ISSN 1870-4069

Utilización de un sistema en tiempo real para la predicción de contaminación del aire

Carlos Lino-Ramírez, Rogelio Bautista-Sánchez, Sandra P. Bombela-Jiménez

Tecnológico Nacional de México en León, Guanajuato, México
carloslino@itleon.edu.mx, rogeliobautistasanchez@outlook.com, sandy.bombela@gmail.com

Resumen. En la actualidad existen varias ciudades afectadas con altos índices de contaminación, debido a varias cosas entre ellas, el crecimiento vehicular, empresas que arrojan contaminantes a la intemperie debido a sus procesos industriales, basura e incendios forestales, entre otras. En este artículo se presenta la propuesta de un sistema en tiempo real que puede estar monitoreando variables ambientales en varios puntos de la ciudad y hacer una predicción del comportamiento de dichas variables. Los datos provienen de datos proporcionados por el Sistema Estatal de Información de Calidad del Aire (SEICA) en el estado de Guanajuato. Se aplican técnicas de rellenado de datos para completar los valores perdidos, se realiza el etiquetado de la calidad del aire de acuerdo al semáforo proporcionado por el SEICA y, se realiza una transformación a la base de datos para el entrenamiento de la red neuronal que se utiliza para la predicción de la calidad del aire. Los datos adquiridos se normalizan, se agrupan, y con ellos se estructuran los componentes de predicción, y se hace también el análisis estadístico y de estructura.

Palabras clave: tiempo real, predicción, contaminación, clasificación.

Use of a Real-Time System to Predict Air Pollution

Abstract. Currently there are several cities affected with high levels of pollution, due to several things among them, vehicle growth, companies that throw pollutants out in the open because of their industrial processes, garbage and forest fires, among others. In this article we present the proposal of a real-time system that can be monitoring environmental variables in several points of the city and make a prediction of the behavior of these variables. The data comes from data provided by the State Air Quality Information System (SEICA) in the state of Guanajuato. Data filling techniques are applied to complete the missing values, the air quality labeling is done according to the semaphore provided by the SEICA and a transformation is made to the database for the training of the neural network that is used for the prediction of air quality. The acquired data are normalized, grouped, and with them the prediction components are structured, and the statistical and structure analysis is also made.

Keywords: real time, prediction, pollution, classification.

1. Introducción

1.1. Contaminación ambiental

Uno de los principales problemas que enfrenta la humanidad actualmente, es la contaminación, debido a que genera gran preocupación para las grandes ciudades, ya que de éste se desprenden diversos problemas para la salud humana [1]. El incremento de la contaminación en las ciudades, se da debido a varios factores, como por ejemplo la industria energética, la movilización urbana, entre otros. El utilizar recursos no renovables, como lo es el petróleo o el carbón, para la producción de energía genera emisiones contaminantes como el dióxido de azufre (SO2), monóxido de carbono (CO), entre otros. A nivel mundial, se utilizan 38 % derivados del petróleo para generar energía, 28 % utiliza carbón, 21 % gas natural, 6 % energía nuclear y solo el 7 % se centra en generar energía con recursos renovables [2]. Otra fuente de contaminación son los medios de transporte, de los que una gran parte de contaminantes es emitida por los automóviles [2]. Según las Naciones Unidas, hay alrededor de 7 billones de personas actualmente en el mundo [3], esto representa una enorme fuente de emisiones contaminantes, agravando el problema cada vez más, dado que las personas tienden a migrar a grandes ciudades, buscando mejores oportunidades de empleo, vivienda, servicios, entre otros. Esto provoca que las urbes continúen su expansión, y conlleva a que estén generando grandes emisiones de contaminación, lo que deteriora la calidad de aire y con ello acarrea problemas de salud. Según datos del Consejo Nacional de Población (CONAPO), el 72.3% de la población, en México, vive en zonas metropolitanas. Y según la ONU, en los próximos 10 años, las poblaciones rurales comenzarán a disminuir. Todo esto, genera que la salud de las personas de las grandes ciudades, se deteriore cada vez más. Las personas que viven en lugares con altos índices de contaminación, son más propensas a adquirir enfermedades del tipo respiratorias, como el asma o alergias [4].

1.2. Procesamiento Big Data

Se puede considerar que la era del Big Data" nace con el desarrollo de MapReduce y Hadoop como las primeras "tecnologías Big Data". Estas tecnologías se centraban en un enfoque de Batch Processing. Es decir, el objetivo era acumular todos los datos que se pudieran, procesarlos y producir resultados que se "empaquetaban" por lotes. Con este enfoque, Hadoop ha sido la herramienta más empleada. Es una herramienta realmente buena para almacenar enormes cantidades de datos y luego poder escalarlos horizontalmente mientras vamos añadiendo nodos en nuestro clúster de máquinas, ver figura 1.

