Implementación de una plataforma basada en IoT e IA para la detección deanomalías en aplicaciones de monitorización de la calidad de aire: Caso de estudioMaterial particulado Campus Bogotá Universidad Sergio Arboleda

Felipe Olivares, Giovanni Gonzaléz, Juan Aranda, Camilo Rodriguez

***Resumen*—Identifying anomalies in sensory networks is an activity that currently provides multiple advantages, such as preventive maintenance at the nodes, reliability of the captured data, scalability to any type of variable, among many others. Currently, artificial intelligence is at the forefront, as the technology designated to carry out this type of research. This article presents a methodological and implementation proposal to identify anomalies using data engineering, artificial intelligence techniques and IoT networks, focusing on PM 2.5 particulate matter in the city of Bogotá as case study.**

***Palabras cláve:*—**IoT, Artificial Intelligence, Artificial Intelligence of things

1. INTRODUCCIÓN

Con el avance de la tecnología se ha buscado ir mejorando distintos componentes a nivel de hardware y de software en varias implementaciones relacionadas con la parte electrónica y de sistemas. Existen diversos temas que combinan los dos componentes, uno de ellos es el despliegue de redes sensoriales y la analítica que se realiza a los datos capturados. Con el fin de profundizar este tema se propone realizar el diseño y despliegue de una red sensorial sobre el campus de la universidad Sergio Arboleda la cual obtenga datos relacionados a la calidad del aire y se pueda realizar un análisis a esta información, para dar un manteamiento preventivo a la red y poder hacer protocolos con base al análisis realizado.

Una de las razones por la que se profundiza en este tema, es debido a la importancia que toman a partir de inicios del siglo XXI, definiendo a las redes inalámbricas sensoriales (WSN, por sus siglas em inglés) como una red de características auto-configurables integrada por un pequeño número de nodos (sensores) distribuidos en un espacio logrando una comunicación entre sí, para luego llevar esta información a un punto central y realizar una actividad específica según el dato almacenado.

Con el fin de explorar la posibilidad de implementar mejoras en cada uno de estos componentes, se hace una investigación en la cual se puede dictaminar que la mayoría de las implementaciones hacen un profundo análisis sobre la red desplegada (nodos) dejando de lado los componentes de puente y nodo final tal como se ve en implementaciones basadas en el diseño de la red [2] [3]. Así mismo, también se hace una investigación enfocada en las diferentes técnicas utilizadas en el tratamiento de los datos. El resultado de esta muestra que conceptos como inteligencia artificial y el aprendizaje automático son usados en este tipo de implementaciones debido a que ofrecen un amplio abanico de posibilidades para usar, como ejemplo se tiene los algoritmos genéticos, las redes neuronales, la lógica difusa y la hiperheurística por el lado de inteligencia artificial. Por el lado del aprendizaje automático se tiene métodos para agilizar procedimientos repetitivos o evaluar la integridad de la información [5] [6].

Todo esto lleva a pensar que existe la posibilidad de recrear una red sensorial capaz de hacer una captura de datos, para luego ser enviados a un nodo final en el cual se usen técnicas de inteligencia artificial y/o aprendizaje automático para generar una acción basada en la información recibida. Cabe resaltar que una red creada bajo estos principios puede traer grandes beneficios para el usuario final tales como, el monitoreo constante de los nodos, aviso temprano de fallas encontradas en la red, bajo consumo de energía para garantizar una mayor funcionalidad, entre otros.

2. MARCO TEÓRICO

En este apartado, se describen las tecnologías representativas para el desarrollo del proyecto, enunciando diversos métodos de análisis y detección de anomalías usadas en diversos proyectos.

2.1. Artificial Intelligence of Things

Siendo el tema central de este documento las anomalías de los datos en un red sensorial haciendo el uso de inteligencia artificial, se hace necesario explicar ciertos conceptos.

El primer concepto es el de internet de la cosas, según Weber es una arquitectura basada en la Internet, que permite la facilidad en el intercambio de bienes y servicios entre redes de suministro, generando un impacto importante en la seguridad y privacidad del usuario [1].

El concepto de inteligencia artificial se define como la capacidad de crear sistemas capaces de aprender y razonar

como un ser humano, en donde puedan aprender y sepan como solucionar un problema a partir de una condiciones dadas, para poder tomar una decisión e implementarla [3].

El uso del internet de las cosas ha venido ha venido creciendo con la necesidad del usuario final de tener los dispositivos y elementos de trabajo capaces de conectarse a internet a la palma de la mano, con el fin de poder sacar provecho de la situación para un determinado uso, que va desde la visualización de datos ya sean sensores, actuadores o controladores, hasta el punto de fusionarlo con inteligencia artificial y lograr la compra de artículos de la canasta familiar dependiendo del modelo que se presente.

Las ventajas que trae consigo la implementación del internet de las cosas se pueden clasificar de la siguiente manera [2]:

1. Velocidad de análisis de datos: Con la toma de una cantidad razonable de datos, se puede tomar una decisión rápida y efectiva. Esto no solo se evidencia en aplicaciones cotidianas si no que también es implementado en las grandes empresas.
2. Facilidad de seguimiento: Se permite tener un seguimiento en cuanto a la calidad y la cantidad de los dispositivos empleados, con el fin de facilitar la logística y la seguridad del usuario.
3. Ahorro en tiempo y dinero: Este parámetro visto desde un ahorro en el capital humano, el dinero y tiempo del usuario, generando una reducción en los riesgos económicos y mecánicos.

