ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELECTRÍCA Y ELECTRÓNICA / DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICACIONES Y REDES DE INFORMACIÓN

DETECCIÓN DE EVENTOS DE CONTAMINACIÓN ACÚSTICA INDUSTRIAL BASADO EN UNA RED DE SENSORES

ALMACENAMIENTO, PROCESAMIENTO DE EVENTOS ACÚSTICOS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TELECOMUNICACIONES

RONALDO ALEXANDER ALMACHI MURILLO

ronaldo.almachi@epn.edu.ec

DIRECTOR: DR. TARQUINO FABÍAN SÁNCHEZ ALMEIDA

tarquino.sanchez@epn.edu.ec

Quito, octubre 2022

CERTIFICACIONES

Yo, RONALDO ALEXANDER ALMACHI MURILLO declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

RONALDO ALEXANDER ALMACHI MURILLO

Certifico que el presente trabajo de integración curricular fue desarrollado por RONALDO ALEXANDER ALMACHI MURILLO, bajo mi supervisión.

DR. TARQUINO FABÍAN SÁNCHEZ ALMEIDA

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

RONALDO ALEXANDER ALMACHI MURILLO

DR. TARQUINO FABÍAN SÁNCHEZ ALMEIDA

DEDICATORIA

Dedico este Trabajo de Integración Curricular
a mis padres, hermanos y amigos, quienes
siempre me apoyaron a lo largo de mi
trayectoria académica de forma incondicional.

AGRADECIMIENTO

Agradezco en primer lugar a mis padres y hermanos, quienes siempre me brindaron su ayuda de diferentes maneras hasta lograr mi objetivo, también agradezco al Dr. Tarquino Sánchez por su paciencia y comprensión durante el desarrollo del presente proyecto, y finalmente agradezco a mis amigos, con quienes compartí gratos momentos.

ÍNDICE DE CONTENIDO

CERTIFICACIONES I
DECLARACIÓN DE AUTORÍA II
DEDICATORIA III
AGRADECIMIENTO IV
ÍNDICE DE CONTENIDO V
RESUMEN VII
ABSTRACT VIII
1 INTRODUCCIÓN 1
1.1 Objetivo general 3
1.2 Objetivos específicos 3
1.4 Marco teórico 5
1.4.1 Red de sensores 5
1.4.2 Parámetros ambientales 5
1.4.3 Bases de datos 9
1.4.4 Aprendizaje Automático 9
1.4.5 Aprendizaje no supervisado 10
1.4.6 Programación 13
1.4.7 Lenguaje de programación: Python 13
2 METODOLOGÍA 15
2.1 Recolección y almacenamiento de datos 16
2.1.1 Recolección de datos 16
2.1.2 Almacenamiento de datos 17
2.2 Tratamiento de datos 18
2.2.1 Interpretación de los datos 18
2.2.2 Histogramas de los parámetros ambientales 21
2.2.3 Descripción de la base de datos 24
2.2.4 Normalización de los datos 26
2.3 Implementación del algoritmo de aprendizaje automático 27
2.3.1 Elección del valor de K óptimo 28
2.3.2 Ejecución del algoritmo K-Means con el valor de K óptimo 30
2.3.3 Gráfico de los clústeres 31
2.4 Análisis de resultados 33
3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES 37

3.1 Resultados 37
3.2 Conclusiones 43
3.3 Recomendaciones 44
4 REFERENCIAS BIBLIOGRÁFICAS 45
5 ANEXOS 46

RESUMEN

En este estudio se presenta el proceso llevado a cabo para clasificar datos de tipo cuantitativo con nivel de medida continuo como el ruido, la temperatura y la humedad, que se encuentran ubicados en un repositorio digital (Opendatasoft), para dicho propósito se usa tecnologías de la información y algoritmos de aprendizaje automático. El primer capítulo del Trabajo de Integración Curricular presenta la fundamentación teórica que nos será de utilidad para el desarrollo del componente, por lo que dicha información está enfocada en: describir los parámetros que se van a usar en la clasificación, investigar normas internacionales referente a los parámetros mencionados, describir los modelos de aprendizaje automático que se van a implementar, realizar una descripción la base de datos, presentar el lenguaje de programación que se va a usar para generar el código y finalmente el método para procesar datos. En el segundo capítulo se hablará enteramente del proceso llevado a cabo para la clasificación de datos, por lo que se ha divido en etapas acorde a lo presentado en la fundamentación teórica. Finalmente, el último capítulo trata acerca de las pruebas de funcionamiento, conclusiones y recomendaciones a las que se llegó una vez culminado todo el componente.

PALABRAS CLAVE: Aprendizaje automático, agrupamiento, clasificación, ruido, temperatura, humedad, Python.

ABSTRACT

This study presents the process carried out to classify quantitative data with a continuous level of measurement such as noise, temperature, and humidity, which are located in a digital repository (Opendatasoft), for this purpose information technologies are used and machine learning algorithms. The first chapter of the Curriculum Integration Work presents the theoretical foundation that will be useful for the development of the component, so that information is focused on: describing the parameters that will be used in the classification, investigating international standards of reference to the parameters, describe the machine learning models that are going to be implemented, make a description of the database, present the programming language that is going to be used to generate the code and finally the method to process data. In the second chapter, the process carried out for data classification will be fully discussed, which is why it has been divided into stages according to what is presented in the theoretical foundation. Finally, the last chapter deals with the performance tests, conclusions and recommendations that were reached once the entire component was completed.

KEYWORDS: Machine learning, clustering, classification, noise, temperature, humidity, Python.

1 INTRODUCCIÓN

Hoy en día los procesos industriales se han desarrollado y expandido alrededor del mundo gracias a los avances tecnológicos, por lo que es muy común encontrar que pequeñas y medianas empresas, de todo tipo, usen procesos industriales a una escala menor de lo que lo haría una gran empresa, sin embargo, sin importar si una empresa es pequeña o grande se va a generar contaminación ambiental. En base a este problema organizaciones como la Organización Mundial de Salud (OMS) han definido una serie de recomendaciones que intentan salvaguardar la integridad tanto de las personas como del medio ambiente, en consecuencia, el gobierno de cada país tiene la potestad de acotar las recomendaciones de dichas organizaciones o generar normar y leyes con cambios acorde a la situación de cada país, con el propósito de ayudar a mitigar efectos de la contaminación ambiental producida por el sector industrial, transporte público, vuelos comerciales, etc.

La contaminación ambiental en la mayoría de los casos es resultado de la actividad humana, como la emisión de gases de efecto invernadero a la atmósfera o la explotación de recursos naturales renovables y no renovables, como consecuencia la temperatura, que es una magnitud física que se ve afectada por la contaminación ambiental, lo que provoca cambios de temperatura en zonas inesperadas, siendo uno de los casos más conocidos el derretimiento de los polos debido al aumento de temperatura. Otro efecto que produce el parámetro antes mencionado es el incremento de la humedad, que, si bien no provoca efectos adversos en la salud de una persona, proporciona un ambiente idóneo para que se generen microorganismos que afecten la integridad de la gente, y finalmente el ruido ambiental que se considera un tipo de contaminación acústica, causa efectos adversos en el oído humano generando una discapacidad casi irreversible, si no se toman medidas preventivas.

Una forma de conocer la contaminación ambiental existente en un lugar en específico es mediante el monitoreo de parámetros ambientales, que se consigue con el uso de sensores que pueden estar implementados como una red, para así cubrir un área mucho más extensa y además que se puedan comunicar entre si de una forma sencilla, dicha red puede estar estructurada de tal manera que se adapte al ambiente a ser monitoreado, para tomar medidas preventivas y correctivas en caso de ser necesario. Es importante usar los dispositivos adecuados para registrar las variaciones de los parámetros ambientales, puesto que dichos datos serán la materia prima para este estudio.

El propósito general del Trabajo de Integración Curricular (TIC) es detectar eventos de contaminación acústica basados en una red sensores, este proceso se ha divido en 3

componentes los cuales consisten en: captura y transporte de datos, procesamiento de datos y visualización de datos correspondientemente, en este caso nos enfocaremos enteramente en el componente 2, para lo cual es necesario conocer la naturaleza de los datos que van a procesar y así poder realizar un análisis basándose en criterios ya establecidos referentes a los parámetros ambientales como temperatura, humedad y ruido, haciendo énfasis en el ruido ya que será la principal variable a ser estudiada.

La temperatura y humedad guardan cierta relación entre sí, pues al aumentar la temperatura también incrementa la humedad, por otro lado, el ruido ambiental no guarda relación con los parámetros mencionados, sin embargo, es importante mencionar que en el caso de la contaminación acústica se debe tomar como referencia el ruido pico o máximo registrado, y el ruido promedio por largos intervalos de tiempo basándose en las recomendaciones de la OMS. El propósito de usar la temperatura, la humedad y el ruido, para el Trabajo de Integración Curricular, es que son magnitudes físicas fáciles de medir y además se puede encontrar gran cantidad de datos generados por equipos de monitoreo ambiental.

En el presente Trabajo de Integración Curricular correspondiente al componente 2, busca generar un nuevo parámetro a partir de la temperatura, humedad y ruido, lo que nos permite clasificar estos datos de una forma más sencilla, donde dicho nuevo valor se encuentra en un rango finito de categorías o grupos, que indican si predomina la temperatura, la humedad o el ruido, o su vez una combinación de 2 parámetros y en el peor de los casos los 3, esto se lo va a conseguir con la ayuda de modelos de aprendizaje automático que pueden ser supervisados y no supervisados, para lo cual se necesita recursos como: bases de datos y lenguajes de programación de alto nivel.

El estudio se ha divido en 3 capítulos los cuales consisten en:

- **Marco teórico:** En esta sección se presentará información referente a: los parámetros que se van a usar (temperatura, humedad y ruido ambiental), recomendaciones por organismos internacionales referentes a las variables mencionadas, bases de datos, entornos de programación, lenguajes de programación de alto nivel y modelos de aprendizaje automático.
- **Metodología:** En este apartado se describe el método llevado a cabo para conseguir representar con un parámetro la temperatura, humedad y ruido, con modelos de aprendizaje automático.

- **Resultados, conclusiones y recomendaciones:** En este capítulo se presentan los resultados obtenidos de la sección anterior donde se le da el respectivo análisis, finalmente se presentan las conclusiones y recomendaciones.

1.1 Objetivo general

Implementar un modelo de aprendizaje automático que permita obtener un parámetro en función de la temperatura, humedad y ruido ambiental, para clasificar dichos valores, considerando aquellos que pueden tener afectación en la salud humana haciendo énfasis en el ruido ambiental con base en los parámetros establecidos por la OMS, de una forma más sencilla.

1.2 Objetivos específicos

1. Buscar bases de datos de sistemas de monitoreo ambiental, y procesar dichos datos para obtener la información más representativa de la base de datos.
2. Generar un parámetro nuevo en función de la temperatura, humedad y ruido ambiental, que se encuentre en un rango finito de grupos, con la ayuda de modelos de aprendizaje automático.
3. Clasificar los parámetros ambientales, con el propósito de buscar tendencias o patrones que afectan la salud de las personas basándonos en el caso del ruido por los estándares de la OMS.
4. Presentar los resultados obtenidos mediante gráficos, haciendo especial énfasis el ruido ambiental y como este parámetro afecta a la salud humana.

1.3 Alcance

La contaminación ambiental es un problema que se ha presentado varios años atrás, pero a medida que el tiempo pasa los efectos son más evidentes y desastrosos, es porque organismos como OMS han desarrollado recomendaciones, para mitigar y prevenir los efectos de la contaminación ambiental y así salvaguardar la salud de las personas. Una forma de conocer que tanto ha cambiado el ambiente es realizar mediciones de magnitudes físicas como la temperatura, humedad, ruido, etc. Y así generar medidas preventivas con mayor precisión y eficacia.

Para dicho propósito primero se debe obtener datos ambientales que van a ser recopilados por Sistemas de Monitoreo Ambiental (EMS por sus siglas en inglés), y partir de estos datos

se puede determinar cómo ha cambiado la temperatura, humedad, ruido ambiental, etc., en un sector en particular.

Para el caso del TIC se ha seleccionado la base de datos generada por un EMS ubicado en Australia, que fue desarrollado por una empresa australiana con sede en Perth, ARCS Group en asociación con la Universidad Tecnológica de Sydney (UTS) como parte del proyecto TULIP, que consiste en la investigación y ejecución de proyectos de ciudades inteligentes por parte de la UTS. La base de datos fue obtenida mediante un repositorio digital (Opendatsoft).

Una vez que se ha seleccionado la base datos, se va a usar los valores de temperatura, humedad y ruido ambiental, para generar un nuevo parámetro, con la ayuda de modelos de aprendizaje automático, que nos permita clasificar los datos acordes un criterio que será desarrollado a lo largo del TIC basándonos en recomendaciones o normas referentes a la contaminación ambiental, por lo que el nuevo parámetro va a estar en un rango finito que se conocerán como grupos o categorías. Para tal propósito, se va a usar lenguajes de programación de alto nivel, tanto para la implementación de los modelos de aprendizaje automático y análisis de resultados mediante gráficos generados a partir de la información resultante.