Como se puede ver en la figura 1, el problema que aparece en este enfoque es que el retraso en tiempo que introduce disponer de un ETL que carga los datos para su procesamiento, no será tan ágil como hacerlo de manera continua con un enfoque de tiempo real. El procesamiento en trabajos batch de Hadoop MapReduce es el que domina en este enfoque. Y lo hace, apoyándose en todo momento de un ETL. Hasta la fecha la gran mayoría de las organizaciones han empleado este paradigma Batch. No era necesaria mayor sofisticación. Sin embargo, como ya comentamos anteriormente, existen exigencias mayores. Los datos, en muchas ocasiones, deben ser procesados en

tiempo real, permitiendo así a la organización tomar decisiones inmediatamente. Esas organizaciones en las que la diferencia entre segundos y minutos sí es crítica. Hadoop, en los últimos tiempos, es consciente de esta economía de tiempo real en la que nos hemos instalado. Por ello, ha mejorado bastante su capacidad de gestión. Sin embargo, todavía es considerado por muchos una solución demasiado rígida para algunas funciones. Por ello, hoy en día, solo es considerado el ideal en casos como cuando no se necesita un cálculo con una periodicidad alta, cuando los cálculos se deban ejecutar solo a final de mes y cuando la generación de informes es con una periodicidad baja.

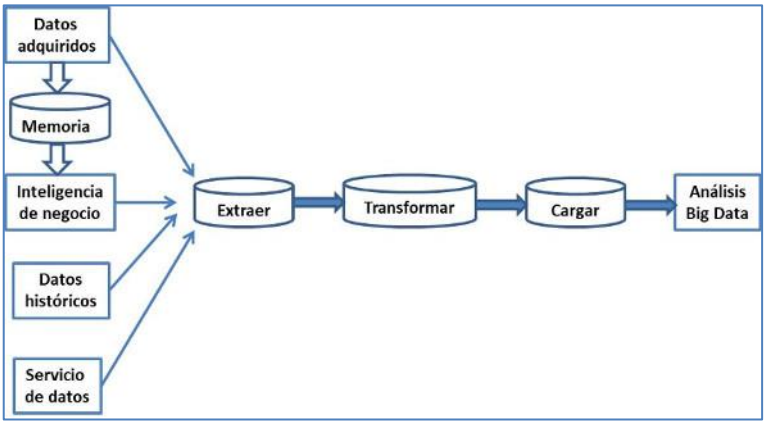

Fig. 1. Procesamiento batch con Big Data.

En los últimos años han surgido una serie de herramientas y tecnologías alrededor de Hadoop para ayudar en esa tarea de analizar grandes cantidades de datos. Para analizar las mismas, lo descomponemos en las cuatro etapas de la cadena de valor de un proyecto de Big Data:

1. Ingesta de datos:

Destacan tecnologías como: Flume: recolectar, agregar y mover grandes cantidades de datos desde diferentes fuentes a un data store centralizado. Comandos HDFS: utilizar los comandos propios de HDFS para trabajar con los datos gestionados en el ecosistema de Hadoop. Sqoop: permitir la transferencia de información entre Hadoop y los grandes almacenes de datos estructurados (MySQL, PostgreSQL, Oracle, SQL Server, DB2 y otras).

2. Procesamiento de datos:

Destacan tecnologías como: MapReduce: El proceso por el cual los datos que se obtienen en tiempo real van siendo capturados temporalmente para un posterior procesamiento. Hive: framework creado originalmente por Facebook para trabajar con el sistemas de ficheros distribuidos de Hadoop (HDFS). El objetivo no era otro que facilitar el trabajo, dado que a través de sus querys SQL (HiveQL) podemos lanzar consultas que luego se traducen a trabajos MapReduce. Pig: herramienta que facilita el análisis de grandes volúmenes de datos a través de un lenguaje de alto nivel. Su estructura permite la paralelización, que hace aún más eficiente el procesamiento de volúmenes de datos, así como la infraestructura necesaria para ello. Cascading: crear y ejecutar flujos

de trabajo de procesamiento de datos en clústeres Hadoop usando cualquier lenguaje basado en JVM (la máquina virtual de Java). De nuevo, el objetivo es quitar la complejidad de trabajar con MapReduce y sus trabajos. Es muy empleado en entornos complejos como la bioinformática, algoritmos de Machine Learning, análisis predictivo, Web Mining y herramientas ETL. Spark: facilita enormemente el desarrollo de programas de uso masivo de datos. Creado en la Universidad de Berkeley, ha sido considerado el primer software de código abierto que hace la programación distribuida accesible y más fácil para "más público" que los muy especializados.

