Propuesta de Automatización de Ventilación en Data Centers mediante Herramientas de Inteligencia Artificial

Federico, D'Angiolo – Fernando, Asteasuain – Ivan, Kwist – Matias, Loiseau
Ingeniería en Informática.
Departamento de Tecnología y Administración.
Universidad Nacional de Avellaneda – Avellaneda, Buenos Aires, Argentina.
{fdangiolo,fasteasuain, ikwist, mloiseau}@undav.edu.ar

Resumen

En la actualidad, la Inteligencia Artificial (IA) se encuentra abarcando muchas áreas, sobre todo la de procesamiento de datos que tiene en cuenta el sensado de variables climáticas. Por esta razón, en el presente trabajo, se propone estudiar cómo se puede modificar la ventilación de un Data Center (DC) mediante el análisis de las variables climáticas que afectan a dicho recinto. Para esto, utilizando IA, se podrán procesar datos como por ejemplo, temperatura, humedad y presión. Para llevar a cabo esta acción, se propone en un comienzo realizar un sensado de dichas variables en el Data Center y luego, mediante IoT (Internet of Things), enviar estos datos a la nube. Con estos datos a disposición, el objetivo de este trabajo es obtener modelos que permitan estimar cuándo activar dicha ventilación de acuerdo a cómo se agrupan los meses del año en términos del clima. Esto se llevará a cabo mediante herramientas de IA, más específicamente, utilizando Machine Learning (ML).

1. Introducción

Hoy en día, dado el avance tecnológico en lo respectivo a Industrias 4.0, en especial, en las áreas de sistemas en tiempo real y de sensores para el muestreo de datos, resulta interesante estudiar cómo se pueden integrar éstas con los algoritmos de Inteligencia Artificial (IA)[1] [2]. Haciendo mayor foco en esta conjugación, es importante estudiar tecnologías como AIoT/IIoT/IoT (Artificial Intelligence of Things/Industrial Internet of Things/Internet of Things) que permiten realizar mejoras en el ámbito que respecta a IA e IoT [3]. Esto último tiene dos aristas, por un lado, mediante IA, se pueden realizar análisis de correlación entre datos para poder construir modelos que permitan predecir ciertos comportamientos y, por el lado de IoT, tenemos la conexión de distintos dispositivos y sensores con otros, mediante Internet. Ambas han crecido de manera exponencial en la industria como en la academia [4]. Atendiendo al caso de los Data Centers, resulta de una gran necesidad monitorear variables como temperatura, humedad y presión para que los equipos que se encuentran dentro, puedan trabajar correctamente [6]. Por esta razón, es conveniente estudiar modelos que permitan, no sólo monitorear al ambiente sino que puedan actuar sobre el mismo dando indicaciones de cuándo encender o apagar ventilaciones. Para llevar a cabo esta labor, en este trabajo se describen conceptos de Aprendizaje Supervisado (AS) y No Supervisado (ANS). Específicamente dentro de AS, algoritmos que permitan realizar clasificaciones tales como Naive Bayes, KNN y Árboles de decisión, mientras que, por el lado del ANS, se comentarán algoritmos como K-Means [13] [14] [15]. Al estudiar estos modelos, se podrá concluir cuál es el mejor a la hora de activar la ventilación dentro del DC, objetivo principal de este desarrollo. Por último, articulando estos algoritmos con la necesidad en el cuidado de un Data Center, surge la idea de Green Data Centers en donde se puede apreciar la idea transformadora de ésta a la hora de reducir el consumo necesario para mantener refrigerado a dicho DC. De esta forma, el objetivo es lograr eficiencia energética y sostenibilidad mediante una gestión térmica [16] [17] [18].

1.1. Conceptos Generales aplicados al caso de estudio

El objetivo de estudiar clasificadores radica en el siguiente caso: si se quiere conocer cuándo activar la ventilación en un momento del año, resulta importante disponer, para ese momento, los valores de temperatura, humedad y presión (pensados como ternas de valores). Con ellos, se puede entonces utilizar un clasificador y obtener el resultado de activar o no una ventilación, siendo que la etiqueta que usen los clasificadores para su funcionamiento, proviene del agrupamiento que realice

previamente K-Means. Para conocer cuál puede ser el mejor modelo de clasificador, se acude a evaluar su precisión (accuracy, en inglés).

Para el caso de la presente descripción, las variables climáticas que se encuentran dentro del Data Center, serán muestreadas mediante sensores de temperatura, humedad y presión, para luego, mediante protocolos de IoT, transmitirlas a ThingSpeak, la cual es una plataforma que permite realizar la visualización y seguimiento en tiempo real de los datos. Desde aquí, se podrán descargar los mismos mediante un dataset el cual contiene datos tomados en tiempo real, para lo cual, resulta importante estudiarlos y realizar una limpieza si fuese necesario. Además, hay que considerar que los mismos no se encuentran etiquetados para que los clasificadores puedan trabajar correctamente. Por esta razón, uno de los primeros pasos es realizar una limpieza de datos para luego poder aplicarles algoritmos de clustering como por ejemplo, K-Means, y poder así estudiar su agrupación. Luego de esta acción, se procederá entonces a entrenar algoritmos de clasificación como los mencionados.

