inicialmente los profesionales de la planta de Filtros y Saneamientos de Agua de Chihuahua perteneciente a las JMAS nos entregaron 3gb de información en archivos perteneciente a los registros del Sistema SCADA 32000 el cual va registrando los datos relacionados a los diferentes sensores de las diferentes instalaciones a lo largo de la Red Hidráulica de Chihuahua.

Tipos de Instalaciones:

- Tanques
- Pozos
- Rebombeos
- Válvulas

Información registrada por tipos de instalación

Tanques

- Gastos de salida
- Gastos de entrada
- Nivel

Pozos

- Presión
- Gasto
- Falla de fase

Rebombeos

- Gasto de entrada

- Gasto de salida
- Presión de succión
- Presión de descarga

Válvulas

- Presión de succión
- Presión de salida

Los datos recopilados de las instalaciones disponibles son del período de 09 de junio 2018 a 12 de junio del 2019. Estos datos fueron entregados en archivos independientes por cada día de bases de datos Access. Debido a la dificultad para el procesamiento de los datos en el formato que fue entregado, lo primero que se realizó fue emplear un *script de Python* para pasar cada uno de los archivos individuales a una base de datos general. Se empleó *PostgreSQL* como gestor de bases de datos para la base de datos general, a la cual se le traspasó la información. Para acceder a esta base de datos, previamente se realizó una solicitud al personal de la JMAS debido a que esta es una información sensible.

3.1.2. VISUALIZACIÓN Y ESTADÍSTICAS BÁSICAS DE LOS DATOS

Una vez cargado los datos, es necesario proveer una herramienta que permita la visualización de los mismos, así como generar estadísticas resumidas de la información visualizada, con el objetivo de brindar a los profesionales de la JMAS herramientas de soporte para la toma de decisiones.

Se desarrolla una aplicación Web para la visualización y cálculo de estadísticas básicas de los datos, para su desarrollo se emplea el framework web Django utilizando la base de datos de PostgreSQL creada anteriormente. A continuación, se muestra una imagen del Módulo de Visualización y estadísticas de los datos.

Figura 8. Módulo de Visualización y estadísticas de los datos.

El usuario selecciona el período de tiempo (*fecha inicio, fecha fin*), la *Instalación* de la red hidráulica y por último el *Atributo* de la instalación que desea monitorear y luego presiona el botón Graficar para la generación de la gráfica correspondiente, mostrándole además estadísticas básicas de la información filtrada.

Estadísticas básicas calculadas:

- Valor mínimo
- Valor máximo
- Promedio
- Varianza
- Desviación estándar

3.1.3. PREPROCESAMIENTO DE LOS DATOS

Debido a que la frecuencia de los datos que se tiene no es constante se debe realizar un procedimiento *data imputation* que permita poder mostrar los datos en una frecuencia escogida ejemplo (cada un minuto, cada media hora, cada una hora etc). Para esto se utilizó la biblioteca Pandas de Python muy popular en el procesamiento de datos debido a las funcionalidades que brinda.

Primeramente, se utilizó la biblioteca de *python* ***Pandas*** para convertir los datos en una serie de tiempo, convirtiendo la fecha de una cadena a una fecha como índice. Esto permite que la interacción con estos objetos sea más eficiente, además que con este formato se puede utilizar los métodos ***resample*** o ***asfreq*** para cambiar la frecuencia de los datos ya sea en minutos, horas, días etc. Del índice obtenido se obtienen informaciones tales como año, mes, día de la semana lo cual que también se adicionará al *dataframe* para ser utilizado en el procesamiento de los datos.

En este caso se utilizó el método ***resample*** de Pandas y la frecuencia en la cual se genera la información fue en minutos. Para realizar este procesamiento se debe indicar un método para rellenar los valores en frecuencias donde no se presentan datos, *Pandas* provee varios como: ***ffill*** propagar último valor válido al próximo valor vacío y ***bfill*** usar el próximo valor válido para rellenar el valor existente. En este caso se usó el método ***ffill***.

3.2. PROCEDIMIENTO DE ETIQUETADO DE ANOMALÍAS

Con el objetivo de poder probar la efectividad y exactitud de los procedimientos y algoritmos que serán empleados para la predicción y detección de anomalías es de vital importancia poder etiquetar de forma parcial o total los datos que fueron obtenidos en eventos anómalos o de las instalaciones de la red hidráulica. Debido a esto surge la necesidad de proveer un módulo que permita el etiquetado de las anomalías por los usuarios expertos de la JMAS. A continuación, se presenta la figura 9, la cual muestra una imagen del módulo para el etiquetado de anomalías.

3.2.1. MÓDULO DE ETIQUETADO DE ANOMALÍAS

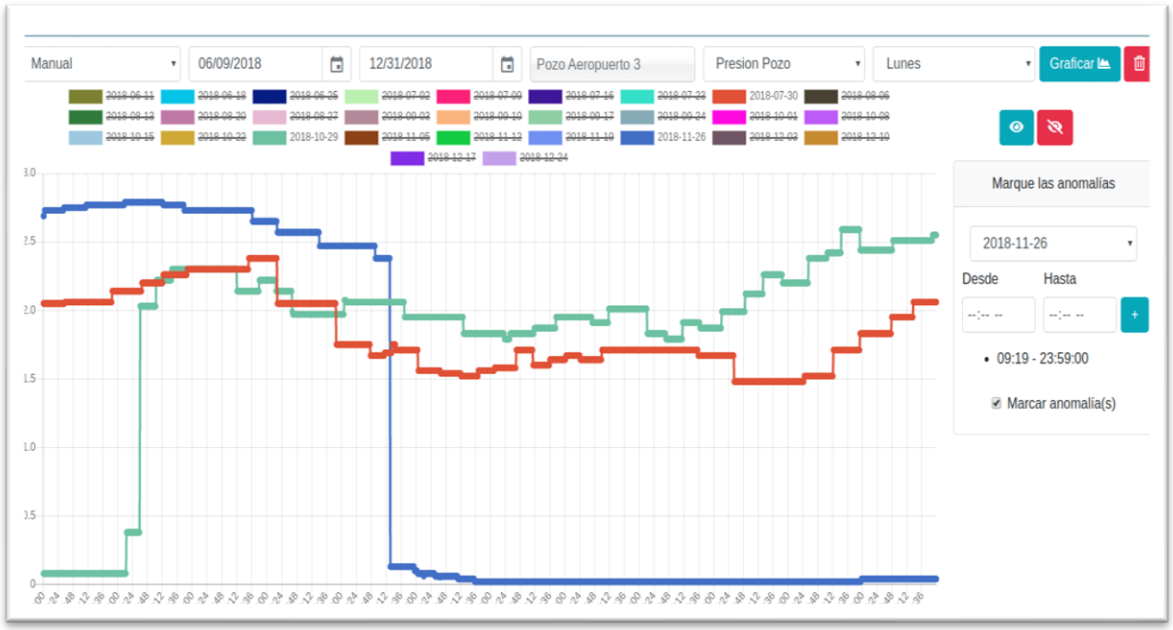

Figura 9. Módulo para el etiquetado de anomalías.

El usuario inicialmente selecciona la *fecha inicial*, *fecha final*, la *instalación*, el *atributo* que se desea monitorear y además se brinda la posibilidad de escoger el *día de la semana* que se desea monitorear (lunes, martes, etc.), esto es necesario debido a que el consumo del agua puede ser diferente dependiendo del día de semana y de esta forma el experto puede tomar una decisión más asertiva en el proceso de etiquetado.

Luego se presiona el botón Graficar y se le muestra al usuario series de tiempo relacionadas a los datos introducidos, de las cuales el usuario puede ocultar y mostrar algunas de las series mostradas para tomar una mejor decisión. Este módulo permite seleccionar una fecha determinada y para ese día el experto puede seleccionar el o los intervalos de tiempo en los cuales el determina que ocurre una anomalía.

3.2.2. DESCRIPCIÓN DEL PROCESO DE ETIQUETADO DE LOS DATOS

Una vez implementado el módulo para el etiquetado de los datos se lleva a cabo un proceso parcial de etiquetado de los datos. Este proceso fue llevado a cabo por un usuario experto, el jefe de turno de la planta de Filtro de Agua de Chihuahua utilizando el módulo implementado.

Se seleccionaron dos instalaciones el Pozo Aeropuerto 3 y el Tanque Loma Larga con el objetivo de etiquetar las anomalías que estuvieran presentes en el período junio 2018 a junio 2019. Se fue filtrando de mes en mes y luego se procedió a filtrar todos los días lunes, martes y demás días hasta culminar y posteriormente los demás meses. A continuación, se muestra una tabla resumen del proceso de etiquetado de los datos.

	Pozo Aeropuerto 3	**Tanque Loma Larga**
Cantidad de Días	387	387
Total, de anomalías	204	235
Total, días anómalos	160	165
Tiempo promedio de duración de anomalías.	6 horas 12 minutos	2 horas 18 minutos

Tabla 1. Resultados del proceso de etiquetado

3.3. PROCEDIMIENTO PARA LA PREDICCIÓN Y DETECCIÓN DE ANOMALÍAS

Una vez que se tienen datos etiquetados se puede comenzar a emplear procedimientos y algoritmos de *Machine Learning* para la predicción y detección de anomalías.

3.3.1. COMO PROBLEMA DE CLASIFICACIÓN

La *Detección de Anomalías* se ha abordado como un problema particular de una tarea de clasificación, que tiene como objetivo distinguir entre observaciones normales y anómalas [6]. Por tanto, luego de obtener datos etiquetados por un usuario experto es posible realizar experimentos para evaluar la detección de anomalías como problema de clasificación.

Los enfoques principales como problema de clasificación se basan en seleccionar ventanas de tiempos deslizantes de longitud *n* de la serie de tiempo general y establecer criterios para determinar si en esa ventana de tiempo hay existencia de anomalías.

Hay distintas formas de establecer estos criterios, debido a que el interés de este trabajo además de la detección de anomalías se enfoca también en la predicción de las mismas se presentan los siguientes enfoques principales para lograr este proceso.

Enfoque Clasificación usando ventanas de tiempo generales

Para este enfoque se agrupará la serie de tiempo general en subseries de tiempo por días. Luego para cada una de esas subseries se seleccionan las ventanas de tiempos de la siguiente forma:

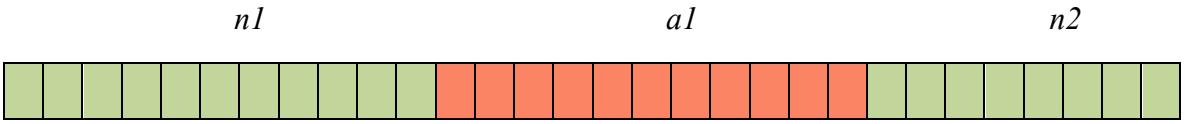

Figura 10. Ventanas de tiempo generales.

La fila anterior representa una subserie de tiempo para un día específico donde las columnas *n1, n2* representan ventanas de tiempos no anómalas y *a1,* una ventana de tiempo anómala. Debido que se quiere predecir las anomalías en tiempo real, de cada una de las ventanas (*n1, n2*) se extraen un vector de características que será los descriptores de esas ventanas. En este caso particular se trata de identificar si dada una ventana de tiempo lo que viene a continuación corresponde a una anomalía, es por eso que los descriptores de la ventana *a2* no se tendrá en cuenta. De esta subserie de tiempo se van extraer los siguientes descriptores que van a conformar los datos que serán pasados a los clasificadores.

	X values	Y values
getDescriptors(n1)	x1	-1
getDescriptors(n2)	x2	1

Tabla 2. Extracción de descriptores de ventanas generales

Los descriptores pueden ser estadísticas básicas tales como, (*media, mediana, desviación estándar*). El vector de características de *n1* se le asigna como etiqueta *-1* debido a que después de esa ventana de tiempo viene una ventana que contiene anomalías no así para la ventana *n2.*

El procedimiento anterior se realiza para todas las subseries de tiempo diarias y se conforma un *dataset* que se utiliza para entrenar clasificadores y luego validar resultados.

Enfoque Clasificación usando ventanas de tiempo deslizantes

Para este enfoque se agrupa la serie de tiempo general en subseries de tiempo por días. Luego en vez de agrupar ventanas de tiempo generales se define un tamaño fijo *n* para la ventana de tiempo y la etiqueta para esa ventana seleccionada es -1 indicando anomalía si el valor *n+1* corresponde a una anomalía de lo contrario es 1. Por ejemplo, para *n*=30 las ventanas de tiempos deslizantes quedarían así:

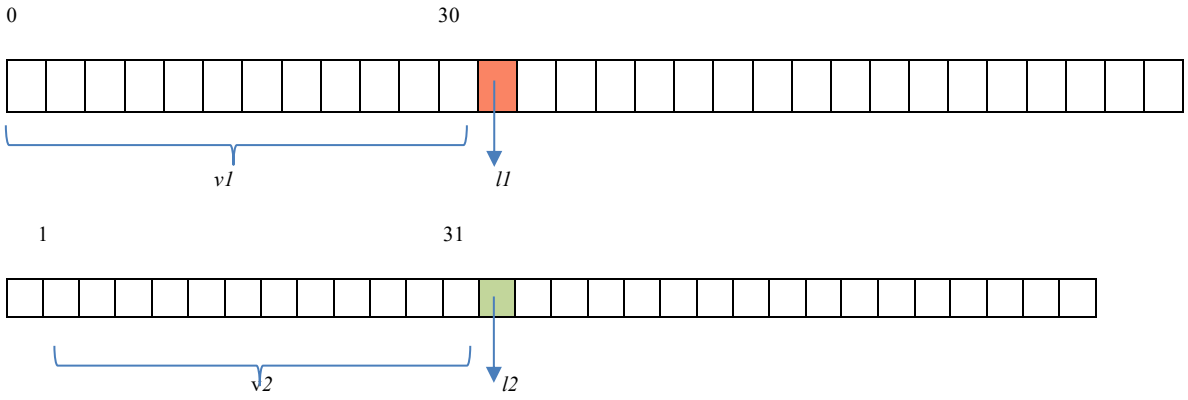

Figura 11. Ventanas de tiempo deslizantes.

	X *values*	Y *values*
getDescriptors(v*1*)	*x1*	y1= si *l1 es anomalía => -1*
getDescriptors(v*2*)	*x2*	y1= si *l1 es normal => 1*

Tabla 3. Extracción de descriptores de ventanas deslizantes

Los descriptores pueden ser estadísticas básicas tales como, (*media, mediana, desviación estándar*). El vector de características de v*1* se le asigna como etiqueta *-1* si el elemento *l1* en la serie de tiempo representa una anomalía de forma contraria a la ventana *v2* se le asigna 1 si el valor *l2* representa un valor normal.

El procedimiento anterior se realiza para todas las subseries de tiempo diarias y se conforma un *dataset* que se utiliza para entrenar los clasificadores y luego validar resultados.

3.3.2. COMO PROBLEMA DE REGRESIÓN

Enfoque de regresión usando la predicción del error

De los enfoques para la predicción y detección de anomalías en series de tiempo, uno de los más populares es viéndolo como un problema de regresión. Este enfoque emplea modelos de generalización explícitos, donde se crea un modelo resumido por adelantado para capturar el comportamiento normal de la instancia monitoreada, y utiliza la desviación entre el comportamiento normal esperado y el comportamiento real como métrica de error para la detección de anomalías [29].

Debido a la popularidad de este enfoque evidenciada en los trabajos [6], [16], [29] [32], [33] se presenta como un paso a seguir por los resultados obtenidos en los trabajos previamente mencionados en la sección 1.1.

Este enfoque emplea un regresor que se entrena en secuencias de datos normales, de modo que aprende el patrón normal de datos. En el momento de la prueba, se espera que las secuencias normales estén bien reconstruidas, mientras que las anómalas no, ya que el modelo no ha visto datos anómalos durante el entrenamiento [6], luego se establece un umbral y el error de predicción será tomado cómo métrica para determinar si valor es anómalo o no.

La popularidad y buenos resultados de algoritmos *Deep Learning* en los últimos años sirven de guía en la investigación para la selección de los mismos. Dentro de estos algoritmos se escoge la *LSTM* para la predicción y detección de anomalías en series de tiempo por la obtención de resultados satisfactorios en su aplicación, principalmente en tareas de secuencia temporal.

De los enfoques presentados en este apartado se obtienen el enfoque y modelos que presentan los mejores resultados y son integrados en un módulo de predicción y detección de anomalías en tiempo real, permitiendo a los profesionales de la JMAS dar soporte para la toma de decisiones.

Para la evaluación y validación de los resultados obtenidos en los experimentos, se utilizan las métricas de evaluación anteriormente expuestas en la sección *2.6.* Estas son

empleadas para comparar los resultados de los diferentes algoritmos utilizados, para así poder escoger la mejor opción a ser utilizada para integrarla en la herramienta web.

3.4. PROCEDIMIENTO PARA LA EXPLORACIÓN DE LOS DATOS DEL SISTEMA DEL CENTRO DE INFORMACIÓN Y SERVICIOS (CIS)

Para la exploración de los datos del sistema *CIS* se deben responder las siguientes interrogantes que permitirá la realización de un análisis de esta información y verificar si es posible utilizar esta información para el proceso de detección y predicción de anomalías en tiempo real.

3.4.1. EXPLICAR EN QUÉ CONSISTEN ESTOS DATOS

En este apartado se explica en qué consisten los datos recopilados por el sistema CIS, y se presenta el diagrama de flujo del procedimiento que origina los estos datos. Además, se da respuesta a las siguientes preguntas: ¿cuál es el catálogo de posibles fallos reportados por los usuarios y cómo sabemos si una queja fue atendida?

3.4.2. EXPLORACIÓN DE LOS DATOS DEL SISTEMA CIS

En este apartado se presenta la exploración de los datos que se tiene actualmente y las siguientes preguntas:

- ¿Cuántos registros existen?
- ¿En qué periodo del año se registran más quejas?
- Reportes por colonias, zonas
- Principales respuestas proporcionadas a los reportes de usuarios.

3.5. INTEGRAR RESULTADOS EN APLICACIÓN WEB

Construir un prototipo funcional de una herramienta Web que hace uso de la integración de los experimentos realizados, puede apoyar en la toma de decisiones a los profesionales de las JMAS para brindar un mejor servicio de la administración del agua. La misma debe permitir una visualización de los datos, permitir el etiquetado manual de los datos y predecir anomalías en tiempo real de las instalaciones de la Red Hidráulica.