3. Almacenamiento de datos

Destacan tecnologías como: HDFS: sistema de archivos de un cluster Hadoop que funciona de manera más eficiente con un número reducido de archivos de datos de gran volumen, que con una cantidad superior de archivos de datos más pequeños. HBase: permite manejar todos los datos y tenerlos distribuidos a través de lo que denominan regiones, una partición tipo Nodo de Hadoop que se guarda en un servidor. La región aleatoria en la que se guardan los datos de una tabla es decidida, dándole un tamaño fijo a partir del cual la tabla debe distribuirse a través de las regiones. Aporta, así, eficiencia en el trabajo de almacenamiento de datos.

4. Servicio de datos

En esta última etapa, en realidad, no es que destaque una tecnología o herramienta, sino que destacaría el "para qué" se ha hecho todo lo anterior. Es decir, qué podemos ofrecer/servir una vez que los datos han sido procesados y puestos a disposición del proyecto de Big Data [5].

1.3. Sistemas en tiempo real

A la hora de procesar grandes volúmenes de datos existen dos principales enfoques: procesar una gran cantidad de datos por lotes o bien hacerlo, en pequeños fragmentos, y en "tiempo real". Parece, así, bastante intuitivo pensar cuál es la idea del paradigma en tiempo real que se tratará. Este enfoque de procesamiento y análisis de datos se asienta sobre la idea de implementar un modelo de flujo de datos en el que los datos fluyen constantemente a través de una serie de componentes que integran el sistema de Big Data que se esté implantando. Por ello, se lee como procesamiento streaming o de flujo. Así, en tiempos muy pequeños, procesamos de manera analítica parte de la totalidad de los datos, y, con estas características, se superan muchas de las limitaciones del modelo batch. Por otro lado, una cosa es denominarlo tiempo real y otra es realmente pensar que esto se va a producir en verdadero tiempo real, ver figura 2.

Las limitaciones aparecen por que:

- Se debe disponer de suficiente memoria para almacenar entradas de datos en cola. Fíjense en la diferencia con el paradigma batch, donde los procesos de MapReduce podrían ser algo lentos, dado que escribían en disco entre las diferentes fases.
- La tasa de productividad del sistema debería ser igual o más rápida a la tasa de entrada de datos. Es decir, que la capacidad de procesamiento del sistema sea más ágil y eficiente que la propia ingesta de datos. Esto, de nuevo, limita bastante la capacidad de dotar de instantaneidad al sistema.

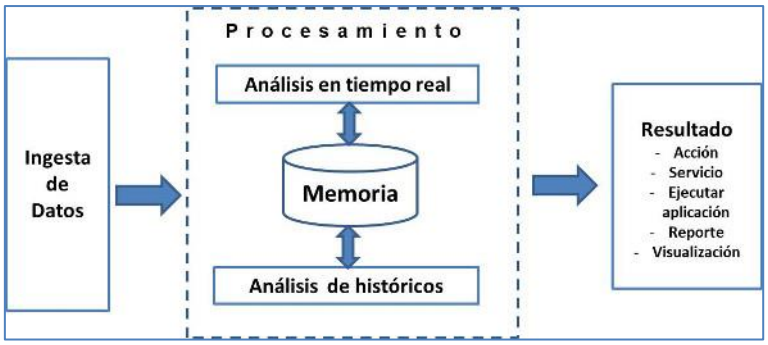

Fig. 2. Procesamiento en tiempo real.

Con la estructura mostrada en la figura 2 podremos procesar flujos de datos en tiempo real con sistemas big data, con algunos de los actuales modelos de procesamiento existentes para esta tecnología.

1.4. Sistema para la predicción de contaminación del aire

En este artículo se hace la propuesta de un sistema para predecir la contaminación del aire haciendo uso de un ambiente big data que permite la recolección de los datos en tiempo real y que nos permita predecir de la mejor manera posible la contaminación del aire, administrando de manera correcta desde la ingesta de datos, el procesamiento, el almacenamiento y el servicio de los datos gestionado con herramientas tecnológicas para Big Data.

2. Trabajos relacionados

A través de determinados algoritmos y empleando las variables y los dispositivos adecuados se pueden detectar desde problemas graves de salud pública hasta tendencias de mercado, pasando por los rendimientos de una cosecha con sistemas Big Data. Lucila Ballarino, responsable global de Transformación Digital de Fundación Telefónica, asegura que también existen ejemplos de gestión de datos climatológicos, de salud o de criminalidad. Pueden predecirse desastres naturales en forma de terremotos y tsunamis; analizarse rutas de propagación de virus en tiempo real para tomar decisiones rápidas y poder controlarlo lo antes posible, evitando pandemias; o incluso anticiparse a la realización de un crimen. Por otra parte, también desde el punto de vista medioambiental el Big data se está empleando para realizar simulaciones de cuál va a ser la evolución climatológica, de las corrientes del mar, de cómo va a afectar eso a la fauna y flora, comenta que permite poner en marcha acciones preventivas para controlarla y minimizar ese impacto. Desde el punto de vista medioambiental el Big data se está empleando para realizar simulaciones de cuál va a ser la evolución climatológica, de las corrientes del mar, de cómo va a afectar eso a la fauna y flora, contaminación ambiental, etc. A nivel local, estas herramientas pueden ser muy útiles para el trabajo de campo de las ONG. Pueden aprovecharse de las redes sociales, por ejemplo, para poder identificar y

