2019

CONAIISI

VII Congreso Nacional de Ingeniería Informática - Sistemas de Información

14 y 15 de Noviembre de 2019

RIIITIC

DIIT
Departamento de Ingeniería e Investigaciones Tecnológicas

confedi
Consejo Federal de Decanos de Ingeniería Argentina

Universidad Nacional de La Matanza

Algoritmo de KNN Aplicado al Mantenimiento de un Datacenter

Federico Gabriel D'Angiolo[1], Iván Federico Kwist[1], Matias Loiseau[1], David Exequiel Contreras[1], Gregorio Oscar Glas[1]
[1]Universidad Nacional de Avellaneda, Avellaneda, Argentina. Departamento de Tecnología y Administración. Ingeniería en Informática.
fdangiolo@undav.edu.ar, ivankwist@hotmail.com.ar, matiasloiseau@gmail.com, anddcp@gmail.com, gglas@undav.edu.ar

Resumen

El presente trabajo describe la aplicación del Algoritmo KNN al estudio del comportamiento climático de un Datacenter. Este análisis permite comprender cómo varían la temperatura y la humedad dentro del ambiente durante las distintas estaciones del año, con el objetivo de modificar la ventilación dentro del recinto. Esto último permite mantener operativos a los servidores y sistemas de cómputo que se encuentran dentro. Si bien esta aplicación resulta específica para el Datacenter, el concepto se puede aplicar a cualquier habitación o recinto que necesite tener un control de sus variables.

1. Introducción

En la actualidad se tiene acceso a una gran cantidad de datos que facilitan la automatización de sistemas y agilizan las tareas que de otra manera resultan laboriosas. Con el caudal de datos que se puede conseguir a través de distintos sensores y el avance de la tecnología en el área de las comunicaciones, resulta necesario el procesamiento de éstos con el objetivo de obtener conclusiones sobre un comportamiento en especial o determinar, como se describe más adelante en este trabajo, el grado de ventilación que necesita una habitación. Por esta razón, los algoritmos de Inteligencia Artificial (IA) obtienen una gran relevancia en el análisis de grandes datos. Ejemplo de esto se da en el Trabajo "A Review at Machine Learning Algorithms Targeting Big Data Challenges", de los autores Abhinav Rathor., Manasi Gyanchandan [1], donde se describe un análisis comparativo de los algoritmos de Aprendizaje Automático basados en grandes cantidades de datos (Big Data).

Dentro de la IA, uno de los campos que tiene gran aplicación es el Aprendizaje Automático o Machine Learning, el cual, a su vez, se encuentra dividido en tres ramas importantes: Aprendizaje Supervisado, Aprendizaje No Supervisado y Aprendizaje Reforzado. Un ejemplo de lo comentado, se da en el Trabajo "Machine Learning for Engineering", del autor Jeff Dyck [2], en donde se estudian las técnicas de Apendizaje Automático para resolver problemas de producción dentro de la automatización.

En particular, en el presente trabajo, se estudia el comportamiento dinámico que tienen las variables como temperatura y humedad dentro de un Datacenter y para esto se recurre entonces a los algoritmos de Aprendizaje Supervisado, particularmente, al algoritmo de clasificación de KNN o K-Nearest Neighbor (o, traducido, K Vecinos Cercanos), con el objetivo de poder estudiar cómo varían la temperatura y la humedad según las estaciones del año. Este análisis permite obtener conclusiones sobre la ventilación adecuada que debiera haber.

Resulta valioso poder conocer cómo se adecúan la temperatura y la humedad dentro de un Datacenter, dado que los equipos que se encuentran dentro resultan de vital importancia para mantener las comunicaciones. En el trabajo "The benefits of data center temperature monitoring", de los autores Truúcă, M., Soran, A [3], se comenta de forma clara que los equipos informáticos trabajan correctamente en un rango comprendido entre 20 °C y 23 °C, lo cual genera una de las premisas de este trabajo, es decir, cuál es el rango de temperatura que se debe mantener controlado y, por ende, monitoreado. La otra premisa de partida es el de monitoreo de humedad.