Para cada instalación de la Red Hidráulica se debe entrenar un modelo de *ML* que sea capaz de detectar los comportamientos anómalos de esa instalación, esto es necesario debido a que incluso instalaciones del mismo tipo pueden comportarse de manera diferente.

La herramienta debe permitir el entrenamiento de cada uno de estos modelos, así como guardar el estado de estos modelos para poder ser aplicados posteriormente sin tener que ser entrenados en el momento de su uso y para que no se afecte el rendimiento de la aplicación.

Además, debe emplear un *regresor* para predecir los valores futuros de los atributos de las instalaciones de la red hidráulica y luego detectar si en esas predicciones hay comportamientos anómalos usando los algoritmos entrenados con los datos ya etiquetados. De esta forma esta herramienta podría predecir y detectar anomalías en tiempo real que servirán de apoyo para la toma de decisiones a los profesionales de la JMAS.

CAPÍTULO 4: EXPERIMENTACIÓN Y RESULTADOS

En este capítulo se muestra una descripción de los diferentes experimentos realizados en la investigación, así como los resultados obtenidos en los mismos. Estos experimentos sirven como base para la selección de los enfoques y algoritmos de ML que son utilizados para la predicción y detección de anomalías en tiempo real.

4.1. EXPERIMENTO *BASELINE*

Este experimento muestra la necesidad de emplear algoritmos de Machine Learning para la predicción y detección de anomalías en tiempo real en las series de tiempo de las instalaciones de la red hidráulica. El objetivo principal del experimento es demostrar que estas series de tiempo univariadas representan un reto para su clasificación y que no se puede lograr el objetivo de detectar y predecir anomalías en tiempo real con un simple umbral.

4.1.1. DESCRIPCIÓN DEL EXPERIMENTO

1. Partiendo del conjunto de datos sin aplicar ninguna transformación a los mismos. En este caso se toma la serie de tiempo donde se tiene la fecha y hora de la medición y el valor del sensor en ese instante para la instalación.
2. Se escoge un umbral fijo y los valores por encima de este umbral son marcados como anomalías. El umbral fijo es seleccionado por un usuario experto de la JMAS, el cual determina para cada instalación un umbral según el cual los valores por encima se consideran valores anómalos para esa instalación. Esto indica que los valores por encima del umbral conllevan a un comportamiento anómalo de la instalación.
3. Se emplean métricas para evaluar los resultados obtenidos y se visualizan los mismos. Las métricas empleadas son las descritas en la sección 2.6 del marco teórico.

Este experimento, como todos los realizados en este capítulo se realizan para las dos instalaciones de la Red Hidráulica a las que se le etiquetaron los datos usando la herramienta desarrollada y un usuario experto de las JMAS.

4.1.2. BASELINE INSTALACIÓN POZO AEROPUERTO3

La instalación Pozo Aeropuerto3 es una de las instalaciones que fue etiquetada por el usuario experto, de la Tabla 1 se puede apreciar que el total de anomalías etiquetadas de esta instalación es 204. El atributo que se muestra es la *presión* de agua de esta instalación durante el período junio 2018 a julio 2019. A continuación, se muestra la visualización de la *presión* durante el período expuesto con una frecuencia de 1 minuto, además se muestra con círculos rojos los valores que fueron etiquetados como anomalías.

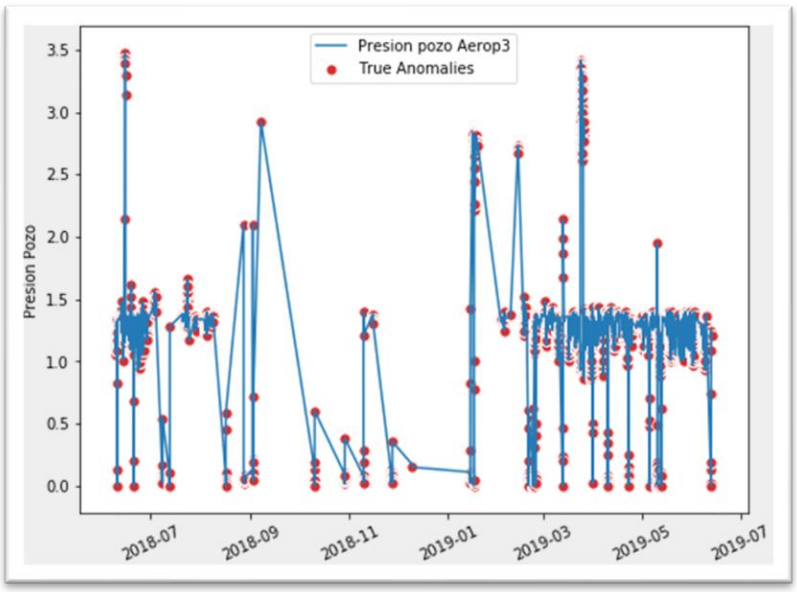

Figura 12. Visualización de los datos reales y anomalías del pozo Aeropuerto 3.

Luego se escoge un umbral fijo que será utilizado para detectar anomalías que supere este umbral. Para este caso se fue probando con diferentes umbrales proporcionados por un usuario experto de la JMAS para escoger el umbral que proporciona los mejores resultados según las métricas de la evaluación. En el caso de esta instalación el umbral seleccionado fue de **1.4.**

Resultados

Se muestra los valores calculados para cada una de las métricas y la curva ROC.

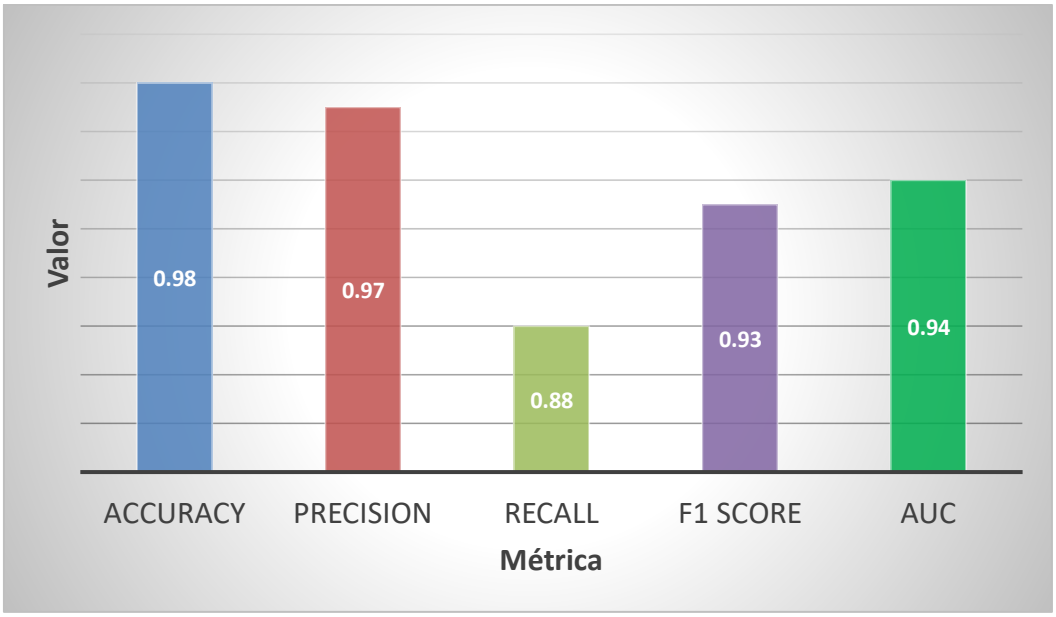

Figura 13 Resultados del experimento *baseline* para el pozo Aeropuerto 3.

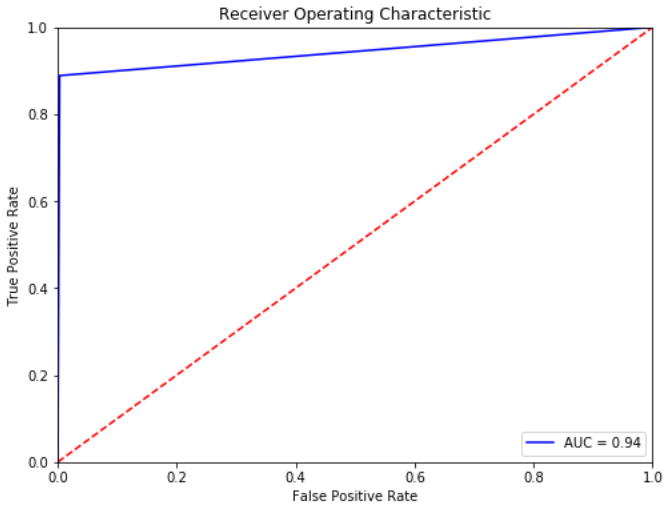

Figura 14. Curva ROC de resultados del experimento baseline para el pozo Aeropuerto 3.

4.1.3. BASELINE INSTALACIÓN TANQUE LOMA LARGA

La instalación Tanque Loma Larga es la otra instalación que fue etiquetada por el usuario experto, de la Tabla 1 se puede apreciar que el total de anomalías etiquetadas de esta instalación es 235. El atributo que se muestrea es el *nivel del agua* de esta instalación durante el período junio 2018 a julio 2019. A continuación, se muestra la visualización del *nivel*

durante el período expuesto con una frecuencia de 1 minuto, además se muestra con círculos rojos los valores que fueron etiquetados como anomalías.

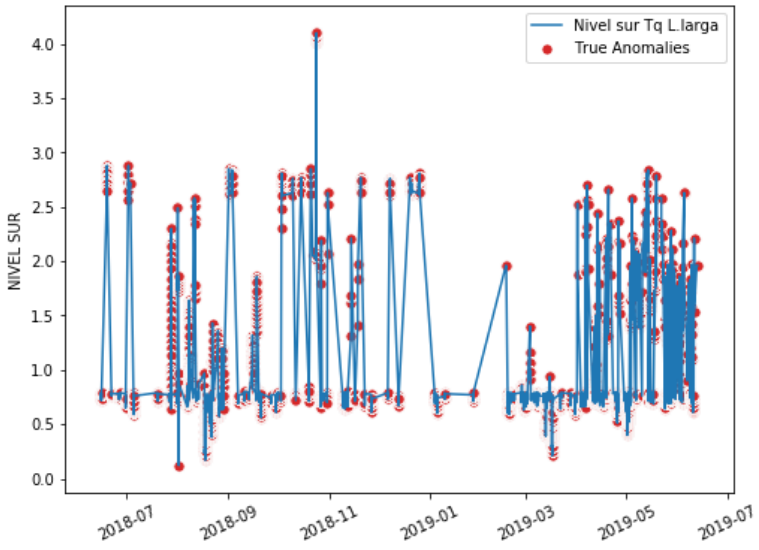

Figura 15. Visualización de los datos reales y anomalías para el tanque Loma Larga.

Luego se escoge un umbral fijo que será utilizado para detectar anomalías que supere este umbral. Para este caso se fue probando con diferentes umbrales proporcionados por un usuario experto de la JMAS para escoger el umbral que proporciona los mejores resultados según las métricas de evaluación. En el caso de esta instalación el umbral seleccionado fue de **0.8**.

Resultados

Se muestra los valores calculados para cada una de las métricas y la curva *ROC*.

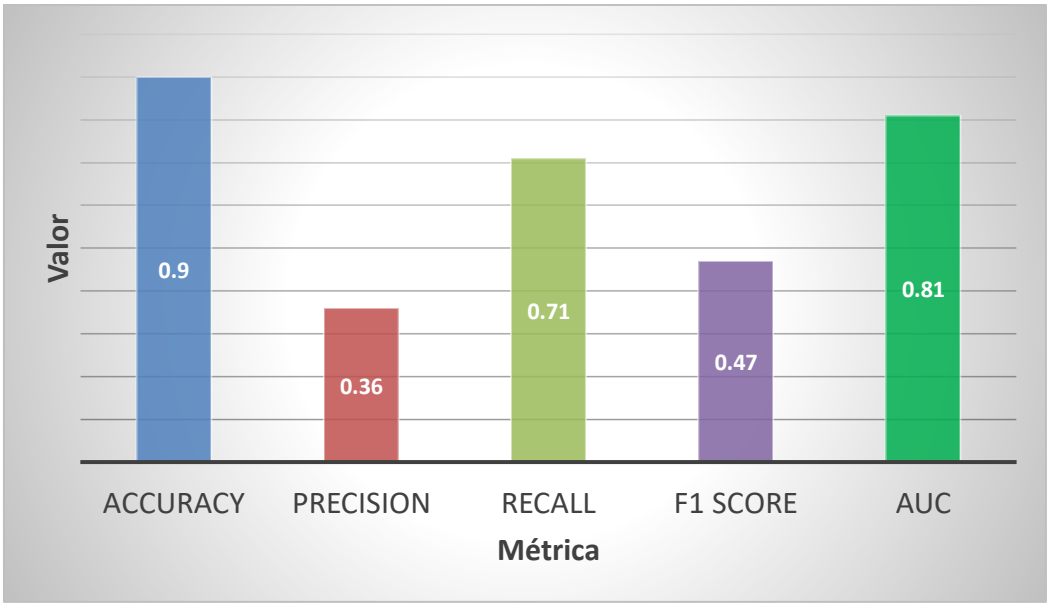

Figura 16. Resultados del experimento *baseline* para el tanque Loma Larga.

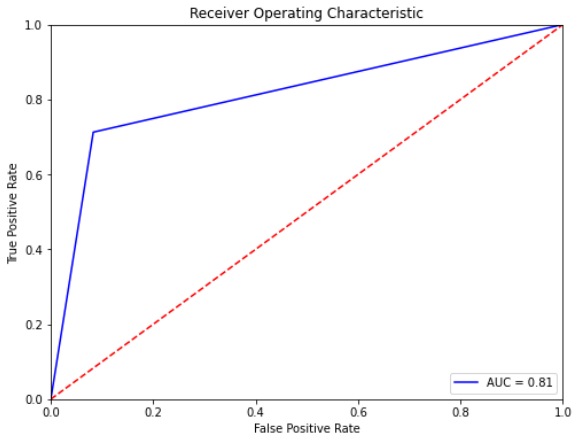

Figura 17. Curva ROC de resultados del experimento para el tanque Loma Larga.

Los resultados anteriormente expuestos deben ser mejorados por los próximos experimentos para evidenciar el uso de enfoques y métodos de *ML,* aunque es importante aclarar que este trabajo no solo se enfoca en detectar las anomalías si no poder predecirlas en tiempo real y este método no cumple estas expectativas.

4.2. EXPERIMENTO CLASIFICACIÓN CON VENTANAS GENERALES

En este experimento se trata predecir si dada una serie de tiempo, lo próximo a esa serie representa una anomalía o no.

DESCRIPCIÓN DEL EXPERIMENTO

1. Se agrupan las series de tiempo por cada día a partir de la serie general.
2. Ser normalizan los datos usando el *StandardScaler* de *Scikit-learn*.
3. Se aplica el procedimiento explicado en sección 3.3.1 del capítulo de metodología relacionado al experimento de clasificación de ventanas generales para la extracción de vectores de características y conformar el *dataset*.
4. Una vez conformado el *dataset* se entrenan un conjunto de clasificadores dividiendo el mismo en 75% para training y el 25% para testing.

INSTALACIÓN POZO AEROPUERTO3

Resultados

Se muestra resultados para cada uno de los clasificadores utilizados.

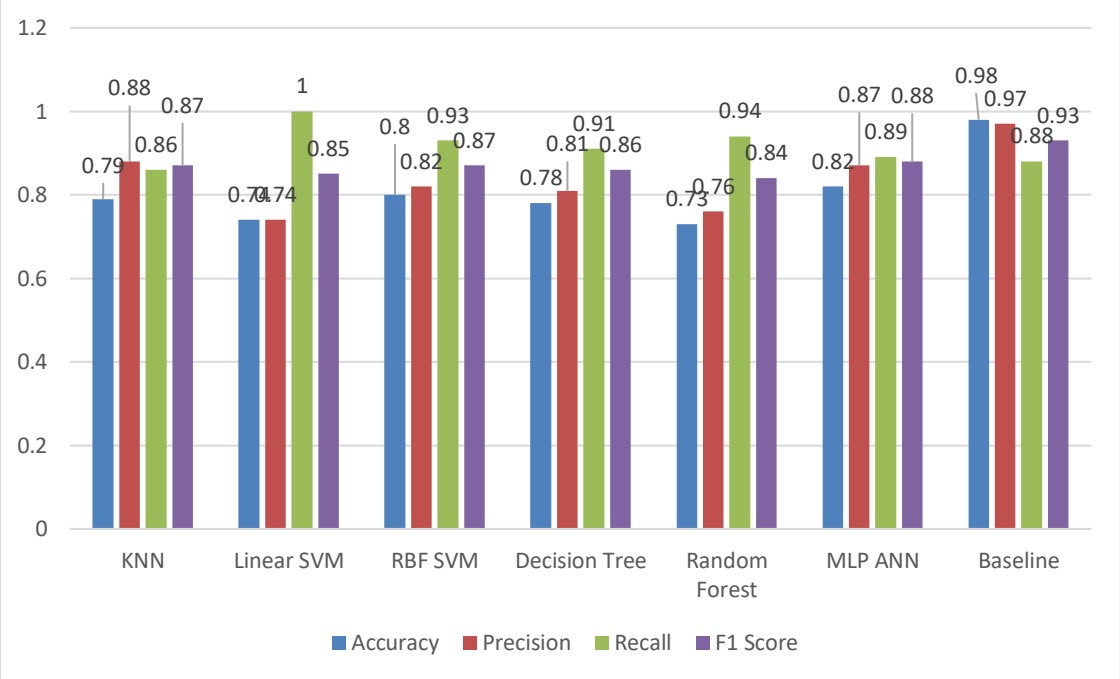

Figura 18. Resultados del experimento clasificación con ventanas generales para el pozo Aeropuerto 3.

INSTALACIÓN TANQUE LOMA LARGA

Resultados

Se muestra resultados para cada uno de los clasificadores utilizados.