localizar a aquellas personas en riesgo de exclusión que necesitan ayuda y que se resisten a pedirla, por ejemplo. Otro ejemplo es el de la segmentación de campañas para hacer llegar a los donantes correctos las causas que de verdad le interesan, favoreciendo así las donaciones. O aplicar esa segmentación para captar a nuevos voluntarios para que participen en campañas enfocadas en sus gustos [6].

En la literatura de proyectos de investigación sobre este tema encontramos actualmente una gran variedad atacando de distintas maneras este tipo de problema, como los autores Donnelly, Misstear y Broderick que en su artículo presentan un modelo para producir pronósticos de calidad del aire en tiempo real con alta precisión y alta eficiencia computacional. Las variaciones temporales en los niveles de dióxido de nitrógeno (NO2) y las correlaciones históricas entre la meteorología y los niveles de NO2 se utilizan para estimar la calidad del aire 48 horas antes.

La regresión no paramétrica del núcleo se utiliza para producir factores linealizados que describen las variaciones en las concentraciones con la velocidad y dirección del viento y, además, para producir factores estacionales y diurnos. La base del modelo es una regresión lineal múltiple que utiliza estos factores junto con los parámetros meteorológicos y la persistencia como predictores.

El modelo fue calibrado en tres sitios urbanos y un sitio rural, y el modelo ajustado final logró valores R de entre 0.62 y 0.79 para pronósticos por hora y entre 0.67 y 0.84 para pronósticos máximos diarios. La validación del modelo utilizando cuatro parámetros de evaluación modelo, un índice de concordancia (IA), el coeficiente de correlación (R), la fracción de valores dentro de un factor de 2 (FAC2) y el sesgo fraccional (FB) arrojaron buenos resultados. La IA para pronósticos de 24 horas del NO2 por hora estuvo entre 0.77 y 0.90 en sitios urbanos y 0.74 en el sitio rural, mientras que para pronósticos máximos diarios estuvo entre 0.89 y 0.94 para sitios urbanos y 0.78 para el sitio rural. Se observaron valores de R de hasta 0,79 y 0,81 y valores de FAC2 de 0,84 y 0,96 para las predicciones máximas por hora y diarias, respectivamente. El modelo solo requiere datos de entrada simples y muy pocos recursos computacionales. Resultó ser un medio preciso y eficiente para producir pronósticos de calidad del aire en tiempo real [7].

En otro artículo realizado por Adams y Kanaroglou abarcan este tema desde el uso de sistemas de mapeo de riesgo ya que estos sistemas proporcionan datos de concentración absoluta o las concentraciones se utilizan para derivar un índice de calidad del aire, que proporciona el riesgo de contaminación del aire para una mezcla de contaminantes del aire con un valor único. Cuando la información de riesgo se presenta como un valor único para una región entera, informa sobre la variación espacial dentro de la región. Sin una comprensión de la variación local, los residentes solo pueden tomar una decisión parcialmente informada al elegir las actividades diarias. El valor único generalmente se proporciona debido a un número limitado de unidades de monitoreo activas en el área. En su trabajo superan ese problema aprovechando las técnicas móviles de control de la contaminación del aire, la información meteorológica y la información sobre el uso de la tierra para cartografiar los riesgos para la salud de la contaminación del aire en tiempo real. Proponen un enfoque que pueda proporcionar información mejorada sobre el riesgo para la salud al público mediante la aplicación de modelos de redes neuronales dentro de un marco inspirado en la regresión del uso de la tierra. Las campañas móviles de monitoreo de la contaminación del aire se modelaron con una serie de variables de predicción que incluían información sobre las características del

uso de la tierra circundante, las condiciones meteorológicas, las concentraciones de contaminación del aire de los monitores de ubicación fija, e información de tráfico durante el tiempo de recolección. Se modelan partículas finas y dióxido de nitrógeno. Durante el proceso de adaptación del modelo, reservan el veinte por ciento de los datos para validar las predicciones. Los rendimientos de los modelos se miden con un coeficiente de determinación de 0.78 y 0.34 para PM2.5 y NO2, respectivamente. Aplican una medida de importancia relativa para identificar la importancia de cada variable en la red neuronal para superar parcialmente los problemas de la caja negra de los modelos de redes neuronales [8].