2.2. Internet de las cosas: Redes sensoriales

Al ser la tecnología IoT algo muy nuevo a nivel de implementación, no se tiene una idea generalizada sobre el tipo de arquitectura que se deba implementar, por lo que a continuación se muestran diferentes tipos de arquitecturas que se usan [8] .

Estas arquitecturas se pueden dividir en tres:

- Nivel tres.
- Nivel cinco.
- Basados en Cloud y Fog

En la arquitectura de nivel tres se tienen las siguientes divisiones: percepción, red, aplicación. En el nivel de percepción se tiene la parte física, en donde se encuentran los sensores que tienen como función la recopilación de datos que se tienen alrededor. El nivel de aplicación es hacer el puente entre el nivel de percepción y los servidores, para poder transmitir y procesar los datos obtenidos de los sensores.Finalmente el nivel de aplicación es el que se encarga de ejecutar el despliegue de la solución IoT que se realizo [8] .

En la arquitectura de nivel cinco se tiene la siguiente división: percepción, transporte, proceso, aplicación y negocio. Donde el nivel de percepción y aplicación son los mismos que en la arquitectura de nivel tres. En el capa de transporte hace referencia a la transferencia de datos por medio de redes como 3G, LAN, Bluetooth, RFID y NFC.

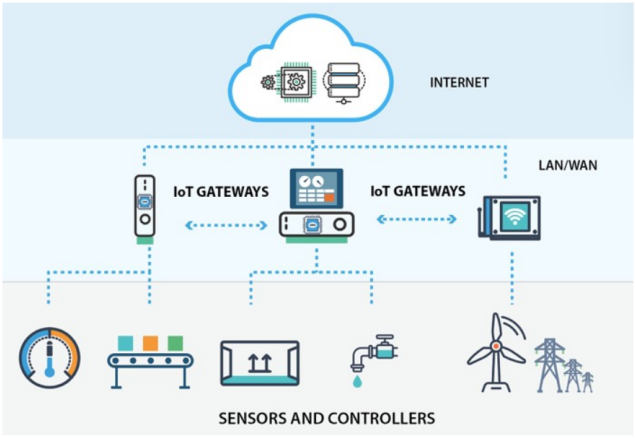

Figura 1: Arquitectura nivel 3. Tomado de [14]

La capa de proceso almacena y analiza una gran cantidad de datos, para un mejor funcionamiento se utilizan tecnologías como bases de datos, cloud computing y big data. En el nivel de negocio se gestiona las aplicaciones, el modelo de negocio y la privacidad.

En la arquitectura basada en cloud computing y fog. En la parte de cloud computing se tiene que se basa en el procesamiento de datos en infraestructura en la nube, manteniendo el procesamiento en el centro, la aplicación arriba y los dipositivos abajo. En la arquitectura de fog se tiene que el procesamiento de procesos va a ser realizado en el mismo dispositivo, para no tener que enviar los datos a la nube.

2.3. Inteligencia Artificial: Técnicas de detección de anomalías

Se debe mencionar que cada una de las técnicas a analizar tiene su correspondiente implementación en Python, además de tener un modelo base en Scikit-Learn a implementar.

- SVM

  El objetivo del algoritmo de máquina de vectores de soporte es encontrar un hiperplano en un espacio N-dimensional (N - el número de características) que clasifica claramente los puntos de datos para separar las dos clases de puntos de datos, hay muchos hiperplanos posibles que podrían elegirse. El objetivo radica en encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre puntos de datos de ambas clases. Maximizar la distancia del margen proporciona cierto refuerzo para que los puntos de datos futuros se puedan clasificar con más confianza. [21]

- k-Nearest Neighbor

  Es un método que simplemente busca en las

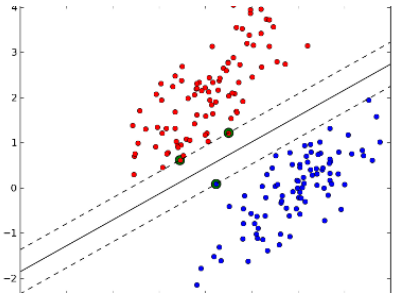

Figura 2: Arquitectura SVM. Tomado de [21]

observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean. [21] Funciona en tres pasos:

- Calcular la distancia entre el ítem a clasificar y el resto de items del dataset de entrenamiento.
- Seleccionar los «k» elementos más cercanos (con menor distancia, según la función que se use)
- Realizar una «votación de mayoría» entre los k puntos: los de una clase/etiqueta que ¡¡dominen¿¿ decidirán su clasificación final. [21]

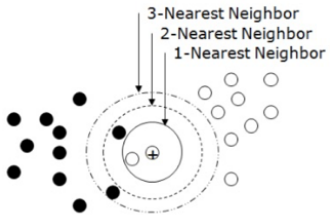

Figura 3: Arquitectura KNN. Tomado de [21]