1.4 Marco teórico

1.4.1 Red de sensores

Una red de sensores se define como una infraestructura de comunicaciones o grupo de transductores enfocados en el monitoreo, registro y respuesta ante algún fenómeno en un lugar en específico. Un sensor usualmente puede monitorear parámetros físicos como la temperatura, humedad, presión, ruido, velocidad del viento, intensidad luminosa, etc [1].

Para que los sensores de una red se puedan comunicar entre ellos, se necesita de infraestructura y protocolos de comunicación, que pueden afrontar varios retos como recursos limitados, altos costos y lugares inhóspitos propensos a causar fallas en el sensor. Sin embargo, esto se ha podido solucionar con el avance tecnológico mediante sensores de bajo costo, bajo consumo energético, pequeños, multifuncionales, resistentes entre otras características [1].

La red de sensores esta conformada por nodos sensores que se caracterizan por ser pequeños, portátiles y livianos, que se comunican entre sí, uno o más nodos pueden funcionar como sumideros que son los encargados de comunicarse directamente con el usuario a través de un medio de comunicación guiados (cables de cobre, fibra óptica) o no guiados (bluetooth, wifi, satélite) [1].

Actualmente, las redes de sensores inalámbricas se han vuelto muy populares debido a que pueden desempeñar funciones de una red de sensores tradicional, sin su alto coste y dificultad de implementación [2].

Para el TIC, vamos a usar bases de datos que registran parámetros como la temperatura, humedad y ruido, estos datos se obtienen a partir de sistemas de monitoreo ambiental que se comunican entre sí, que conformar una red de sensores de EMS, dicha red mantiene sus comunicaciones mediante protocolos como LoRaWAN, Wifi, etc.

1.4.2 Parámetros ambientales

Los parámetros ambientales nos permiten conocer las condiciones de ambiente de un sector o lugar en específico, dichos valores son conseguidos mediante sensores que pueden determinar diferentes magnitudes físicas como: temperatura, humedad, ruido, gas en el aire, etc.

Para el presente caso del TIC, nos enfocaremos y usaremos en los parámetros de temperatura, humedad y ruido, esto ya que son valores bien definidos, fáciles de medir y sencillos de interpretar con la herramienta adecuada.

Temperatura

La temperatura es comúnmente conocida como una medida de calor y frio, siendo esta definición dada en base a las sensaciones fisiológicas de las personas, sin embargo, no es posible asignar valores numéricos basándonos únicamente en sensaciones. Afortunadamente existen materiales que presentan cambios de temperatura repetibles y predecibles, lo que nos puede ayudar a establecer una base para la medición precisa de la temperatura; uno de los materiales más usados, hasta la invención de termómetros electrónicos, fue el mercurio el cual se expande y contrae según varie la temperatura. [3]

Las escalas nos proveen una base común para las mediciones de temperatura, actualmente se usa el SI con la escala Celsius (ºC), que va desde 0 a 100 ºC, y la escala Fahrenheit (ºF), que va desde 32 a 212 ºF, esto se conoce como escala de dos puntos, dado que los valores de temperatura van entre los rangos ya mencionados.[3]

La temperatura en escala Celsius (ºC) y en escala Fahrenheit (ºF), se relacionan como se indica en la ecuación 1.1.

$$T(\text{ºF}) = 1.87 * T(\text{ºC}) + 32 \tag{1.1}$$

Humedad

La humedad es un parámetro que mide la cantidad de vapor de agua que hay en el aire. Otro parámetro muy usado en la mayoría de los sistemas de monitoreo ambiental es la humedad relativa, valor que mide la cantidad de agua en el aire, pero en relación con la cantidad máxima de vapor de agua (humedad). Es importante mencionar que esta variable guarda relación la temperatura ya que esta al ser mayor, el vapor de agua en el aire también aumenta.[4]

La humedad relativa (HR), al ser una variable dependiente de la temperatura se puede expresar mediante una fórmula como se indica en la ecuación 1.2 [5].

$$HR = \frac{e(T)}{e_S(T)} * 100\ \% \tag{1.2}$$

$HR =$ humedad relativa en %

$e(T) =$ presión parcial real del vapor de agua en aire húmedo, en P_a

$e_S(T) =$ presión parcial del vapor de agua en aire húmedo saturado, en P_a

Ruido

En este caso primero se debe dar un concepto al sonido, que se define como un fenómeno producido por perturbaciones mecánicas que se propagan como un movimiento ondulatorio en el aire u otro medio, que se caracteriza por tener velocidad de sonido (c) , frecuencia (f) y longitud de onda (λ), como se describe en la ecuación 1.3 [6].

$$\lambda = \frac{c}{f} \tag{1.2}$$

En base a la definición de sonido, el ruido es conocido como una clase de sonido no deseado, que dependiendo de la situación puede afectar el bienestar fisiológico y psicológico de las personas.[6]

El tipo de ruido que va a ser objeto de análisis en el presente TIC es el ambiental, este es emitido por todo tipo de fuentes como tráfico vial, aéreo y ferroviario, construcciones, sector industrial, vecindarios (discotecas, cafeterías, etc.) y obras públicas, excepto por el ruido producido en el ambiente laboral, que se conoce ruido industrial y que es estudiando dentro de la empresa. [6]

Para conocer el nivel de ruido debemos medir el nivel de presión de sonora (L_p), que se define como la relación que existe entre la presión sonora del sonido más intenso y la del sonido más débil, lo que ha llevado a usar una escala logarítmica en unidas de dB, como se indica en la ecuación 1.4.

$$L_P = 20 \log\left(\frac{P}{P_{ref}}\right) \tag{1.4}$$

$P_{ref} =$ Presión de referencia a un tono audible

$P =$ Presión sonora

Además, para realizar dichas mediciones se toma como referencia el nivel de ponderación A, el cual se encuentra estandarizado para el oído humano con un rango de frecuencias de 20 Hz a 20 kHz, a una distancia que va de 0.5 a 1 metro de la fuente de ruido, siendo la unidad dBA la cual va a ser usada para el TIC [6].

Es de suma importancia analizar los efectos causados por el ruido, puesto que la percepción del sonido por parte de las personas es fundamental en las interacciones sociales, además que ruidos muy altos pueden afectar la integridad física de las personas [6].

La OMS ha presentado un documento con directrices para el ruido ambiental, siendo que estas pueden ser de utilidad para realizar diversas gestiones como la medición y prevención de la contaminación acústica a nivel mundial. Actualmente tenemos múltiples fuentes de ruido ya sea en áreas urbanas o rurales, pero es un problema mucho mayor en áreas urbanas debido a la gran cantidad de fuentes de contaminación acústica, es por eso que mediante un estudio y la colaboración de varios participantes se ha desarrollado la Tabla 1, donde se puede encontrar valores de referencia para el ruido en ambientes específicos [6].

Para medir el ruido en el ambiente nos basaremos en dos criterios el primero es el nivel de presión sonora L_P, descrito anteriormente, y el nivel de presión sonora equivalente $L_{p_{eq}}$ que tiene un principio similar, pero por un periodo T de tiempo.

Tabla 1. Valores de referencia para el ruido en entornos específicos [6].

Ambientes Específicos	**Efecto (s) crítico (s) en la salud**	$L_{p_{eq}}$ **Equivalente (dBA)**	T **Tiempo base (horas)**	L_P **Instantánea (dBA)**
Sala de estar al aire libre	Molestias graves, diurnas y nocturnas.	55	16	-
Viviendas en interiores	Inteligibilidad del habla y molestias moderadas, durante el día y la noche. Alteración del sueño, durante la noche.	30-35	8-16	45
Viviendas en exteriores	Alteración del sueño, ventana abierta (valores exteriores.)	45	8	60
Salas de clases en escuelas y preescolares, en interiores	Inteligibilidad del habla, alteración en la comunicación.	35	Durante clase	-
Dormitorios del preescolar, en interiores	Alteración del sueño	30	Tiempo de sueño	45
Patio de la escuela	Molestia (fuente externa)	55	Durante la actividad	-
Hospital: salas de espera y habitaciones	Alteración del sueño, durante la noche. Alteración del sueño, durante el día y la noche.	30	8-16	40
Sector industrial, áreas comerciales y de tráfico, en interiores y exteriores	Discapacidad auditiva.	70	24	110
Ceremonias, festivales y eventos	Discapacidad auditiva (para personas expuestas al menos 5 veces al año).	100	4	110

Vía pública	Discapacidad auditiva.	85	1	110
Música y otros sonidos a través de auriculares o audífonos.	Discapacidad auditiva.	85	1	110
Sonidos de juegos mecánicos, fuegos artificiales y armas de fuego.	Discapacidad auditiva	-	-	120 (niños) 140 (adultos)

1.4.3 Bases de datos

Las bases de datos son un método para almacenar datos estructurados, donde dicha información puede provenir de distintas fuentes, este tipo de tecnologías nos permiten asegurar la integridad de los datos y facilitar la manipulación de los mismos[7]. Además, las bases de datos se usan en todo tipo de organizaciones (gobiernos, empresas, tiendas, hogar, etc.), en donde las actividades cotidianas tienen contacto en algún momento con las distintas bases de datos, que pueden ser directa o indirectamente [8].

Existen distintas bases de datos, pero una de las más conocidas y usadas es el modelo relacional, que se basa en la noción matemática de la relación que puede ser representado mediante una tabla o arreglo bidimensional. Dicha tabla se encuentra conformada por filas y columnas, donde las columnas cuentan con un nombre del atributo que representan y los valores que se registran en cada columna. Por otro lado, las filas de las tablas nos permiten relacionar los diferentes valores dados por las columnas y así presentar un dato asociado con diferentes parámetros [8].

En el Trabajo de Integración Curricular se utiliza una base de datos relacional, donde dicha información fue obtenida a partir de una de red de sensores del Sistema de Monitoreo Ambiental (EMS) [9]. Estos sensores fueron desarrollados por una empresa australiana con sede en Perth, ARCS Group, en asociación con la University of Tecnology Sydney, como parte de proyecto TULIP, proyecto enfocado en el desarrollo de ciudades inteligentes [9].

Los base de datos obtenida por los sensores fue recopilada usando la red LoRaWAN de IoT comunitario del Consejo: The Things Network [9].

1.4.4 Aprendizaje Automático

El Aprendizaje Automático, o mejor conocido como Machine Learning en inglés, es un subcampo de la informática que se encarga del desarrollo de algoritmos, que se basan en un conjunto de datos que representan algún fenómeno, donde dichos datos pueden

provenir de la naturaleza (temperatura, humedad, emisión de gases, etc.) o ser generados por el humano (datos personales, datos íntimos, datos bancarios, etc.). Otra definición del machine learning se conoce como el proceso enfocado a resolver un problema, mediante la recopilación datos y construcción de un modelo estadístico basado en los datos recolectados, lo cual nos puede ayudar a solucionar un problema [10].

Tipos de aprendizaje

El aprendizaje puede ser: supervisado, no supervisado y semi supervisado.

Aprendizaje supervisado

Es una técnica que nos permite deducir un método para etiquetar datos, a partir de un conjunto de datos previamente etiquetado, es importante mencionar que este tipo de métodos se usa cuando se tiene un conjunto de datos y sus etiquetas, pero no se tiene una regla clara para asignarle dicha etiqueta [10].

Aprendizaje no supervisado

Es una técnica que nos permite tomar un conjunto de datos y agruparlos acorde a un criterio, lo que nos ayuda a reducir la dimensión de los datos que pueden tener mucha o poca relación entre ellos [10].

Aprendizaje semi supervisado

Es una técnica que nos permite deducir un método para etiquetar datos, al igual que en el aprendizaje supervisado, pero con conjuntos de datos etiquetados y sin etiquetar, con la diferencia que en este caso se busca producir un mejor modelo [10].

Para este trabajo se utilizar el aprendizaje no supervisado debido a que los datos de temperatura, humedad y ruido ambiental no están etiquetados.

1.4.5 Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica que nos permite agrupar datos, sin etiquetas, basados en un análisis, como el agrupamiento o clustering de datos, que tiene como objetivo encontrar subgrupos homogéneos dentro de un conjunto de datos, para dicho propósito se han desarrollado algoritmos que se caracterizan por basarse en la distancia entre observaciones. Los algoritmos más usados son: agrupamiento K-Medias y agrupamiento jerárquico [11].

Para el TIC, se basa en el algoritmo de agrupamiento K-Medias o mejor conocido en inglés K-Means, esta decisión será explicada a lo largo del documento.

Algoritmo de agrupamiento K-Medias o K-Means

Se define como un algoritmo del tipo particional, ya que divide los objetos en un número K de clústeres previamente especificados, sin tomar en consideración una estructura jerárquica, donde fácilmente se puede aplicar a problemas de agrupación por similitud, en consecuencia puede ayudar a la compresión cualitativa y cuantitativa de grandes conjuntos de datos de N dimensiones [12].

Este algoritmo funciona de manera iterativa, es decir, divide de una forma óptima al conjunto inicial de datos en K clústeres, parámetro indicado anteriormente, que se basa en la minimización de la distancia interna, ecuación 1.5, que propone una especificación matemática que describe el proceso llevado a cabo por el algoritmo K-Means [12].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1.5}$$

$d(x,y) =$ distancia entre dos objetos

$n =$ dimensión del conjunto

$i =$ número de iteración

Proceso del algoritmo K-Means [12].