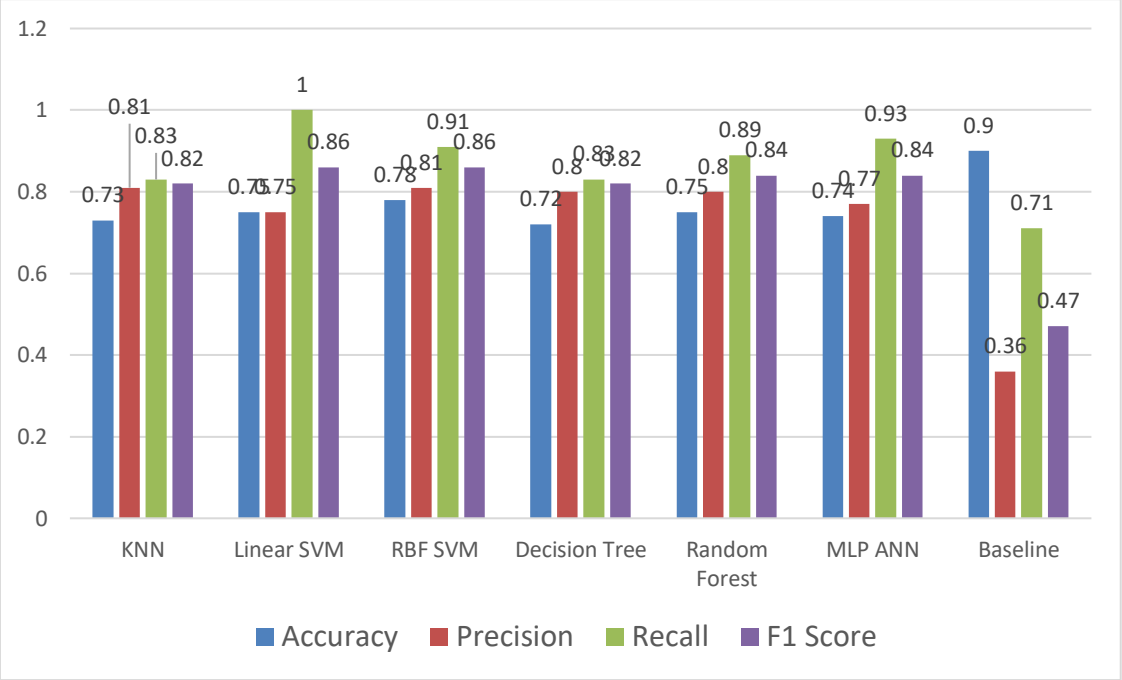

Figura 19. Resultados del experimento clasificación con ventanas generales para el tanque Loma Larga.

Se puede apreciar en los resultados anteriormente expuestos que para el caso de la Instalación Aeropuerto 3, el experimento presentado no supera al *baseline* pues el mejor resultado obtenido fue con *MLP ANN* alcanzando un 0.88 de *F1-score* mientras que *baseline* alcanzó un 0.93. En cambio, para la Instalación Tanque Loma Larga dos clasificadores *LinearSVM* y *RBFSVM* sobrepasaron ambos con 0.86 de *F1-score* al *baseline* con 0.84.

4.3. EXPERIMENTO CLASIFICACIÓN CON VENTANAS DESLIZANTES

En este experimento se trata predecir si dada una serie de tiempo, lo próximo a esa serie representa una anomalía o no.

DESCRIPCIÓN DEL EXPERIMENTO

1. Se agrupan las series de tiempo por cada día a partir de la serie general.
2. Ser normalizan los datos usando el StandardScaler de Scikit-learn.
3. Se aplica el procedimiento explicado en la sección 3.3.1 del capítulo metodología relacionado al experimento de clasificación ventanas deslizantes con un tamaño de la ventana de 30 para la extracción de vectores de características y conformar el *dataset*.
4. Una vez conformado el *dataset* se entrenan un conjunto de clasificadores dividiendo el mismo en 75% para *training* y el 25% para *testing*.

INSTALACIÓN POZO AEROPUERTO3

Resultados

Se muestra resultados para cada uno de los clasificadores utilizados.

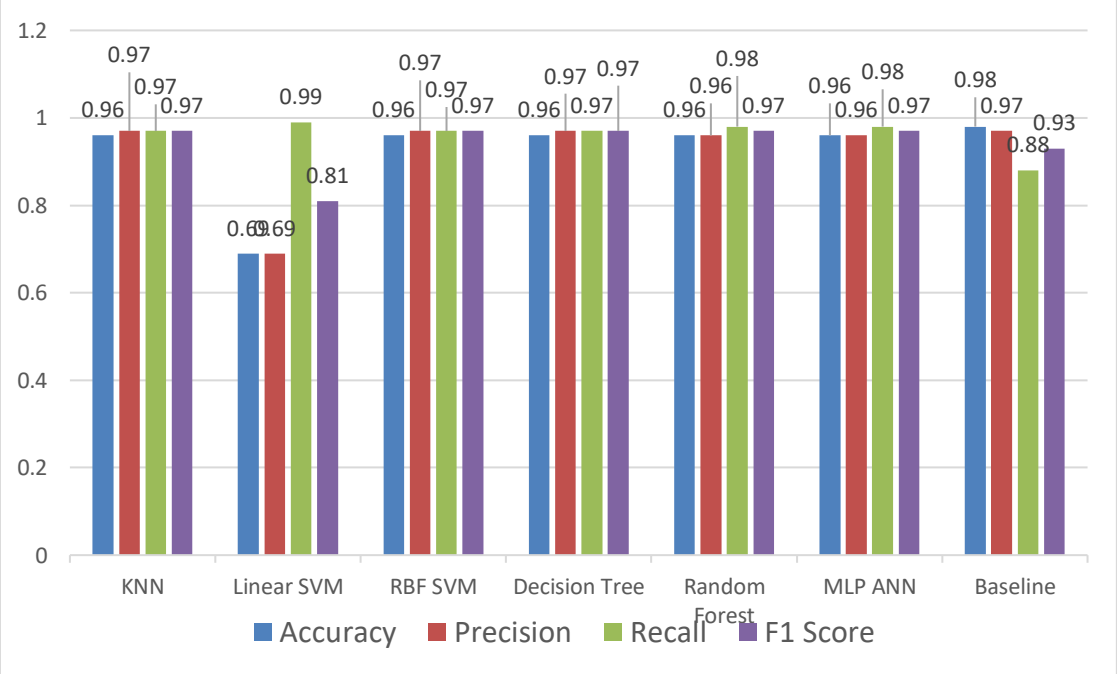

Figura 20. Resultados del experimento clasificación con ventanas deslizantes para el pozo Aeropuerto 3.

INSTALACIÓN TANQUE LOMA LARGA

Resultados

Se muestra resultados para cada uno de los clasificadores utilizados.

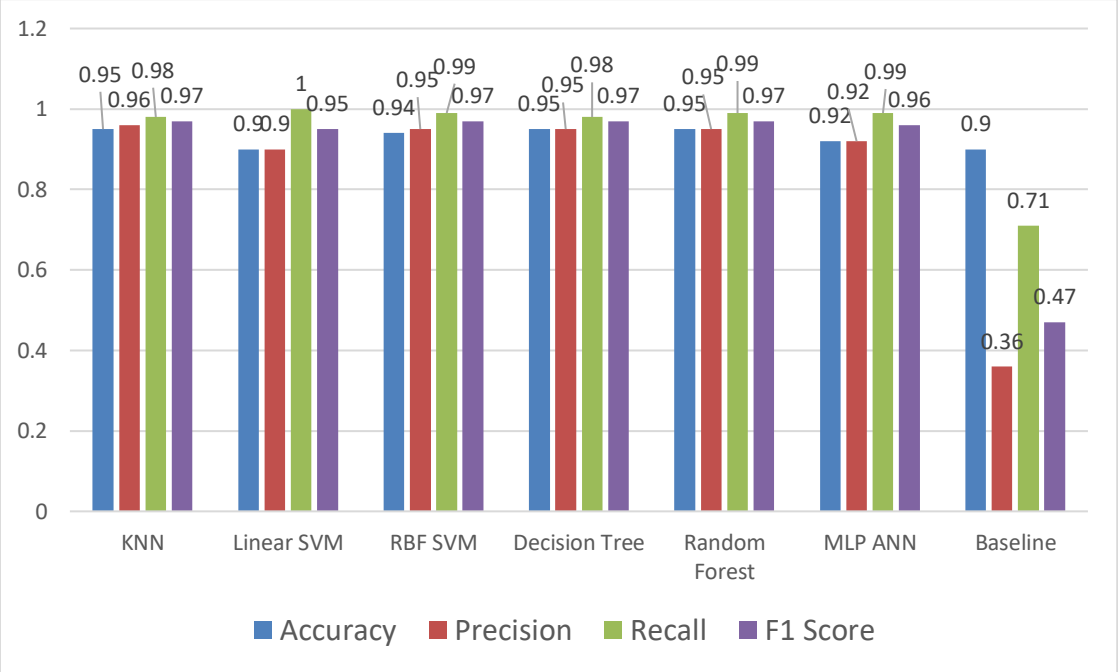

Figura 21. Resultados del experimento clasificación con ventanas deslizantes para el pozo tanque Loma Larga.

Se puede apreciar en los resultados anteriormente expuestos que, para ambas instalaciones, el experimento presentado supera de forma significativa al *baseline* en la mayoría de los clasificadores. Es importante aclarar que el tiempo de entrenamiento y prueba es superior al enfoque de ventanas generales, pero considero que debido a los resultados obtenidos es aceptable.

4.4. EXPERIMENTO LSTM COMO REGRESOR + UMBRAL FIJO

En este experimento se trata de detectar y predecir anomalías a partir de un histórico de datos, usando una *DNN* como regresor en este caso una *LSTM* y utilizar el error predicción para detectar las anomalías una vez seleccionado un umbral fijo.

La arquitectura de la *LSTM* es la siguiente:

```
model.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_2 (LSTM)                (None, 64)                16896
_________________________________________________________________
dropout_2 (Dropout)          (None, 64)                0
_________________________________________________________________
repeat_vector_1 (RepeatVecto (None, 30, 64)            0
_________________________________________________________________
lstm_3 (LSTM)                (None, 30, 64)            33024
_________________________________________________________________
dropout_3 (Dropout)          (None, 30, 64)            0
_________________________________________________________________
time_distributed_1 (TimeDist (None, 30, 1)             65
=================================================================
Total params: 49,985
Trainable params: 49,985
Non-trainable params: 0
_________________________________________________________________
```

Figura 22. Arquitectura de la LSTM.

DESCRIPCIÓN DEL EXPERIMENTO

1. Se agrupan las series de tiempo por cada día a partir de la serie general.
2. Ser normalizan los datos usando el *StandardScaler* de *Scikit-learn*.
3. Se aplica el procedimiento explicado en el apartado de metodología relacionado al enfoque de regresión de ventanas deslizantes con un tamaño de la ventana de 30, para la transformación de los datos y conformar el *dataset*.
4. Una vez conformado el *dataset* se entrenan una red neuronal profunda *LSTM* usando los datos transformados sin anomalías, esto se realiza con el objetivo de que la *LSTM* al predecir un valor que representa una anomalía sea mayor el error de predicción al no haber visto antes estos datos. La *LSTM* se entrena con datos de junio 2018 a abril 2019 y luego se utiliza para predecir el período de abril 2019 a julio2019.
5. Una vez que se entrena la *LSTM* es utilizada para realizar la predicción.
6. Al obtener las predicciones se utiliza el error de predicción para detectar anomalías. Para esto se establece un umbral y los valores que superen ese umbral se marcan como anomalías.

INSTALACIÓN POZO AEROPUERTO3

A continuación, se muestra la serie de tiempo real que se desea predecir por la *LSTM* y luego la predicción y la real juntas:

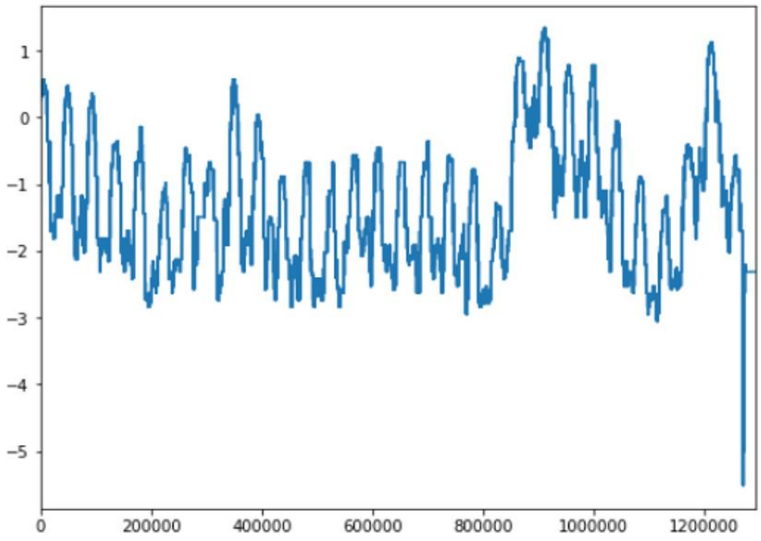

Figura 23. Serie de tiempo real del pozo Aeropuerto 3.

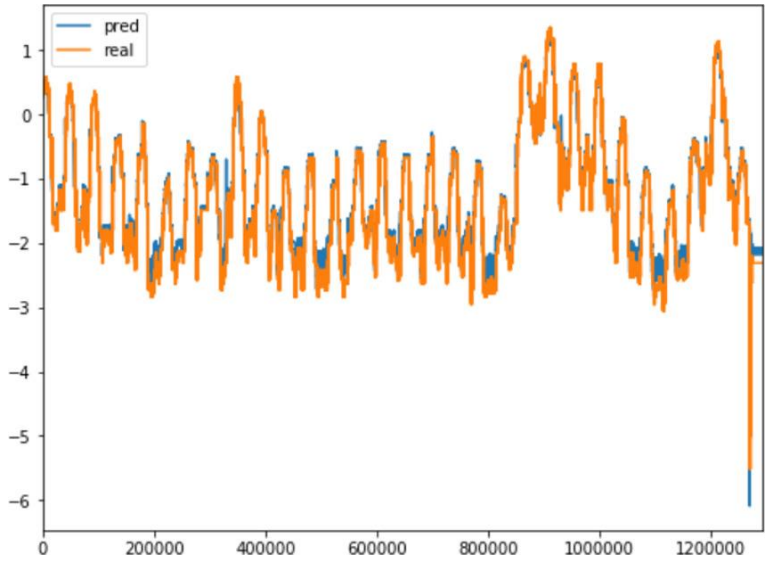

Figura 24. Serie de tiempo real y la predicción del pozo Aeropuerto 3.

Una vez que se tienen las predicciones se calcula la predicción del error y se establece un umbral fijo, para este caso el umbral que se escoge luego de realizar diferentes pruebas es de **0.033625.**

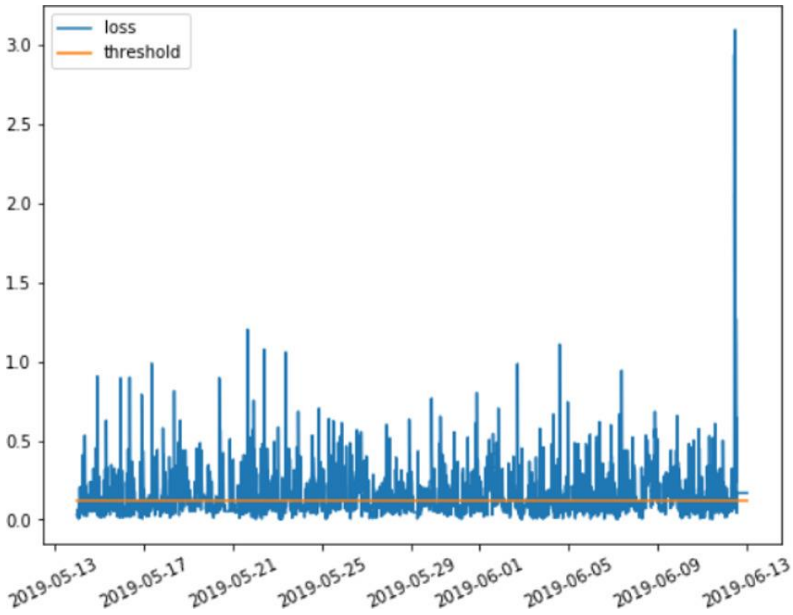

Figura 25. Predicción del error y umbral para el pozo Aeropuerto 3.

Luego se utiliza este umbral para detectar las anomalías sobre las predicciones hechas por la *LSTM*.

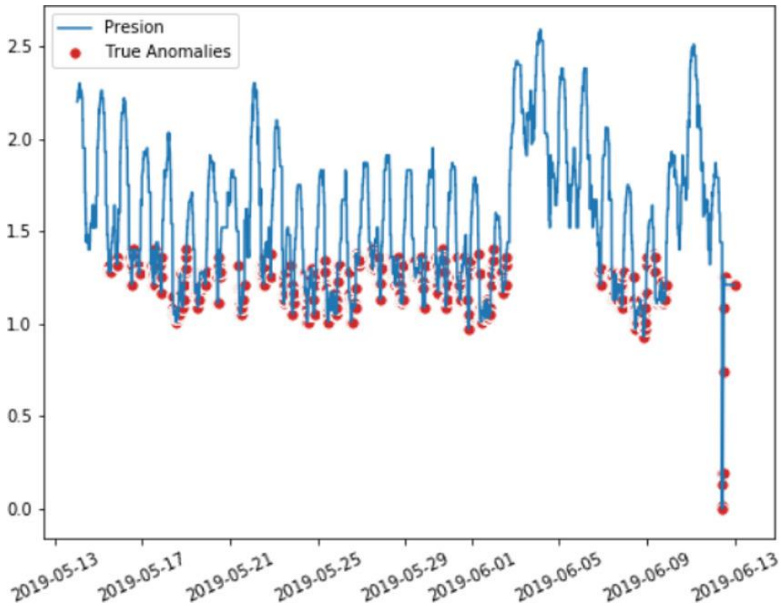

Figura 26. Serie de tiempo real con anomalías reales para el pozo Aeropuerto 3.