- Kernel-PCA

PCA es un método lineal. Es decir, solo se puede aplicar a conjuntos de datos que son linealmente separables. Hace un excelente trabajo para los conjuntos de datos, que son linealmente separables. Pero, si lo usamos para conjuntos de datos no lineales, podríamos obtener un resultado que puede no ser la reducción de dimensionalidad óptima. Kernel PCA utiliza una función de kernel para proyectar el conjunto de datos en un espacio de características de dimensiones superiores, donde es linealmente separable. Es similar a la idea de Support Vector Machines. [25]

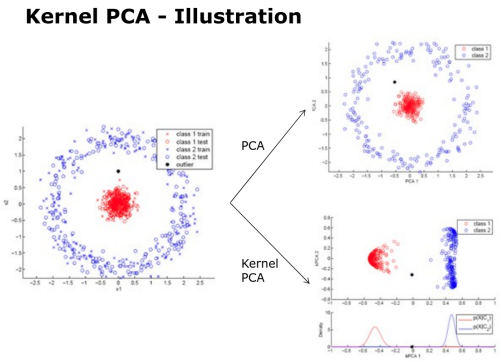

Figura 4: Arquitectura KPCA. Tomado de [20]

3. Planteamiento detección de anomalías

3.1. Metodología de detección

En primera instancia es preciso mencionar que la identificación de una anomalía en un dato tan volátil como el material particulado de 2.5 micras es sumamente difícil de detectar, esto debido a que tiende a cambiar sus valores a grandes escalas en pocas horas. Como valor agregado, se considera importante realizar un análisis correspondiente a los factores en los cuales se podría evidenciar el incremento de esta medida mencionada con anterioridad, entre los más frecuentes se encuentran: El paso de un vehículo que genere una fuerte contaminación en ese instante, un incendio cercano, una falla en el sensor que se encuentre tomando la medida, e inclusive una descalibración total de la estación en un momento dado del día.

Dadas estas razones, se propone una metodología de detección, que podría brindar una forma de realizar una detección temprana de una anomalía en una estación de monitoreo, esto generaría un sin fin de beneficios, por mencionar algunos: exactitud y veracidad en los datos recopilados por la estación, alertas sobre mantenimiento preventivo de los equipos que se encargan de tomar las mediciones, y posiblemente una metodología general de detección aplicable a cualquier tipo de dato que brinde un sensor.

La idea planteada en el presente documento radica a grandes rasgos en realizar una confrontación entre variables al momento de analizar un dato como anómalo o no. Para el presente caso, como se expuso con anterioridad, se cuenta con los datos obtenidos de las doce estaciones presentes en la ciudad de Bogotá, estos datos fueron procesados con distintas herramientas con el fin de obtener unos datos limpios y generar un dataset con las variables necesarias para el análisis propuesto. Con exactitud el proceso de desarrollo y tratamiento de datos se expone de la siguiente manera y bajo las siguientes etapas: Análisis de variables, descarga de datos desde http://201.245.192.252:81/Report/stationreport, creación de flujo en la herramienta para análisis de datos Knime, creación de scripts en “Python” para correcciones finales sobre los datos y la posterior carga del dataset final a diversos scripts en la herramienta “Jupyter Notebook”,

con el fin de realizar un análisis sobre las gráficas y la distribución de datos obtenida. Se plantea un posible esquema para rellenar los datos faltantes en cada estación, y posteriormente un análisis de detección de anomalías con técnicas de inteligencia artificial, en este caso una máquina de soporte vectorial de una clase OC-SVM, el dataset suministrado para este método contiene como variables " PM2.5 µg/m3, Vel Viento m/s, Dir Viento Grados y Temperatura °C ". Principalmente porque son variables que se relacionan directamente con la medición, no obstante los resultados obtenidos no son los más fiables al determinar como anomalía una medición obtenida desde la estación, debido a esto, la idea de tener una red sensorial de bajo costo que se encontrara captando mediciones constantemente funcionaria como una fuente de datos mayor, para ampliar estas variables y lograr una correcta identificación de una anomalía. Estableciendo hipótesis particulares, tales como: Si la velocidad del viento es alta, la concentración de material particulado debería disminuir en la estación, de no ser así, estamos en frente de una posible anomalía en la medición, que podría corresponder a qué pasó un vehículo contaminante, se encuentra un incendio relativamente cerca o para nuestro favor una anomalía en la medición, esto podría ser determinado, verificando el estatus de la red sensorial planteada con anterioridad.

3.2. Requerimientos técnicos

Dada la contingencia actual, se establecen requerimientos tanto técnicos como físicos para plantear una situación ideal, en la que la metodología de detección propuesta podría funcionar de una manera eficiente.

4. Modelo de datos

Con el fin de generar la construcción y el análisis de un dataset, que proporcione y abarque los requerimientos planteados con anterioridad, se realizo una ingeniería de datos, en distintas plataformas y mediante distintos procesos que serán explicados a continuación, inicialmente, se procedió a descargar los datos presentes en la pagina de la RMCAB, es importante resaltar que son 12 estaciones a analizar y la pagina no permite descargar datos de mas de 5 estaciones durante un periodo anual, a raíz de esto, se generaron 21 archivos en formato xlsx, que comprenden las 12 estaciones en el periodo 2013 - 2019. Estos 21 archivos fueron cargados en la herramienta para análisis de datos KNIME como se evidencia en la figura 6, con el fin de lograr mejores rendimientos en cuento a tiempos de desarrollo y facilidad de manejo de la data presente. El esquema del flujo general presente en KNIME puede ser visto en la figura 5, comprende 4 metanodos, de los cuales 3 son los encargados de realizar todo el cargue de información a la plataforma y el restante se ocupa de realizar los últimos cambios mediante expresiones regulares.