En otro artículo los autores Nath y Patil hacen uso de la variable MH (mixing height) para representar la profundidad de dispersión de la capa límite atmosférica, es un parámetro de entrada crucial en los modelos de contaminación atmosférica. Sin embargo, existe una enorme incertidumbre en su estimación ya que no es una variable directamente medible. En general, la MH se calcula a partir de las mediciones de radiosonda dos veces al día desde la estación meteorológica más cercana, especialmente en los países en desarrollo. Sin embargo, estos valores extrapolados causan errores severos en la predicción ya que MH depende del sitio y del tiempo. En su trabajo aplican un modelo de crecimiento de altura de mezcla in situ simple (IMG, in situ mixing height growth), que puede estimar los valores de MH en tiempo real in situ a partir de mediciones superficiales de viento y temperatura fácilmente disponibles, a algunos modelos de predicción de contaminación atmosférica comúnmente utilizados. Los modelos de caja (BM, box models) los utilizan para predicciones a gran escala, pero suponen una altura tope constante, aunque su precisión depende en gran medida de su variación. IMG lo aplicaron a un modelo de caja fotoquímica, ya que la formación de ozono depende en gran medida de la insolación y está controlada por valores de MH en tiempo real. Las concentraciones de ozono estimadas por IMG-BM mostraron una mejora del 13% en comparación con las estimadas a partir de los valores de radiosonda extrapolados habituales. La aplicación de IMG a GDM para las industrias mostró que el modelo IMG mejora considerablemente la precisión de predicción y puede utilizarse de manera rentable [9].

El autor Elbir propone el modelo meteorológico CALMET y su modelo de dispersión de soplo CALPUFF y se utilizaron para predecir la dispersión de las emisiones de dióxido de azufre de las fuentes de calor industriales y domésticas en Izmir, la tercera provincia más grande de Turquía. El dominio de modelado cubrió un área de 80 × 100 km centrada en el área metropolitana de Izmir con un espaciado de malla de 1000 m. Los análisis estadísticos se llevaron a cabo para evaluar el rendimiento del modelo mediante la comparación de las series temporales pronosticadas y medidas de las concentraciones de dióxido de azufre en cuatro estaciones de monitoreo utilizando dos métodos principales: raíz del error cuadrático medio (RMSE) y un índice de acuerdo (d). El índice de acuerdo varió de 0,51 a 0,77 en cuatro estaciones de monitoreo y el RMSE total varió de 0.36 a 0.66 para el año 2000. El rendimiento general del modelo para cuatro estaciones de monitoreo se encontró bueno con una precisión de aproximadamente 68%. El acuerdo de las predicciones y medidas del modelo fue mejor para dos estaciones de monitoreo urbano Karsiyaka y Bornova, en comparación con las otras estaciones urbanas Alsancak y Konak [10].

Los autores Kim et al., presentan en su artículo una propuesta para controlar la calidad del aire interior (IAQ) en las estaciones de metro, las estrategias de control basadas

en el modelo predictivo que no tiene el efecto de la temperatura debido a variaciones estacionales, se han utilizado actualmente. En este trabajo, se proponen modelos dependientes de la temporada para el monitoreo y la predicción de IAQ, que se ocupan de los cambios estacionales.

Los datos en tiempo real de diversos contaminantes (la concentración de la plataforma PM10 y PM2.5, la temperatura, la humedad y la concentración de nitrógeno) durante marzo de 2008 a febrero de 2009 se obtienen de la estación de metro de Seúl. La prueba MANOVA se ha llevado a cabo para conocer la medida cuantitativa de las diferencias entre diferentes conjuntos de datos de tres estaciones (primavera y otoño, verano e invierno).

Los métodos de regresión de PCA y PLS se aplican en conjuntos de datos de un año (para desarrollar un modelo global) y cuatro temporadas (para desarrollar modelos estacionales) para monitorear y predecir el IAQ. Los resultados de este estudio muestran que los modelos estacionales pueden predecir los datos futuros de PM10 y PM2.5 antes que el modelo global [11].

En otro artículo relacionado los autores Han et al. diseñan un marco para recopilar y almacenar diversos dominios de datos sobre las causas de las enfermedades cardiovasculares, y construyen una base de datos integrada de big data. Una variedad de bases de datos de código abierto se integró y migraron a dispositivos de almacenamiento distribuidos. La base de datos integrada estaba compuesta de datos clínicos sobre enfermedades cardiovasculares, encuestas nacionales de salud y nutrición, información estadística, censos de población y vivienda, datos de administración meteorológica y datos del servicio de evaluación y evaluación del seguro médico.

El marco se compone de datos, velocidad, análisis y capas de servicio, todos almacenados en dispositivos de almacenamiento distribuidos. Finalmente, propusieron un marco para un sistema de predicción de enfermedades cardiovasculares basado en la arquitectura lambda para resolver los problemas asociados con los análisis en tiempo real de big data. Este sistema puede usarse para ayudar a predecir y diagnosticar enfermedades, como enfermedades cardiovasculares [12].