En base a lo comentado este trabajo propone, mediante el uso de herramientas de Machine Learning (ML), la construcción de modelos que permitan predecir cuándo sería conveniente encender o apagar la ventilación dentro de un Data Center. Esta activación de la ventilación dependerá de los cambios climáticos internos al DC, es decir, las variaciones en temperatura, humedad y presión debido a la apertura y cierre de puertas, al intercambio de calor con el exterior, y la cantidad de personas que ingresen o egresen. De esta forma, la selección de los hiperparámetros de los algoritmos de ML se realizará en base a la premisa de que los equipos dentro del Data Center, no sobrecalienten debido a la acción climática del ambiente. Es importante aclarar que este trabajo describe la acción comentada y que, como trabajo a futuro, se pretende validar la acción final de cuantificar cómo cambian las variables climáticas al accionar la ventilación, es decir, mediante un lazo cerrado que incluya la acción de ventilación.

1.2 Trabajos relacionados

En la actualidad existen diversos estudios, muchos de ellos resultan importantes a la hora de incrementar la eficiencia en la ventilación de Data Center puesto que así se reduce el consumo eléctrico del recinto. Es aquí entonces donde dela contribución de IA resulta de gran aporte ya que se pueden crear modelos predictivos que contribuyan con el consumo eléctrico. Es decir, mediante la recolección de datos, se pueden crear modelos para mejorar la ventilación del DC [5]. En este sentido, actualmente resulta muy interesante la investigación en el área de “Green Data Centers” donde se pueden encontrar objetivos como: optimizar la eficiencia de energía en el hardware utilizado, lograr la integración de energías renovables y estudiar cómo mejorar los sistemas de refrigeración. Todos estos temas requieren de algoritmos de Inteligencia Artificial para realizar estas mejoras [6] [7] [8] [9] [10].

Existen diversos trabajos que relacionan Data Centers con Machine Learning como por ejemplo, aquellos donde se crean modelos de predicción para conocer el consumo y la velocidad que debe tener la ventilación para mejorar el rendimiento de Data Centers, es decir, mediante el muestreo de datos, se crean modelos basados en ML para poder saber cómo afectará el rendimiento de la sala [11]. Por otro lado, se pueden encontrar trabajos donde, mediante técnicas de ML, más específicamente, usando clustering, se pueden encontrar debilidades y fortalezas de las características térmicas del Data Center. Esto se realiza tomando datos extraídos del mismo recinto. Aquí se puede ver cómo mediante este tipo de técnicas se pueden encontrar formas de gestionar y mejorar la ventilación del ambiente citado [12].

Teniendo en cuenta los trabajos descriptos, en este trabajo se pretende obtener modelos que permitan estimar cuándo debe activarse la ventilación dentro del DC para mitigar efectos de temperatura y humedad en los distintos meses del año. Para esto, se estudiarán algoritmos de Clustering y Clasificadores.

Este trabajo se distribuye de la siguiente forma: En la sección 2, se comentarán brevemente los Algoritmos de Aprendizaje Supervisado y No Supervisado. Luego, en la sección 3, se describe la aplicación de estos algoritmos al caso de estudio para que a continuación, en la sección 4, se puedan estudiar los análisis y resultados obtenidos. Finalmente, en la sección 5 se describen las conclusiones con los posibles trabajos a futuro.

2. Machine Learning: Conceptos preliminares.

Para poder llevar a cabo la labor de estimar cuándo apagar o encender la ventilación del DC, a continuación se realizará una breve descripción de cada uno de los algoritmos utilizados. Se comienza con la explicación de K-Means, un algoritmo de Aprendizaje No Supervisado que será utilizado para generar etiquetas (explicado en la sección 3). Luego se comentan los algoritmos de Aprendizaje Supervisado, más precisamente, de clasificación. En este caso, se hace referencia a Naive Bayes, KNN y Árboles de decisión, que permitirán realizar la predicción en cuanto a la activación o no de la ventilación.

2.1. K-Means

El algoritmo K-Means es un modelo de ANS que se usa en casos donde se tiene un conjunto de datos no

etiquetados. Este modelo se encuentra basado en centroides y, mediante el cálculo de distancias entre el centroide y los puntos contenidos en el conjunto de datos, se pueden crear clusters o grupos de datos.

Para esto, primero se toma un valor de K y luego se divide a los datos en K categorías, de forma que exista similitud entre los datos y poder así, generar grupos con las mismas características [14].

Cuando se habla de distancias, existen generalmente dos tipos: Euclidiana y de Manhattan. Es de gran importancia estudiar en algunos casos, cuál de las dos conviene utilizar.