Los tres primeros metanodos presentes en la caja amarilla de la figura 5, se descomponen en los nodos presentes de la figura 6.

Posteriormente el metanodo presente en la caja naranja de la figura 5, se descompone en los 4 nodos de la figura 7. No obstante, fue necesario realizar dos procesos de limpieza y estructuración adicionales mediante dos scripts en python, los cuales tenian la función de normalizar los datos presentes en cada estación y determinar el porcentaje de datos vacios o nulos en cada una de estas, gracias a esto, se llego a la conclusión que aproximadamente entre un 1 15 - 17 % de los datos obtenidos de cada estación, se encuentra vacíos o en su defecto nulos. Esto indica que cabe la posibilidad de realizar un llenado de datos, como se propone en [28]

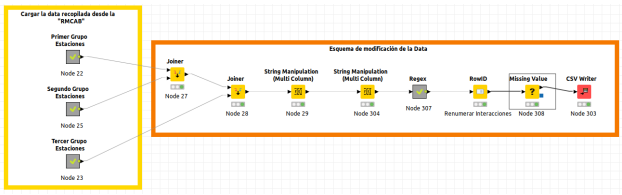

Figura 5: Flujo general de KNIME que abarca todo el proceso de ETL.

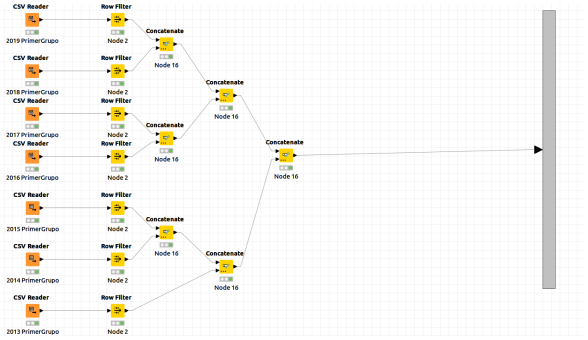

Figura 6: Parte de flujo en Knime que contempla la carga y unificación de los archivos descargados de la RMCAB

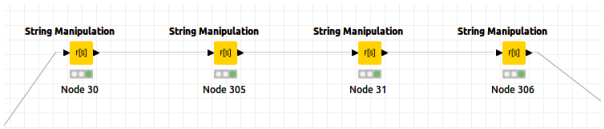

Figura 7: Flujo final de Knime que expone los nodos con expresiones regulares para el tratamiento final de los datos.

Como se mencionó anteriormente, la data recopilada desde la Red de Monitoreo de Calidad del Aire de Bogotá "RMCAB", presenta inconsistencias en torno a los valores captados en distintos momentos por cada una de las estaciones presentes en Bogotá. Por inconsistencias se hace referencia a valores presentes en la data como faltantes o nulos, en promedio 15 % de los datos, a consecuencia de errores del instrumento, clima desfavorable, perturbaciones en la infraestructura, entre otras. Alterando de forma significativa la representatividad de los mismos, con el fin de proponer una solución a esta situación se evidencian varias opciones al realizar un ETL de los datos, principalmente se presenta la

opción de eliminar o remover estos datos, sin embargo, esto generaría una significativa disminución del volumen de los datos recopilados, por consiguiente se plantean dos métodos de rellenar los valores faltantes en el dataset provenientes de [28], uno hace referencia al llenado de datos por medio de una regresión lineal, el siguiente se enfoca en el llenado de datos por medio de una red neuronal, por último pero no menos importante hace referencia al llenado de datos mediante la técnica ARIMA.

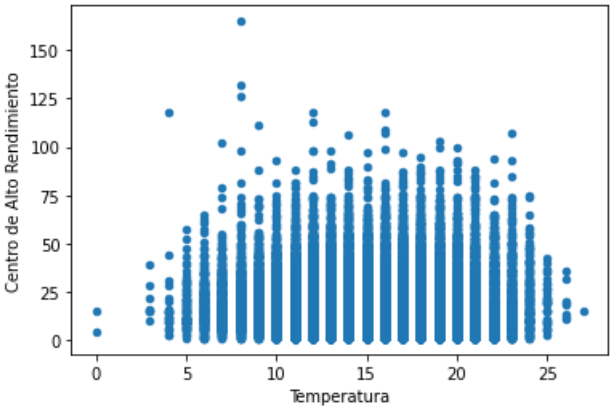

Figura 8: Dispersión de datos entre el valor de PM 2.5 y la Temperatura.