A diferencia de los trabajos que se describen en esta sección, con este proyecto se pretende sentar las bases para establecer estaciones de sensores que puedan estar continuamente monitoreando las principales partículas contaminantes a través de un sistema big data que recolecte toda la información en un sistema central e inmediatamente se procese la información para predecir por algunos periodos de tiempo establecidos el comportamiento de la contaminación en base a los registros actuales e históricos.

3. Componentes de predicción

Existe una distinción y separación de las partes de un conjunto de datos hasta llegar a conocer sus elementos basados en su entorno. Cada uno de los datos tiene características propias que los determinan y distinguen claramente de los demás. De ellos se eligen uno entre otros con el objetivo de predecir la magnitud física que mide la sensación subjetiva de calor o frío. El poder generar un aviso previo de un hecho que va a suceder para poder evitar una proximidad de un daño o peligro ordenado por clases. Antes de entrar en concreto con términos de la predicción es necesario procesar la información en diferentes etapas, para las cuales los datos que describen el comportamiento del set

de datos que fueron generados y que son importantes junto con las características seleccionadas para los procedimientos de descubrimiento de escenarios en el entorno de trabajo. En seguida se da una explicación de los métodos que se están usando:

Bdscan. Para la estimación del número de escenarios para la identificación de multiescenarios. Este tiene un enfoque basado en densidad, modelando los clusters como cúmulos de alta densidad de puntos. Por lo cual, si un punto pertenece o no a un clúster, debe estar cerca de un conjunto de otros puntos de dicho clúster. Haciendo uso de las características seleccionadas, hará el descubrimiento de los diferentes escenarios que se encuentran en un entorno de trabajo. Esta sub etapa tiene la función también de poder detectar outliers que pueden afectar el descubrimiento de estos multiescenarios, los cuales son descartados para repetir el proceso con el conjunto de datos limpio. Como resultado de esta sub etapa se tiene un conjunto de datos representativos de las variables por cada sensor exentos de outliers y un parámetro de estimación del número de clusters identificados.

K-means. Para la partición del entorno por agrupamientos. Este tiene como objetivo la partición de un conjunto de *n* observaciones (representaciones estadísticas de las variables promedio por día de cada sensor en su ubicación geográfica original) en k grupos en el que cada observación pertenece al grupo cuyo valor medio es el más cercano. El parámetro k es el valor resultante de la estimación del número de clusters identificados.

Chi-Square. Para la selección de características. Es evidente que el uso de todas las características de cada escenario permiten describirlo completamente, analizando el comportamiento se puede entender que no es necesario tener toda la información para poder obtener una buena representación del entorno, además el uso de esfuerzo computacional se puede ahorrar cuando se seleccionan los mejores representantes de un escenario del entorno de trabajo. Es por ello que cada agrupamiento es enviado a esta sub etapa para la selección de las mejores características (sensores con su información) que representar mejor al escenario.

LSTM. Para la predicción de las variables en estudio. Para la tarea de predicción se emplean redes de tipo Long Short Term Memory (LSTMs), siendo éstas un caso especial de redes neuronales tradicionales y de tipo RNN. Este tipo de especial de redes neuronales son ampliamente utilizadas en problemas de predicción en series temporales debido a que su diseño permite recordar la información durante largos períodos y facilita la tarea de hacer estimaciones futuras empleando períodos de registros históricos.

4. Flujo de datos

Para poder procesar el set de datos, es necesario seleccionar el mejor conjunto que tenga un periodo con la mayor cantidad de datos, este periodo seleccionado se vuelve el principal insumo para cada proceso. El siguiente paso es la realización de una reconstrucción de datos usando la técnica MICE (Multivariate Imputation by Chained Equations), con la información completa se realiza una prueba de normalidad para descubrir si el conjunto de datos tiene una distribución normal, posteriormente el conjunto de datos completo ahora puede ser analizado para entender mejor el comportamiento

que tiene obteniendo la información estadística del mismo, dichos resultados estadísticos, proveen la información necesaria para el descubrimiento de los escenarios que se encuentran en el entorno de trabajo.

En términos de los componentes de predicción, el uso de DBSCAN se aplica con 3 características representativas de los diferentes conjuntos de datos generados por cada sensor, para lo cual es indispensable tener la ubicación de los sensores que captan la información y el mejor representante de la información por día, usando DBSCAN con estos parámetros se pueden identificar un numero N aproximado de agrupamientos, que a continuación se ratifican usando K-Means.

Hasta este punto el conjunto de datos no ha cambiado, el conjunto de datos completo es necesario para las etapas más importantes, el conjunto de datos estadísticos representativos del conjunto de datos general que se usa en esta etapa y permite generar un mejor desempeño para el procesamiento de la información más adelante.