1. Como primer paso para cada ejemplo de x_k se calcula el prototipo más cercano A_g y que se puede incluir en la lista del prototipo, ecuación 1.6.

$$A_g = argmin\{d(x_k, A_i)\} \quad \forall_i = 1 \ldots . n \tag{1.6}$$

2. Después de introducir todos los ejemplos, cada uno de los prototipos A_k, tiene un conjunto de elementos, ecuación 1.7.

$$(A_k) = \{x_{k1}, x_{k2} \ldots \ldots . x_{km}\} \tag{1.7}$$

3. A continuación, el prototipo se desplaza hasta el centro del conjunto de ejemplos, ecuación 1.8.

$$A_k = \frac{\left(\sum_{i=1}^{m} x_{k_i}\right)}{m} \tag{1.8}$$

4. Como siguiente paso, el procedimiento es repetido hasta que ya no haya desplazamiento de prototipos. Las entradas k, se dividen en regiones y el prototipo de cada región estará en el centro de cada una para reducir las distancias cuadráticas euclidianas para cada uno de los patrones de entrada y el centro más cercano. Lo que minimiza el valor de J, ecuación 1.9.

$$J = \sum_{i=1}^{k} . \sum_{i=1}^{k} M_{i,n} d_{EUCL}(x_n - A_i)^2 \quad (1.9)$$

5. Finalmente, se tiene que m es el conjunto de patrones, d_{EUCL} es la distancia euclidiana, x_n el ejemplo de entrada, A_i el prototipo de la clase i y $M_{i,n}$ la función que indica la pertenencia del ejemplo n a la región i. Siendo que si la función vale 1 el prototipo A_i es el más cercano al valor de x_n, y en el caso contrario la función llega a tener un valor de 0, ecuación 2.0.

$$M_{i,n} = \begin{cases} 1 & si \quad d_{EUCL}(x_n - A_i) < d(x_n - A_s) \quad \forall_s \neq i, s, = 1,2, \dots, k \\ 0 & Dado\ el\ caso\ contrario \end{cases} \quad (2.0)$$

Método del Codo

Una forma de asegurarnos que el valor de los clústeres K es óptimo, es usar el método del codo que tiene el propósito de explicar y verificar que el número K es apropiado tomando como referencia la función de costo descrita en la ecuación 2.1.

$$J = \sum_{i=1}^{k} . \sum_{x \epsilon C_i} |x - C_i|^2 \quad (2.1)$$

$J =$ Función de costo

$x =$ Elemento de los clústeres C_i

$K =$ Es el número de clústeres de $|C_i|$

Cuando el valor de K, para la función de costo, es menor que el número de clústeres reales dicha función tendrá un valor pequeño, lo ideal es que para estos casos la función de costo no sea un valor alto; y a medida que K aumente y se acerca al valor de clúster óptimo, por lo que el valor de la función de costo disminuirá drásticamente y mantendrá la misma tendencia para valores K superiores, es mediante este patrón que se puede reconocer el número de clústeres óptimos [13].

Análisis de Componentes Principales (PCA)

El propósito del PCA, es la reducción de dimensiones de un espacio de observación de dos objetos que son parte del estudio, usualmente esto se da con N objetos a dimensiones 2 o 3 componentes. Esta reducción se obtiene en base a la creación de combinaciones lineales de variables que caracterizan algún fenómeno o estudio, además el PCA debe satisfacer una serie de condiciones matemáticas y estadísticas [14].

El método del codo y el PCA, son métodos matemáticos que nos ayudan a corroborar la correcta ejecución de algoritmo K-Means, en este TIC no se les da especial énfasis a los métodos matemáticos usados ya que todos van a ser implementados en Python, que tiene librerías que nos permiten implementar todos estos métodos con una función.

1.4.6 Programación

La programación se conoce como el proceso de tomar un algoritmo (conjunto ordenado de operaciones sistémicas que nos permite encontrar la solución a un problema) y codificarlo con la ayuda de algún lenguaje de programación, para que sea ejecutado por una computadora. Es importante mencionar que para la elaboración de un algoritmo se necesita el planteamiento de un problema que requiera una solución mediante una serie de pasos que en su mayoría pueden ser propuestos como problemas matemáticos [15].

Para la implementación del algoritmo K-Means, nos vamos a valer de la programación, usando lenguajes de programación de alto nivel y entornos de programación que nos faciliten la implementación de modelos de aprendizaje automático.

1.4.7 Lenguaje de programación: Python

Los lenguajes de programación nos permiten ejecutar instrucciones en lenguaje de máquina para realizar una tarea en específico, este lenguaje puede ser de bajo nivel, el cual consiste en una serie de pasos predefinidos que nos ayudan a cumplir un objetivo en específico con la característica de que es un tanto extenso y difícil de escribir, que puede estar propenso a errores. Por otro lado, los lenguajes de alto nivel son mucho más sencillos de escribir, esto gracias a que las instrucciones para cumplir una tarea son mucho más sencillas y cortas de implementar, pero con la desventaja de que se tiene que llevar a cabo un proceso adicional, lo que produce un retraso en su ejecución[16].

Python es considerando un lenguaje de programación de alto nivel con semántica dinámica y además se encuentra orientado a objetos. Cuenta con estructuras de datos integradas de alto nivel que se complementan con escritura y enlaces dinámicos, esto facilita su uso para el rápido desarrollo de aplicaciones incluso nos permite conectar diferentes componentes que usen Python. La sintaxis con la que trabaja es sencilla y fácil de aprender ya que se enfatiza en la legibilidad, por lo que se reduce el costo de mantenimiento del programa, además admite módulos y paquetes, lo que permite modularidad del programa y reutilización del código. Python cuenta con un intérprete y una extensa biblioteca estándar que se encuentran disponibles en código fuente y en sistema binario disponible para los principales sistemas operativos con libre distribución[17].

Uno de los principales atractivos que ofrece Python a los programadores es que aumenta la productividad, esto gracias a que no hay ningún paso de compilación, puesto que el ciclo de edición-prueba-depuración es sumamente rápido. Esto implica que depurar programas en Python es sencillo por lo que un error o una entrada incorrecta no generan una falla de

segmentación. Por otro lado, si el intérprete descubre un error genera una excepción y si no se detecta una excepción el intérprete imprime un seguimiento de la pila. Un depurador de nivel de fuente permite la inspección de variables locales y globales, la evaluación de expresiones arbitrarias, el establecimiento de puntos de interrupción, el paso a través del código una línea la vez, etc. El poder introspectivo de Python se puede apreciar ya que el depurador está escrito en Python[17].

2 METODOLOGÍA

La contaminación ambiental es un factor que ha aumentado a lo largo del tiempo debido a múltiples razones, siendo una de estas la actividad humana, es por lo que hoy en día se tiene la necesidad de llevar un control mucho más riguroso de los fenómenos físicos que se ven afectados por la contaminación ambiental como la humedad, temperatura y ruido. Los parámetros mencionados pueden ser monitoreados por sensores que registran las variaciones de dichos fenómenos físicos, acorde a un estándar que se usa a nivel internacional.

El monitoreo de parámetros como la temperatura, humedad y ruido, se lo realiza mediante sistemas de monitoreo ambiental que registran variaciones de estos valores, ya sea de forma instantánea o almacenados en bases de datos para un posterior análisis. Si bien se puede llevar un monitoreo en tiempo real del cómo cambian dichos parámetros, también es recomendable poder reconocer y entender patrones que indiquen un cambio de estas variables, para así poder generar medidas preventivas o en el peor de los casos mitigar efectos producidos por estos cambios.

Los modelos de aprendizaje automático nos permiten realizar diversas funciones como, técnicas para etiquetar datos, agrupar datos acordes a un criterio, reducir dimensión de datos, reconocer patrones en grandes conjuntos de datos, reconocimiento de imágenes, etc.

Para el presente TIC, vamos a implementar un algoritmo para agrupar datos, que es el algoritmo K-Means, algoritmo que nos permite encontrar grupos en conjuntos de datos mediante el principio de la mínima distancia entre dos objetos, mientras que se usa el método del codo para escoger un número óptimo de clústeres o grupos.

Los modelos de aprendizaje automático actualmente son desarrollados en lenguajes de alto nivel como Python, por tanto, son muy sencillos de implementar con la librería adecuada, así mismo una vez implementado estos modelos, se puede realizar un análisis basado en los resultados obtenidos, donde dichos resultados pueden ser presentados mediante gráficos que nos muestren información útil y comprensible, con el propósito de tomar medidas preventivas y correctivas bajo ciertos criterios, todo esto refiriéndonos a fenómenos físicos los cuales han cambiado su comportamiento debido a la contaminación ambiental.

Para cumplir con lo propuesto en los párrafos anteriores se va a seguir con el siguiente proceso:

- Recolección y almacenamiento de datos
- Tratamiento de datos
- Implementación del algoritmo de aprendizaje automático
- Análisis de resultados
- Presentación de análisis

2.1 Recolección y almacenamiento de datos

2.1.1 Recolección de datos

Parámetros como la temperatura, humedad y ruido, son magnitudes físicas que pueden ser percibidas por las personas mediante los sentidos, pero si queremos darle un valor un cuantitativo se debe usar materiales que reaccionen a estos tres parámetros de una forma predecible y repetitiva. Los sensores de temperatura, humedad y ruido nos permiten registrar estos cambios basándonos en un estándar o recomendaciones desarrolladas por organizaciones como la OMS, ISO, etc. De tal forma que puedan ser representados mediante un valor y una unidad de medida, en el caso de la temperatura se usa los grados Celsius, para el caso de la humedad tenemos la HR representada en % y finalmente para el ruido se tiene los dBA, unidad estandarizada para el oído humano.

Una red que integra múltiples sensores puede ser denominada sistema de monitoreo, y si este dispositivo tiene el propósito de registrar magnitudes físicas del medio ambiente, se la conocerá como Sistema de Monitoreo Ambiental (EMS), dicho sistema puede ser más o menos complejo, pues todo depende del alcance que se le quiera dar al EMS.

Se utiliza la base de datos relacionales que ha sido obtenida por el EMS desarrollado por la University of Tecnology Sidney, ya que cuenta con una gran cantidad de EMS distribuidas por el territorio australiano a lo largo de Newcastle, como se indica en la Figura 1, estos sensores nos permiten registrar parámetros como hora, fecha, huso horario, nombre de la estación de monitoreo, ubicación, temperatura, humedad, ruido pico, ruido, promedio y concentración de gases en el aire, además esta red de monitoreo ambiental ha funcionado desde el 7 de agosto del 2019 hasta el 31 de julio del 2022 con un total de 179898 de registros.

Es importante mencionar que los EMS recopilaron los datos usando la red LoRaWAN de IoT comunitario del consejo: The Things Network.

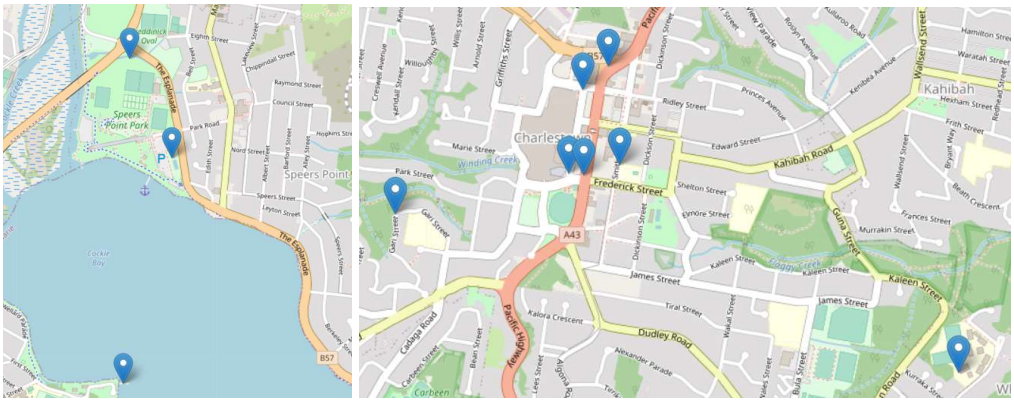

Figura 1. Ubicación de los EMS desarrollado por la UTS. Fuente: El Autor

2.1.2 Almacenamiento de datos

Una vez que las diferentes estaciones de monitoreo registran los parámetros ambientales, estos deben ser almacenados de tal manera que sean sencillos de acceder y manipular, por lo que la red de EMS guarda dichos valores en bases de datos del tipo relacional, es decir los diferentes valores son almacenados en forma de tablas, donde las columnas tienen un nombre que describen el parámetro ambiental y valor numérico, además de la hora, fecha, ubicación y nombre de la estación donde se registró la medición realizada, por otro lado las filas relacionan cada uno de los valores presentados en las columnas asignándoles un valor único que es el tiempo, Figura 2.