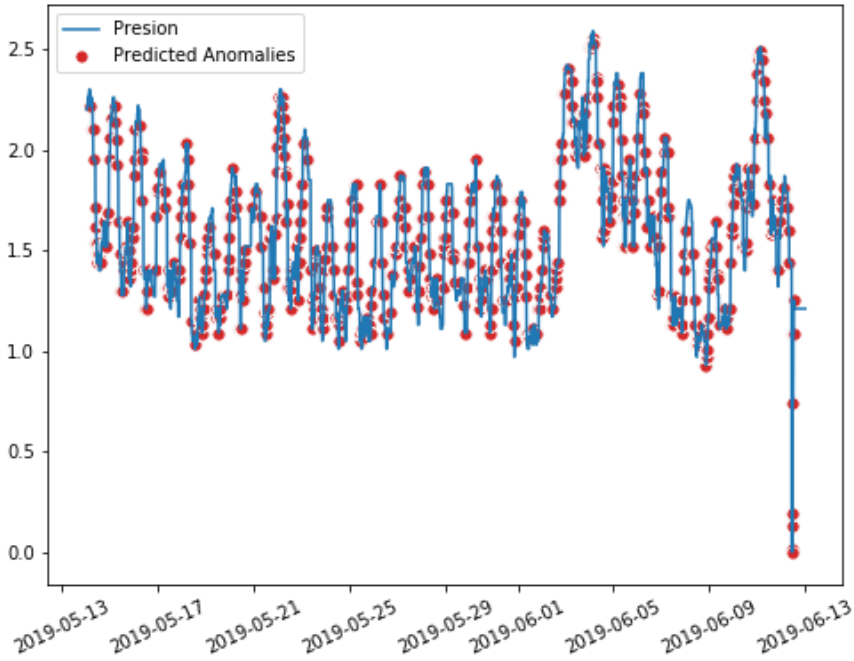

Figura 27. Serie de tiempo real con anomalías predichas por la *LSTM* para el Aeropuerto 3 del experimento *LSTM* como regresor + umbral fijo.

Resultados

Evaluación del regresor.

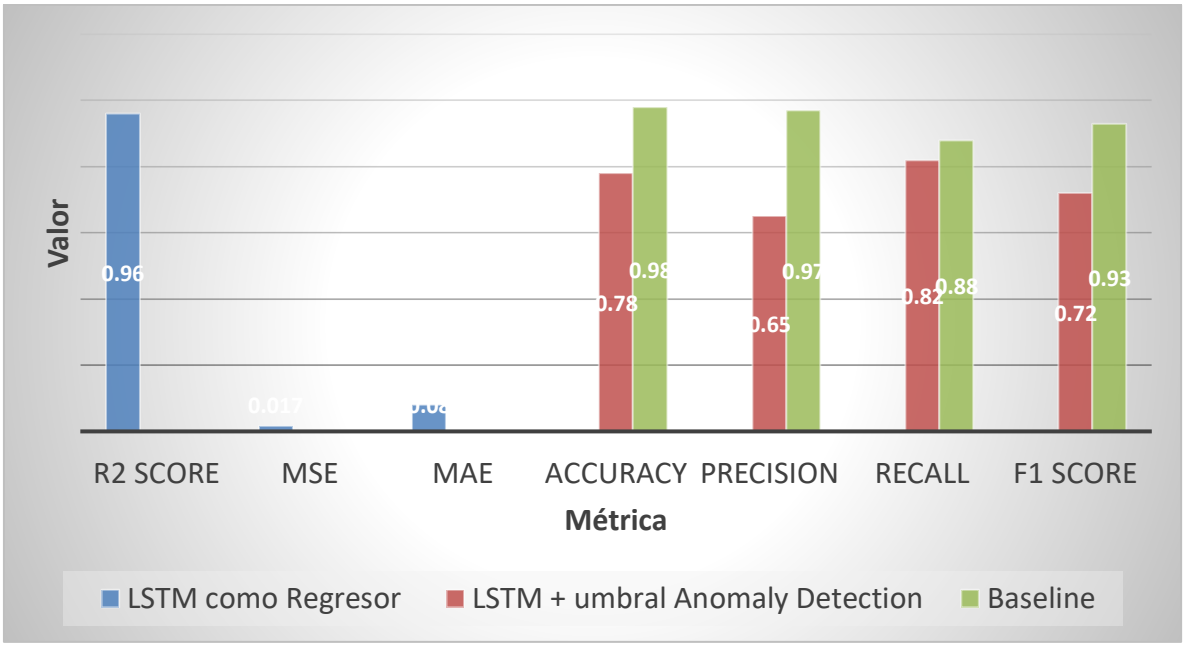

Figura 28. Resultados de evaluación de LSTM como regresor y LSTM + umbral para la detección y predicción de anomalías en el pozo Aeropuerto 3.

INSTALACIÓN TANQUE LOMA LARGA

A continuación, se muestra la serie de tiempo real que se desea predecir por la *LSTM* y luego la predicción y la real juntas:

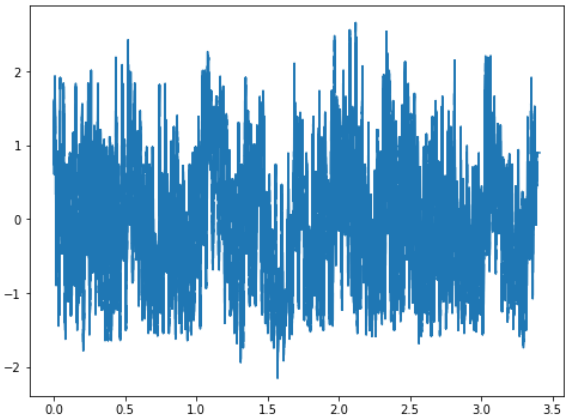

Figura 29. Serie de tiempo real del tanque Loma Larga.

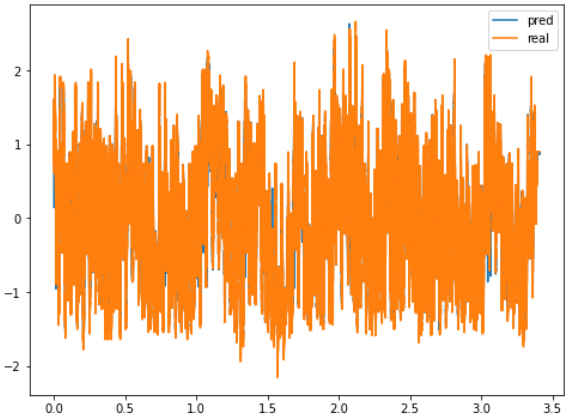

Figura 30. Serie de tiempo real y la predicción del tanque Loma Larga.

Una vez que se tienen las predicciones se calcula la predicción del error y se establece un umbral fijo, para este caso el umbral que se escoge luego de realizar diferentes pruebas es de **0.35.**

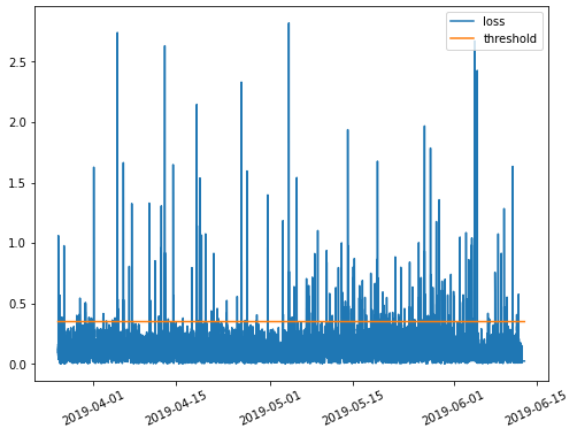

Figura 31. Predicción del error y umbral para el tanque Loma Larga.

Luego se utiliza este umbral para detectar las anomalías sobre las predicciones hechas por la *LSTM*.

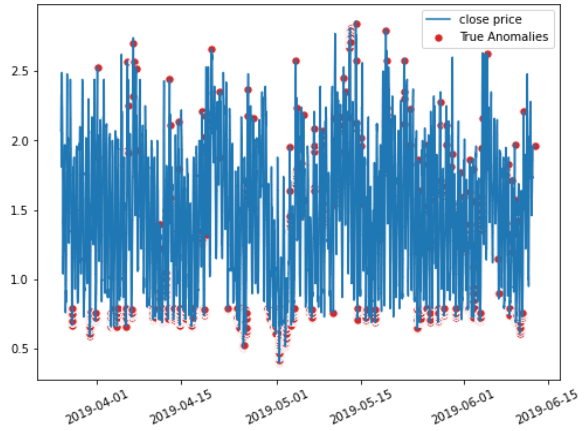

Figura 32. Serie de tiempo real con las anomalías reales para el tanque Loma Larga.

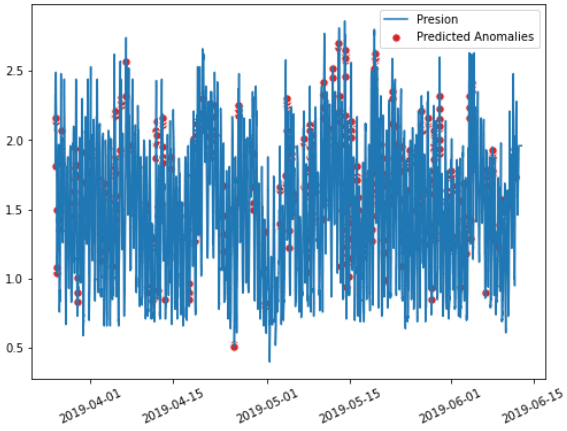

Figura 33. Serie de tiempo real con anomalías predichas por la *LSTM* para el tanque Loma Larga del experimento *LSTM* como regresor + umbral fijo.

Resultados

Evaluación del regresor.

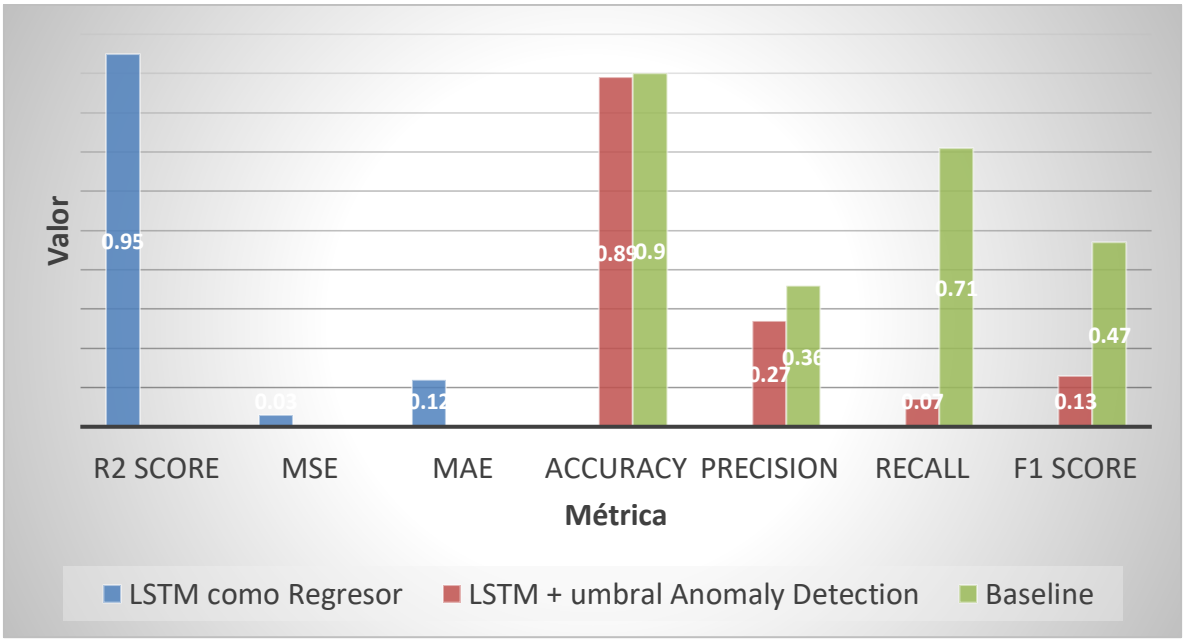

Figura 34. Resultados de evaluación de LSTM como regresor y LSTM + umbral para la detección y predicción de anomalías en el tanque Loma Larga.

Como se puede apreciar la *LSTM* presenta muy buenos resultados como regresor para la predicción de la serie de tiempo. Sin embargo, el enfoque de usar la predicción del error de la *LSTM* como métrica para detectar y predecir las anomalías, no exhibe resultados favorables, además de que no se supera al *baseline* en ninguno de los dos casos. Por lo tanto, se debe probar con otro enfoque.

4.5. EXPERIMENTO LSTM COMO REGRESOR + CLASIFICADOR

Debido a los buenos resultados obtenidos tanto en el experimento de clasificación usando ventanas deslizantes y la *LSTM* como regresor se realiza un nuevo experimento utilizando la *LSTM* para predecir la serie de tiempo y un clasificador entrenado usando el mismo procedimiento del experimento de clasificación de ventanas deslizantes para determinar de la serie de tiempo predicha cuales son anomalías o no.

DESCRIPCIÓN DEL EXPERIMENTO

1. Mismo procedimiento del experimento anterior en la sección 4.5 hasta obtenidas las predicciones de las *LSTM*.

2. Al obtener las predicciones se utiliza un clasificador entrenado según experimento de ventas deslizantes para detectar cuál de los valores predichos son anomalías o no.

INSTALACIÓN POZO AEROPUERTO3

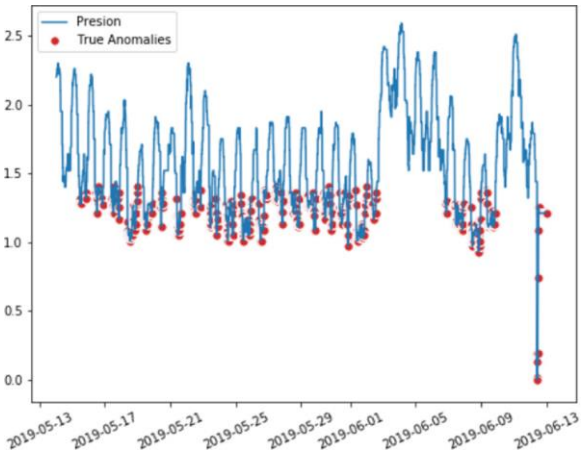

Figura 35. Serie de tiempo real con las anomalías reales para el pozo Aeropuerto 3.

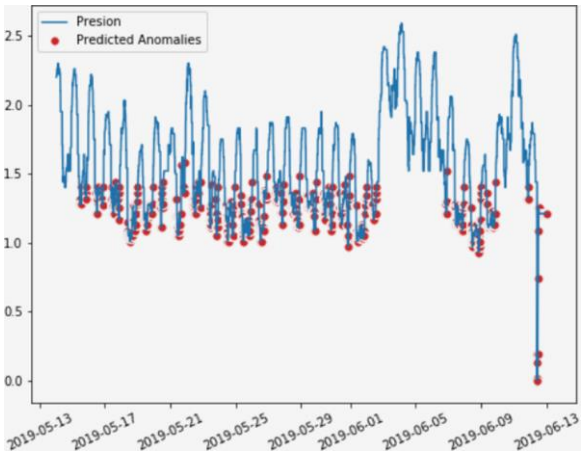

Figura 36. Serie de tiempo real con anomalías predichas por el enfoque usando la LSTM como regresor + DecistionTree como clasificador.

Resultados

Evaluación del proceso de predicción de anomalías.

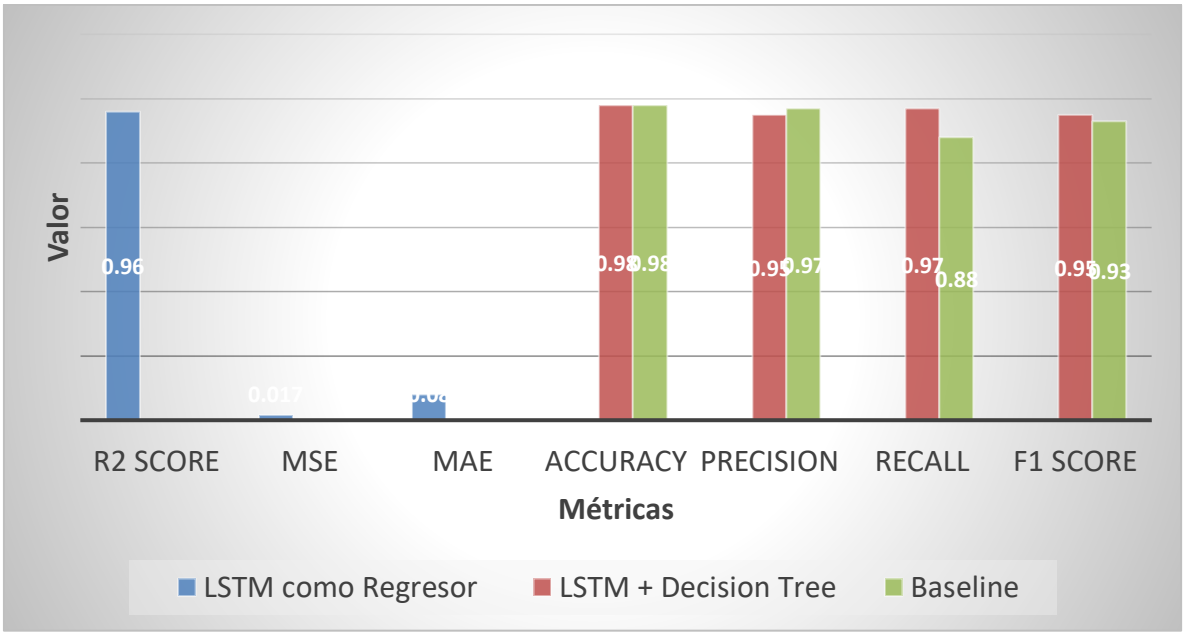

Figura 37. Resultados de evaluación de LSTM como regresor y LSTM + DecisionTree para la detección y predicción de anomalías en el pozo Aeropuerto 3.