Para aplicar este algoritmo, previamente se realizó un Dataset con valores de temperatura y humedad, los cuales fueron obtenidos mediante sensores ubicados dentro del Datacenter. Estos valores fueron muestreados durante siete meses (octubre 2018 – mayo 2019, sin contar enero, donde hubo una caída irreparable de la red) para lograr un conjunto de datos robusto. Con estos datos, el objetivo es realizar tres estudios distintos, basados en el algoritmo de KNN: el primero consiste en

obtener un modelo que clasifique pares de datos de temperatura y humedad entre dos grupos, uno de ellos entre los meses octubre-febrero y el otro entre marzo-mayo. El segundo estudio toma tres grupos para la clasificación, uno de ellos entre los meses octubre-noviembre-diciembre, otro entre febrero-marzo y otro último entre abril-mayo. Finalmente, el tercer estudio, consiste en dividir los datos en siete grupos siendo cada uno de estos el respectivo mes: octubre, noviembre, diciembre, febrero, marzo, abril o mayo. Con esto, lo que se busca es analizar cuál es la eficiencia del algoritmo al clasificar con distintos tipos de grupos. Es decir, este trabajo no se enfoca directamente en encontrar el valor óptimo de **K,** sino en el análisis de división de grupos que permita obtener la mejor eficiencia para la clasificación. Las conclusiones obtenidas podrán demostrar cuál sería la mejor forma de aplicar este algoritmo a una habitación que tenga un comportamiento similar.

Esta clasificación comentada no tiene en cuenta solamente la distribución de temperatura y humedad producto de las estaciones del año, sino también a las perturbaciones que puede haber dentro del Datacenter. Este recinto resulta ser utilizado por personas que trabajan dentro, lo cual deviene en perturbaciones tales como la apertura de las puertas o incluso el impacto de la cantidad de personas que puedan encontrarse trabajando. Por el contrario, es importante analizar cómo se distribuye la temperatura y la humedad los fines de semana o momentos en los que no hay nadie, como puede ser la madrugada, para compararlos con los momentos de máxima actividad.

Resulta conveniente agregar que, aparte de sensar variables como la temperatura y la humedad los cuales resultan importantes a los efectos del Datacenter, también se muestreó la presión del ambiente para analizar, luego de las experiencias comentadas, si resulta conveniente tener un parámetro más como soporte para las dos utilizadas (temperatura y humedad).

Para el desarrollo de este trabajo se estudiaron casos relacionados con la variación de temperatura en Datacenters y casos de aplicación del algoritmo de KNN. Ejemplo de lo mencionado se da en el trabajo “The benefits of data center temperature monitoring”, en donde se analiza la monitorización de la temperatura en un Datacenter [3]. También resulta importante el trabajo “Environmental Parameters Control in Datacenter”, de los autores Truúcă, M., Soran, A., Abrudean, M [4], donde se muestra el control y monitorización de los parámetros ambientales de un Datacenter. En cuanto al algoritmo KNN, se estudió el trabajo “Un Modelo de Detección de Anomalías en una LAN usando K-NN y Técnicas de Computación de Alto Desempeño”, de los autores Gil Costa, V., Errecalde, Marcelo., Taranilla, María Teresa [5], en el cual se presenta una forma de pre-procesar datos para identificar anomalías mediante el algoritmo de clasificación KNN. También resulta interesante como base de este trabajo lo comentado en “Learning to detect spam messages”, de los autores Barrionuevo, M., Lopresti, Mariela., Miranda, Natalia., Piccoli, Fabiana [6], donde se investiga el rendimiento del método KNN en tareas de detección de spam. Teniendo en cuenta estos últimos trabajos, se puede ver la gran amplitud de aplicación que tiene el algoritmo de KNN a la hora de clasificar, teniendo en cuenta sobre todo, la eficiencia que tiene. Por esta razón, dado que en el presente trabajo resulta necesario tener una clara diferenciación de los datos, es que se acude a este algoritmo.

Teniendo en claro todos estos casos de aplicación, se procedió a realizar los estudios comentados sobre los datos de temperatura y humedad para obtener conclusiones y efectuar la correlación correspondiente con estos casos de estudio.

El presente trabajo se divide de la siguiente forma: la sección número dos comienza describiendo al algoritmo KNN, luego en la sección tres se comenta cómo se aplica este algoritmo al trabajo realizado sobre el Datacenter y por último, en la sección cuatro, se presentan los resultados obtenidos sobre los cuales se podrán evaluar las respectivas conclusiones. En particular, en la sección 4.3, se propone una mejora a la clasificación, teniendo en cuenta una tercera variable: la presión del ambiente. Es decir, si bien el trabajo describe cuál sería la mejor forma de agrupar datos con dos dimensiones, también se analiza el caso del agregado de una tercera variable con el objetivo de optimizar dicha clasificación.