2.2. Naive Bayes

Este tipo de clasificador tiene un gran desempeño y permite determinar la probabilidad de un suceso dado un conjunto de condiciones, mediante el teorema de Bayes. El enfoque es sencillo desde su utilización, entendiendo que el adjetivo "ingenuo" (Naive) se atribuye a la independencia condicional de las causas. Esto, en algunos casos, puede ser difícil de aceptar inmediatamente, dado que en muchos contextos la probabilidad de una característica particular está estrictamente correlacionada con otra. Es decir, la aparición de una causa no suele ser independiente de la presencia de otras. Sin embargo, en Zhang H., "La optimalidad de Naive Bayes", AAAI 1, no. 2 (2004): 3, el autor demostró que bajo condiciones particulares, diferentes dependencias se eliminan entre sí, con lo cual un clasificador de Naive Bayes puede lograr un rendimiento muy alto incluso si su ingenuidad se viola.

2.3. KNN

El algoritmo de clasificación KNN (K-Nearest Neighbor), o algoritmo de vecinos más cercanos, es de gran utilización donde la idea central es clasificar nuevos datos de categoría desconocida, en base a la distancia euclidiana (u otra forma de medir distancia), respecto a los datos previamente conocidos, los cuales, se pueden encontrar en un dataset.

El primer paso de este algoritmo es extraer las características del dato a clasificar y compararlas con las de cada dato de categoría conocida que se encuentran en el conjunto de prueba (dataset). A continuación, el segundo paso es determinan los ***K*** vecinos más cercanos y realizar una votación para determinar la categoría a la que pertenecerá el dato de muestra.

Este algoritmo tiene algunas ventajas, como por ejemplo, ser fácil de comprender y tener un buen efecto de clasificación. Sin embargo, como desventajas, es útil mencionar el costo en tiempo y espacio [15].

2.4. Árboles de decisión

Este algoritmo, ampliamente usado en acciones de clasificación, se basa en nodos y ramas, como en un árbol. Cada uno de los nodos representa características en una categoría que será clasificada. Debido a su simplicidad y precisión en diversos conjuntos de datos, los árboles de decisión son ampliamente utilizados en distintas aplicaciones [13].

3. Aplicación al caso de estudio

Previamente al estudio en cuanto a la activación de las distintas ventilaciones en el Data Center, resulta conveniente describir cómo se realizó el sensado de las variables climáticas en el ambiente. Para esto, en la figura 1, se muestra una vista superior del Data Center y cómo se distribuyen los distintos servidores.

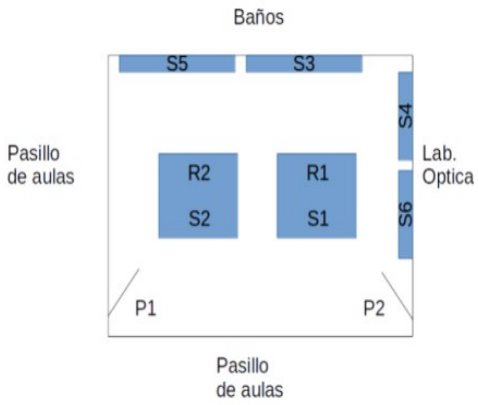

Figura 1. Layout del Data Center

La distribución espacial que se propone en la figura 1, tiene en cuenta a los Racks denotados mediante R1 y R2 los cuales alojan un servidor cada uno. Luego los módulos S1,...,S6, construidos por los autores para el desarrollo del presente trabajo, son los encargados de sensar temperatura, humedad y presión del ambiente, los cuales se encuentran basados en módulos WiFi ESP8266 y sensores DS18B20 y DHT22. Estos módulos se encargan de tomar la información climática del medio y la envían, mediante WiFi, a un servidor externo (Thingspeak) el cual almacena todos los datos. Como resguardo, los datos de temperatura sensados se envían a los propios servidores del Data Center. En esta etapa, es importante resaltar que la toma de datos se realiza cada quince segundos dado que es el período de muestreo permitido por el servicio de la plataforma de datos, ThingSpeak. Además, el sensado se realizó entre septiembre y marzo, con lo cual, no se pudo sensar a los meses de invierno.

Además es importante tener en cuenta que el Data Center se encuentra rodeado de otros laboratorios, es decir,

no se encuentra sujeto a variaciones climáticas externas importantes. Aparte, como se puede ver en la figura 1, el DC cuenta con dos puertas denotadas con P1 y P2, las cuales se abren y se cierran cada vez que ingresan o egresan personas, lo que puede producir cambios en el flujo de calor. Esto es adecuado mencionar pues la apertura de las puertas, puede provocar variaciones en la temperatura, humedad y presión del lugar, sobre todo en épocas de trabajos intensos. El lugar no cuenta con ventanas hacia el exterior.

En algunas ocasiones, ocurren caídas en la red de WiFi o cortes de luz momentáneos, los cuales producen un reseteo de los módulos y, al producirse esto, se emiten valores que no se condicen con el comportamiento del ambiente hasta el momento. Todo lo descripto hasta aquí, reviste de gran importancia, dado que así se conforma el conjunto de datos, el cual resulta ser la entrada a los algoritmos de ML. Por lo tanto, si los mismos no se encuentran revisados adecuadamente, los algoritmos clasificarán de forma incorrecta.