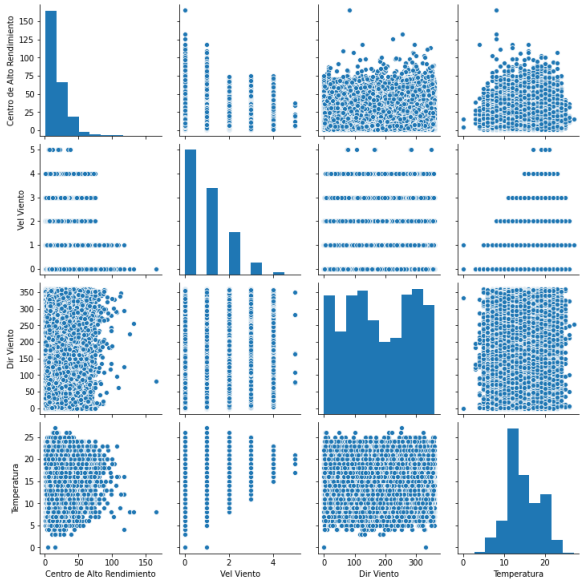

Figura 9: Visualización de comportamiento de linealidad y distribución de las variables.

4.0.1. Regresión Lineal. El análisis de una regresión lineal consiste en generar una ecuación (modelo) que, basándose en la relación existente entre ambas variables, permita predecir el valor de una a partir de la otra. [referencia]. El método lineal de mínimos cuadrados (Ecuación 1) pretende

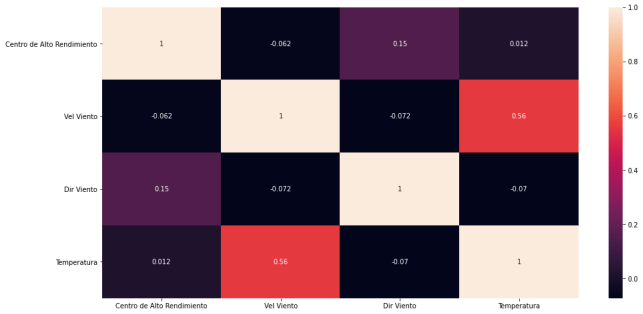

Figura 10: Evaluación de correlación de las variables, la matriz de correlación permite observar que importancia hay de incidencia de cambio y comportamiento entre las variables, esta se mide de 0-1, además se puede entender como porcentaje, entre mas porcentaje se tenga mas relación hay entre las features.

determinar la función continua que mejor se ajuste al comportamiento del conjunto de los datos y su relación entre sí. Su objetivo es minimizar la suma de cuadrados de la diferencia entre los puntos dados por la función y los de los datos. [31].

4.1. Técnicas de AI

4.1.1. OC-SVM. En primera instancia se pretende demostrar los conceptos básicos detrás de una SVM (Suport Vector Machine)

Dado que el problema que se desea solucionar es linealmente no separable, como técnica de inteligencia artificial, se propone utilizar OC-SVM (One Class - Suport Vector Marchine), de acuerdo al método por parte de Tax and Duin (SVDD), este método adopta un enfoque esférico, diferente al método por parte de Schölkopf que se enfoca en un hiperplano. El algoritmo de Tax and Duin obtiene un límite en forma de esfera, presente en el espacio de características, alrededor de los datos. El volumen de esta hiperesfera se minimiza, para minimizar el efecto de incorporar valores atípicos o anormales en la solución, básicamente el fin de esta técnica consiste en separar los valores normales y anómalos, encerrando en una esfera los datos normales (véase la figura 10)

Con el fin de realizar lo mencionado con anterioridad, la hiperesfera resultante se caracteriza por un centro a y un radio R mayor a 0 como distancia desde el centro hasta cualquier vector de soporte en el límite, del cual se minimizará el volumen R$\hat{2}$. El centro a es una combinación lineal de los vectores de soporte (que son los puntos de datos de entrenamiento para los cuales el multiplicador de Lagrange no es cero). Al igual que la formulación tradicional, podría requerirse que todas las distancias desde los puntos de datos xi al centro sean estrictamente menores que R, pero para crear un soft margin nuevamente se utilizan variables de holgura Ei con el parámetro de penalización C, donde C es un parámetro ajustable, la idea de integrar las variables de holgura y el parámetro de penalización C se basa en

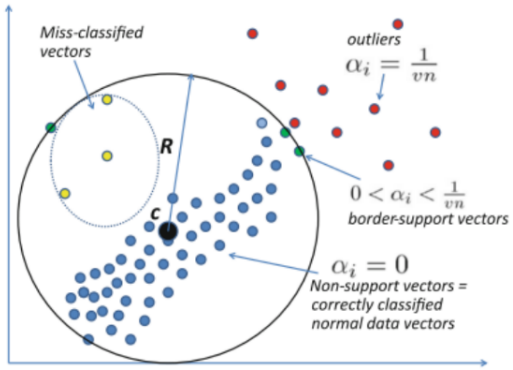

Figura 11: Espacio de características definido por una hiper-esfera OC-SVM, tomado de [30]

indicar cómo es posible aceptar puntos mal clasificados, pero penalizando su error de clasificación a través de una combinación lineal con la función objetivo, es una noción denominada soft margin la cual permite tratar datos más realistas, por ende una forma de añadir el coste de los errores a la función objetivo es intentar resolver el siguiente problema de minimización:

$$\min_{R,\,\mathbf{a}} R^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to:}$$

$$\begin{aligned} \|x_i - \mathbf{a}\|^2 &\leq R^2 + \xi_i && \text{for all } i = 1,\ldots,n \\ \xi_i &\geq 0 && \text{for all } i = 1,\ldots,n \end{aligned}$$