2. Algoritmo KNN (K- Nearest Neighbor)

Este algoritmo se encuentra dentro del grupo de algoritmos de clasificación, siendo éstos a su vez parte del área de Machine Learning. El objetivo que tiene KNN es el de tomar una muestra con ciertas características y clasificarla dentro un grupo de datos que contenga las mismas características. Previamente a la clasificación, resulta necesario definir los distintos grupos con sus respectivas características para que, al clasificar la muestra, la misma pueda ser comparada. Cabe aclarar que cada grupo está conformado por un conjunto de datos los cuales configuran un *dataset*. Para llevar a cabo la clasificación se calcula la distancia entre la muestra a clasificar y cada uno de las valores del dataset, y se toman las **K** muestras con menor distancia, siendo **K** una variable predeterminada. Para este cómputo se utiliza la *distancia euclidiana*, la cual, para el caso en el que los datos tengan dos dimensiones, se calcula como:

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_1 - q_1)^2} \qquad (1)$$

Siendo:

p: muestra elegida p = (p_1,p_2)

p_1: primera coordenada de la muestra

p_2: segunda coordenada de la muestra

q: vecino de la muestra q = (q_1,q_2)

q_1: primera coordenada del vecino de la muestra

q_2: segunda coordenada del vecino de la muestra

Una vez definidos los **K** vecinos de la muestra, se procede a clasificarla. El algoritmo realiza ésto calculando la cantidad de vecinos perteneciente a cada grupo y asignándole a la muestra el grupo mayoritario. En caso de que haya un empate, se elige un grupo aleatoriamente. Por esta razón, resulta importante tener en cuenta el valor que tiene **K** para que el algoritmo tenga pueda clasificar eficientemente. Por ejemplo, si **K** tuviera un valor K = 100, se tendría un amplio conjunto de valores para decidir si la muestra pertenece a uno u otro grupo mientras que, por ejemplo, con un **K** = 1, la muestra quedaría definida unívocamente por el grupo de su vecino más cercano. Es por esto que resulta valioso estudiar el *Score* (o puntuación) que tiene el algoritmo para cada valor de **K**.

Lo anterior descripto se puede resumir de la siguiente forma:

1) Separar el conjunto de datos en distintos grupos, según las características. En este caso los grupo serán separados según los meses del año.

2) Calcular el valor óptimo de **K** que permite obtener el máximo *score* del algoritmo.

3) Tomar una muestra (que no sea parte del conjunto de datos), que en este caso pertenece a cualquier día del año, y clasificarla mediante el algoritmo de KNN. [7]

En el siguiente apartado se estudiará cómo se aplica este algoritmo al presente trabajo y para eso será importante analizar cuál es el valor de **K** que mejor se ajusta para obtener un resultado óptimo. Como base de este estudio se tomó el trabajo “Ischemic Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps”, en donde se estudia el valor óptimo que debe tener **K** [7].

3. Aplicación de KNN al caso de estudio

Previamente a la descripción de la aplicación, resulta fundamental comentar cómo se dispusieron los distintos sensores en el Datacenter, con el objetivo de estudiar cuál es el punto espacial del recinto bajo estudio.

En la Fig. 1 se observa la distribución de equipos y sensores dentro del Datacenter.

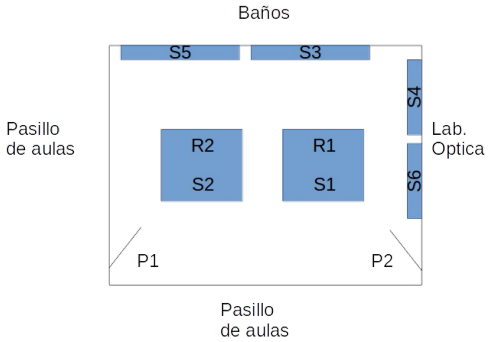

Figura 1. Distribución de equipos y sensores dentro del Datacenter

Siendo:

R1: Rack 1, donde se encuentra alojado el servidor N.º 1

R2: Rack 2., donde se encuentra alojado el servidor N.º 2

P1: Puerta que hacia el pasillo.

P2: Puerta que hacia el Laboratorio de Óptica.