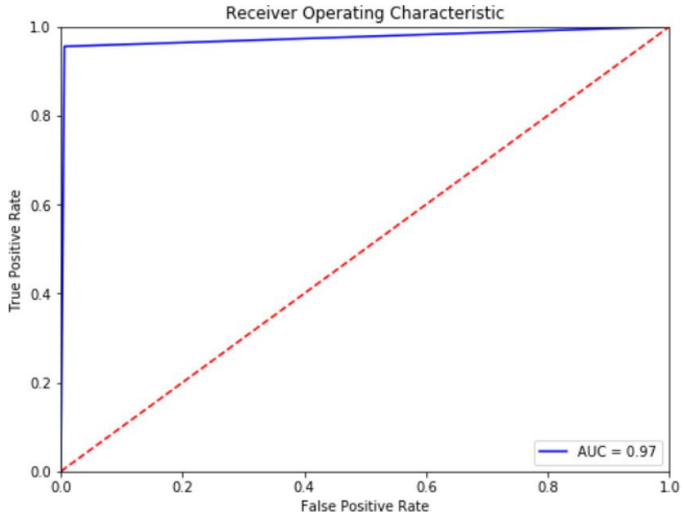

Figura 38. Curva ROC de resultados para el experimento LSTM como regresor + DecistionTree como clasificador para el pozo Aeropuerto 3.

Se puede apreciar muy buenos resultados para detección de anomalías usando este enfoque combinado el cual pretende ser una opción viable para la predicción y detección de anomalías en tiempo real. Se debe extender este experimento a la otra instalación y presentar resultados similares.

INSTALACIÓN TANQUE LOMA LARGA

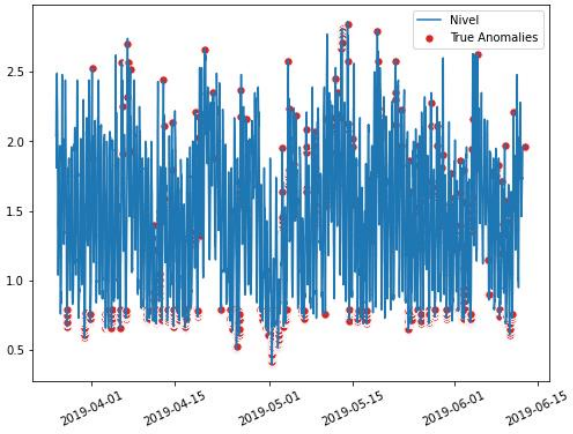

Figura 39. Serie de tiempo real con las anomalías reales para el tanque Loma Larga.

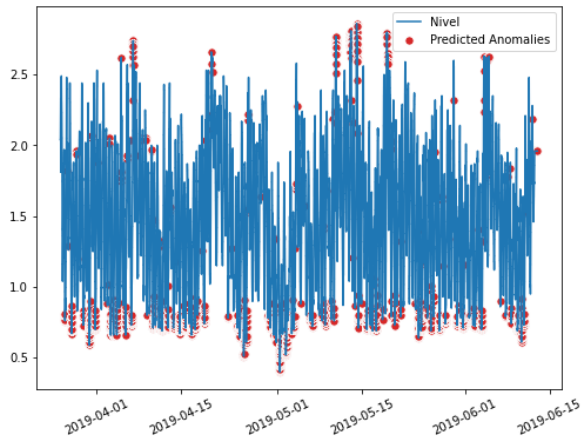

Figura 40. Serie de tiempo real con anomalías predichas por el enfoque usando la LSTM como regresor + DecistionTree como clasificador para el tanque Loma Larga.

Resultados

Evaluación del proceso de predicción de anomalías.

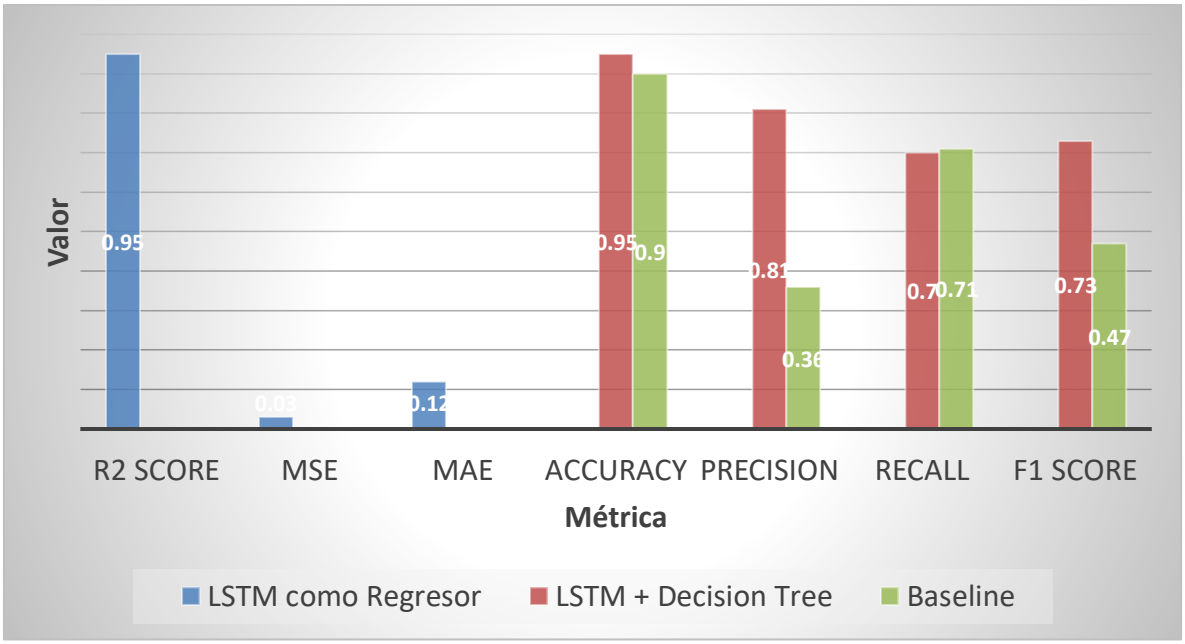

Figura 41. Resultados de evaluación de LSTM como regresor y LSTM + DecisionTree para la detección y predicción de anomalías en el tanque Loma Larga.

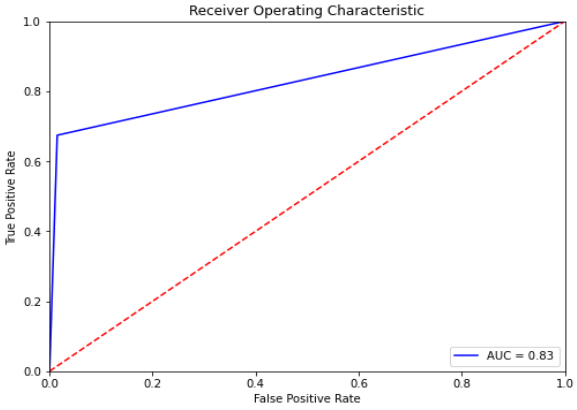

Figura 42. Curva ROC de resultados para el experimento LSTM como regresor + DecistionTree como clasificador para el tanque Loma Larga.

Como se puede apreciar este enfoque los resultados presentados superan al propuesto como baseline por lo tanto puede ser considerado como enfoque a usar en la herramienta de predicción y detección de anomalías en tiempo real.

CONCLUSIONES DEL CAPÍTULO

Luego de analizar los experimentos anteriores realizados se aprecia que la arquitectura de la red neuronal profunda brinda muy buenos resultados en la tarea de

regresión para predecir los valores reales tanto de la presión como del nivel de las instalaciones de la red hidráulica. En el caso de predecir las anomalías en tiempo real el enfoque que presenta los mejores resultados es el de utilizar la red neuronal como regresor y *DecisionTree* como clasificador. En el último experimento realizado se muestra además que mientras el regresor ofrezca mejores resultados también mejora el proceso de predicción de las anomalías en tiempo real.

CAPÍTULO 5: EXPLORACIÓN DE DATOS DEL SISTEMA CIS

5.1. EXPLICAR EN QUÉ CONSISTEN ESTOS DATOS

Estos datos son recopilados mediante el reporte de faltas de agua, fugas que los usuarios comunican vía telefónica a través del Centro de Información y Servicio (CIS) de la JMAS Chihuahua, durante el período junio 2018 a julio 2019. Este centro brinda atención personalizada las 24 horas los 365 días del año a través del número único de telefonía 073. Una vez que los usuarios registran el reporte se le entrega a los mismos un número de folio a través del cual los usuarios pueden darle seguimiento al estado del reporte emitido en la dirección http://www.jmaschih.gob.mx/

En cada reporte de usuario se registra la siguiente información:

- No. folio
- Tipo de problema: problema planteado por el usuario
- Contrato usuario: número de contrato del usuario
- Estado del folio: estado del reporte (Asociado, Cancelado, Terminado)
- Zona: zona referente al municipio de Chihuahua (zona 1 a zona 6).
- Nombre del usuario
- Teléfono
- Dirección
- Colonia: colonia del municipio de Chihuahua
- Fecha
- Solución/Respuesta
- Observaciones
- Operador recibe llamada
- Responsable zona

Adicionalmente a los reportes, el CIS ofrece otra información útil y precisa a la comunidad entre los que se encuentran:

- Consultas de horarios de suministro de agua en las distintas colonias de la ciudad.
- Saldo en el recibo de agua
- Requisitos para trámites de contratos.
- Cambios de titular o dictámenes de factibilidad.
- Recomendaciones para el ahorro de agua.
- Difusión de las redes sociales de la JMAS y el portal gubernamental http://www.jmaschih.gob.mx. En este portal se pueden realizar servicios de pago en línea, generar facturas, y darle seguimiento al folio de atención del CIS [34].

DIAGRAMA DE FLUJO DEL PROCEDIMIENTO QUE LOS ORIGINA

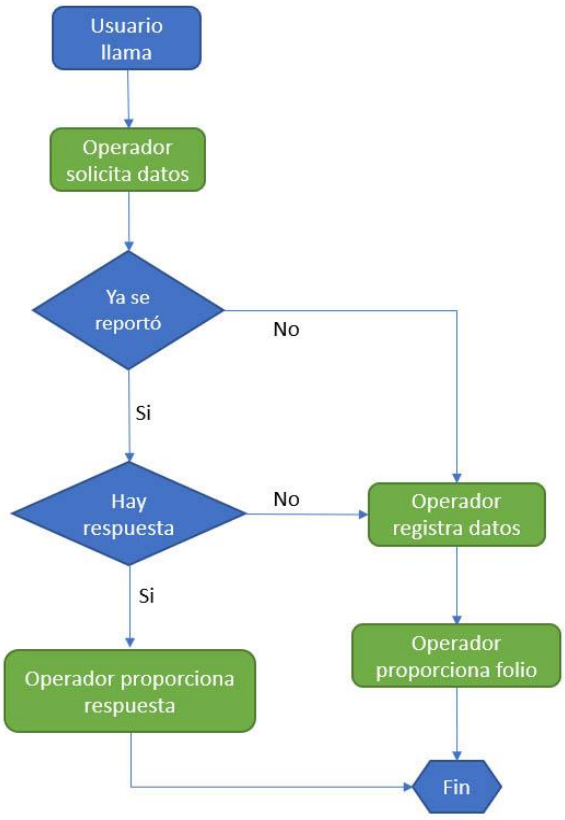

Figura 43. Diagrama de flujo de cómo se originan los datos del sistema CIS.

CATÁLOGO DE POSIBLES FALLOS REPORTADOS POR LOS USUARIOS

El sistema CIS gestiona diversos tipos de fallos reportados por los usuarios, pero los datos que se tienen actualmente presentan reportes de Falta de Agua, Fuga en la Calle, Fugas de Agua recuperada reportados por los usuarios en el período de junio 2018 a julio 2019.

CÓMO SABEMOS SI UNA QUEJA FUE ATENDIDA

Como se presenta en el diagrama de flujo una vez que lo usuarios reportan el problema se le es entregado un número de folio del reporte, mediante el cual el usuario puede revisar el estado del reporte en la dirección http://www.jmaschih.gob.mx y examinar la respuesta otorgada al reporte.

5.2. EXPLORACIÓN DE LOS DATOS

Se tienen reportes de *faltas* y *fugas de agua* en el período junio 2018 a julio 2019.

¿CUÁNTOS REGISTROS EXISTEN?

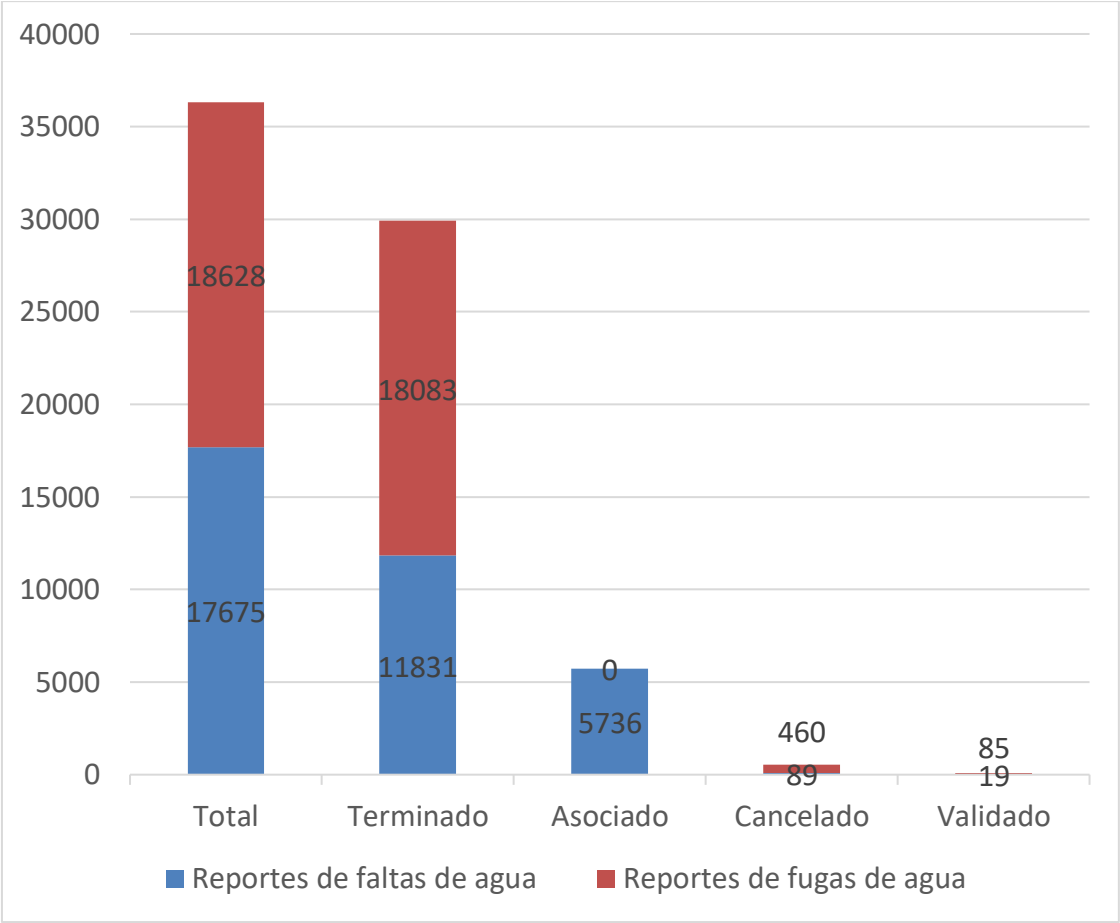

Figura 44. Cantidad de reportes que se tienen en el CIS.

Como se muestra en la figura 43 se tiene un total de 18628 reporte de fuga de agua y 17675 de falta de agua, de los cuales han pasado a estado terminado o sea que se les dio respuesta a los usuarios, 18083 en el caso de la fuga de agua y 11831 respecto a la falta de agua.

¿EN QUÉ PERIODO DEL AÑO SE REGISTRAN MÁS QUEJAS?

En el caso de reportes de falta de agua se puede apreciar en la figura 44 que se registran más quejas durante los meses de mayo y junio mientras que en los meses de octubre y noviembre son menos los reportes. En el caso de los reportes de fugas de agua se puede apreciar que en el mes de enero es donde se reportan más quejas y el menor número de quejas sobre fugas de agua se presenta en el mes de junio, lo que tiene sentido pues este mes es

donde más falta de agua se presenta, manteniendo valores semejantes en los otros meses para las fugas.

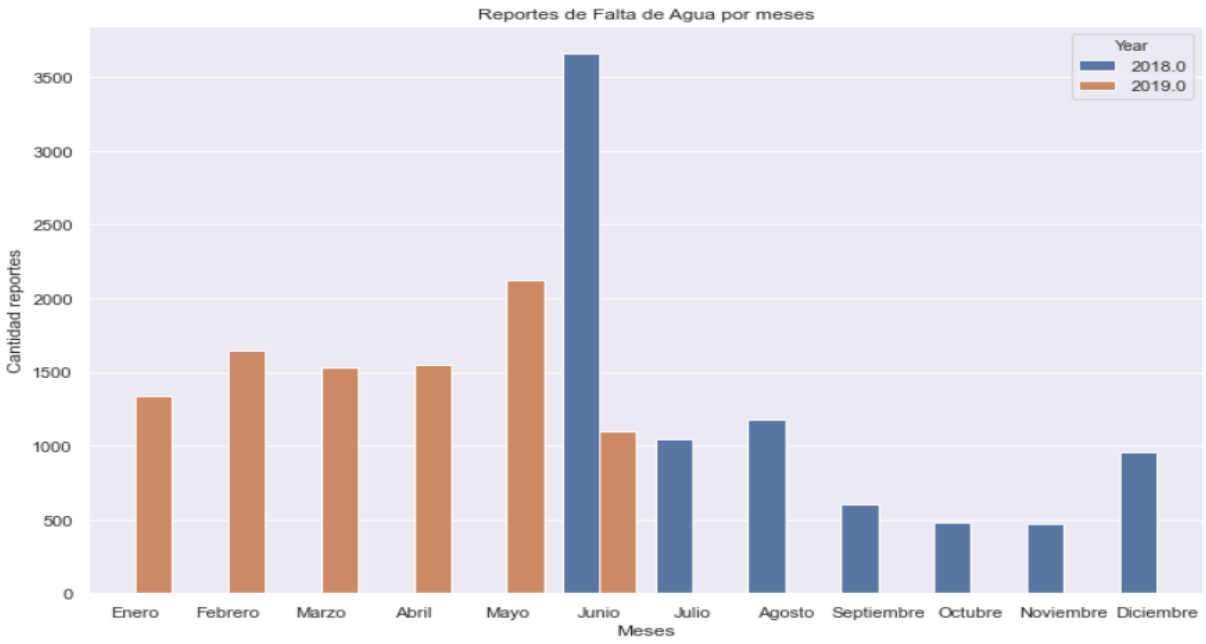

Figura 45. Reportes de falta de agua por meses del sistema CIS.