Como resultado del descubrimiento de los multiescenarios, el conjunto de datos completos se particiona según los multiescenarios descubiertos generando diferentes subsets de datos. A continuación, se hace uso del método de selección por Chi Square, tomando, de cada subset generado, los mejores representantes, reduciendo así la cantidad de información de cada sensor en el periodo de tiempo seleccionado. Cada subset reducido de cada cluster ahora está listo para poder ser procesado por el modelo de predicción, el cual toma la información para entrenar el modelo, una vez entrenado el modelo, ya está listo para poder procesar la información del entorno en términos reales, generando ventanas de predicción de un intervalo de tiempo en el futuro definido por el usuario.

Como resultado de esta predicción se genera el catálogo de documentos de registros de predicción, que es enviado a un predictor de riesgos que evalúa cada documento dentro del catálogo para identificar si pertenece a un conjunto que se considera de riesgo o a un conjunto que se considera dentro de los intervalos establecidos como normales. El resultado de este procesamiento final de la información es la probabilidad existente en el catálogo de documentos de registro de predicción de que en algún momento en el futuro se registre un evento de riesgo.

5. Modelo de predicción

La base de datos fue etiquetada de acuerdo al semáforo de calidad del aire del SEICA que se divide en cinco categorías (Buena, Satisfactoria, No satisfactoria, Mala y Muy mala), para cada una de las instancias se verifica a que categoría pertenece cada uno de los cinco contaminantes obteniendo así cinco posibles categorías, después la categoría asignada a la instancia la peor de estas cinco. Los límites para definir a que categoría pertenece cada contaminante se encuentran especificados en la tabla 1. En los años 2014 al 2017 no existieron registros que pertenecían a la categoría "Mala" y "Muy Mala".

Tabla 1. Clasificación de las cinco variables contaminantes.

Contaminante	**PM_{10}**	**O_3**	**SO_2**	**NO_2**	**CO**
Unidad de medida	**Ug/m^3**	**Ppb**	**Ppb**	**Ppb**	**Ppm**
Bueno	0-54	0-64	0-99	0-198	0-9
Satisfactoria	55-74	65-69	100-109	190-209	9-10

No satisfactoria	75-174	70-130	110-174	210-315	11-15
Mala	175-274	131-184	175-239	316-420	16-22
Muy mala	>275	>185	>240	>420	>22

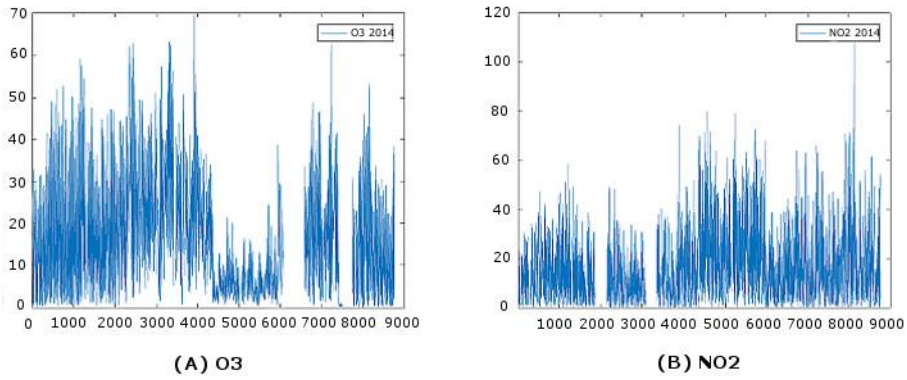

Fig. 3. Comportamiento del O3 y del NO2 en 2014.

Las bases de datos cuentan con un histórico del año 2014, 2015, 2016 y 2017. Cada año tiene registrado el mes, el día y la hora (de la hora 0 hasta la hora 23), es decir que en un año tenemos aproximadamente entre 8757 registros y 8781 (por ser año bisiesto). En cada una de las variables censadas de todos los años, existen valores perdidos como se muestra en la figura 3 (inciso a) O3 e inciso b) NO2), donde se puede mostrar que el eje x representa el periodo (las horas) y en el eje y el valor que toma cada uno de los contaminantes.

Se realizó el mismo método de graficado en todos los años para ver el comportamiento de cada una de las variables. También se realizó un estudio más minucioso para ver la distribución de los datos.

6. Resultados

Después de haber extrapolado los datos faltantes, se aplica una operación de retraso para transformar la base de datos [13]. Es decir, para predecir la calidad del aire de la próxima hora se tienen que incluir en el procesamiento no solo los datos actuales sino también datos pasados, los últimos dos o tres registros pasados como se muestra en la tabla 2 donde la transformación de la base de datos para predecir la etiqueta (calidad del aire) de la instancia 4, se utilizan los atributos de las instancias 1, 2 y 3 como se muestra en la tabla 3.

Tabla 2. Base de datos antes de ser transformada.