S1, S2, S3, S4, S5 y S6: Módulos de temperatura, humedad y presión, integrados

Como se puede observar en la Fig. 1, existen diversos módulos que son los encargados de sensar la temperatura, humedad y presión, dependiendo su ubicación dentro del recinto. Para el caso del presente trabajo se toma como referencia el módulo tres (S3 en la Fig.1), dado que puede sensar cómo afectan las variables que circundan al Servidor N. °1 (R1 en la Fig.1), sin sufrir perturbaciones directas debido a la apertura de las puertas.

Los módulos comentados (S1, S2, S3, S4, S5 y S6) fueron construidos a partir de sensores de temperatura y humedad conectados a placas de desarrollo con WiFi integrado, de manera que cada dato tomado por dichos sensores sea enviado a través de una red de este tipo a una plataforma donde puedan ser estudiados [11]. Se utilizó la Plataforma ThingSpeak, ya que desde ésta se pueden analizar los datos vía web y de forma local. Esta última se refiere a la instalación de ThingSpeak en uno de los propios servidores del Datacenter para resguardar los datos ante un posible problema con la plataforma web. Para lograr la transmisión a través de la red de WiFi, se utilizaron protocolos de IoT, los cuales son de gran aplicación para transmisión de datos de sensores, ya que reducen el ancho de banda consumido. Si bien esto no es objeto de estudio del presente trabajo, resulta conveniente comentarlo para aclarar cómo fueron obtenidos los datos desde los sensores.

Una vez guardados los datos, resulta importante hacer un análisis exhaustivo de estos para comprobar que estén sincronizados y no sean errados. El concepto de sincronización refiere al caso de que los datos de temperatura y humedad sean del mismo momento ya que luego de analizarlos se observó que había algunos

valores de temperatura que fueron enviados con unos segundos de retraso respecto a los de humedad. Esto demandó hacer un análisis de datos para desechar aquellos cuyo retraso fuera considerable, ya que de otra forma no sería útil para su análisis. Luego, cuando se trata el concepto de errado, se refiere a tener valores que no sean coherentes con los que debe haber en un Datacenter, es decir, si un valor de temperatura en un momento dado, fue de 81 °C y diez segundos después se ve que el mismo sensor registra 23 °C, es evidente que fue un error en la toma de datos del sensor o en la comunicación, ya que resulta poco probable que en un recinto como el tratado en el presente trabajo la temperatura cambie abruptamente. Esto también resultó necesario desde el punto de vista del análisis de datos.

Con lo mencionado anteriormente, se quiere aclarar que resulta fundamental hacer un análisis de datos previo a la utilización del algoritmo de KNN, ya que de lo contrario, si se tuviera un caudal de datos importante que no cumpla con los criterios mencionados anteriormente, la clasificación se tornaría errada y además perturbaría a la forma de puntuación elegida para conocer si el algoritmo resulta eficiente a la hora de clasificar.

Con estos datos sensados se construye un dataset con valores de temperatura y humedad que van desde los 20°C hasta los 33°C y de 20% a 90%, respectivamente. Luego, para comprobar la efectividad del algoritmo KNN, se procede a dividir el dataset en distintos grupos para comprobar cuál es la forma más eficiente de separarlos. Para esto, como se comentó anteriormente, se realizarán tres experimentos: el primero consiste en separar al dataset en dos grupos y obtener el valor óptimo de **K** que permite obtener la mejor eficiencia (a través del score). Luego, el segundo experimento tiene el mismo objetivo que el anterior pero separando al dataset en tres grupos. Por último, se repite el mismo experimento pero tomando siete grupos. Hecho esto, se procederá comparar tanto el **K** obtenido en cada experimento como su *score*, para poder realizar una comparación efectiva. El criterio para separar en grupos deviene de las épocas del año donde la temperatura se nota elevada (ya que resultan críticas para el funcionamiento de los equipos), es decir, se puede tomar como grupo de “altas temperaturas” a la época comprendida entre los meses de febrero y marzo o, también, los meses de noviembre, diciembre, febrero, marzo y abril, dado que las temperaturas elevadas suelen extenderse algunos meses más. Por esta razón es que no resulta conveniente tomar grupos que contengan mayor cantidad de meses, dado que se estaría entrando en meses donde la temperatura disminuye y la clasificación se tornaría errónea.