En primera instancia, dado que los datos obtenidos no se encuentran etiquetados, debido a que se obtienen directamente de los sensores, resulta importante aplicar un algoritmo de clustering que permita conformar grupos (clusters). Por esta razón, se procederá a utilizar K-Means dada su versatilidad. Luego, se procederá a usar distintos algoritmos de clasificación.

3.1. Datos obtenidos

Los datos obtenidos a través de los sensores necesitan ser procesados dado que, en algunas circunstancias, debido a cortes de luz o de WiFi, los mismos no son del todo correctos. Para solucionar esto, se utilizan los Boxplots ya que estos permiten visualizar valores atípicos, sobre todo en datos obtenidos a través de mediciones, los cuales pueden tener valores fuera de lo esperado. De esta manera, con esta herramienta, se pueden detectar datos atípicos y eliminarlos, en caso de que sea necesario. A modo de ejemplo, en la figura 2a se observa cómo quedan los datos de temperatura cuando son obtenidos de forma cruda y en la figura 2b se muestran los datos procesados. De forma similar se procede con los datos de humedad y presión.

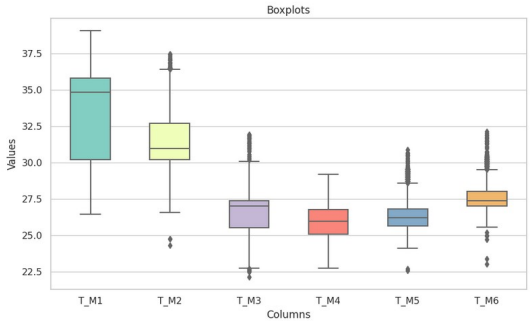

Figura 2a. Boxplots de temperatura: datos crudos.

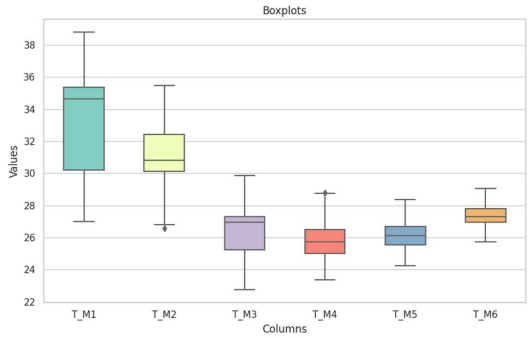

Figura 2b. Boxplots de temperatura: datos procesados

En los gráficos anteriores, TM1,...,TM6, representan los valores de temperatura obtenidos de cada módulo instalado en el Data Center. De las figuras anteriores, se puede ver que el caso donde se filtran los datos (figura 2b), representa un mejor conjunto de datos para luego ser utilizado por los algoritmos de ML.

3.2. Limitaciones del estudio

Ante lo comentado hasta aquí, mas específicamente en la sección 1.1, a futuro se pretende agregar la acción de ventilación para observar los resultados. Este trabajo se concentra en muestrear las variables ambientales para que, mediante algoritmos de IA, se pueda conocer cuándo sería conveniente encender o apagar la ventilación dentro de un Data Center.

Además, ante los datos obtenidos en el período mencionado, como trabajo a futuro, se pretende cuantificar los meses restantes para obtener un año completo de sensado; al día de hoy, se puede considerar una limitación del estudio pero que a futuro será contemplada.

4. Análisis y resultados

Dada la cantidad de datos obtenidos de los sensores, el primer paso es la utilización de K-Means para conformar grupos de datos y, para esto, se utilizarán los datos provenientes de la periferia de los servidores, es decir, de

los módulos S3,...,S6. Es importante aclarar que se usan los datos de la periferia dado que pueden muestrear cómo se comporta el clima dentro del DC. Con esto se quiere decir lo siguiente: si además de muestrear la periferia, tomamos los módulos S1 y S2, estos últimos se verán perturbados por la temperatura de los servidores dado que se encuentran sobre dichos servidores. Por esta razón, se declinó en el uso de los módulos S1 y S2.

De esta forma se podrán conformar grupos de datos que tengan en cuenta la hora, día y mes de la toma de datos para así, conformar grupos de meses donde resulte necesaria la activación de la ventilación. De esta forma cada cluster o grupo, indicará cuándo se debe ventilar el Data Center, teniendo en cuenta la variación de temperatura, humedad y presión, basados en las distintas variaciones que pueda haber en la sala.

Previamente al análisis de los datos, se procederá a comentar el hardware, el software y el conjunto de datos utilizados. Luego se procederá a estudiar los resultados obtenidos mediante algoritmos de clustering y clasificación.

4.1. Hardware, Software y Conjunto de datos.

En cuanto al Hardware, se utilizó un procesador Intel Core I7 de 2.40GHz con 16 GB de Memoria RAM. Luego, en cuanto al Software, se desarrolló en Python mediante bibliotecas de Scikit learn.

En base a los datos obtenidos mediante la toma de datos con sensores se dividió al conjunto en dos partes: una para entrenamiento y otra para validación, siendo esta relación de 70/30, respectivamente.

Con respecto al conjunto de datos, el mismo contiene como columnas, a los datos de temperatura, humedad y presión de cada módulo y en sus filas se puede encontrar el dia, el mes y el año como así también el valor numérico de cada una de las variables [19]. El código se puede encontrar en [20], el cual se encuentra en modificación continua.