	Time	Device Name	Temperature (°C)	Humidity (%)	Average Sound (dBA)	Peak Sound (dBA)	Location
0	2020-07-31T23:55:19+10:00	EMS Five Islands Roundabout Speers Point	6.5	100.0	56.0	84.0	-32.95804,151.61749
1	2020-07-31T23:53:20+10:00	EMS Charlestown Public School	11.2	100.0	58.0	73.0	-32.965146,151.696981
2	2020-07-31T23:51:49+10:00	EMS Charlestown Square North Piazza Bus Stop	11.3	100.0	61.0	71.0	-32.962313,151.695294
3	2020-07-31T23:48:38+10:00	EMS Charlestown Square South Piazza	10.4	100.0	60.0	70.0	-32.965482,151.694651
4	2020-07-31T23:46:14+10:00	EMS Gari Street Charlestown	9.1	100.0	57.0	67.0	-32.967,151.68681

Figura 2. Base de datos del EMS con los 5 primeros valores. Fuente: El Autor

Actualmente, la base de datos del EMS de la UTS se encuentra disponible en un repositorio digital de manera gratuita en Opendatasoft (https://n9.cl/ruidoambiental) [9], donde inclusive se provee de herramientas para visualizar los datos mediante gráficos, esta sección es de suma importancia ya que la base de datos será la materia prima para el TIC, ya que la implementación de los algoritmos será tomando como referencia la base de datos del EMS.

2.2 Tratamiento de datos

Para esta sección se usará el lenguaje de programación de alto nivel, como es el Python, en conjunto con el entorno de programación, Google Colab, debido a que estas herramientas están enfocadas a la ciencia de datos, pues Python permitirá implementar modelos de aprendizaje automático de una forma sencilla usando funciones preestablecidas, mientras que Google Colab cuenta con paquetes para ciencia de datos preinstalados, lo que hace a estas herramientas idóneas para el desarrollo de este trabajo, además es importante mencionar que Google Colab, ofrece una entorno de programación online con recursos propios.

2.2.1 Interpretación de los datos

Como primer punto debemos guardar la base de datos en una variable para sea fácil de manipular, esto se logrará con ayuda de librerías pandas, es importante designar el símbolo que será reconocido como separación de las columnas, en este caso ";", por otro lado, para que el programa reconozca los decimales se ha asignado el carácter "." , como se indica en la Figura 3.

```
#Lectura de la base de datos con pandas
contaminacion = pd.read_csv('/content/drive/MyDrive/Noveno Semestre/Diseño de TIC/Bases de datos/environmental-monitoring-system-historical-to-31-july-2020.csv', sep=';', decimal='.')
```

Figura 3. Lectura de la base de datos. Fuente: El Autor

Con la ayuda de la función describe(), vamos a tener los siguientes resultados.

- count: Este parámetro nos indica la cantidad de valores no vacíos, es decir que tienen un número.
- media: Este valor indica la media de la columna analizada.
- std: Con este parámetro se indica la desviación estándar.
- min: Indica el mínimo valor de una columna.
- 25%: Indica el percentil considerando el 25%
- 50%: Indica el percentil considerando el 50%.
- 75%: Indica el percentil considerando el 75%.
- Max: Indica el máximo valor de una columna.

Esto aplicado a la base de datos nos da el siguiente resultado, como se indica en la Figura 4.

```
contaminacion.describe()
```

	Temperature (°C)	Humidity (%)	Average Sound (dBA)	Peak Sound (dBA)
count	179898.000000	179898.000000	10133.000000	160943.000000
mean	20.745944	83.553430	65.069081	79.455093
std	5.701885	19.528507	6.441112	10.879190
min	2.100000	0.000000	52.000000	0.000000
25%	17.000000	74.000000	60.000000	71.000000
50%	20.800000	90.000000	64.000000	80.000000
75%	24.100000	100.000000	70.000000	88.000000
max	51.000000	100.000000	89.000000	122.000000

Figura 4. Resultado de la función describe en la base de datos. Fuente: El Autor

Para el caso del Average Sound y Peak Sound, es importante notar que el número de valores no vacíos es menor a la Temperatura y Humedad, cuando deberían ser iguales, lo que indica que en esos valores no vacíos se ha registrado un parámetro de tipo nulo, por lo que se debería reemplazar estos valores de nulo por cero.

Usando el comando select_dtypes('object').nunique(), vamos a determinar si algún argumento que no sea de valor numérico se repite a lo largo de la base de datos, como se indica en la Figura 5.

```
contaminacion.select_dtypes('object').nunique()
```

```
Time           178781
Device Name        10
Location           10
dtype: int64
```

Figura 5. Valores no numéricos repetidos. Fuente: El Autor

El resultado presentado en la Figura 4, indica que los valores de tiempo son únicos para cada caso, mientras que el nombre de los EMS, y ubicación se repiten en un grupo de 10, es decir hay 10 EMS ubicados en diferentes lugares, lo que corrobora lo visto en la Figura 1.

Con el comando groupby(), se puede determinar el número de muestras totales en la base de datos para cada caso, como se indica en la siguientes secciones de código y Figuras 6 y 7.

```
sensores = contaminacion.groupby('Device Name').size()
sensores = sensores.sort_values(ascending=False)
print(sensores)
```

```
Device Name
EMS Charlestown Public School                      25073
EMS Charlestown Square North Piazza Bus Stop       24614
EMS Gari Street Charlestown                        24568
EMS Charlestown Square South Piazza                24511
EMS Five Islands Roundabout Speers Point           23628
EMS Pacific Hwy Crn Charlestown Road               19484
EMS Lake Macquarie Art Gallery Jetty               18861
EMS Pacific Hwy Crn Frederick Street Charlestown   14423
EMS Speers Point Pool                               4116
EMS Whitebridge High School                          620
dtype: int64
```

Figura 6. Número de muestras para cada uno de los dispositivos. Fuente: El Autor

En este caso se puede apreciar, el nombre del dispositivo y la cantidad de muestras que tomo cada EMS en orden descendente.

```
localizacion = contaminacion.groupby('Location').size()
localizacion = localizacion.sort_values(ascending=False)
print(localizacion)
```

```
Location
-32.965146,151.696981    25073
-32.962313,151.695294    24614
-32.967,151.68681        24568
-32.965482,151.694651    24511
-32.95804,151.61749      23628
-32.961403,151.696462    19484
-32.973012,151.617175    18861
-32.965521,151.695336    14423
-32.962639,151.619823     4116
-32.973021,151.712341      620
dtype: int64
```

Figura 7. Número de muestras acorde a cada una de las ubicaciones. Fuente: El Autor

Para este caso el resultado es igual al anterior, solo que se toma como referencia la ubicación y no el nombre del dispositivo, y con esto ya se puede conocer la ubicación de cada sensor.

2.2.2 Histogramas de los parámetros ambientales

En el caso de los parámetros de temperatura, humedad y ruido, los valores varían en un rango más amplio, por lo que aplicar el comando anterior es poco práctico, ya que no se pueden visualizar todos los datos de una manera resumida. Es por lo que se ha optado por usar histogramas para representar dichos valores y así poder visualizar la frecuencia con la se repiten los valores dentro de cada parámetro que va a ser objeto se nuestro análisis.

Se define al histograma como la representación gráfica, generalmente en formas de barras, que sirve para determinar la distribución de un conjunto de datos, con respecto a características cualitativas y cuantitativas.

En la Figura 8, tenemos el histograma de la temperatura, el cual nos indica que hay una mayor frecuencia entre los 20 y 30 °C, con mayor incidencia a un valor aproximado de 25 °C lo que se aproxima a la temperatura ambiente, sin embargo, empieza a existir mediciones de temperatura superiores a los 30 °C, lo que puede afectar directamente al parámetro de la humedad.

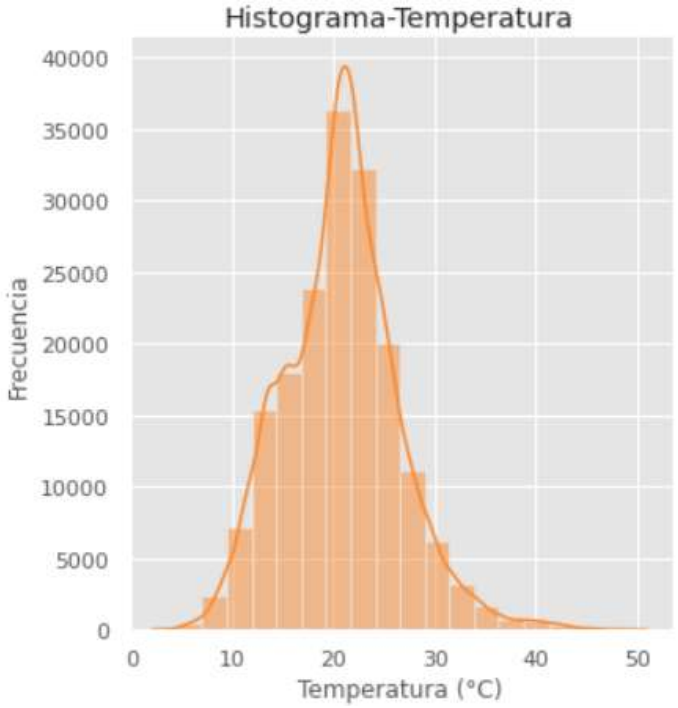

Figura 8. Histograma de la temperatura. Fuente: El Autor

En el caso del histograma de la humedad, se puede apreciar que esta aumenta de 0 a 100 %, esto puede ser un problema ya que es recomendable que la humedad no sobrepase el 70 % de HR (Humedad Relativa), puesto que el aumento de la HR crea un ambiente idóneo para la generación de microorganismos que pueden afectar la salud de una persona, este comportamiento se puede ver evidenciado en la Figura 9.

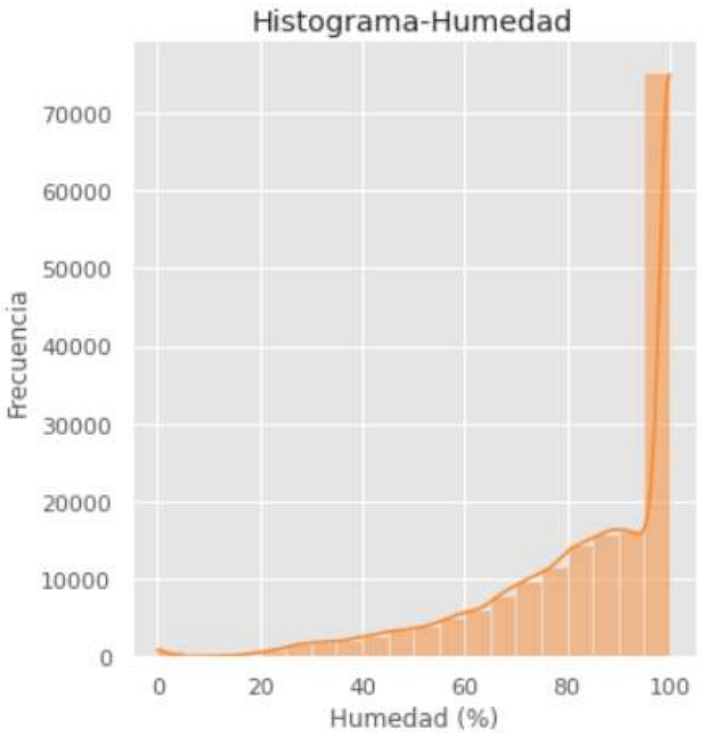

Figura 9. Histograma de la humedad. Fuente: El Autor

Para el ruido ambiental, se tiene un umbral de 83 dBA, el cual nos sirve de referencia para determinar si hay componentes de ruido peligrosas que pueden afectar a la salud humana. En la Figura 10, se puede observar el histograma para el ruido pico o instantáneo, para que este se considere peligroso debe sobrepasar el umbral y mantenerse activo por largos periodos de tiempo, y en base al gráfico se puede decir que hay cierta tendencia a tener un ruido alto, por lo que se le debe prestar especial atención a este parámetro.

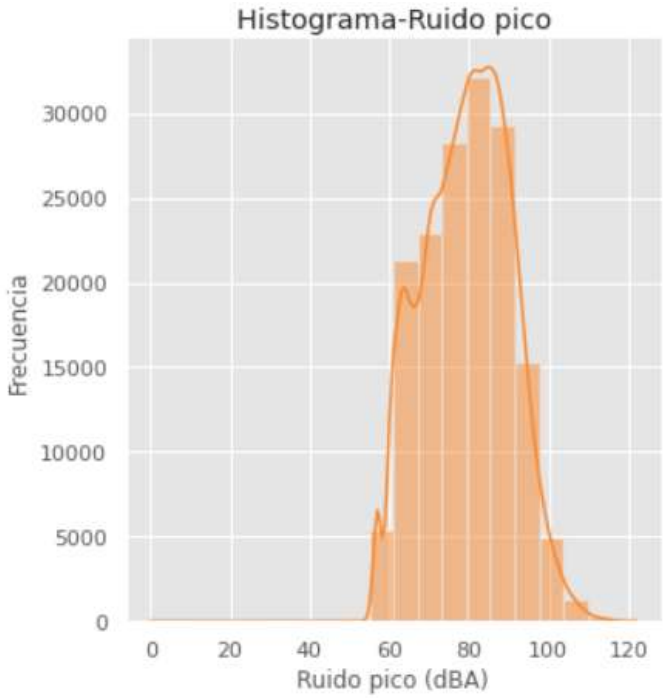

Figura 10. Histograma de ruido pico o instantáneo. Fuente: El Autor

Para el caso del ruido promedio, se toman mediciones aproximadamente durante 1 minuto mientras haya ruido, y se obtiene el promedio de estos, en base al histograma de la Figura 11, se puede apreciar que hay mayor incidencia de ruido entre los 50 y 60 dBA, que además cuenta con varios puntos mínimos, esto puede deberse a que el ruido puede cambiar abruptamente sin necesidad de ir escalando de paso a paso como la temperatura y humedad. Por otro lado, hay muy pocas mediciones sobre los 83 dBA, por lo que se podría considerar dentro de los valores recomendados.