Figura 12: Problema de minimización tomado de [32]

Después de resolver esto mediante la introducción de multiplicadores de Lagrange ai, se puede probar un nuevo punto de datos Z para estar dentro o fuera de la clase. Se considera en clase cuando la distancia al centro es menor o igual que el radio, utilizando el Kernel gaussiano (véase figura 10) como una función de distancia sobre dos puntos de datos

$$\|z - \mathbf{x}\|^2 = \sum_{i=1}^{n}\alpha_i \exp\left(\frac{-\|z - x_i\|^2}{\sigma^2}\right) \geq -R^2/2 + C_R$$

Figura 13: Ecuación para probar un nuevo punto de datos. tomado de [32]

Resultados evaluación OC-SVM, con features ”Temperatura - PM 2.5recopilados y analizados en la estación correspondiente al Centro de Alto Rendimiento. Es importante determinar que los resultados obtenidos de la implementación no fueron los deseados, esto se sustenta dado diversos factores, inicialmente la cantidad de variables a analizar no son suficientes, se tuvo que realizar una investigación de variables que tuviesen una relación directa con la medición del material particulado, posterior a este desarrollo se eligieron como variables, ”Vel Viento m/s”, ”Temperatura C”, ”Direccion del Viento”, y finalmente el valor del material particulado en ese instante ”PM 2.5”.

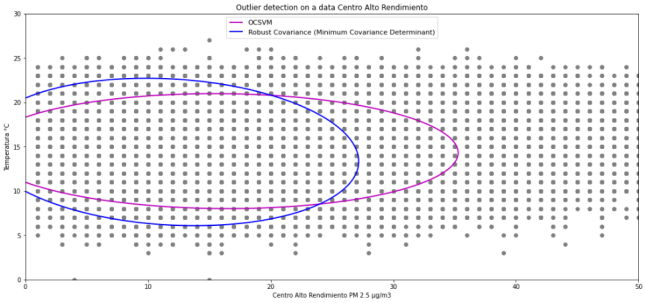

Figura 14: Resultados obtenidos evaluación de OC-SVM, linea azul hace referencia a la frontera de desición diseñada por el algoritmo suministrado los datos de el valor de PM2.5 y la temperatura correspondiente en ese instante, la morada hace referencia a la Covarianza Robusta, como metodo de detección identificado en sklearn, no osbtante es muy sensible a la presencia de valores atípicos en el conjunto de datos.

5. Diseño del sistema de adquisición y comunicación de datos.

A continuación, se muestra y explica la arquitectura propuesta para la implementación del sistema. Para un mejor entendimiento de la estructura y funcionamiento, se realiza un diagrama de bloques, actividades y secuencia. En cada uno de ellos se plasma una parte importante sobre la arquitectura del sistema.

Se realiza primero un diagrama de bloques con el fin de ver el funcionamiento del sistema seccionado por bloques, donde se refleja las entradas y salidas que se tienen en cada uno de ellos y cómo se relacionan entre sí. Toda la arquitectura se puede dividir en tres bloques generales: adquisición, transmisión y procesamiento y visualización.

El bloque de adquisición tiene la función de capturar datos mediante un sensor especial de material particulado y un módulo que almacena la información capturada de manera temporal, luego se realiza un procesamiento del dato para que pueda ser enviado al siguiente bloque. La siguiente etapa es la transmisión que es conformada por el gateway, donde los datos son guardados de manera temporal hasta lograr conexión con la siguiente fase. Por último, está el procesamiento y visualización de la información en el que se realiza un análisis y de acuerdo a esto se visualiza en el front end web, para que el usuario tome una decisión frente a la información que aparece en pantalla. La figura 15 muestra mencionado:

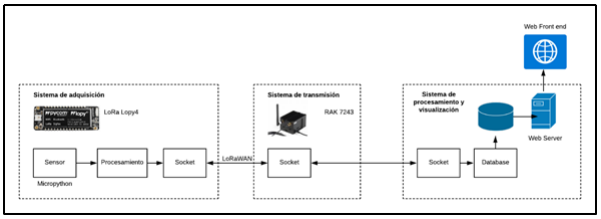

Figura 15: Diagrama de bloques

El diagrama de actividades permite ver de manera particular cada elemento mencionado en la figura 15, de este modo se ve la función que cumple dentro del bloque y las tecnologías que se usan. La figura 16 muestra dicho proceso de manera más detallada.