Habiendo construido el dataset, resulta importante aclarar que todos los datos que lo componen tienen dos componentes o coordenadas: temperatura y humedad. Por ejemplo, si se toma un dato perteneciente a cualquier día del año, el mismo puede contener los valores T_1 = 26,6 °C y H_1 = 55%. De esta forma, los gráficos presentados en la siguiente sección serán exhibidos mediante dos ejes, en el vertical se muestra la humedad y en el horizontal, la temperatura

4. Análisis y resultados obtenidos

En base a lo comentado sobre la distribución de grupos, el primer experimento a realizar consiste en tomar los datos de temperatura y humedad dentro del período entre octubre del 2018 y mayo del 2019 (este período conforma al dataset), y separar a estos en dos grupos. El objetivo de esto es poder, ante una nueva muestra de temperatura y humedad, evaluar el desempeño del algoritmo KNN como se comentó anteriormente.

4.1 Clasificación mediante dos grupos de temperatura y humedad.

Los datos tomados entre octubre del 2018 y mayo del 2019 se dividirán en dos grupos: uno denominado “Periodo 0”, correspondiente a los meses de octubre, noviembre, diciembre y febrero, y otro grupo denominado “Periodo 1”, correspondiente a los meses de marzo, abril y mayo. El “Período 0” representa a los meses del año donde suele haber mayor temperatura y el “Período 1” a los meses donde la temperatura frecuenta ser menor. Esto se observa en la Fig. 2

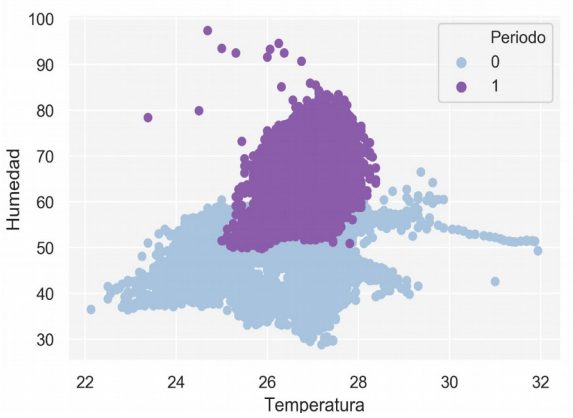

Figura 2. Gráfico del Dataset

En la Fig. 2 se grafica el dataset en un eje de coordenadas humedad-temperatura, asignándole el color celeste a los puntos pertenecientes al “Período 0” y el color violeta a los pertenecientes al “Período 1”. En la Fig.2 se observa que hay datos pertenecientes al “Período 0” que se solapan con datos del “Período 1” dado que entre los meses de marzo y mayo, la temperatura y la humedad no suelen mantenerse constantes. Además, dada la selección de meses del “Período 0”, se nota una gran variación en la temperatura pero no así en la humedad, en cambio, en el “Período 1” se nota mayor variación en cuanto a humedad pero no tanto en temperatura.

Teniendo en cuenta la distribución de los datos, el experimento se completa al tomar una nueva muestra y

observar si esta pertenece al “Período 0” o al “Período 1” y ver así, en qué época del año se encuentra el Datacenter.

De la Fig. 1 se puede observar que existe una nube de datos que se solapan entre sí (datos de color violeta por encima de los de color celeste), al intentar clasificar una nueva muestra resulta importante analizar el valor óptimo de **K** para lograr el máximo score a la hora de clasificar. Para esto, se realiza un gráfico del score que tiene el algoritmo en función de los valores de **K** el cual se puede ver en la Fig.3

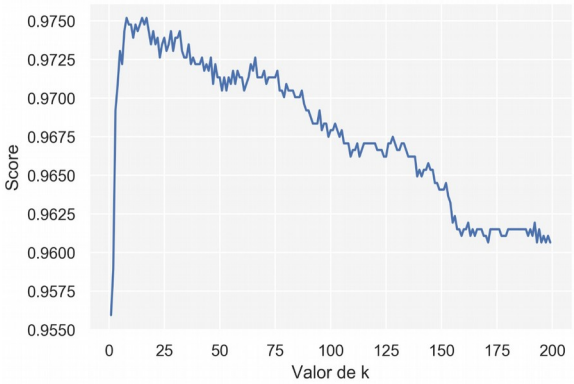

Figura 3. Score vs K

De la figura anterior se ve que a medida que el valor de **K** aumenta, disminuye la puntuación (score), por eso se resulta conveniente elegir el valor de **K** que implique un mayor score del algoritmo. Este valor, observando la Fig.3, da como resultado K = 8 para un score de 0,975.