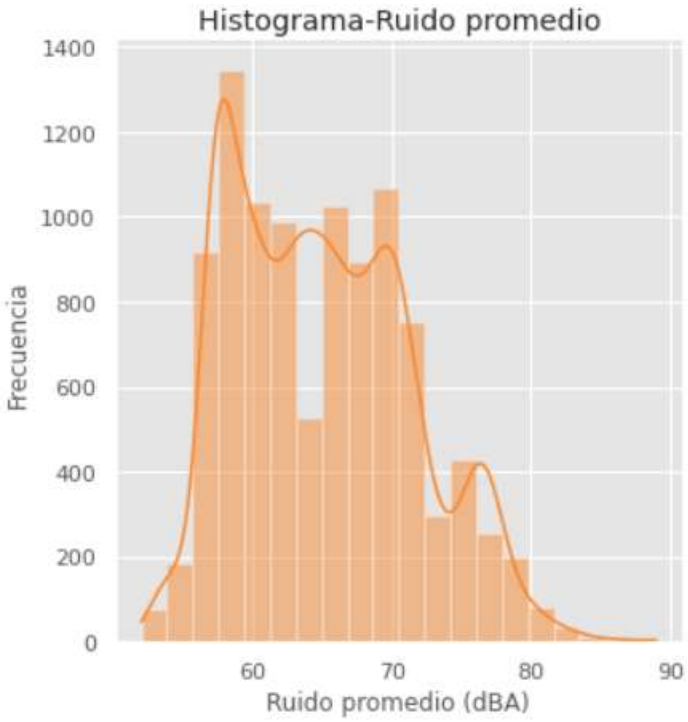

Figura 11. Histograma de ruido promedio. Fuente: El Autor

Es importante mencionar que las figuras desde la 8 hasta la 11, fueron generadas con Python en el entorno de programación Google Colab.

2.2.3 Descripción de la base de datos

La base datos que se va a usar tiene un tamaño total de 17989 de filas y 13 columnas, siendo que tres de estas columnas corresponde al tiempo, nombre del dispositivo y ubicación de este, mientras que las otras corresponden a mediciones de los parámetros físicos ya mencionados. En muchos casos los sensores están implementados de tal manera que registran datos cuando detectan una variación en el parámetro destinado a ser monitoreado, pero si los cambios son casi inexistentes el sensor para evitar consumir recursos registra valores de tipo nulo, que indica que no hubo cambios apreciables en el ambiente y se pueden registrar como valores de tipo Null o Na. Este tipo de datos puede llegar a ser un problema al momento de implementar un algoritmo de aprendizaje automático, ya que estos no están preparados para manejar dichas variables, es por lo que como medida preventiva vamos a cambiar todos los valores de tipo Null o Na a cero, con la siguiente sección de código.

```
contaminacion = contaminacion.fillna(0)
```

Como bien se mencionó la base de datos cuenta con 13 columnas, de las cuales 4 son de nuestro interés como lo son la temperatura, humedad, ruido pico y ruido promedio, ya que estos parámetros serán usados en el algoritmo K-Means para encontrar un número óptimo clústeres, es por por lo que se va a definir variables auxiliares, donde se guardara el valor de los parámetros que nos serán de utilidad como se indica en la siguiente sección de código, el primer comando elimina las columnas que se encuentran escritas dentro del paréntesis, mientras que la segunda línea de código toma solo las columnas que están escritas dentro del argumento, como se indica en la Figura 12.

```
contaminacion=contaminacion.drop(['Carbon Monoxide CO (ppm)','Ozone O
3 (ppm)','Nitrogen Dioxide NO2 (ppm)','PM10 Particulate Matter (µg/m3
)','PM2.5 Particulate Matter (µg/m3)','PM1 Particulate Matter (µg/m3)
'],axis=1)
contaminacion_variables=contaminacion[['Average Sound (dBA)','Peak So
und (dBA)','Temperature (°C)','Humidity (%)']]
```

Figura 12. Eliminación de columnas y asignación de variables auxiliares. Fuente: El Autor

Con la ayuda del comando describe(), se puede determinar si los datos han sufrido algún cambio después que se han ejecutado las secciones de código previamente descritas, tal y como se indica en la Figura 13, en este punto es importante mencionar el número de datos no vacíos es igual en todos los casos, situación que era diferente en la Figura 4, que puede deberse a que los valores de Na fueron cambiados por 0, lo que implica un cambio en los demás parámetros calculados por la función describe().

	Average Sound (dBA)	Peak Sound (dBA)	Temperature (°C)	Humidity (%)
count	179898.000000	179898.000000	179898.000000	179898.000000
mean	3.665105	71.083286	20.745944	83.553430
std	15.079453	26.476118	5.701885	19.528507
min	0.000000	0.000000	2.100000	0.000000
25%	0.000000	67.000000	17.000000	74.000000
50%	0.000000	78.000000	20.800000	90.000000
75%	0.000000	87.000000	24.100000	100.000000
max	89.000000	122.000000	51.000000	100.000000

Figura 13. Descripción de la base de datos con los cambios realizados. Fuente: El Autor

2.2.4 Normalización de los datos

Normalizar datos es un requisito que se debe cumplir para poder usar algoritmos de aprendizaje automático, ya que si las características individuales de cada columna no se parecen a los datos estándar normalmente distribuidos (como valores gaussianos con media 0 y varianza unitaria) pueden llegar a tener malos resultados [18].

En la Figura 14 se indica como se normalizan los datos.

```
scaler = StandardScaler()
scaler.fit(contaminacion_variables)
contaminacion_norm = scaler.transform(contaminacion_variables)
pd.DataFrame(contaminacion_norm).describe()
```

Figura 14. Normalización de los datos. Fuente: El Autor

1. La primera línea de código creo un objeto que tiene como propósito dar una media de cero y varianza unitaria que servirá como punto de referencia para la normalización.
2. La segunda línea código calcula la media y la desviación estándar que se usará para el escalado tomando como referencia a los datos guardados en la variable contaminacion_variables.
3. En la tercera línea de código, una vez almacenados los valores de la media y desviación estándar, procedemos a normalizar los datos de la variable contaminacion_variables para que tengan una media de cero y varianza unitaria.
4. En la cuarta línea de código, se comprueba que los datos han sido normalizados para lo cual se va a usar la función describe(), sin embargo, debido a que el tipo de variable cambio, procedemos a realizar un cast en la variable contaminacion_norm para que vuelva a ser de tipo DataFrame, esto se lo puede apreciar en la Figura 15.

	0	1	2	3
count	1.798980e+05	1.798980e+05	1.798980e+05	1.798980e+05
mean	-5.055613e-17	1.415572e-16	5.561175e-16	-3.033368e-16
std	1.000003e+00	1.000003e+00	1.000003e+00	1.000003e+00
min	-2.430536e-01	-2.684815e+00	-3.270146e+00	-4.278548e+00
25%	-2.430536e-01	-1.542257e-01	-6.569677e-01	-4.892057e-01
50%	-2.430536e-01	2.612442e-01	9.480433e-03	3.301117e-01
75%	-2.430536e-01	6.011742e-01	5.882380e-01	8.421850e-01
max	5.659034e+00	1.923124e+00	5.305989e+00	8.421850e-01

Figura 15. Descripción de los datos normalizados. Fuente: El Autor

2.3 Implementación del algoritmo de aprendizaje automático

Existen distintos tipos de algoritmos de aprendizaje automático con diversos propósitos, la elección de este se hará en base a la aplicación que se le quiera dar, para el presente TIC se busca reducir la dimensión de datos y agruparlos en distintas categorías, para lo cual se ha optado por usar algoritmos de aprendizaje no supervisado.

Los algoritmos de aprendizaje no supervisado nos permiten agrupar datos, sin etiquetar, basados en un criterio. Uno de los algoritmos más usados es el K-Means, el cual va a ser implementado en el presente estudio, que es del tipo particional tomando un número K de clústeres o grupos predeterminados.

Para comenzar con el desarrollo del algoritmo K-Means, primero se debe realizar un diagrama de flujo para representar la secuencia de actividades del programa, como se indica en la Figura 16.

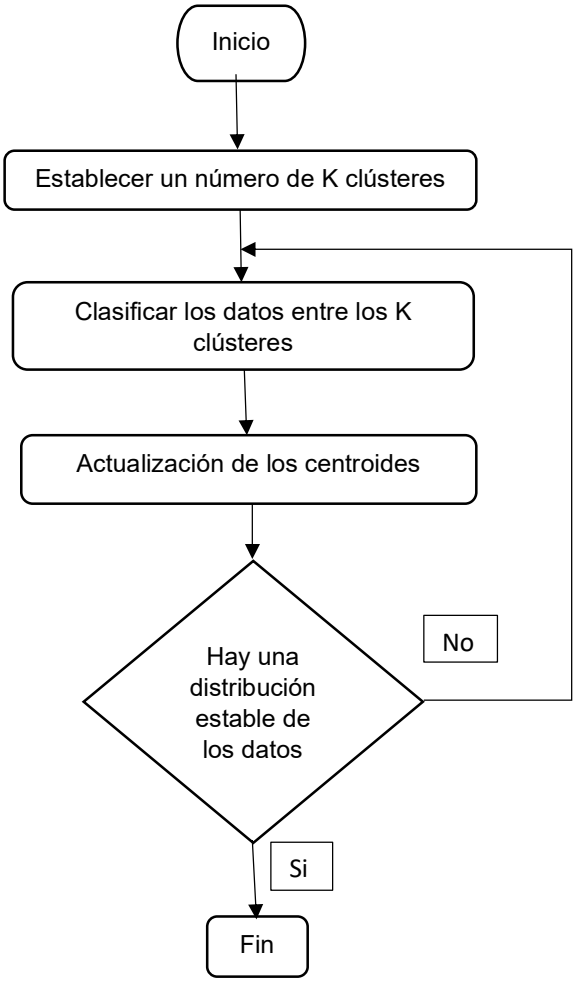

Figura 16. Diagrama de flujo del algoritmo K-Means. Fuente: El Autor

El proceso descrito en el diagrama de flujo de la Figura 16, se encuentra implementado en la función conocida en Python como KMeans(), por lo que en este capítulo nos basaremos en la descripción de otros aspectos como: elegir del valor de K clústeres óptimos, ejecutar del algoritmo K-Means con el valor de K óptimo y graficar los clústeres para comprobar que la distancia media intra-clusters es la adecuada.

2.3.1 Elección del valor de K óptimo

Como primer punto se debe asignar un valor K de clústeres, los cuales pueden estar en un rango a partir del 2, esto ya que se busca generar particiones que cuenten con un centroide y una clasificación de datos estables, donde dicha estabilidad será corroborada con la ayuda del Análisis de Componentes Principales (PCA, por sus siglas en inglés).

Para determinar el valor de K clústeres óptimos, nos vamos a basar en el método del codo (elbow´s method), que usa la distancia media entre un objeto y su centroide, tomando como

referencia las distancias intra-cluster, es decir que mientras más grande es el número de clústeres K, la varianza intra-cluster tiende a ser menor. Si la distancia intra-cluster es menor es mejor, puesto que los clústeres son compactos. Por tanto, en esencia el método de codo busca un valor de K, donde no se mejore en gran medida la distancia media intra-intra-cluster.

Con la ayuda del lenguaje de programación Python se puede ejecutar la tarea descrita en el párrafo anterior, como se indica en la Figura 17.

```
valores_k1 = list(range(1,15))
lista_distorsion1 = []

for i in range(len(valores_k1)):
  kmeans1 = KMeans(n_clusters=valores_k1[i])
  kmeans1.fit(contaminacion_norm)
  lista_distorsion1.append(sum(np.min(cdist(contaminacion_norm, kmean
s1.cluster_centers_,'euclidean'), axis=1)) / contaminacion_norm.shape
[0])
```

Figura 17. Elección del clúster óptimo. Fuente: El Autor

En primer lugar, se crea un vector con N valores para los K clústeres, y también se crea una lista para guardar valores de distorsión, después con la ayuda de un bucle for ejecutamos el algoritmo K-Means para cada valor de K, el uso de este algoritmo es muy sencillo ya que únicamente se debe asignar un valor de K a la vez, en la función KMeans(n_clsuters=K), a continuación la función fit() nos permite determinar los centroides para los datos ambientales que previamente fueron limpiados y normalizados, todo esto para un valor de K, y finalmente con la ayuda de la distancia euclidiana podemos determinar la distorsión para cada caso, los resultados de la ejecución del código mencionado pueden ser representados mediante un gráfico como se indica en la Figura 18.