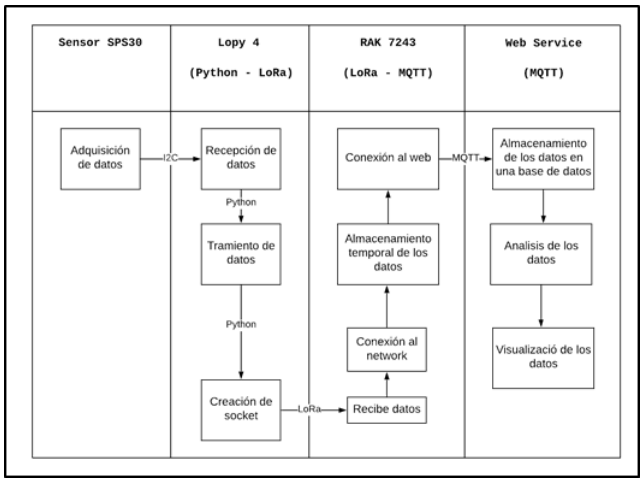

Figura 16: Diagrama de actividades

El último diagrama que se realiza es el de secuencia cuyo objetivo es mostrar el intercambio de datos con los diferentes objetos de la arquitectura para cumplir con una sentencia. La figura 17 muestra el comportamiento del sistema a través del tiempo.

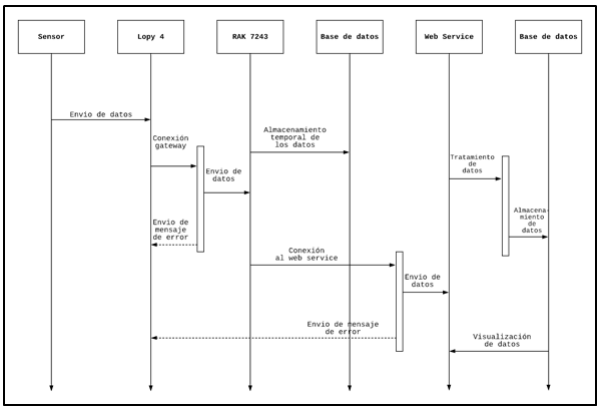

Figura 17: Diagrama de secuencias

Finalmente, con base en la figura 15 y los procesos descritos en las figuras 16 y 17 se concluye que la arquitectura del sistema será compuesta por tres elementos que son: un nodo que representa el proceso de adquisición de datos, un gateway que hace referencia función del bloque de transmisión y un servidor web que contiene el procesamiento de datos y visualización.

6. CONCLUSIONES

- Dado el uso de componentes de bajo consumo de energía, se logra plantear un sistema que pueda ser funcional las 24 horas del día los 7 días de la semana. Con el fin de poder capturar datos en diferentes franjas horarias para lograr un dataset que sea robusto.
- La ingeniería de datos realizada, indica que si bien no es posible determinar con exactitud una anomalía en la medición dados los datos y variables que se poseen, con el planteamiento y requerimientos establecidos se podría llegar generar un nuevo dataset, que satisfaga estas falencias e inclusive se permita realizar un taggeo del dataset determinando la evaluación de una anomalía o no, con base a las demás variables categóricas presentes.
- Se puede concluir que para planteamientos de área extensa lo mejor es emplear protocolos de comunicación que estén enfocados al internet de las cosas, debido a que estos permiten tener una carga a nivel de procesamiento menos pesada que otros protocolos y a nivel de consumo energético estos están diseñados para tener un consumo bajo.
- Es correcto precisar que la forma de llenar la serie de datos proveniente de la RMCAB"por medio de regresiones lineales, ARIMA, y herramientas de aprendizaje automático es validada por medio de parámetros estadísticos, dado que valores como, la correlación, el RMSE y el BIAS, presentan valores aceptables dado el criterio establecido en [1] y en [2].
- El uso de protocolos enfocados al internet de las cosas permiten una flexibilidad de uso para la visualización de los datos, ya que estos cuentan con plataformas propias con diferentes herramientas para percibir los datos capturados. Además de contar con la posibilidad de usar servidores propios que se ajustan a las necesidades del usuario.
- Como pronostico de series temporales ARIMA, funciona bien a corto plazo obteniendo un promedio de RMSE por estación en 13.688 g m3, sin embargo, en grandes volúmenes de datos el RMSE, tiende a llegar casi a 25.868 g m3 y en ocasiones hasta 74.868 g m3, esto indica que si se desea realizar el llenado de datos mediante un pronostico de series temporales con ARIMA, lo ideal seria realizar por año, o inclusive determinar el periodo de tiempo especifico y ajustar los valores del metodo para esta aproximación.

- Las redes neuronales son consideradas como cajas negras, por ende es demasiado complicado calcular o extraer información útil, como por ejemplo la importancia de las características, no obstante, es posible realizar una aproximación en cuanto a que variable categórica representa un peso mayor en el momento de realizar una inferencia, estoy es posible hacerlo gracias a el método de la biblioteca sklearn "Feature Importance".

Referencias

[1] Salazar, J., Silvestre, S. (n.d.). Internet de las cosas. Czech Republic: České vysoké učení technické v Praze.

[2] Master-internet-of-things.com. (2019). Ventajas y desventajas del uso de IoT – Máster en Internet of Things. [online] Available at: https://www.master-internet-of-things.com/ventajas-desventajas-del-uso-iot

[3] AuraPortal. (2019). Qué es la Inteligencia Artificial • Definición, ejemplos y casos de uso.. [online] Available at: https://www.auraportal.com/es/que-es-la-inteligencia-artificial

[4] Máster en Deep Learning : Universidad de Alcalá - Madrid. (2019). Ventajas y Desvenajas del uso de Inteligencias Artificiales - Máster en Deep Learning : Universidad de Alcalá - Madrid. [online] Available at: https://master-deeplearning.com/ventajas-desventajas-inteligencia-artificial

[5] Iberdrola. (2019). ¿Somos conscientes de los retos y principales aplicaciones de la Inteligencia Artificial?. [online] Available at: https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial

[6] Lin, Y. W., Lin, Y. B., Liu, C. Y. (2019). AItalk: a tutorial to implement AI as IoT devices. IET Networks, 8(3), 195-202.