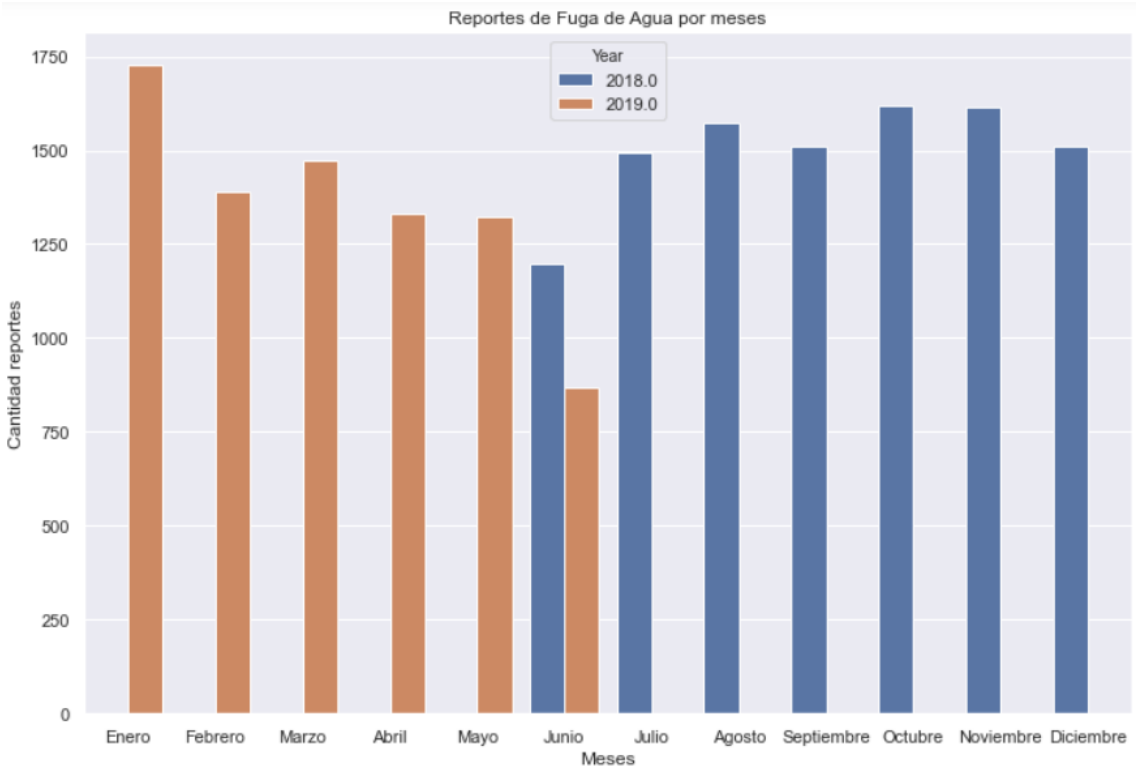

Figura 46. Reportes de fuga de agua por meses del sistema CIS.

¿SE REPORTAN FUGAS Y FALTAS TODOS LOS DÍAS?

Se realiza un análisis de la información recopilada para detectar si existe algún día que no se hayan reportados faltas o fugas de agua, a continuación, se presentan los meses con mayor y menor cantidad de reportes. Las gráficas se visualizaron de forma tal que pudiera presenciarse los reportes por día de cada mes, se muestra el mes de junio 2018 mes con mayor cantidad de reportes y el mes de septiembre 2018 uno de los meses con menor cantidad de reporte, en este último se pueden ver los días con menor cantidad de reportes los cuales son el día 8 con dos reportes y el día 23 con dos reportes. Luego del análisis realizado se concluye que en los datos que se tienen no existe un día en el que no se haya reportado un problema.

Figura 47. Distribución de reportes por día para el mes de junio 2018.

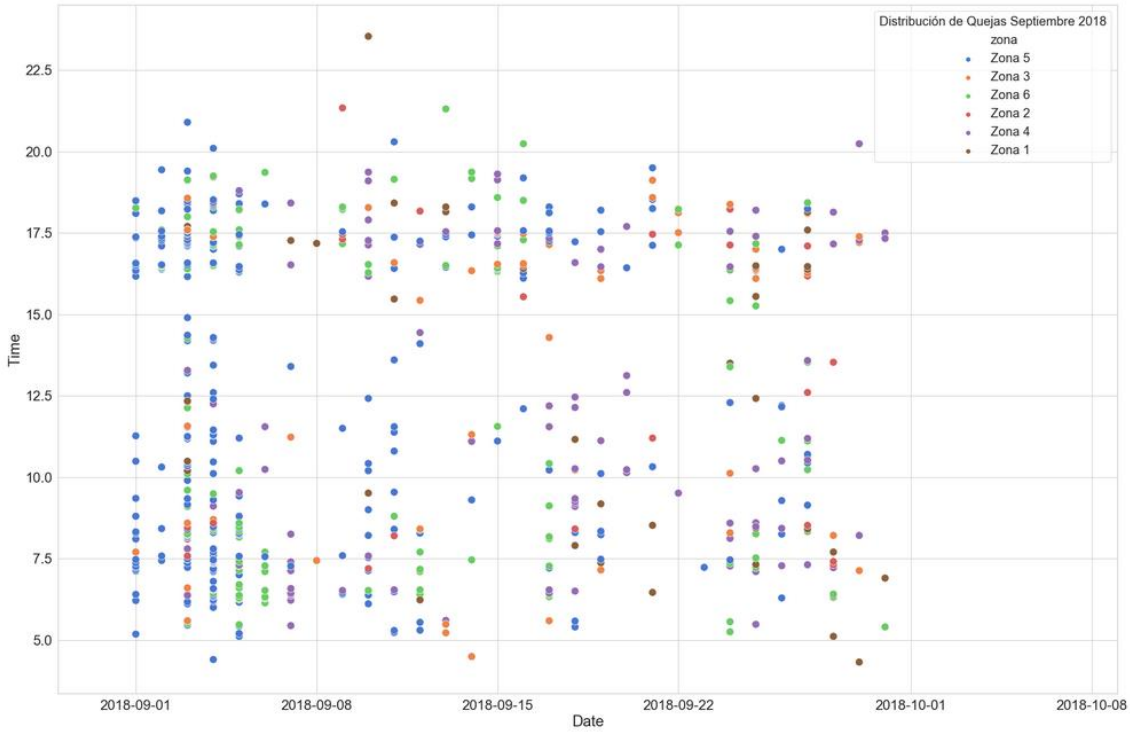

Figura 48. Distribución de reportes por día para el mes de septiembre 2018.

¿EXISTEN USUARIOS QUE CONSTANTEMENTE ESTEN REPORTANDO QUEJAS? ¿O TODOS SON ÚNICOS USUARIOS?

A continuación, se muestran gráficas donde se puede apreciar los 35 usuarios que más reportes han realizado en cuanto a la falta y fugas de agua correspondientemente. Esto demuestra que hay hogares que constantemente están reportando problemas relacionados a la falta y fugas de agua, aunque también que gran parte de los reportes registrados son de forma anónima en el caso de las fugas de agua.

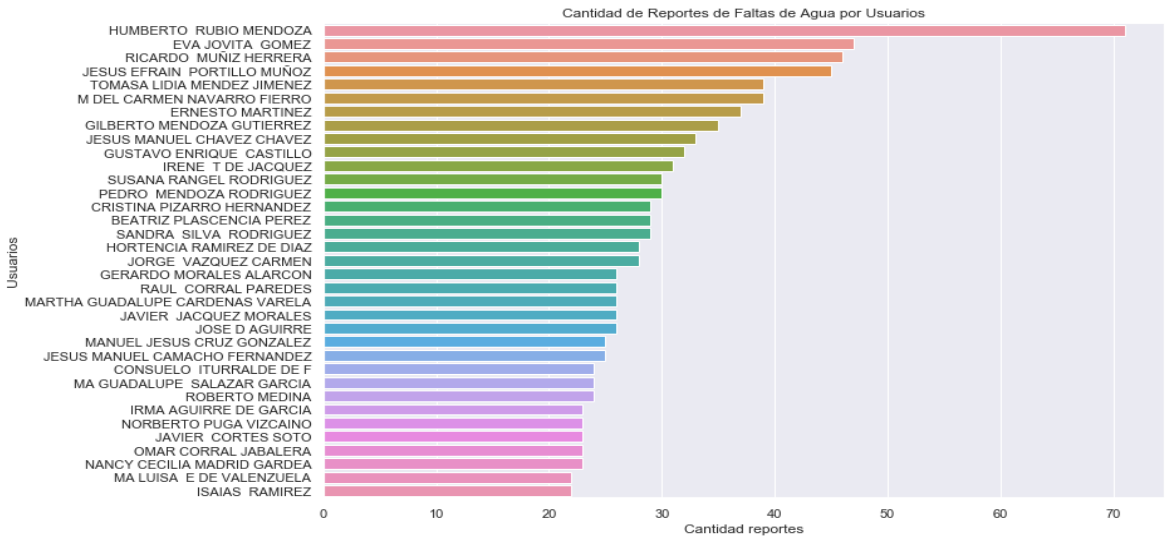

Figura 49. Cantidad de reportes de faltas de agua por usuarios.

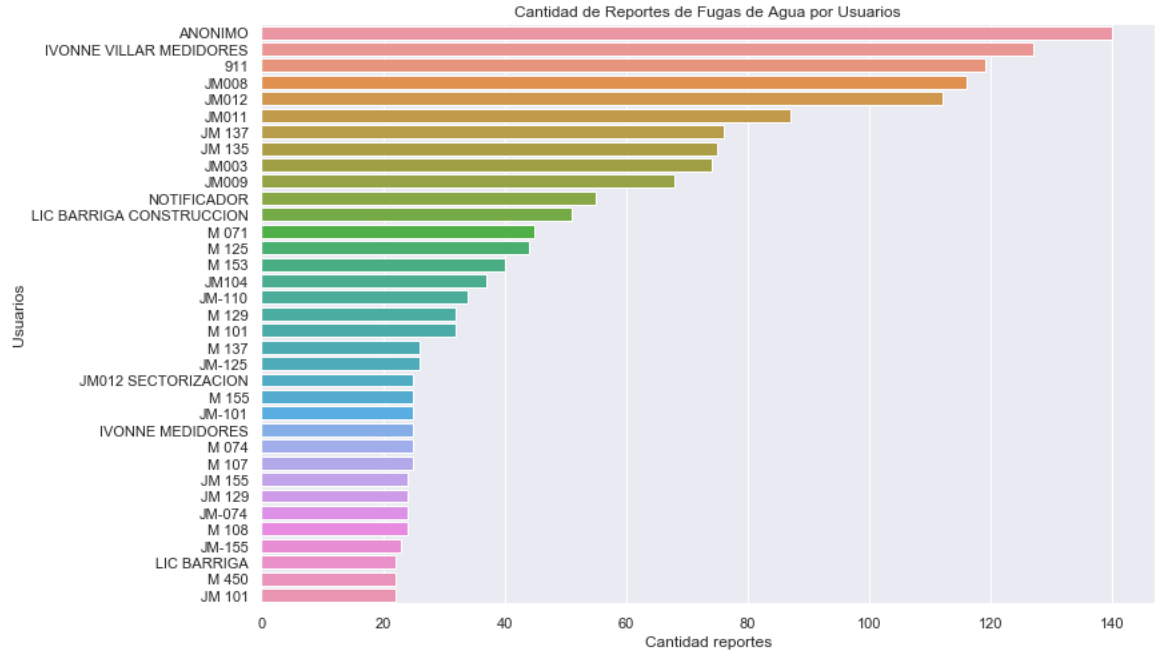

Figura 50. Cantidad de reportes de fuga de agua por usuarios.

¿CUÁLES SON LAS ZONAS QUE MÁS QUEJAS PRESENTA?

Se puede apreciar que tanto para el caso de reporte de falta de agua como para el caso de fugas de agua la zona con más reportes es la zona 5, con más de 6500 reportes de falta de agua y más de 2000 reportes respecto a fugas.

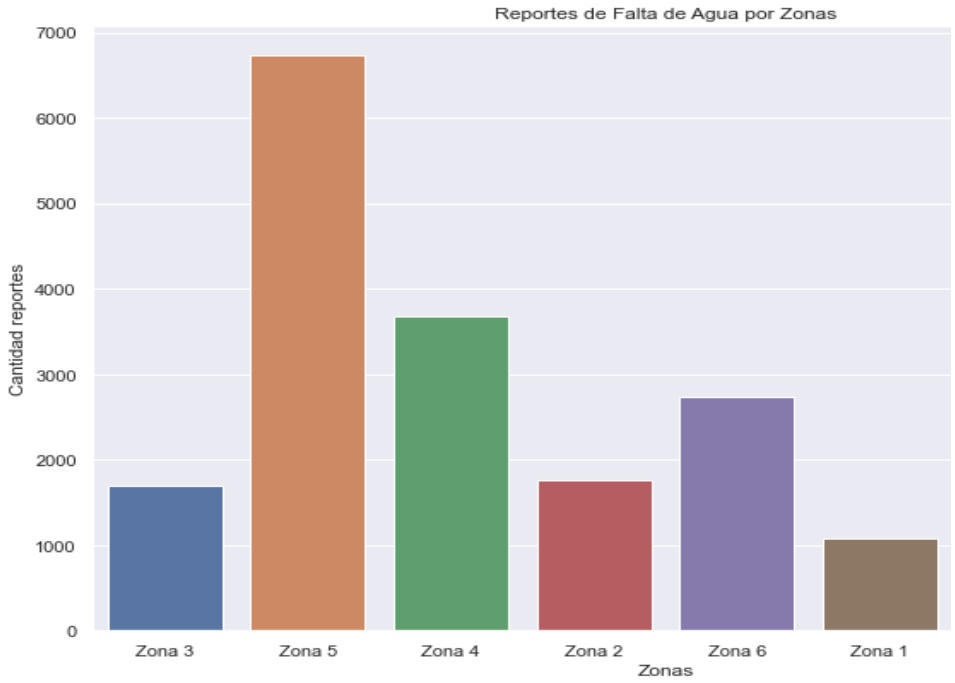

Figura 51. Reportes de falta de agua por zonas.

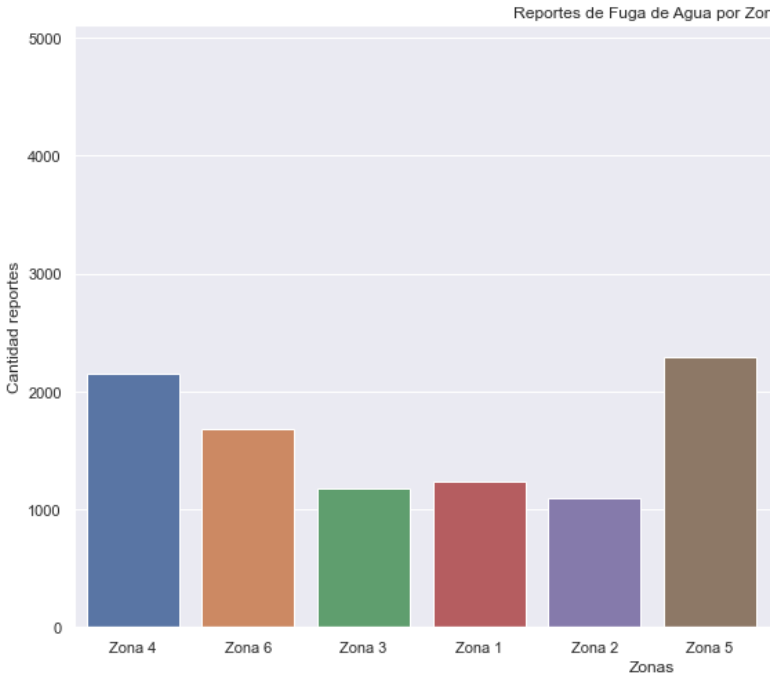

Figura 52. Reportes de fuga de agua por zonas.

¿CUÁLES SON LAS COLONIAS QUE MÁS QUEJAS PRESENTAN?

A continuación, se muestran las 35 colonias con más reportes tanto de faltas y fugas de agua correspondientemente. En cuanto a la falta de agua la colonia que más reportes presenta es Santa Rosa y en el caso de fugas de agua es la colonia Riberas de Sacramento.

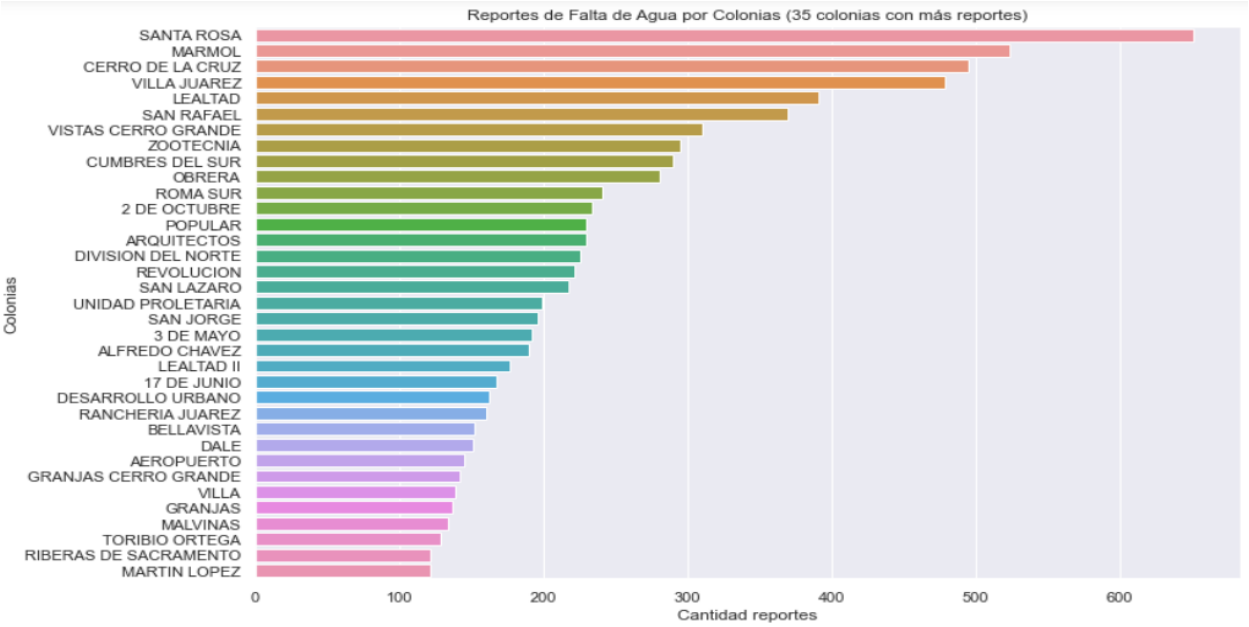

Figura 53. Reportes de falta de agua por colonias.