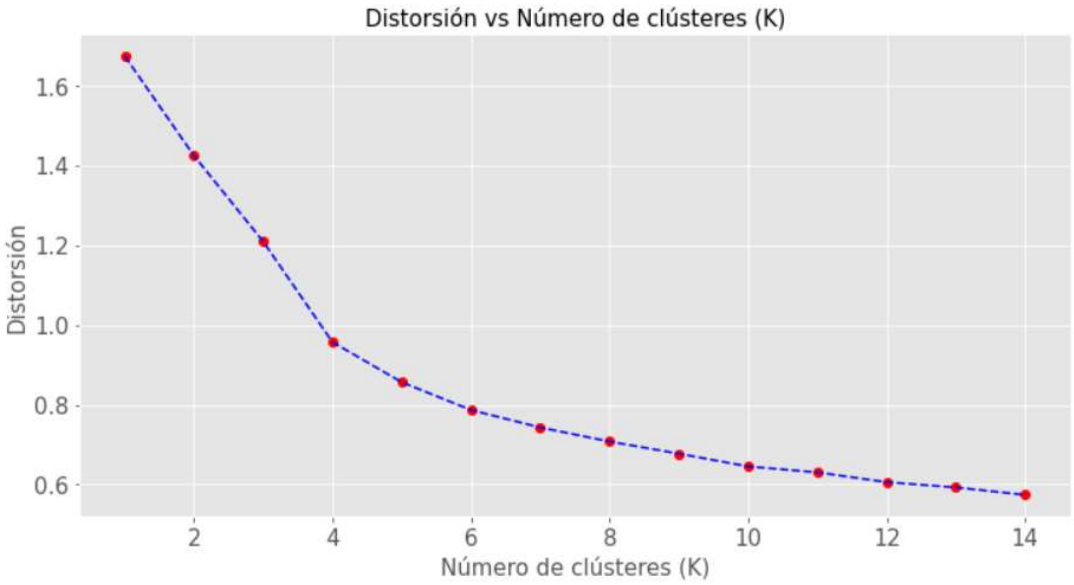

Figura 18. Distorsión vs Número de clústeres (K). Fuente: El Autor

En base al gráfico presentado en la Figura 18, usando el criterio del método del codo se determina que el valor de K clústeres óptimo es 4, esto ya que cuando el valor de K llega a ser 4 presenta una distorsión aproximada de 0.95, y partir de este punto para valores de K superiores a 4 la distorsión no cambia en gran medida.

2.3.2 Ejecución del algoritmo K-Means con el valor de K óptimo

Una vez escogido el valor de K óptimo, procedemos a ejecutar el programa, pero únicamente con el valor de K=4, para esto se reutiliza el código presentado en el caso anterior, como se indica en la Figura 19.

```
nc=4
clustering = KMeans(n_clusters=nc)
clustering.fit(contaminacion_norm)
```

Figura 19. Ejecución del algoritmo K-Means con el valor de K óptimo. Fuente: El Autor

La forma en la que se agrupan los datos es mediante la asignación de una etiqueta que ha sido determinada por el número de clústeres K, que en este caso va de desde un valor de 0 hasta 3, es decir a cada fila le corresponde un etiqueta que representa un grupo, esta información se almacena en la variable clustering.labels_ que puede ser integrada a la base de datos con el comando que se presenta a continuación.

```
contaminacion['KMeans_Clusters']=clustering.labels_
```

Con la ayuda del comando head(), nos podemos asegurar de que las etiquetas se encuentren disponibles en la base de datos, como se presenta en la Figura 20.

Time	Device Name	Temperature (°C)	Humidity (%)	Average Sound (dBA)	Peak Sound (dBA)	Location	KMeans_Clusters
2020-07-31T23:55:19+10:00	EMS Five Islands Roundabout Speers Point	6.5	100.0	56.0	84.0	-32.95804,151.61749	2
2020-07-31T23:53:20+10:00	EMS Charlestown Public School	11.2	100.0	58.0	73.0	-32.965146,151.696981	2
2020-07-31T23:51:49+10:00	EMS Charlestown Square North Piazza Bus Stop	11.3	100.0	61.0	71.0	-32.962313,151.695294	2
2020-07-31T23:48:38+10:00	EMS Charlestown Square South Piazza	10.4	100.0	60.0	70.0	-32.965482,151.694651	2
2020-07-31T23:46:14+10:00	EMS Gari Street Charlestown	9.1	100.0	57.0	67.0	-32.967,151.68681	2

Figura 20. Base de datos con la etiqueta obtenida a partir del algoritmo K-Means. Fuente: El Autor

Una vez que las etiquetas se han agregado a la base de datos, de puede determinar cuantas muestras hay para cada categoría, esto con la ayuda del comando groupby(), como se indica en la Figura 21, en base a esto se puede deducir que las etiquetas que predominan los datos es la 0, después la 1, la 2 y la 3 en orden descendente, para poder determinar alguna tendencia debemos representar estos resultados con herramientas como PCA.

```
KMeans_Clusters
0    106642
1     44153
2     18970
3     10133
```

Figura 21. Número de muestras por etiqueta. Fuente: El Autor

2.3.3 Gráfico de los clústeres

Para representar el conjunto de datos de una manera adecuada vamos a usar el Análisis de Componentes Principales (PCA, por sus siglas en inglés). PCA es conocido como un método matemático usado en el aprendizaje no supervisado, que consiste en extraer información útil en base a las variables originales, esto nos ayuda a simplificar el análisis de datos representándolos con 2 o 3 dimensiones, conocidos como componentes 1, 2 y 3. Dicho proceso esta indicado en la Figura 22.

```
pca=PCA(n_components=2)
pca_contaminacion=pca.fit_transform(contaminacion_norm)
pca_contaminacion_df = pd.DataFrame(data=pca_contaminacion, columns =
['Componente_1', 'Componente_2'])
pca_nombres_contaminacion= pd.concat([pca_contaminacion_df, contamina
cion[['KMeans_Clusters']]], axis=1)
pca_nombres_contaminacion
```

Figura 22. Análisis de Componentes Principales a los resultados obtenidos. Fuente: El Autor.

Para este caso vamos a usar 2 componentes, esto lo realizamos en la primera línea de código con la ayuda de la función PCA(n_components=2).

En la segunda línea de código, tenemos las función fit_transform(), que nos ayuda a ajustar el modelo con los valores normalizados de la variable contaminacion_norm y aplicamos una reducción de dimensionalidad con respecto a la misma variable.

Para la tercera línea de código guardamos las variables de las componentes 1 y 2, obtenidas anteriormente y realizamos un cast para guardar dichas variables como DataFrame.

En la cuarta línea de código, le agregamos el valor de las etiquetas generas a partir de los clústeres, y con la ayuda comando head(), podemos visualizar el resultado de ejecutar esta sección de código, como se indica en la Figura 23.

	Componente_1	Componente_2	KMeans_Clusters
0	-3.716704	-0.237151	2
1	-3.222627	-0.140503	2
2	-3.298926	-0.099101	2
3	-3.374790	-0.007536	2
4	-3.438258	0.188596	2

Figura 23. PCA con componentes 1 y 2. Fuente: El Autor

Una vez que se ha obtenido las PCA con 2 componentes, las vamos a representar gráficamente, como se indica en la Figura 24. En base a este gráfico se puede decir que los centroides están bien definidos ya que datos pertenecientes a un mismo grupo se mantienen agrupadas en un mismo sector, además que se tiene claras lineas de decisión

por lo que se puede decir que el número de clústeres K y la distancia media intra-clusters es estable y adecuada.

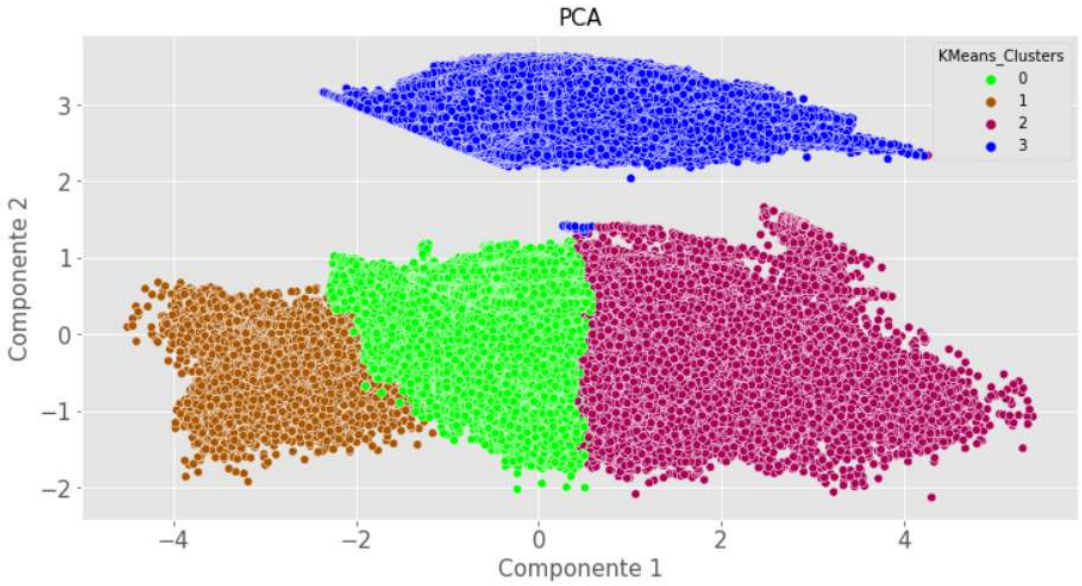

Figura 24. PCA representados gráficamente. Fuente: El Autor

2.4 Análisis de resultados

En la Figura 25, se puede apreciar que hay 4 grupos que tienen centroides bien definidos, cada categoría corresponde a un clúster que va entre los valores de 0, 1, 2 y 3, etiquetas que fueron obtenidas a partir de la ejecución del algoritmo K-Means con K=4, dicha agrupación se realizó con base al criterio de la distancia media intra-clusters, en donde los parámetros involucrados fueron la temperatura, humedad, ruido pico y ruido promedio.

Una vez que se ha etiquetado todo el conjunto de datos se asigna un significado a cada clúster, y esto se logra relacionando los parámetros utilizados como: temperatura, humedad y ruido, con su respectiva etiqueta como se indica en la Figura 24.

Para visualizar la tendencia que tiene cada clúster, se realiza un análisis basado en las recomendaciones o normas internacionales a las que se rigen los parámetros de temperatura, humedad y ruido.

Para el clúster 0

Como se indica en la Figura 25, se tiene un valor de ruido pico que varia entre 50 y 120 dBA, variación que es considerada peligrosa tomando como referencia las recomendaciones dadas por la OMS, donde se indica que valores iguales o superiores 110 dBA pueden causar discapacidad auditiva, ver Tabla 1.

Si bien el ruido pico es un parámetro que debe ser considerado de riesgo, es poco común que tales niveles de ruido se presenten con mucha frecuencia, sin embargo, acorde a las recomendaciones dadas por la OMS una exposición prolongada de tiempo a ciertos niveles de ruido, que son menores al caso anterior, pueden afectar de igual manera pero con la desventaja de que una persona no lo percibiría como una amenaza, es así que el ruido promedio durante un periodo de tiempo afecta la salud de una persona considerando niveles de ruido no tan altos que apenas superan el umbral establecido por la OMS, pero que se repiten con mucha frecuencia en intervalos de tiempo como: una exposición a un nivel de ruido promedio de 85 dBA por el periodo de 1 hora, puede causar discapacidad auditiva, y de la misma forma para valores de ruido más bajos, ver Tabla 1.

Con respecto a la temperatura, se puede decir que no hay variación que pueda ser considerada peligrosa ya que apenas supera la temperatura ambiente, por otro lado, la concentración de humedad HR, es alta, bordea el 100 %, lo que debe ser un parámetro para tener muy en encuenta.

Para el clúster 1

Como se puede apreciar en la Figura 25, se le debe prestar mucha atención al ruido ambiental, ya que se tiene un ruido pico que supera los 85 dBA, además que el ruido promedio varia entre los 50 y 85 dBA, lo que puede ser una señal de que, si hay largos intervalos de tiempo con un ruido alto, se puede producir efectos adversos en la salud.

Para el parámetro de la temperatura, tenemos cambios que van desde los 5 ºC hasta los 23ºC, estos valores no son considerados peligrosos. En el caso de la humedad esta varia desde el 40 % hasta el 100% de HR, siendo una tendencia repetitiva y que se debe prestar especial atención para tomar medidas preventivas.

Para el clúster 2

Como se indica en la Figura 25, el ruido pico y ruido promedio, mantienen una tendencia similar al primer caso.

Además, en la temperatura ya existe un cambio que se debe considerar como alerta, pues las temperaturas llegan casi a los 55 ºC, puesto que es la máxima temperatura que puede soportar una persona, por otro lado, la humedad cambia desde los valores más bajos hasta los más altos, esto como consecuencia de las altas temperatura.

Para el clúster 3

Como se indica en la Figura 25, el ruido ambiental no presenta peligro alguno, puesto que el ruido alto es casi inexistente, tanto para el ruido instantáneo como para el ruido promedio.