4.2. Análisis mediante clustering.

Dado que el primer paso es aplicar K-Means, se accede a llevar a cabo la curva de Elbow para observar cuál es la cantidad de clusters adecuada. En la figura 3 se puede ver esto:

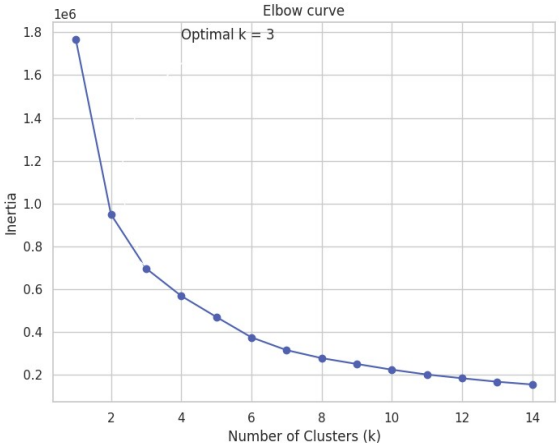

Figura 3. Elbow para datos de la periferia de los servidores

En la figura 3 se puede observar que el número de clusters es K = 3, el cual surge tomando el “codo” de la curva. Algorítmicamente esto se produce tomando el punto de inflexión mediante la segunda derivada.

Dado que el valor resultante es K = 3, esto genera la idea de que habrá tres grupos de datos que serán establecidos entre los distintos meses. Para comprender y visualizar esto, se propone la siguiente tabla:

Tabla 1. Tabla de asignación de clusters. K vs Meses

K vs Meses		
Septiembre	Octubre-Febrero	Marzo
Cluster 0	Cluster 1	Cluster 2

En la tabla 1 se puede observar que K-Means agrupa las épocas del año en tres, lo cual genera la idea de que existirán tres etiquetas que serán utilizadas luego, por los clasificadores. Dado que K-Means tiene como salida la cantidad de clusters (K), se debe adoptar una política de ventilación para los meses del año según este valor de K. Los valores de la tabla 1, se obtuvieron de acuerdo a cómo se agrupan los labels según las fechas, por ejemplo, para el caso del Cluster 0, la fila correspondía al mes de septiembre, para del Cluster 1, las filas correspondían a los meses de octubre-febrero y para el Cluster 2, las filas seleccionadas se asignaron al mes de marzo.

Estos agrupamientos obtenidos, podrán ser utilizados como etiquetas (labels, en inglés), para la utilización de los distintos clasificadores y saber así, cuándo activar la ventilación. Por ejemplo, dado que la cantidad de Clusters varía entre 0 y 2, una posible política de ventilación podría ser que la misma sea activada para cuando la etiqueta se encuentra en el Cluster 1 y, para cuando las etiquetas valen 0 o 2 apagar la ventilación. Esto se puede lograr mediante clasificadores.

A modo descriptivo, a continuación se muestra en la tabla 2, los centroides de cada cluster y luego la justificación de por qué se eligió al Cluster 1 como

elección de ventilación. Dado que son tres clusters, se encuentran tres centroides con sus respectivas características (features, en inglés). Cada uno de estos contienen los valores de cada variable, es decir, analizando la tabla 2, los cuatro primeros valores pertenecen a la variable de humedad, seguidos de los cuatro valores de temperatura y los últimos cuatro pertenecientes a la presión (cada centroide en columna).

Tabla 2. Tabla de asignación de centroides

Centroides		
Cluster 0	Cluster 1	Cluster 2
42.81	51.51	60.67
41.03	38.56	47.98
39.68	41.70	50.00
38.60	39.25	47.48
26.31	27.26	26.36
25.54	26.74	25.92
26.62	26.11	25.98
27.51	27.86	27.58
1019.95	1007.03	1013.18
1021.11	1008.16	1014.35
1021.01	1008.10	1014.28
1019.71	1006.78	1012.98

De la tabla anterior, se puede observar cómo se distribuyen los centroides. La elección de activar la ventilación en el Cluster 1, proviene de cómo se comporta la temperatura. Si se analiza la tabla 2, en el Cluster 1 se ven los valores de temperatura un poco más altos en comparación con los otros clusters (filas 5 a 8), por eso la elección. También puede implementarse otra política de ventilación en base a la distribución de los centroides.

4.3 Comparación entre clasificadores.

Dado que en el análisis anterior se obtuvieron las distintas etiquetas o labels, a continuación se propone estudiar la precisión de los clasificadores para que, en base a valores tomados en un determinado momento de temperatura, humedad y presión, se pueda saber cuál resulta en un mejor modelo para activar la ventilación del DC.