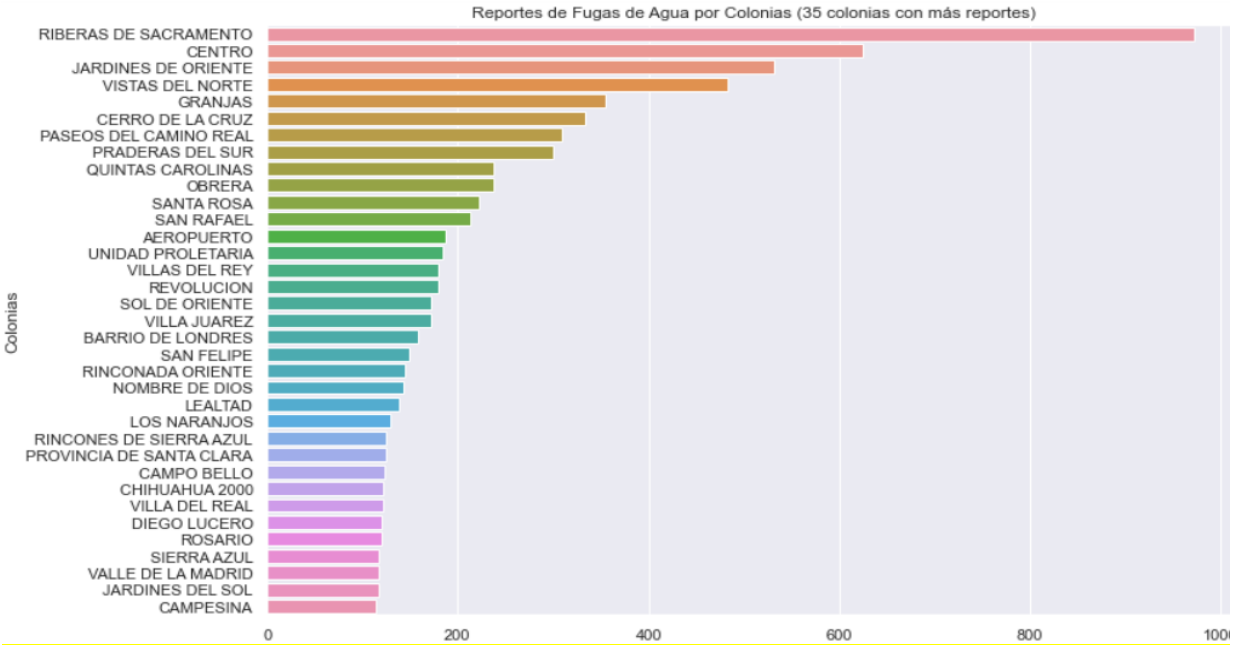

Figura 54. Reportes de fuga de agua por colonias.

¿CUÁLES SON LAS RESPUESTAS MÁS COMUNES?

Se muestra a continuación las respuestas más comunes proporcionadas a los usuarios respecto a los reportes realizadas de faltas y fugas de agua.

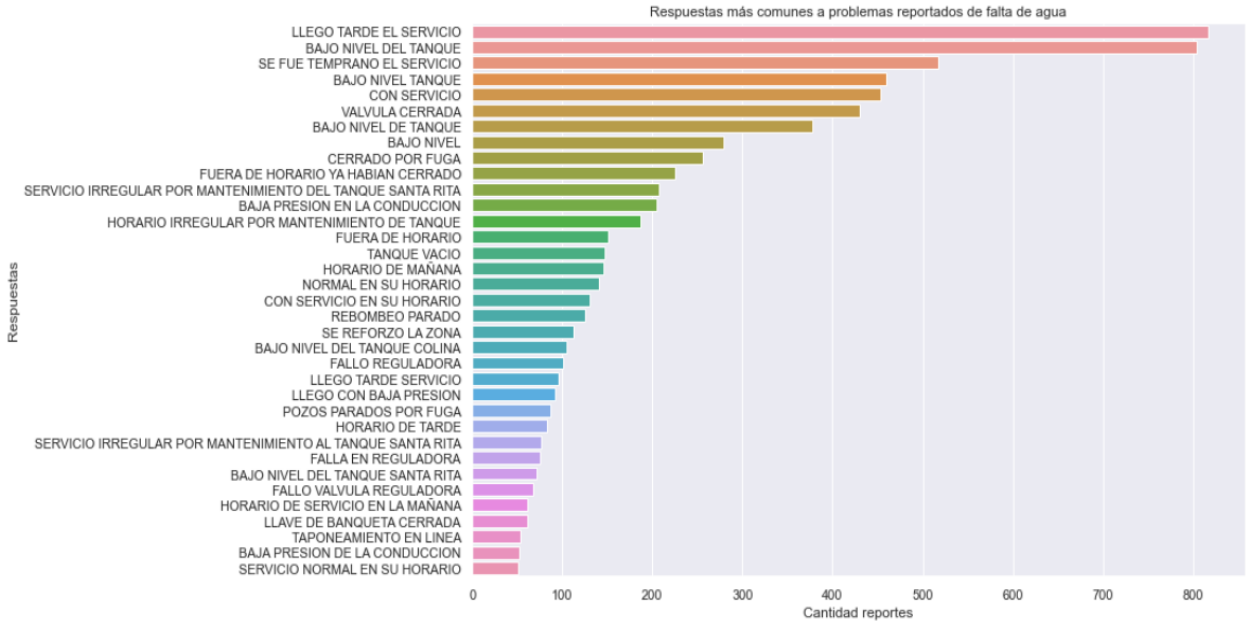

Figura 55. Respuestas más comunes a reportes de falta de agua a los usuarios.

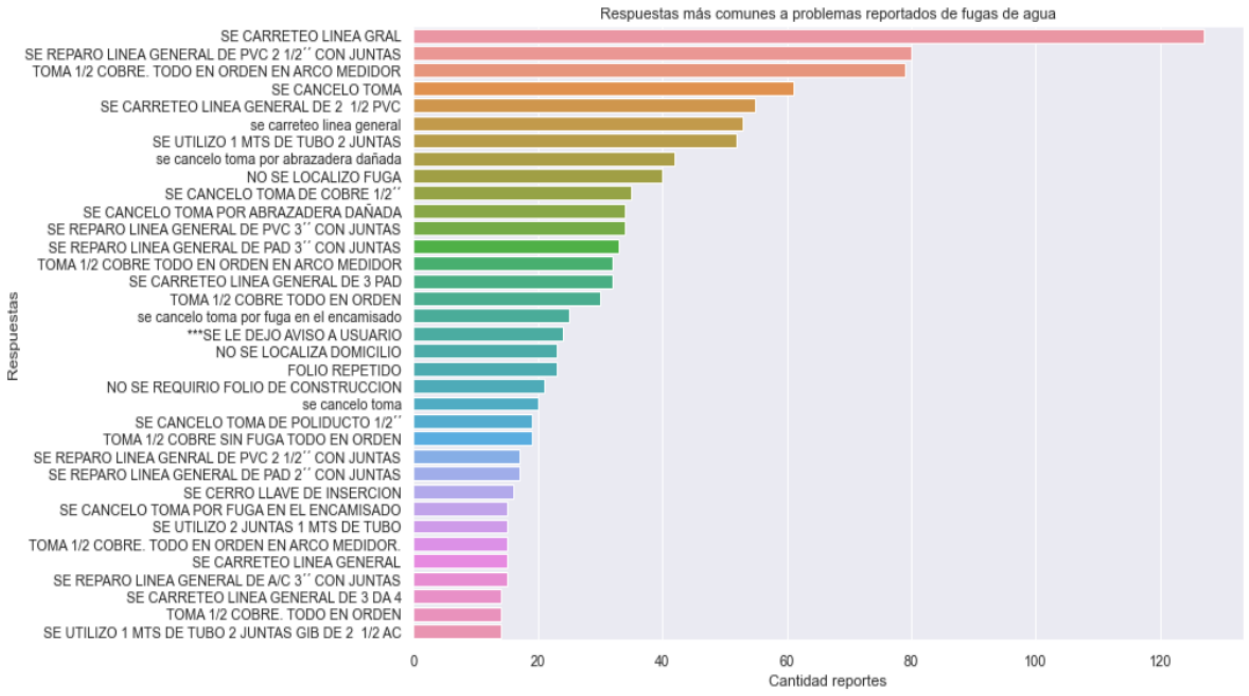

Figura 56. Respuestas más comunes a reportes de falta de agua a los usuarios.

5.3. CONCLUSIONES DEL CAPÍTULO

Al análisis realizado a la información obtenida del sistema *CIS* permite dar respuestas a las interrogantes planteadas logrando un mayor entendimiento sobre, en qué consisten estos datos, cuál es el proceso que los origina, cuáles son las colonias y zonas que más reportes presentan, con qué frecuencia se realizan estos reportes y en qué período del año se registran más reportes relacionados a estas quejas. Para la interrogante planteada relacionada con emplear la información del sistema *CIS* para el proceso de detección y predicción de anomalías no se obtuvo una respuesta satisfactoria; debido a que para esto se necesita tener la relación de a qué zona pertenece y a las zonas donde tienen incidencia cada una de las instalaciones de la red hidráulica. Una vez detectado esto se inició un proceso de recopilación de esta información, pero debido al tiempo y a la dependencia de los profesionales de la JMAS no se logró la obtención de esta información a tiempo, aunque la solicitud de la misma se realizó y se espera que sea entregada lo que permite que se pueda emplear en investigaciones futuras.

CAPÍTULO 6: INTEGRACIÓN DE RESULTADOS OBTENIDOS A LA HERRAMIENTA WEB

En este capítulo se describe el proceso de integración de los resultados obtenidos en un prototipo funcional de una herramienta Web, que permita además de la visualización y etiquetado de los datos el entrenamiento de los algoritmos de *ML* con mejor resultado en los experimentos para cada una de las instalaciones de la red hidráulica tanto para las instalaciones que ya presentan datos etiquetados como para aquellas que se vayan etiquetando en el futuro.

Para la integración de los resultados fue necesario proveer dos nuevos módulos a la herramienta web los cuales son detallados a continuación.

6.1. MÓDULO PARA ENTRENAR MODELOS DE MACHINE LEARNING PARA CADA INSTALACIÓN DE LA RED HIDRÁULICA

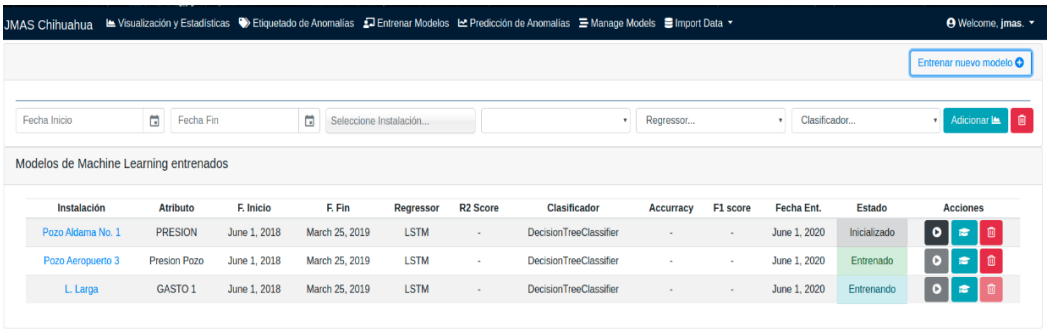

Figura 57. Módulo para entrenar modelos de *ML*.

Este módulo provee las funcionalidades necesarias para el entrenamiento de los modelos de *ML* con mejores resultados, seleccionados a partir de los experimentos realizados. Para esto se siguen los siguientes pasos:

1. El usuario selecciona una *fecha de inicio* y *fecha fin* del período que se tomará para el entrenamiento.
2. Posteriormente selecciona una *instalación* para la cual se tienen datos etiquetados, de lo contrario el entrenamiento no puede ser realizado debido a que se emplea un enfoque supervisado.

3. Luego de seleccionar la instalación se continua con la elección de los *algoritmos* de *regresión y clasificación* que será utilizados para el entrenamiento; en este caso solo se tiene como regresor la *LSTM* y el *DecisionTree* para el proceso de *clasificación,* que fueron los algoritmos con mejores resultados en los experimentos para el proceso de detección y predicción de anomalías.

Debido a que las tareas de entrenar estos algoritmos se pueden tardar dependiendo del período de tiempo seleccionado fue necesario crear un mecanismo para que estas tareas se fueran ejecutando en un segundo hilo de ejecución. De esta forma se evita que se tenga que esperar a que termine de entrenar un modelo para poder continuar con la operación de la herramienta. De esta forma se permite que dependiendo del poder de cómputo de la computadora donde se despliegue, se pueda ir entrenando varios modelos para las instalaciones de la red hidráulica. Para su correcto funcionamiento la interfaz debe ir actualizando cada cierto intervalo de tiempo el estado del proceso de entrenamiento de estos modelos. Además, se permite guardar el entrenamiento realizado para poder ser utilizado en tiempo real para la detección predicción de anomalías.

6.2. MÓDULO PARA DETECTAR Y PREDECIR ANOMALÍAS EN TIEMPO REAL

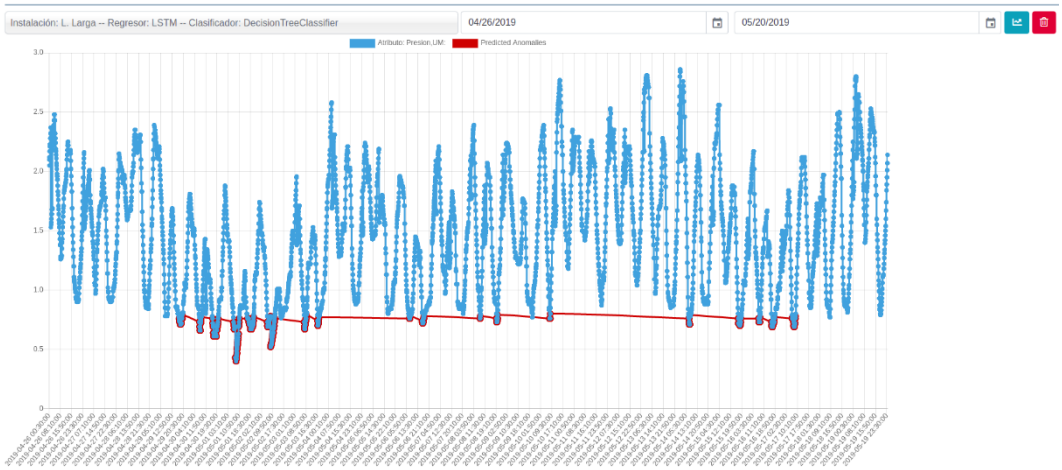

Figura 58. Módulo para visualizar las predicciones de anomalías.

Este módulo provee las funcionalidades para mostrar las predicciones de las anomalías para aquellas instalaciones que tengan previamente modelos entrenados. Para esto

inicialmente el usuario selecciona un modelo entrenado para cierta instalación, luego selecciona un período de tiempo (*fecha inicio* y *fecha fin*) para el cual el sistema de proporcionar una predicción tanto de la serie de tiempo, así como de las anomalías (marcadas de color rojo) que son predichas para esa instalación. El usuario puede ver cada valor específico que donde ocurren estas anomalías tanto como el atributo que se está analizando (*presión, nivel, etc.*) posicionando el cursor encima del punto que desee conocer.

El módulo es desarrollado para que el usuario pueda estar analizando las predicciones de varias instalaciones a la misma vez con la limitante del poder de cómputo de la computadora donde se despliega esta aplicación.

6.3. CONCLUSIONES DEL CAPÍTULO

Con el desarrollo de esta herramienta web se proporciona a los profesionales de la JMAS un mecanismo para que puedan monitorizar las diferentes instalaciones de la red hidráulica y que brinde predicciones sobre comportamientos anómalos en las mismas en tiempo real. Este proceso les permitirá tomar acciones preventivas para mitigar o disminuir el impacto de estas anomalías en el proceso de distribución del agua. Con la explotación de esta esta herramienta se podrían disminuir los eventos de falta y fugas de agua reportados por los usuarios al CIS lo cual tendría un impacto positivo para la gestión del preciado líquido en el municipio de Chihuahua, puesto que debido a la región la administración del agua es un tema crítico.

CAPÍTULO 7: DISCUSIÓN APORTES Y CONCLUSIONES GENERALES

En este capítulo se presentan los aportes brindados por esta investigación y las conclusiones generales, mediante un análisis de la solución a las preguntas de investigación que fueron guiando el trabajo investigativo. Luego se exhiben los trabajos futuros y recomendaciones para la extensión o realización de trabajos relacionados con esta investigación.

7.1. DISCUSIÓN DE PREGUNTAS DE INVESTIGACIÓN

¿Cómo presentar los datos obtenidos de forma que puedan ser utilizados por la comunidad de Machine Learning?

En respuesta a esta pregunta de investigación representa el primer aporte de la tesis pues se proporciona una biblioteca de Python: ´***jmas_water_data***´ la cual contiene los siguientes *datasets* con los datos de las instalaciones de la red hidráulicas listos para ser usados por la comunidad de *ML*:

***full_dataset*:** *pandas dataframe* que contiene los todos datos de los sensores para cada una de las instalaciones de la red hidráulica de Chihuahua.

labeled_dataset: *pandas dataframe* que contiene los datos etiquetados de anomalías en dos instalaciones de la red hidráulica, el etiquetado fue realizado por un usuario experto de la JMAS.