La temperatura y humedad se mantiene una tendencia similar a los casos anteriores, sin embargo, es importante notar que la humedad tiene una misma tendencia, es decir llega hasta concentración de 100 % de HR, esto puede atribuirse a su ubicación geográfica y a los cambios de temperatura.

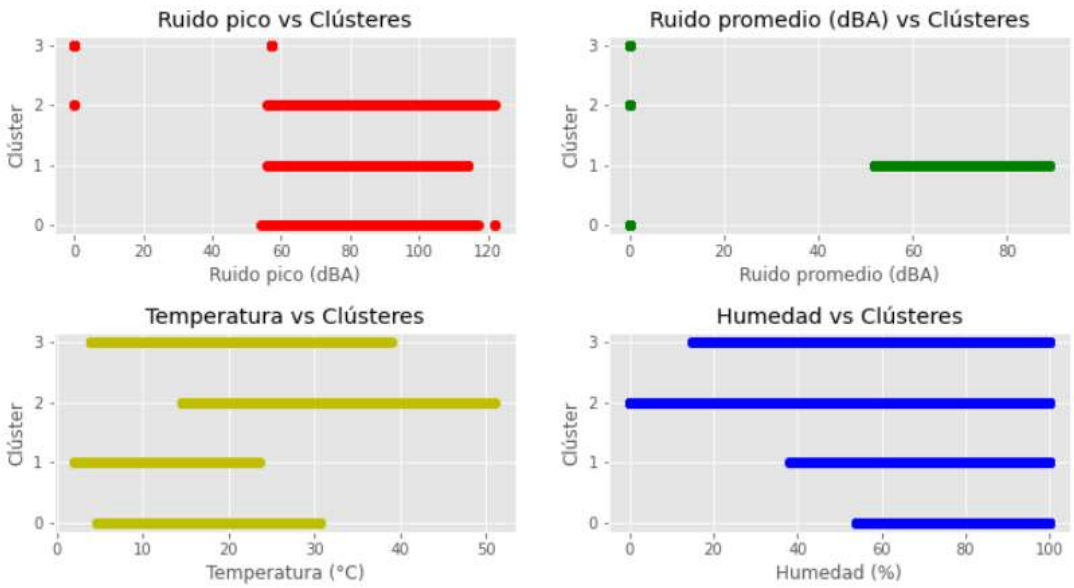

Figura 25. Parámetros ambientales vs clústeres. Fuente: El Autor

En la Tabla 2, tenemos un resumen del análisis realizado para cada uno de los clústeres.

Tabla 2. Resumen del análisis de los clústeres

Clúster	Temperatura	Humedad	Ruido
0	No hay cambios peligrosos	Concentración alta, tener cuidado en lugares cerrados y con poca ventilación	Niveles altos, que no son constantes, pero pueden causar discapacidad auditiva
1	No hay cambios peligrosos	Concentración alta, tener cuidado en lugares cerrados y con poca ventilación	Niveles altos, que se repiten con frecuencia y puede causar daños irreversibles al oído humano.
2	Cambios peligrosos llegando al umbral	Concentración alta, tener cuidado en lugares cerrados y	Niveles altos, que no son constantes, pero pueden causar

	de tolerancia humano 55ºC	con poca ventilación	discapacidad auditiva
3	No hay cambios peligrosos	Concentración alta, tener cuidado en lugares cerrados y con poca ventilación	Niveles seguros

Nota: Ver Anexo I

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

Uno de los objetivos del TIC, correspondiente al componente 2, es implementar un modelo de aprendizaje automático, que nos permita reducir la dimensión de los parámetros ambientales de temperatura, humedad y ruido, para representarlos y clasificarlos con un nuevo valor, que en este caso se conoce como clúster. Esto se ha logrado con éxito usando el algoritmo de aprendizaje no supervisado conocido como K-Means.

Con la ayuda del algoritmo K-Means, se ha agrupado en conjuntos de datos con base al criterio de la distancia intra-clústeres, por lo que la etiqueta resultante de este proceso guarda un significado, como se indica en la Tabla 2, que nos puede ayudar a interpretar los parámetros ambientales de temperatura, humedad y ruido de una manera más sencilla.

Para la presentación de los resultados se va a tomar las mediciones realizadas durante un día y un mes cualesquiera, basándonos enteramente en los clústeres y en el valor de ruido ambiental.

Monitoreo para el día 2020-01-04

La base de datos usada para el presente estudio tiene gran cantidad de información, por lo que se ha estratificado en secciones para su análisis, con este criterio se ha tomado los valores de temperatura, humedad y ruido, registrados durante el día 2020-01-04, esto permitió presentar información útil y fácil de entender, puesto que fue uno de los días donde se registran niveles de ruido altos en la mayoría de EMS.

Como primer punto del análisis vamos a presentar alertas cuando se exceda un valor de ruido ambiental, tomando como referencia el valor más alto registrado durante ese día, dicha información al provenir de un EMS nos ofrece datos adicionales como: la hora de la medición, la temperatura, la humedad, nombre de la estación de monitoreo que registro la alerta, su ubicación y el clúster al que pertenece, todo esto se puede apreciar en la Figura 26.

```
-------------------------Alerta-------------------------------
Time                                              06:03:19
Device Name            EMS Lake Macquarie Art Gallery Jetty
Temperature (°C)                                      39.1
Humidity (%)                                          48.0
Average Sound (dBA)                                    0.0
Peak Sound (dBA)                                     114.0
Location                                -32.973012,151.617175
KMeans_Clusters                                          2
Name: 117833, dtype: object
--------------------------------------------------------------------
-------------------------Alerta-------------------------------
Time                                              01:20:15
Device Name            EMS Lake Macquarie Art Gallery Jetty
Temperature (°C)                                      39.4
Humidity (%)                                          45.0
Average Sound (dBA)                                    0.0
Peak Sound (dBA)                                     114.0
Location                                -32.973012,151.617175
KMeans_Clusters                                          2
Name: 117960, dtype: object
--------------------------------------------------------------------
```

Figura 26. Alertas generadas el 2020-01-04. Fuente: El Autor

Una vez conocida la ubicación del EMS donde se registro la alerta, se procede a representarlo con la ayuda de los mapas interactivos de Python, Figura 27, con esto se puede deducir que las alertas se produjeron en zonas poco pobladas, por lo que el riesgo para la población en general no tiene un nivel de peligro tan alto.

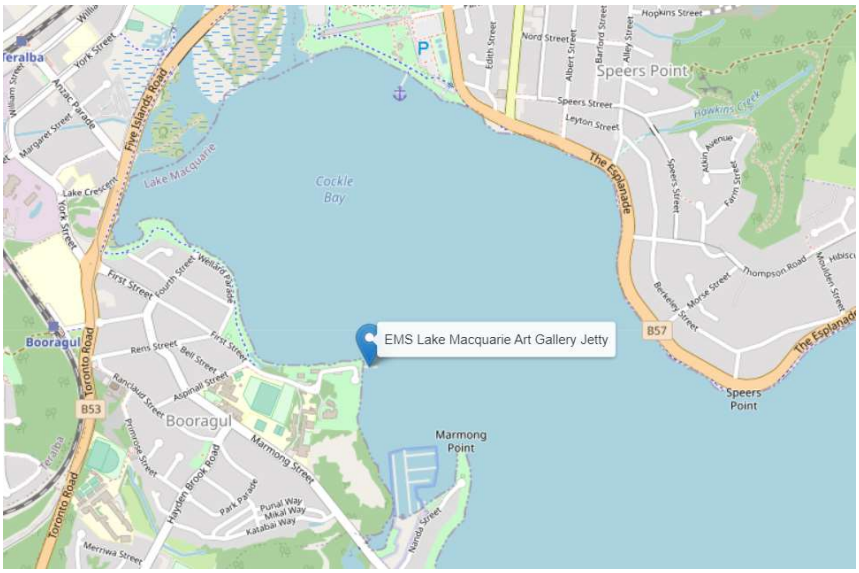

Figura 27. Ubicación del EMS donde se produjo la alerta el día 2020-01-04. Fuente: El Autor

La red de EMS está ubicada geográficamente en lugares cercanos, por lo que debería registrar un número de mediciones similares, esto se puede corroborar con en la Figura 28,

aquí se puede notar que todas las estaciones de monitoreo ambiental registran alrededor de 90 mediciones el día 2020-01-04.

Muestras tomadas en el día por cada EMS

Device Name	
EMS Charlestown Public School	95
EMS Charlestown Square North Piazza Bus Stop	96
EMS Charlestown Square South Piazza	96
EMS Five Islands Roundabout Speers Point	92
EMS Gari Street Charlestown	93
EMS Lake Macquarie Art Gallery Jetty	92
EMS Pacific Hwy Crn Charlestown Road	87

Figura 28. Número de mediciones registrada por el EMS durante el día 2020-01-04. Fuente: El Autor

Uno de los propósitos del presente estudio, es clasificar los datos con ayuda de los clústeres generados por el algoritmo K-Means, en la Figura 29, se puede ver que clústeres tienen mayor incidencia al terminar el día, en este caso el clúster 2 tiene un mayor número de mediciones por lo que en este día se debe tener especial cuidado con los cambios de temperatura y además presenta incidencia de ruido altamente peligroso.

Etiquetadas registradas en el día

KMeans_Clusters	
0	108
2	543

Figura 29. Incidencia de los clústeres durante el día 2020-01-04. Fuente: El Autor

Para saber con más detalle, la distribución de la temperatura, humedad y ruido, a largo del día, se ha generado un gráfico basado en el clúster y la hora de medición de los 3 parámetros ya mencionados, Figura 30, aquí se puede apreciar que la tendencia vista en la sección anterior se mantiene y se produce en horas de la mañana y tarde, con mayor incidencia en la mañana.

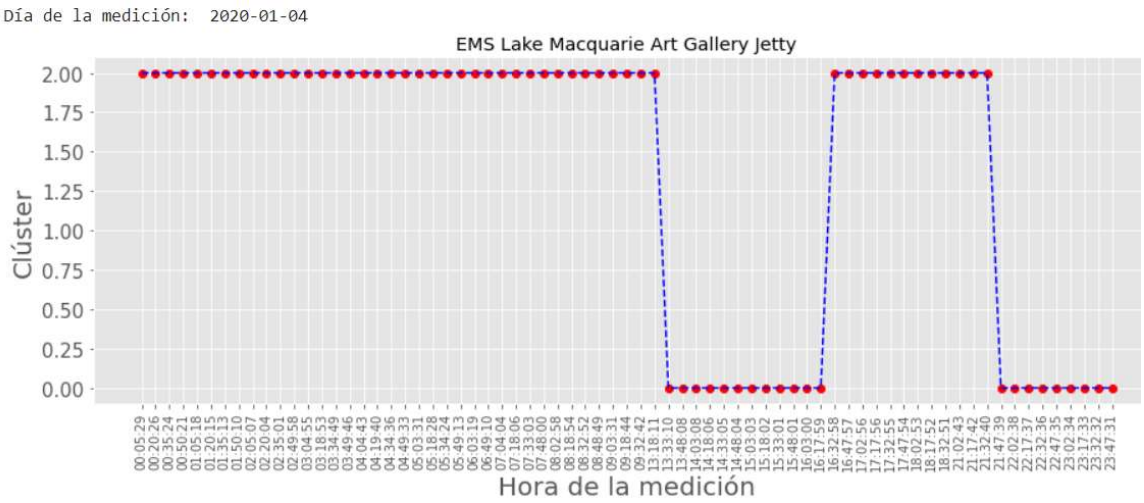

Figura 30. Clústeres vs Hora de medición en el día 2020-01-04. Fuente: El Autor

Monitoreo para el mes de mayo del 2020

Para comprobar la eficacia del análisis, se optado por realizar un proceso similar al anterior, pero para un periodo de tiempo de un mes; para generar una alerta se ha usado el mismo criterio que en el caso anterior, por tanto, se indica el valor de ruido más alto y se describen sus demás parámetros, como se indica en la Figura 31, en este caso el valor de ruido más alto registrado es 114 dBA, valor que puede causar discapacidad auditiva, perteneciente al clúster 2.

```
---------------------------------Alerta-------------------------------------
Time                                                  2020-05-11T23:44:44
Device Name           EMS Pacific Hwy Crn Frederick Street Charlestown
Temperature (°C)                                                     20.4
Humidity (%)                                                         62.0
Average Sound (dBA)                                                   0.0
Peak Sound (dBA)                                                    114.0
Location                                          -32.965521,151.695336
KMeans_Clusters                                                         2
Name: 36596, dtype: object
----------------------------------------------------------------------------
```

Figura 31. Alerta generada para el mes de mayo del 2020. Fuente: El Autor

La ubicación del EMS donde se registro la alerta se presenta en la Figura 32, esta zona es un lugar concurrido, por lo que se debe prestar principal atención en tomar medidas preventivas y de mitigación con respecto al ruido ambiental.