A continuación se muestra una tabla de comparación entre los distintos clasificadores para observar la puntuación (accuracy), que los diferencia:

Tabla 3. Tabla ejemplo

Comparación de Clasificadores		
Clasificador	Parámetros	Accuracy
Naive Bayes	-	97.20
KNN	5	98.74
Árboles de Decisión	3	96.43

En el caso de Naive Bayes, no se utilizaron hiperparámetros. En cambio, para KNN y Árboles de Decisión, se utilizaron ***K*** (este ***K*** se refiere a KNN) y max depth, los cuales se explican a continuación:

En el caso de KNN, el valor de ***K*** = 5 se tomó en base a la iteración entre distintos valores de ***K***, observando cómo varía su accuracy. Para comprender por qué se eligió este valor, se procede a graficar el accuracy vs ***K***, en la figura 4:

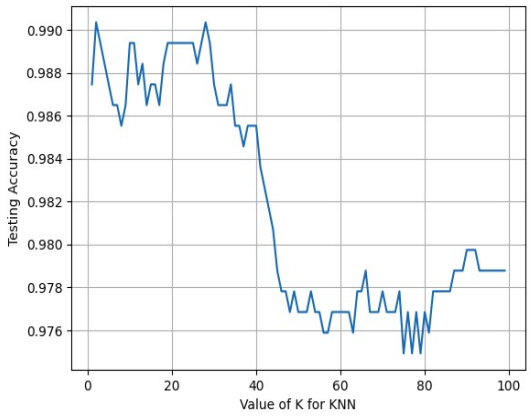

Figura 4. Accuracy vs K. KNN

De la figura 4 se puede ver que para valores de ***K*** entre 1 y 40, aproximadamente, el accuracy varía entre un 98,6% y un 99%, es decir, un 0,4%. De aquí se puede inferir que, si bien es bastante oscilante su accuracy, tomando en cuenta un error del 0,4%, se podría tomar cualquier valor entre 1 y 40 para ***K***. En este caso se prefirió utilizar ***K*** = 5. De la misma figura, es evidente que tomar valores mayores a ***K*** = 40, sería perjudicial para el modelo ya que se incurre en un menor valor de accuracy.

Para el caso del árbol de decisión, se iteró con la profundidad del mismo (max depth) evaluando su accuracy, de forma similar al caso anterior. En la figura 5 se puede ver cómo varía dicho accuracy el cual, es bastante oscilante. De forma similar al caso de KNN, se puede ver que el accuracy varía entre un 98% y un 98,7% como casos extremos, pero, si consideramos una franja acotada entre un 98,2% y un 98,6% para tener un error de 0,4%, entonces podríamos elegir un valor para la

profundidad de 3. De esta forma, estaríamos acotando el accuracy a valores más predecibles.

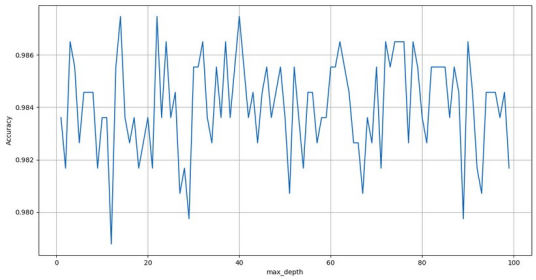

Figura 5. Accuracy vs max depth. Árboles de decisión

Finalmente, en la tabla 3 se puede ver que los tres clasificadores poseen puntuaciones similares; sin embargo, KNN parece ser el mejor para este caso.

Para contrastar el análisis anterior, se pueden utilizar otras métricas importantes, como por ejemplo, las tablas de confusión. Estas permiten hacer un análisis entre el valor verdadero y el valor predicho por cada clasificador. A continuación se muestran dichas tablas para Naive Bayes, KNN y Árboles de Decisión:

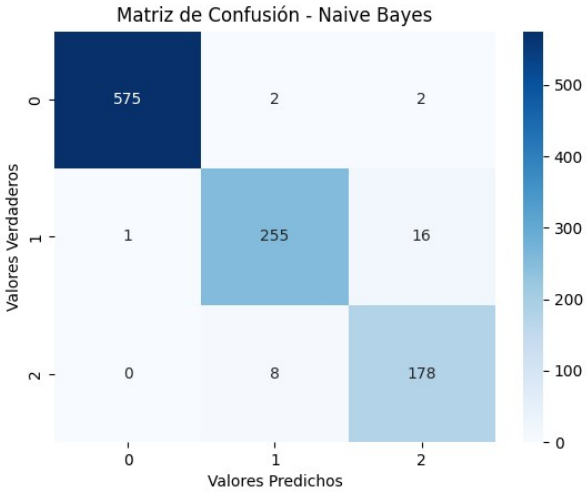

Figura 6. Tabla de Confusión para Naive Bayes.

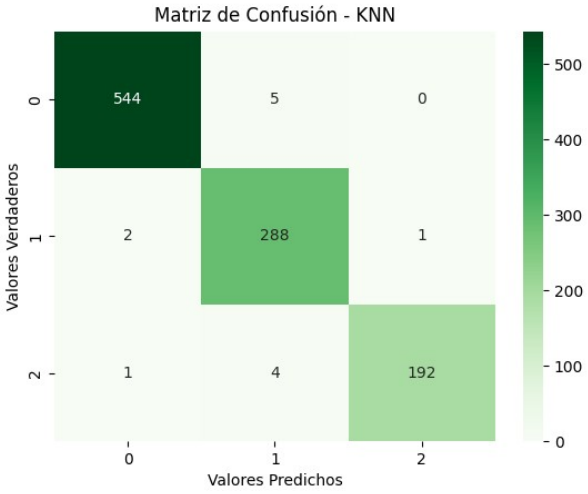

Figura 7. Tabla de Confusión para KNN.