[7] Knickerbocker, J., Budd, R., Dang, B., Chen, Q., Colgan, E., Hung, L. W., ... Narayanan, R. (2018, May). Heterogeneous integration technology demonstrations for future healthcare, IoT, and AI computing solutions. In 2018 IEEE 68th Electronic Components and Technology Conference (ECTC) (pp. 1519-1528). IEEE.

[8] González García, A. J. (2017). IoT: Dispositivos, tecnologías de transporte y aplicaciones.

[9] del Valle, B., David, J. IoT: Tecnologías, usos, tendencias y desarrollo futuro.

[10] T. M. Tatarnikova and I. N. Dziubenko, "Wireless Sensor Network Clustering Model,"2018 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), St. Petersburg, 2018, pp. 1-4.

[11] Sinha, R. S., Wei, Y., Hwang, S. H. (2017). A survey on LPWA technology: LoRa and NB-IoT. Ict Express, 3(1), 14-21.

[12] Machado González, M. L. (2019). Estudio de NB-IoT y comparativa con otras tecnologías LPWAN.

[13] Saiz Miranda, J. (2019). Estudio en detalle de NB-IoT: comparación con otras tecnologías LPWAN considerando diferentes patrones de tráfico.

[14] Archila Córdoba, D. M., Santamaría Buitrago, F. A. (2013). ESTADO DEL ARTE DE LAS REDES DE SENSORES INALÁMBRICOS. Tecnología Investigación Y Academia, 1(2). Recuperado a partir de https://revistas.udistrital.edu.co/index.php/tia/article/view/4437

[15] M. Zennaro, Ïntroducción a las Redes de Sensores Inalámbricos", ed, 2010.

[16] Azar, M. A., Tapia, M., García, J. L., Pérez, A. J. M. (2019, June). Inteligencia artificial de las cosas. In XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan)..

[17] Shana Pearlman. (2019) What is data integrity [online] Available at: https://www.talend.com/resources/what-is-data-integrity/

[18] Salesforce. (2018) Machine Learning y Deep Learning: aprende las diferencias [online] Available at: https://www.salesforce.com/mx/blog/2018/7/Machine-Learning-y-Deep-Learning-aprende-las-diferencias.html

[19] Witten, I. H., Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record, 31(1), 76-77.

[20] Hand, D. J. (2006). Data Mining. Encyclopedia of Environmetrics, 2.

[21] Wu, X., Zhu, X., Wu, G. Q., Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

[22] Rivero Pérez, Jorge Luis. (2014). Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), 52-73. [Online] Available at: http://scielo.sld.cu/scielo.php?script=sci$_arttextpid = S2227 - 18992014000400003lng = estlng = es.$

[23] Torres-Domínguez, O., Sabater-Fernández, S., Bravo-Ilisatigui, L., Martin-Rodríguez, D., García-Borroto, M. (2019). Anomalies detection for big data. Revista Facultad De Ingeniería, 28(50), 62-76. https://doi.org/10.19053/01211129.v28.n50.2019.8793

[24] Apache Kafka. (2017) Introduction to Apache Kafka [Online] Available at: https://kafka.apache.org/intro

[25] ML — Introduction to Kernel PCA (2019) [Online] Available at: https://www.geeksforgeeks.org/ml-introduction-to-kernel-pca/

[26] Vargas, F. A., Rojas, N. Y. (2010). Composición química y reconstrucción másica del material particulado suspendido en el aire de Bogotá. Ingenieria e Investigacion, 30(2), 105-115

[27] Casallas, A., Celis, N., Ferro, C. et al. Validation of PM10 and PM2.5 early alert in Bogotá, Colombia, through the modeling software WRF-CHEM. Environ Sci Pollut Res (2020). https://doi.org/10.1007/s11356-019-06997-9

[28] Casallas, A., Celis, N., Ferro, C. et al. WORKING PAPER Llenado de series de datos de 2014 a 2019 de PM2.5 por medio de una red neuronal y una regresión lineal.(2020).

[29] Jeonghun Choi, Seung Jun Lee. (2020) Consistency Index-Based Sensor Fault Detection System for Nuclear Power Plant Emergency Situations Using an LSTM Network. Sensors 20:6, pages 1651.

[30] N. Shahid, I. H. Naqvi, and S. B. Qaisar, "One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments," Artif. Intell. Rev., vol. 43, no. 4, pp. 515–563, 2015.

[31] Ahn, J., Shin, D., Kim, K., Yang, J. (2017). Indoor Air Quality Analysis Using Deep Learning with Sensor Data. Sensors (Basel, Switzerland), 17(11), 2476. https://doi.org/10.3390/s17112476

[32] Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107(44):776-782. doi:10.3238/arztebl.2010.0776