Esta biblioteca podrá ser instalada haciendo uso del comando de consola ´*pip install jmas_water_data*´ y una vez instalada se pueden importar los *datasets* anteriormente presentados listos para aplicar algoritmos de *ML*. Esto también permitirá la reproducción de los experimentos y resultados presentados en la investigación.

¿Como permitir que un usuario experto pueda identificar y registrar comportamientos que pueden llevar a la falta o fugas de agua en las instalaciones de la Red Hidráulica?

En el capítulo #3 se presenta los módulos de visualización y etiquetado de datos de la herramienta web la cual proporciona un mecanismo para que un experto pueda visualizar los

datos de las instalaciones de la red hidráulica. Además, este usuario experto puede registrar los períodos de tiempos en los cuales determine que ocurren comportamientos anómalos que puede llevar a la falta o pérdidas de agua. La herramienta proporcionada permite el registro de esta información, así como exportar los mismos para ser utilizados de forma externa. Al proporcionar estas prestaciones se considera que se da respuesta a la pregunta de investigación planteada.

¿Los datos actuales son suficientes para predecir y detectar anomalías en tiempo real?

El análisis realizado en el capítulo #4 de experimentos y resultados, evidencia que con los datos etiquetados de las dos instalaciones se puede predecir y detectar las anomalías en tiempo real de las instalaciones, mediante el enfoque de emplear una LSTM como regresor de los valores de los sensores y emplear un clasificador como *DecissionTree* para detectar las anomalías en las predicciones usando ventanas deslizantes. En el caso de la instalación Pozo Aeropuerto 3 se obtuvo un 98% de *accuracy,* 95% de *precisión,* 97% de *recall y* 95% de *f1-score*, la instalación Tanque Loma Larga se obtuvo 95% de *accuracy*, 81% de *precisión, 70%* de recall y 73% de *f1-score*, ambos resultados mejorando el *baseline*. Con estos resultados podemos dar respuesta satisfactoria a la pregunta de investigación, pero sería incorrecto generalizar que se cumple para las demás instalaciones que no se tienen datos etiquetados.

¿Qué información aporta los datos recopilados por el sistema CIS?, esta información puede ser utilizada en el proceso de predecir y detectar anomalías en tiempo real?

En el capítulo #5 se realiza la exploración de los datos del sistema CIS, en este análisis se obtuvo respuestas a las siguientes interrogantes relacionadas a la información que aportan los datos del sistema CIS:

- ¿Cómo se originan los datos
- La cantidad de registros que existen
- Período del año donde se presentan más reportes de usuarios.
- Zonas y colonias de Chihuahua con más reportes de usuarios.
- Respuestas más comunes dadas a los usuarios.

Aunque esta información puede ser de utilidad para demostrar la existencia de problemas relacionados al servicio y distribución de agua en Chihuahua, no se encontró

información en estos datos que pueda ser utilizada en el proceso de predicción y detección de anomalías en las instalaciones de la red hidráulica.

¿Cómo permitir que los profesionales de la JMAS puedan predecir y detectar anomalías en tiempo real?

En el capítulo #6 se describe el proceso de integración de experimentos realizados e implementación de una herramienta Web que incluya tanto los módulos de visualización y etiquetado de los datos, así como un módulo para la predicción y detección de anomalías en tiempo real. El uso de esta herramienta, los profesionales de la JMAS podrán predecir con cierto grado de exactitud comportamientos anómalos en instalaciones de la red hidráulica que puedan afectar el proceso de distribución del agua en Chihuahua.

La herramienta provee formas de entrenar los modelos de *ML* de forma automática con el paso del tiempo o a petición del usuario de forma manual tanto para las instalaciones que ya se encuentran etiquetadas, o una vez que nuevas instalaciones sean etiquetadas se puedan entrenar modelos para la predicción y detección de anomalías en estas. Además, el entrenamiento de esos modelos puede ser guardados para que la aplicación sea más eficiente al momento de operación.

Esta herramienta les permitirá a los profesionales de la JMAS tomar acciones preventivas para erradicar o disminuir el impacto de comportamientos anómalos en instalación de la red hidráulica en el proceso de distribución del agua en Chihuahua. Se considera que con esto se da respuesta a la pregunta de investigación relacionada, cumpliendo con la solución de los problemas planteados en la investigación

7.2. TRABAJOS FUTUROS

En trabajos futuros sería importante validar que el enfoque propuesto en esta investigación aporta buenos resultados para las demás instalaciones de la red hidráulica luego del etiquetado de nuevas instalaciones por expertos de la JMAS.

Debido a que la gran cantidad de instalaciones de la red hidráulica que se encuentra sin etiquetar las anomalías y al tiempo y trabajo que lleva este proceso, sería invaluable encontrar un enfoque No supervisado para el proceso de detección de anomalías en tiempo real, que brinde resultados competitivos respecto a los obtenidos en esta investigación.

se necesita tener la relación de a qué zona pertenece y a las zonas donde tienen incidencia cada una de las instalaciones de la red hidráulica.

Registrar la relación entre las zonas en la que está dividida Chihuahua y las instalaciones de la red hidráulica. Esta relación de comunicar a que zona pertenece y a las zonas donde tienen incidencia cada una de las instalaciones de la red hidráulica. Esta información se solicitó a los profesionales de la JMAS, pero no fue entregada en el período de esta investigación. Con la recopilación de esta información se podría lograr relacionar la información del sistema CIS en el análisis de predicción y detección de anomalías pues si una zona A está relacionada directamente con otra zona B, si ocurre una anomalía de alto impacto en la zona A también afectará a la zona B, lo cual se puede predecir y actuar en consecuencia para disminuir el impacto de la misma, esto también pasa en el caso de las colonias.

7.3. RECOMENDACIONES

Se recomienda modificar el experimento de LSTM como regresor más umbral fijo para la predicción de anomalías en tiempo real, empleando un método para variar el umbral dependiendo del histórico, con el objetivo de verificar si se obtienen resultados competitivos puesto que este enfoque es no supervisado, lo cual ahorraría tiempo y esfuerzo necesario para el etiquetado de los datos.

Se recomienda analizar otras variables relacionadas a las instalaciones de la red hidráulica que se tienen actualmente en los datos, pero que por inconsistencias se descartaron en esta investigación, para esto sería necesario trabajar en conjunto con la JMAS para obtener información más completa y consistente de los sensores de la red hidráulica.

CAPÍTULO 8: APÉNDICES (ANEXOS)

En este capítulo se muestran los anexos de la presente investigación principalmente mostrando en modo de tablas los resultados de los experimentos presentados en la investigación.

Métrica	Valor
Accuracy	0.98
Precision	0.97
Recall	0.88
F1 score	0.93
AUC	0.94

Tabla 4. Resultados experimento baseline para el pozo Aeropuerto 3.

Métrica	Valor
Accuracy	0.90
Precision	0.36
Recall	0.71
F1 score	0.47
AUC	0.81

Tabla 5. Resultados del experimento baseline para el tanque Loma Larga.

Clasificador	Accuracy	Precision	Recall	F1 Score
KNN	0.79	0.88	0.86	0.87
Linear SVM	0.74	0.74	1.0	0.85
RBF SVM	0.8	0.82	0.93	0.87
Decision Tree	0.78	0.81	0.91	0.86
Random Forest	0.73	0.76	0.94	0.84
MLP ANN	0.82	0.87	0.89	**0.88**

Tabla 6. Resultados del experimento clasificación con ventanas generales para el pozo Aeropuerto 3.

Clasificador	Accuracy	Precision	Recall	F1 Score
KNN	0.73	0.81	0.83	0.82
Linear SVM	0.75	0.75	1.0	**0.86**
RBF SVM	0.78	0.81	0.91	**0.86**
Decision Tree	0.72	0.80	0.83	0.82
Random Forest	0.75	0.80	0.89	0.84
MLP ANN	0.74	0.77	0.93	0.84

Tabla 7. Resultados del experimento clasificación con ventanas generales para el tanque Loma Larga.

Clasificador	Accuracy	Precision	Recall	F1 Score
KNN	0.96	0.97	0.97	**0.97**
Linear SVM	0.69	0.69	0.99	0.81
RBF SVM	0.96	0.97	0.97	**0.97**
Decision Tree	0.96	0.97	0.97	**0.97**
Random Forest	0.96	0.96	0.98	**0.97**
MLP ANN	0.96	0.96	0.98	**0.97**

Tabla 8. Resultados del experimento clasificación con ventanas deslizantes para el pozo Aeropuerto 3.

Clasificador	Accuracy	Precision	Recall	F1 Score
KNN	0.95	0.96	0.98	**0.97**
Linear SVM	0.90	0.90	1.0	0.95
RBF SVM	0.94	0.95	0.99	**0.97**
Decision Tree	0.95	0.95	0.98	**0.97**
Random Forest	0.95	0.95	0.99	**0.97**
MLP ANN	0.92	0.92	0.99	0.96

Tabla 9. Resultados del experimento clasificación con ventanas deslizantes para el pozo tanque Loma Larga.

Regresor	R2 score	MSE	MAE
LSTM	**0.96**	**0.017**	**0.08**

Tabla 10. Resultados de evaluación de la LSTM como regresor para el Aeropuerto 3.

Enfoque de Predicción de Anomalías	Accuracy	Precision	Recall	F1 score
LSTM + umbral	**0.78**	**0.65**	**0.82**	**0.72**

Tabla 11. Resultados del proceso de detección y predicción de anomalías del experimento LSTM como regresor + umbral fijo para el Aeropuerto 3.

Regresor	R2 score	MSE	MAE
LSTM	**0.95**	**0.03**	**0.12**

Tabla 12 . Resultados de evaluación de la LSTM como regresor para el tanque Loma Larga Evaluación del proceso de predicción de anomalías.

Enfoque de Predicción de Anomalías	Accuracy	Precision	Recall	F1 score
LSTM + umbral	**0.89**	**0.27**	**0.07**	**0.13**

Tabla 13. Resultados del procedo de detección y predicción de anomalías del experimento LSTM como regresor + umbral fijo para el tanque Loma Larga.

Enfoque de Predicción de Anomalías	Accuracy	Precision	Recall	F1 score
LSTM + Decision Tree	**0.98**	**0.95**	**0.97**	**0.95**

Tabla 14. Resultados del procedo de detección y predicción de anomalías del experimento LSTM como regresor + DecisitionTree como clasificador para el Aeropuerto3.

Enfoque de Predicción de Anomalías	Accuracy	Precision	Recall	F1 score
LSTM + Decision Tree	**0.95**	**0.81**	**0.70**	**0.73**

Tabla 15. Resultados del proceso de detección y predicción de anomalías del experimento LSTM como regresor + DecisitionTree como clasificador para el tanque Loma Larga.

	Total	**Terminado**	**Asociado**	**Cancelado**	**Validado**
Reportes de faltas de agua	17675	11831	5736	89	19
Reportes de fugas de agua	18628	18083	-	460	85

Tabla 16. Total, de reportes que se tienen del sistema CIS.

BIBLIOGRAFÍA

[1] M. Orlović and A. Krajnović, "WATER MANAGEMENT - AN IMPORTANT CHALLENGE FOR MODERN ECONOMICS," in *DIEM: Dubrovnik International Economic Meeting, At Dubrovnik, Croatia*, 2015, p. 24.

[2] Ioana Şerban, "IoT Water Innovation – Use Cases and Benefits - Euro IT Group." [Online]. Available: https://www.euroitgroup.com/iot-water-innovation-use-cases-benefits/. [Accessed: 03-Jun-2019].

[3] G. Gupta, "Monitoring Water Distribution Network using Machine Learning," KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ELECTRICAL ENGINEERING, 2017.

[4] S. M. Zaidi Arshad *et al.*, "Machine learning for energy-water nexus : challenges and opportunities," *Big Earth Data*, vol. 2, no. 3, pp. 1–40, 2018, doi: 10.1080/20964471.2018.1526057.

[5] I. A. Washington, "Big Data and the Future of Water Management," pp. 1–3, 2012.

[6] J. P. Pereira Cardoso, "Unsupervised Anomaly Detection in Time Series Data using Deep Learning," Eindhoven University of Technology, 2019.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection : A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[8] M. Samuelsson, "Anomaly Detection In Time Series Data," CHALMERS UNIVERSITY OF TECHNOLOGY, 2016.

[9] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection $," *Signal Processing*, vol. 99, pp. 215–249, 2014, doi: 10.1016/j.sigpro.2013.12.026.

[10] P. Galeano, D. Peña, and R. S. Tsay, "Outlier detection in multivariate time series by projection pursuit," *J. Am. Stat. Assoc.*, vol. 101, no. 474, pp. 654–669, 2012, doi: 10.1198/016214505000001131.

[11] J. A. Quinn and C. K. I. Williams, "Known Unknowns : Novelty Detection in Condition Monitoring," in *The Iberian Conference*, 2007, p. 6, doi: 10.1007/978-3-540-72847-4_1.

[12] G. A. Barreto and L. Aguayo, “Time series clustering for anomaly detection using competitive neural networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5629 LNCS, pp. 28–36, 2009, doi: 10.1007/978-3-642-02397-2_4.

[13] S. R. Mounce, R. B. Mounce, and J. B. Boxall, “Novelty detection for time series data analysis in water distribution systems using support vector machines,” pp. 672–686, 2011, doi: 10.2166/hydro.2010.144.

[14] D. Vries, B. Van Den Akker, E. Vonk, W. De Jong, and J. Van Summeren, “Application of machine learning techniques to predict anomalies in water supply networks,” *Water Sci. Technol. Water Supply*, vol. 16, no. 6, pp. 1528–1535, 2016, doi: 10.2166/ws.2016.062.

[15] B. R. Kiran *et al.*, “Deep learning for anomaly detection in multivariate time series data,” *Ninth Int. Conf. Pervasive Patterns Appl. Defect 2017*, vol. 13, no. 1, pp. 1–12, 2017, doi: 10.3390/jimaging4020036.

[16] M. Wielgosz, A. Skoczeń, and M. Mertik, “Recurrent Neural Networks for anomaly detection in the Post-Mortem time series of LHC superconducting magnets,” 2017.

[17] T. Wen and R. Keyes, “Time Series Anomaly Detection Using Convolutional Neural Networks and Transfer Learning,” 2019.

[18] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, “Anomaly detection for a water treatment system using unsupervised machine learning,” *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2017–Novem, pp. 1058–1065, 2017, doi: 10.1109/ICDMW.2017.149.

[19] S. Kanarachos, S. R. G. Christopoulos, A. Chroneos, and M. E. Fitzpatrick, “Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform,” *Expert Syst. Appl.*, vol. 85, pp. 292–304, 2017, doi: 10.1016/j.eswa.2017.04.028.

[20] V. H. Alves Ribeiro and G. Reynoso Meza, “Online Anomaly Detection for Drinking Water Quality Using a Multi-objective Machine Learning Approach,” *Conf. Genet. Evol. Comput. Conf. Companion*, pp. 1–2, 2018, doi: 10.1145/3205651.3208202.

[21] O. Karaahmetoglu, F. Ilhan, I. Balaban, and S. S. Kozat, “Unsupervised Online

Anomaly Detection On Irregularly Sampled Or Missing Valued Time-Series Data Using LSTM Networks," pp. 1–11, 2020.

[22] B. Babić, A. Dukić, and M. Stanić, "Managing water pressure for water savings in developing countries," *Water SA*, vol. 40, no. 2, pp. 221–232, 2014, doi: 10.4314/wsa.v40i2.4.

[23] J. Estrada, "La Jornada: Chihuahua: alertan sobre desabasto de agua potable en la capital," 2018. [Online]. Available: https://www.jornada.com.mx/2018/08/28/estados/029n2est. [Accessed: 16-Jul-2020].

[24] M. Bourquin, "Why Make Water Systems Smart? | Water Finance & Management," 2016. [Online]. Available: https://waterfm.com/why-make-water-systems-smart/. [Accessed: 25-Apr-2019].

[25] A. A. Patel, *Hands-On Unsupervised Learning Using Python*. O'Reilly Media, Inc, 2019.

[26] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *J. Inf. Telecommun.*, vol. 3, no. 3, pp. 294–307, 2019, doi: 10.1080/24751839.2019.1565653.

[27] D. Ramotsoela, A. Abu-Mahfouz, and G. P. Hancke, "A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study," *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–24, 2018, doi: 10.3390/s18082491.

[28] S. Tariq Sangyup Lee, M. Shin Lee Jung Okchul, D. Chung, Y. Shin, and S. S. Woo, "Detecting Anomalies in Space using Multivariate Convolutional LSTM with Mixtures of Probabilistic PCA," pp. 2123–2133, 2019, doi: 10.1145/3292500.3330776.

[29] T. Sandra Buda, B. Caglayan, and H. Assem, "DeepAD: A Generic Framework Based on Deep Learning for Time Series Anomaly Detection," in *Advances in Knowledge Discovery and Data Mining*, vol. 1, no. February, Dublin: Springer International Publishing, 2018, pp. 577–588.

[30] T. Kieu, B. Yang, and C. S. Jensen, "Outlier detection for multidimensional time series using deep neural networks," *Proc. - IEEE Int. Conf. Mob. Data Manag.*, vol. 2018–

June, pp. 125–134, 2018, doi: 10.1109/MDM.2018.00029.

[31] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[32] S. Maya, K. Ueno, and T. Nishikawa, “dLSTM: a new approach for anomaly detection using deep learning with delayed prediction,” *Int. J. Data Sci. Anal.*, vol. 8, no. 2, pp. 137–164, 2019, doi: 10.1007/s41060-019-00186-0.

[33] T. Wen and R. Keyes, “Time Series Anomaly Detection Using Convolutional Neural Networks and Transfer Learning,” 2019.

[34] Gobierno del Estado de Chihuahua, “Promueve JMAS chihuahua número único de atención telefónica 073 | Chihuahua.gob.mx.” [Online]. Available: http://www.chihuahua.gob.mx/PROMUEVE-JMAS-CHIHUAHUA-NÚMERO-ÚNICO-DE-ATENCIÓN-TELEFÓNICA-073. [Accessed: 10-Apr-2020].