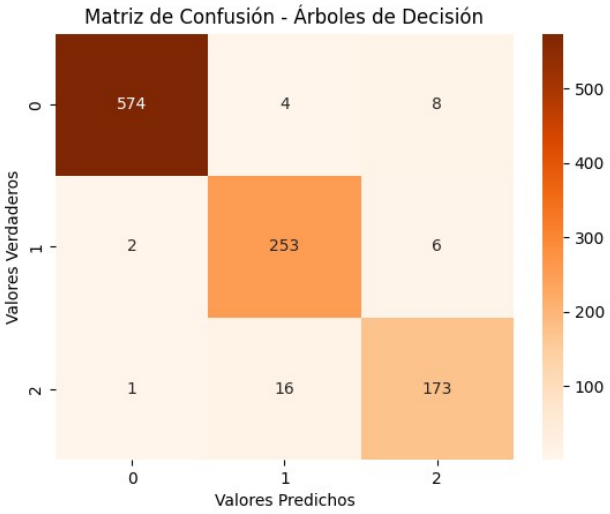

Figura 8. Tabla de Confusión para Árboles de Decisión.

Previamente a realizar el análisis, conviene aclarar que las clases marcadas en cada tabla, en los ejes horizantales y verticales, se refieren a los clusters, es decir, la clase 0, corresponde al Cluster 0 y así sucesivamente.

Observando las tablas de confusión para cada clasificador, se puede observar una gran coincidencia entre los valores verdaderos y valores predichos, lo cual se corresponde con el accuracy obtenido anteriormente. En el caso de KNN, esta coincidencia es más acentuada en la diagonal principal, sobre todo en las clases 1 y 2 pero no tanto así en la clase 0, con respecto a Naive Bayes y a Árboles de Decisión, lo cual valida el análisis anterior mediante accuracy, para elegirlo a este, como el mejor clasificador de los tres.

Por último, para concluir el análisis de métricas sobre los clasificadores, se agregan Precision, Recall y F1-Score, en la siguiente tabla:

Tabla 4. Comparación de métricas para los clasificadores

Métrica	Clase	Naive Bayes	KNN	Árboles de Decisión
Precision	0	1	0,99	0,99
	1	0,96	0,97	0,93
	2	0,91	0,99	0,93
Recall	0	0,99	0,99	0,98
	1	0,94	0,99	0,97
	2	0,96	0,97	0,91
F1-Score	0	1	0,99	0,99
	1	0,95	0,98	0,95
	2	0,93	0,98	0,92

De la tabla 4 se puede observar que, para la clase 0, la métrica Precision es mejor en Naive Bayes pero, para las clases 1 y 2, este clasificador decae frente a KNN. Algo similar ocurre para Recall y F1-Score, donde para dichas clases, KNN resulta ser un mejor clasificador. Para el caso de Árboles de Decisión, se ve que sus valores son menores a Naive Bayes y KNN, por eso se lo descarta. Estos análisis coinciden con el realizado previamente con accurcay y tablas de confusión. De aquí entonces que se elige a KNN como el clasificador más acorde.

De esta forma, tomando un conjunto de valores de temperatura, humedad y presión en cualquier época del año comprendida entre septiembre y marzo, se puede entonces saber si se debe apagar o prender determinada ventilación.

5. Conclusiones y trabajos a futuro

Este trabajo muestra cómo se pueden obtener, mediante algoritmos de IA, modelos que permitan gestionar la ventilación de un Data Center. Para esto, resulta muy importante realizar un procesamiento de los datos obtenidos a través de sensores dado que muchos de ellos pueden presentar casos atípicos. De aquí se puede ver entonces que, aplicando K-Means, se pueden agrupar las épocas del año en tres. Luego, mediante el uso de clasificadores, se puede ver que el mejor de ellos resulta ser KNN aunque Naive Bayes y Árboles de decisión resultan tener un accuracy parecido. Además se utilizaron otras métricas como Precision, Recall y F1-Score, que permiten validar que KNN es el mejor clasificador para este caso. Esta conclusión tiene en cuenta los hiperparámetros de cada clasificador. Es importante aclarar que estos clasificadores fueron elegidos como un primer avance, es decir, si bien existen otros con sus respectivas bondades, se espera a futuro, analizar más clasificadores, como por ejemplo, SVM, Random Forest y XGBoost, entre otros, y compararlos mediante accuracy u otra métrica pertinente como por ejemplo, tablas de confusión, Precision, Recall y F1-Score, entre otros.

Como trabajo a futuro, resulta importante poder estudiar otros modelos de ML, teniendo en cuenta las métricas asociadas. Incluso, se puede incursionar en otras ramas de IA, como Reinforcement Learning o Heurísticas, de manera de tener un mayor aprendizaje sobre cómo mejorar la situación climática del DC.