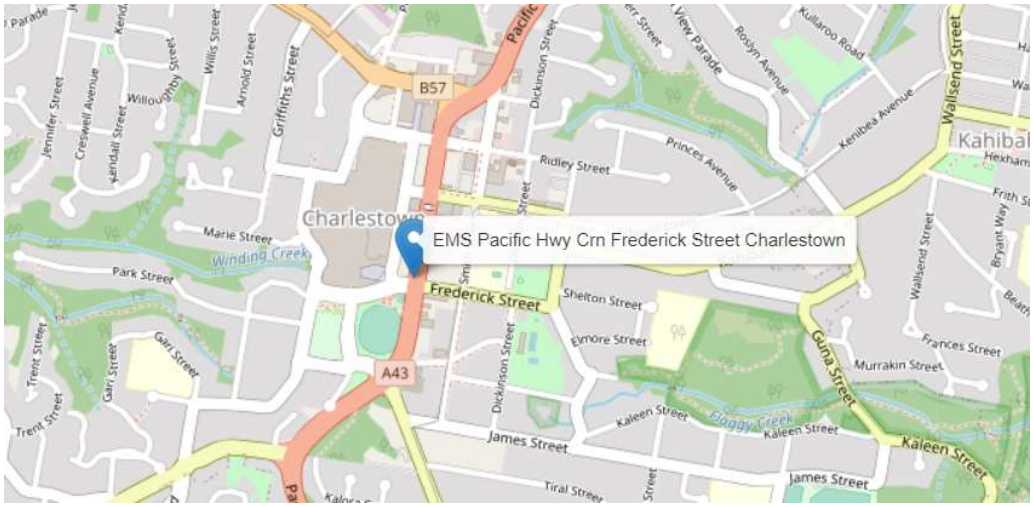

Figura 32. Ubicación del EMS donde se produjo la alerta. Fuente: El Autor

El número de mediciones tomadas durante el mes mantiene una tendencia similar al caso anterior, excepto por el EMS Lake Macquarie Art Gallery Jetty, que presente una cantidad reducida de mediciones, esto puede deberse a que se encuentra en un lugar no muy transitado y con poca población, como se indica en la Figura 33.

Muestras tomadas en el mes por cada EMS

Device Name	
EMS Charlestown Public School	2470
EMS Charlestown Square North Piazza Bus Stop	2399
EMS Charlestown Square South Piazza	2050
EMS Five Islands Roundabout Speers Point	2466
EMS Gari Street Charlestown	2395
EMS Lake Macquarie Art Gallery Jetty	163
EMS Pacific Hwy Crn Charlestown Road	1860
EMS Pacific Hwy Crn Frederick Street Charlestown	1914

Figura 33. Número de mediciones registrada por EMS en el mes de mayo del 2020. Fuente: El Autor

La tendencia de los clústeres en este caso es del grupo 0, lo que indica que durante este mes el ruido no es factor peligroso que se deba considerar para tomar medidas preventivas y correctivas, por otro lado, el grupo 2 indica niveles de ruido alto muy poco frecuentes, pero la incidencia de estos es muy poca. En consecuencia, se puede decir que el mes de mayo del 2020 no presenta niveles peligrosos de ruido y temperatura, es importante mencionar que los niveles de humedad son usualmente altos, pero como ya menciono

anteriormente puede deberse a la ubicación geográfica de las estaciones de monitoreo ambiental, como se indica en la Figura 34.

```
Etiquetadas registradas en el mes 2020-05
KMeans_Clusters
0    14325
2     1392
```

Figura 34. Incidencia de los clústeres en el mes de mayo del 2020. Fuente: El Autor

Representar registros de mediciones durante un mes es poco práctico, es por lo que se ha tomado como referencia el umbral de 100 dBA, para generar la gráfica de la Figura 29, se toma referencia el EMS mencionado en la Figura 29, puesto que fue en esa locación fue donde se registró la alerta, además se puede apreciar una mayor incidencia del clúster 0, y hay pocas muestras del componente 2, lo que corrobora lo ya mencionado para la Figura 35, es decir niveles de ruido y temperatura poco peligrosos, salvo ciertos días donde se registran niveles altos de ruido.

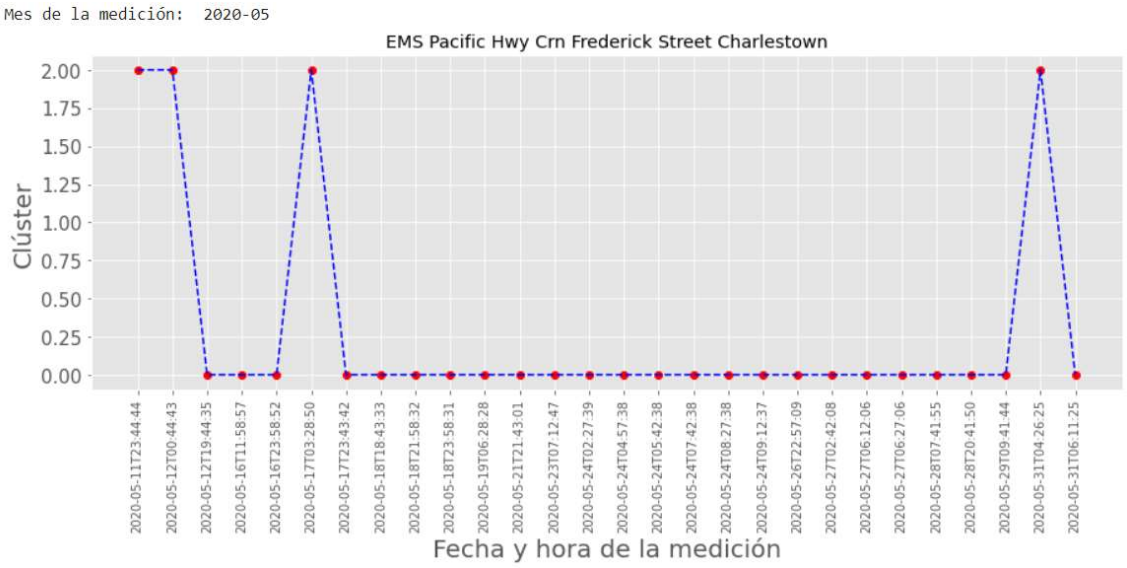

Figura 35. Clústeres vs Fecha y hora de medición para el mes de mayo del 2020. Fuente: El Autor

Nota: Ver Anexo I

3.2 Conclusiones

Las bases de datos actualmente son un recurso muy valioso en todas las áreas de trabajo, puesto que pueden albergar cualquier tipo de información de una forma estructurada, y se pueden manejar fácilmente con las distintas herramientas que existen hoy en día, además, las bases de datos pueden ser compartidas con el resto del mundo mediante el Internet, usando repositorios digitales gratuitos, que pueden ser usadas con propósitos investigación y educación, como en el presente TIC con la base de datos del EMS de la UTS, se ha generado información útil mediante modelos de aprendizaje automático.

Los modelos de aprendizaje automático han renovado la ciencia computacional, ya que se pueden aplicar a distintas áreas como economía, ambiente, tendencias de mercado, etc. Y tener múltiples aplicaciones, como la reducción de dimensiones el cual fue el objeto central del presente estudio, es decir mediante algoritmos de aprendizaje no supervisado, con el algoritmo K-Means, se pudo reducir los 4 parámetros ambientales de temperatura, humedad, ruido promedio y ruido pico, en 1 solo como lo es la etiqueta de KMeans_Clusters, valor que nos puede ayudar a clasificar los datos y reconocer la tendencia de los parámetros de ruido ambiental, temperatura y humedad, para intervalos de tiempo que van desde 1 día hasta 1 mes, basándonos en recomendaciones de la OMS.

El número de clústeres K=4 obtenidos a partir de la implementación del algoritmo K-Means, es el idóneo si nos basamos en el criterio del Método del Codo y el PCA, gracias a esto se puede determinar 4 grupos o categorías en los cuales el ruido es el parámetro que más debe ser considerado de riesgo, puesto que llega a registrar valores que pueden causar discapacidad auditiva, ver Tabla 1, y con la ayuda de estas etiquetas se puede decir al final del día que parámetro ambiental, fue el que más cambios tuvo y si fueron peligrosos para la salud humana, ver Tabla 2. es óptimo

En conclusión, la implementación de un modelo de aprendizaje automático es un proceso que se define con el concepto de ensayo y error, es decir, mediante distintas pruebas con distintos algoritmos se puede determinar cual es el modelo que nos ofrece los mejores resultados, además de adaptarse a la aplicación que se le quiere dar, en este caso se quiere agrupar datos por lo que se usa el algoritmo K-Means, mediante el uso de este modelo se pudo determinar un numero de clústeres K=4, que nos ayudan a clasificar los datos de temperatura, humedad y ruido, asignándoles una etiqueta que indica diferentes tendencias, lo que nos ayuda a generar información útil para tomar medidas preventivas y correctivas en lugares con niveles de temperatura, humedad y ruido, peligrosos, haciendo especial énfasis en los niveles de ruido de un ambiente en específico.

3.3 Recomendaciones

Identificar bien el alcance del proyecto, el tipo de datos que se va a usar, la información resultante que se quiere obtener, para así implementar en modelo de aprendizaje automático adecuado, y reducir los tiempos de desarrollo del TIC.

Usar un entorno de programación en línea que tenga recursos propios como Google Colab, puesto que en muchos casos los modelos de aprendizaje automático se basan en operaciones matemáticas repetitivas que pueden llegar a tardar varios minutos, además que Google Colab, cuenta con librerías preinstaladas para la implementación de distintos modelos de aprendizaje automático.

Una vez que se ha implementado un modelo de aprendizaje automático, es recomendable validar sus resultados usando herramientas matemáticas como el método del codo y PCA, o usando métricas propias del modelo.

En muchos casos las bases de datos tienden a registrar valores nulos en una celda, por lo que un tratamiento y limpieza de datos es indispensable antes de usar una base de datos en un algoritmo de aprendizaje automático.

Muchas de las funciones para la implementación de modelos de aprendizaje automático, vienen con parámetros activados o desactivados por defecto, por lo que es recomendable buscar en Internet la forma de uso de la función.

4 REFERENCIAS BIBLIOGRÁFICAS

[1] K. C. Rahman, “A survey on sensor network”, Journal of Computer and Information Technology, vol. 1, núm. 1, pp. 76–87, 2010.

[2] J. Yick, B. Mukherjee, y D. Ghosal, “Wireless sensor network survey”, Computer networks, vol. 52, núm. 12, pp. 2292–2330, 2008.

[3] M. A. B. y. Y. A. Çengel, Termodinámica, México: McGraw-Hill, 2012.

[4] “What is humidity? Why measure & what your levels mean | Airthings”. https://www.airthings.com/what-is-humidity (consultado el 18 de julio de 2022).

[5] S. Medrano, “Medición de humedad relativa con psicrómetro”, Boletín periódico del laboratorio de metrología. México, 2003.

[6] B. Berglund, T. Lindvall, D. H. Schwela, y World Health Organization, “Guidelines for community noise”, 1999.

[7] R. Camps Paré, L. A. Casillas Santillán, D. Costal Costa, M. Gibert Ginestà, C. Martín Escofet, y O. Pérez Mora, “Bases de datos: Software libre”, 2007.

[8] C. M. Ricardo, Bases de datos. McGraw Hill Educación, 2009.

[9] “Environmental Monitoring System (Historical to 31 July 2020)”. https://data.opendatasoft.com/explore/dataset/environmental-monitoring-system-historical-to-31-july-2020@lakemac-newcastlenswiar/table/?disjunctive.device_name&sort=time (consultado el 31 de julio de 2022).

[10] G. Box, “All models are wrong, but some are useful”, Robustness in Statistics, vol. 202, núm. 1979, p. 549, 1979.

[11] L. Quituisaca-Samaniego, «Aprendizaje no supervisado: agrupamiento o clustering,» de Numérica Resumiendo, Quito, 2017.

[12] J. H. Cáceres, “Clustering basado en el algoritmo K-means para la identificación de grupos de pacientes quirúrgicos”, Trabajo de investigación, Universidad Santo Tomás, seccional Bucaramanga. Bucaramanga, Colombia, 2015.

[13] F. Liu y Y. Deng, “Determine the number of unknown targets in Open World based on Elbow method”, IEEE Transactions on Fuzzy Systems, vol. 29, núm. 5, pp. 986–995, 2020.

[14] A. Maćkiewicz y W. Ratajczak, “Principal components analysis (PCA)”, Computers & Geosciences, vol. 19, núm. 3, pp. 303–342, 1993.

[15] “1.4. What Is Programming? — Problem Solving with Algorithms and Data Structures”. https://runestone.academy/ns/books/published/pythonds/Introduction/WhatIsProgramming.html (consultado el 31 de julio de 2022).

[16] J. Elkner, A. B. Downey, y C. Meyers, “How to Think Like a Computer Scientist: Learning with Python Documentation”, Release, 2010.

[17] “What is Python? Executive Summary”, *Python.org*. https://www.python.org/doc/essays/blurb/ (consultado el 31 de julio de 2022).

[18] “sklearn.preprocessing.StandardScaler”, *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (consultado el 2 de agosto de 2022).

5 ANEXOS

ANEXO I. Enlace de proyecto donde se implementó del modelo de aprendizaje no supervisado.

https://colab.research.google.com/drive/1W6a_WTAABY5Z2PwpRvGMh4KTgasXUZuv?usp=sharing