Instancia	Atr1	Atr2	Atr3	Atributo
1	A1	B1	C1	E1
2	A2	B2	C2	E2
3	A3	B3	C3	E3
4	A4	B4	C4	E4

Tabla 3. Transformación de la base de datos.

Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Etiqueta
A1	B1	C1	A2	B2	C2	A3	B3	C3	E4

La base de datos fue dividida en cuatro bases de datos (una por cada año) y se entrenaron cuatro modelos de MLP, obteniendo así resultados por cada uno de los años.

La tabla 4 muestra los resultados de precisión por la base de datos de cada año, utilizando las 3 técnicas utilizadas para el rellenado de datos y una de las técnicas con alguna modificación (el reetiquetado), dando como mejores resultados de porcentaje de clasificación la técnica de múltiple imputación con reetiquetado.

Tabla 4. Resultados de la clasificación con diferentes técnicas.

Año	**Instancia**	**Clasificación con extrapolación**	**Clasificación con técnica de promedio móvil**	**Clasificación con imputación**	**Clasificación con imputación reetiquetado**
2014	8757	93.11%	93.20%	93.38%	97.56%
2015	8757	95.98%	93.47%	92.88%	96.25%
2016	8781	94.20%	91.49%	89.65%	96.70%
2017	8757	91.46%	94.17%	90.67%	95.78%

7. Conclusiones

Los resultados obtenidos permiten tener confianza en las predicciones y así poder mantener informada a las personas sobre los cambios en los niveles de la calidad del aire en un futuro cercano, se realizaron ejemplos de cómo obtener información y procesarla en tiempo real para calcular su predicción, esto debido a que se contaba con una estación de monitoreo, pero no con todos los sensores necesarios para calcular la calidad del aire, ya que algunos no se pudieron conseguir, pero con los que se tenía los resultados fueron aceptables como se pudo observar.

Referencias

1. Yang, X., Du, J., Liu, S., Li, R., Liu, H.: Air pollution source estimation profiling via mobile sensor networks. In: 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). https://doi.org/10.1109/CITS.2016.7546456 M4 Citavi (2016)
2. Bose, B.: Global Warming: Energy, Environmental Pollution, and the Impact of Power Electronics. IEEE Industrial Electronics Magazine 4(1), 6–17. https://doi.org/10.1109/MIE.2010.935860 (2010)
3. Guo, D., Zhang, Y., He, L., Zhai, K., Tan, H.: Chebyshev-polynomial neuronet, WASD algorithm and world population prediction from past 10000-year rough data. In: Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015, pp. 1702–1707 https://doi.org/10.1109/CCDC.2015.7162194 (2015)
4. Garrido-Lestache, J.S., Rodríguez García, V.: Las enfermedades alergicas. Libro de las enfermedades alérgicas de la fundación BBVA (2012)

5. Rayón, A.: Tecnologías de ingesta de datos en proyectos "big data" en tiempo real, https://blogs.deusto.es/bigdata/tecnologias-de-ingesta-de-datos-en-proyectos-big-data/. Último acceso: 2019/01/01
6. Albendea, G.L.: Big data: una herramienta de predicción útil para el sector social, https://www.compromisoempresarial.com/innovacion_social/2017/-11/big-data-una-herramienta-de-prediccion-util-para-el-sector-social/ (2017)
7. Donnelly, A., Misstear, B., Broderick, B.: Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmospheric Environment 103(2), 53–65. https://doi.org/10.1016/j.atmosenv.2014.12.011 (2015)
8. Adams, M.D., Kanaroglou, P.S.: Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of Environmental Management, 168, pp. 133–141 https://doi.org/10.1016/j.jenvman.2015.12.012 (2016)
9. Nath, S., Patil, R.S.: Prediction of air pollution concentration using an in situ real time mixing height model. Atmospheric Environment, 40(20), 3816–3822 https://doi.org/10.1016/j.atmosenv.2006.02.034 (2006)
10. Elbir, T.: Comparison of model predictions with the data of an urban air quality monitoring network in Izmir, Turkey. Atmospheric Environment 37(15), 2149–2157. https://doi.org/10.1016/S1352-2310(03)00087-6 (2003)
11. Kim, M., Sankararao, B., Kang, O., Kim, J., Yoo, C.: Monitoring and prediction of indoor air quality (IAQ) in subway or metro systems using season dependent models. Energy and Buildings 46, 48–55. https://doi.org/10.1016/j.enbuild.2011.10.047 (2012)
12. Han, S.H., Kim, K.O., Cha, E.J., Kim, K.A., Shon, H.S.: System framework for cardiovascular disease prediction based on big data technology. Symmetry 9(12), 1–11 https://doi.org/10.3390/sym9120293 (2017)
13. Soares, F.M., Souza, A.M.F.: Neural network programming with java. Second edition, Editor: Packt publishing, 270 pp. (2017)