En esta línea, resulta conveniente estudiar técnicas como Gridsearch para optimizar parámetros de algoritmos, sobre todo, para saber cómo se puede optimizar este proceso teniendo en cuenta el tiempo de ejecución. Algo importante a futuro, es contemplar el hecho de ver cómo afecta la ventilación a las variables, es decir, observar cómo y cuánto cambian la temperatura, humedad y presión en comparación con las aquí analizadas.

Por último, como trabajo a futuro, es importante muestrear los meses faltantes para completar el dataset de todo un año y tener así, una mayor precisión en los resultados.

Referencias

[1] Suresh Neethirajan.: “Artifcial Intelligence and Sensor Innovations: Enhancing Live-stock Welfare with a Human-Centric Approach”. 2024.

[2] Shedrack Onwusinkwue., Femi Osasona., Islam Ahmad Ibrahim Ahmad., Anthony Chigozie Anyanwu., Samuel Onimisi Dawodu., Ogugua Chimezie Obi,. Ahmad Hamdan.: Artificial intelligence (AI) in renewable energy: A review of predictive maintenance and energy optimization. 2024.

[3] Kun Mean Hou., Xunxing Diao ., Hongling Shi ., Hao Ding ., Haiying Zhou., Christophe de Vaulx.: Trends and Challenges in AIoT/IIoT/IoT Implementation. 2023.

[4] Javid Ghahremani Nahr., Hamed Nozari., Mohammad Ebrahim Sadeghi.: Green supply chain based on artificial intelligence of things (AIoT). 2021.

[5] Isaev E.A., Kornilov., Grigoriev A.A.: Data Center Efficiency Model: A New Approach and the Role of Artificial Intelligence. 2023.

[6] Ebirim, W., Usman, F., Olu-lawal, K., Ehiobu, N., Ani, E., Montero, D.: Optimizing energy efficiency in data center cooling towers through predictive maintenance and project management. World Journal of Advanced Research and Reviews, 2024, 21(02), 1782–1790. 2024.

[7] Chidolue, O., Ohenhen, P., Umoh, A., Ngozichukwu, B., Fafure, A., Ibekwe, K.: GREEN DATA CENTERS: SUSTAINABLE PRACTICES FOR ENERGY-EFFICIENT IT INFRASTRUCTURE. Engineering Science & Technology Journal P-ISSN: 2708-8944, E-ISSN: 2708-8952 Volume 5, Issue 1, P.No. 99-114, January 2024.

[8] Toto Widyanto., Widjojo Hardjoprakoso.: Greenship Data Center – A Green Data Centre Standard for Indonesia. Proceeding of International Conference on Multidisciplinary Research for Sustainable Innovation, Vol. 1 No. 1 (2024) DOI to be assigned soon. 2024.

[9] Yaran Lianga., Peng Lib., Wen Sua., Wei Lib., Wei Xub.: Development of green data center by configuring photovoltaic power generation and compressed air energy storage systems 2024.

[10] Verónica Bolón., Laura Morán., Brais Cancela., Amparo Alonso.: A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing 2024.

[11] Yaman M. Manaserh., Mohammad I. Tradat., Dana Bani-Hani., Aseel Alfallah., Bahgat G. Sammakia., Kourosh Nemati., Mark J. Seymour.: Machine Learning Assisted Development of IT Equipment Compact Models for Data Centers Energy Planning. 2021.

[12] Anastasiia Grishina., Marta Chinnici., Ah-Lian Kor., Eric Rondeau., Jean-Philippe Georges.: A Machine Learning Solution for Data Center Thermal Characteristics Analysis. 2020.

[13] Bahzad Taha Jijo., Adnan Mohsin Abdulazeez.: Classification Based on Decision Tree Algorithm for Machine Learning. IT Department, Technical College of Informatics Akre, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. 2021

[14] Mengyao Cui.: Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. Shandong University of Finance and Economics, Jinan, Shandong, China. 2020

[15] Haiyan Wang.,Peidi Xu., and Jinghua Zhao.: Improved KNN Algorithm Based on Preprocessing of Center in Smart Cities. College of Computer Science and Technology, Changchun University, Changchun 130022, China and College of Computer, Jilin Normal University, Siping 136000, China. 2021

[16] Onyinyechukwu Chidolue, Peter Efosa Ohenhen, Aniekan Akpan Umoh, Bright Ngozichukwu, Adetomilola Victoria Fafure, & Kenneth Ifeanyi Ibekwe. Green Data Centers: Sustainable Practices For Eenergy-Eficient IT Infrastructure. 2024.

[17] Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, Amparo Alonso-Betanzos. A review of green artificial intelligence: Towards a more sustainable future. 2024.

[18] M Rambabu, Dr. Kunchanapalli Rama Krishna, Dr Mano Ashish Tripathi, Jyoti Kataria, Priyanka Srivastava, Ajay Dixit. Green Computing: Advancing Energy-Efficient Data Centers With AI. 2025

[19] Dataset: https://gitlab.com/info_investigacion/datasets.git

[20] Código: https://gitlab.com/info_investigacion/code