A continuación, para comprobar la efectividad del algoritmo, se toma una muestra conformada por dos coordenadas (temperatura y humedad), la cual resulta ser de 26 °C y 55 % (esta muestra fue separada del dataset, es decir, no forma parte del mismo). El objetivo entonces es ver a cuál “Período” pertenece, tomando un K = 8, para lo cual se presenta la Tabla 1.

Tabla 1. Tabla de datos con K = 8

Temperatura	Humedad	Período
26.60	55.00	1
25.94	55.10	1
26.12	55.00	1
26.19	55.00	0
26.06	54.80	1
25.94	54.80	1
26.06	54.80	1
25.81	55.10	1

La Tabla 1 se puede representar gráficamente en la Fig.4, donde se observa a la muestra nueva con una estrella blanca y en un tono más oscuro los puntos pertenecientes a los vecinos:

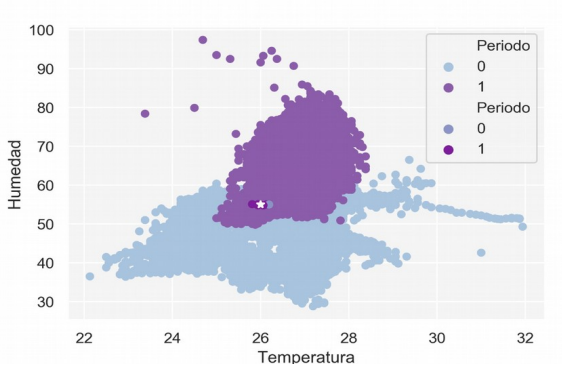

Figura 4. Nube de datos con nueva muestra

Para un mayor detalle de la nueva muestra y sus vecinos, se propone la siguiente figura (Fig.5):

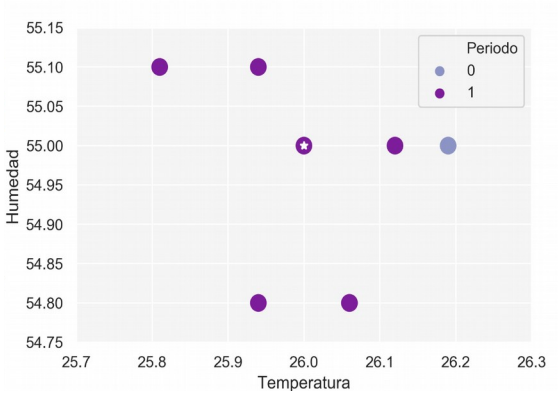

Figura 5. Nube de datos con nueva muestra, teniendo en cuenta los vecinos más cercanos.

En la Fig.5 se puede ver la predominancia de puntos pertenecientes al “Período 1”, de 7 contra 1, para ser más precisos. Esto entonces concluye que la nueva muestra pertenece al período marzo-abril-mayo.

Otra observación fue la de tomar una muestra distinta, por ejemplo, la que que contiene los valores: temperatura 25.5°C y humedad 50% y se repitió el mismo experimento, obteniendo la siguiente Tabla (Tabla 2):

Tabla 2. Tabla de datos con K = 8

Temperatura	Humedad	Período
25.44	50.00	0
25.44	55.00	1
25.50	50.20	0
25.56	50.20	0
25.37	50.20	0
25.50	50.30	1
25.56	49.70	0
25.56	49.70	0

En base a la Tabla 2, se grafican los valores de temperatura y humedad pertenecientes a los dos períodos, y la nueva muestra, obteniendo el siguiente gráfico (Fig.6):

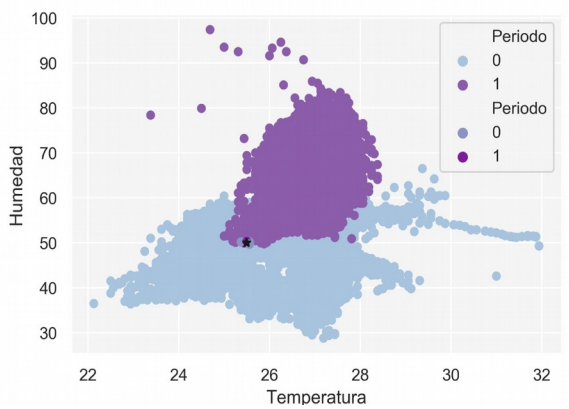

Figura 6. Nube de datos con la nueva muestra

Para un mayor detalle de la nueva muestra y sus vecinos, se propone la siguiente figura:

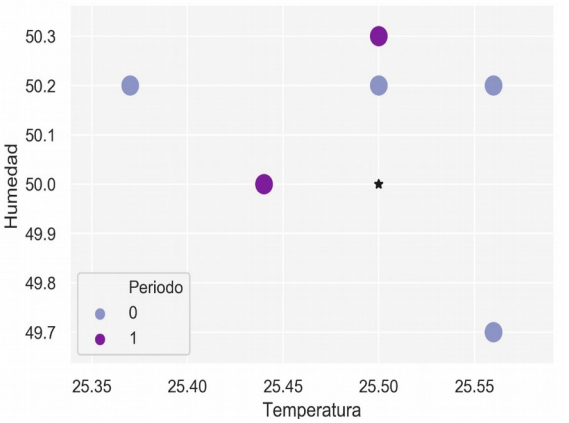

Figura 7. Nueva muestra con K vecinos

En la Fig.7 se puede ver la predominancia de puntos pertenecientes al "Período 0", de 6 contra 2, para ser más precisos. Esto entonces concluye que la nueva muestra pertenece al período octubre, noviembre, diciembre y febrero. Notar que en la Fig. 7 existen datos de temperatura y humedad que se solapan, por ejemplo, el dato T = 25,56 °C y H = 49,70%. En la parte final del ítem 4.3 se propone una posible solución a este inconveniente.

4.2 Clasificación mediante tres grupos de temperatura y humedad.

El próximo experimento consiste en tomar el mismo dataset utilizado hasta ahora (datos de temperatura y humedad en el período octubre 2018 hasta mayo 2019) pero separandolo en tres grupos para evaluar su eficiencia y poder compararla con el caso de dos grupos. Estos grupos son: "Período 0" que toma los meses octubre y noviembre, "Período 1" que abarca los meses diciembre, febrero y marzo y por último, "Período 2" que toma los meses de abril y mayo. Esta distribución se puede observar en la Fig. 8.

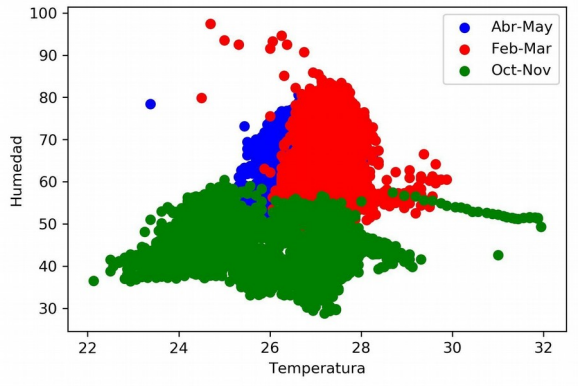

Figura 8. Distribución de Datos para los tres Períodos

En verde se observan los datos del intervalo octubre-noviembre en color verde ("Período 0"), los datos pertenecientes a diciembre, febrero-marzo en color rojo ("Período 1") y en color azul los datos del intervalo abril-mayo ("Período 2").

Es importante notar que hay datos que se solapan unos con otros, dado que hay valores que se parecen según las épocas, por esta razón para visualizar mejor los datos, a continuación se muestra otra figura (Fig. 9) perteneciente al mismo intervalo de meses pero poniendo por encima a las muestras que en la Fig. 8 quedaron por debajo.

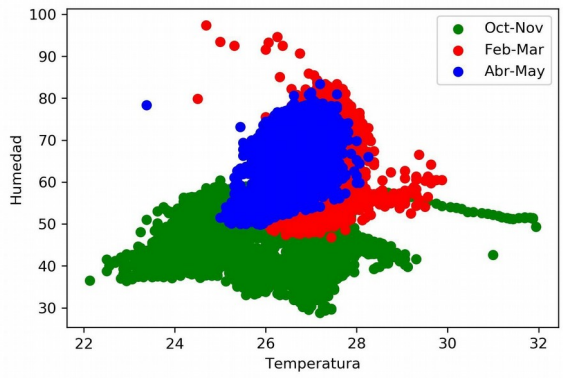

Figura 9. Distribución de Datos para los tres Períodos

Al igual que en el caso anterior, el objetivo es obtener el score del algoritmo para el caso de separar a los datos en tres grupos. Para esto, se grafica el score en función del número de vecinos (**K**), la cual se puede ver en la Fig. 10:

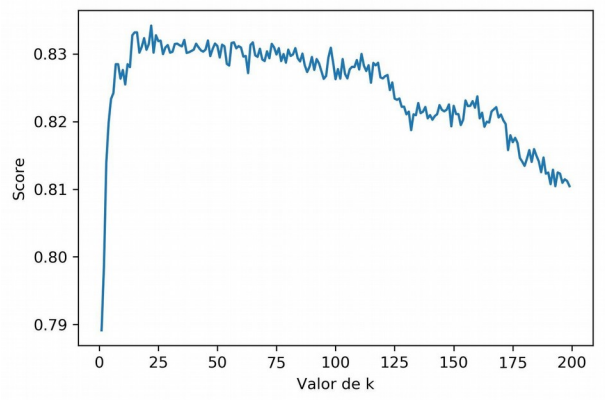

Figura 10. Score vs K

En la Fig.10 se observa que con un valor de K=22 se obtiene una eficiencia de aproximadamente 0,837.

Habiendo encontrado el valor de **K** que mejor se ajusta, el siguiente paso es poner a prueba este modelo con muestras aleatorias de temperatura y humedad. Los colores de los grupos se mantienen como se viene explicando y, para la muestra aleatoria se utiliza el color negro, de forma que se la pueda diferenciar (esta muestra fue separada del dataset, es decir, no forma parte del mismo). La muestra tomada tiene los valores 26°C y 49% de humedad. Esto se observa en la Fig. 11.

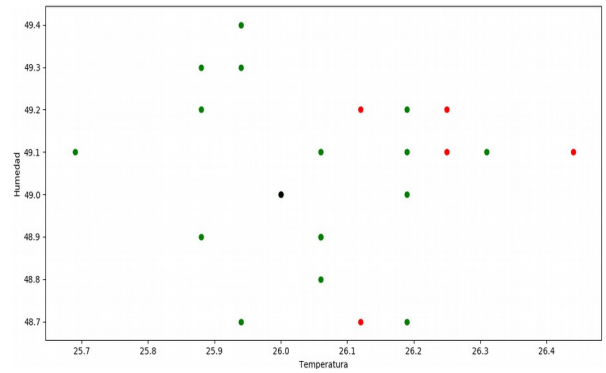

Figura 11. Nueva muestra con K vecinos

En la Fig.11 se puede ver la predominancia de puntos pertenecientes al "Período 0". Esto entonces concluye que la nueva muestra pertenece al "Período 0" (octubre, noviembre).

4.3 Clasificación mediante siete grupos de temperatura y humedad

El último experimento consiste en tomar a cada mes del dataset como un grupo por separado y analizar cuál es el valor óptimo de **K** que mejor se ajusta a la clasificación. Para esto se utilizará el mismo dataset y se divide a cada período de la siguiente forma: "Período 0", octubre; "Período 1", noviembre; "Período 2", diciembre; "Período 3", Febrero; "Período 4", Marzo; "Período 5", Abril; "Período 6", Mayo. Esto se observa en la siguiente imagen (Fig.12).

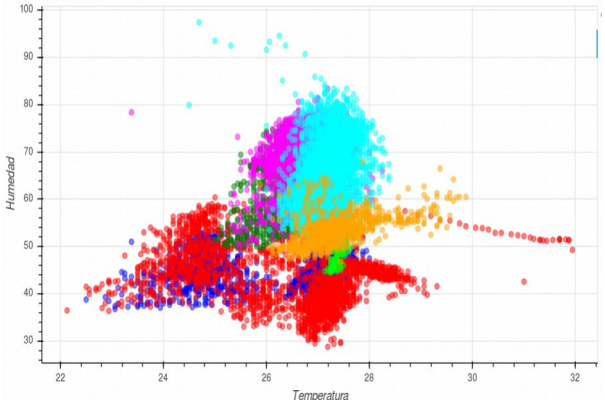

Figura 12. Distribución de Datos para los siete Períodos

En la Fig.12 se tiene la siguiente distribución: lima, octubre; rojo, noviembre; azul, diciembre; naranja, febrero; cyan, marzo; magenta, abril; verde, mayo.

Al igual que en el caso anterior, el objetivo es obtener la eficiencia del algoritmo para el caso de separar a los datos en siete grupos. Para esto, se grafica el score en función del número de vecinos (**K**), lo cual se puede ver en la Fig. 13

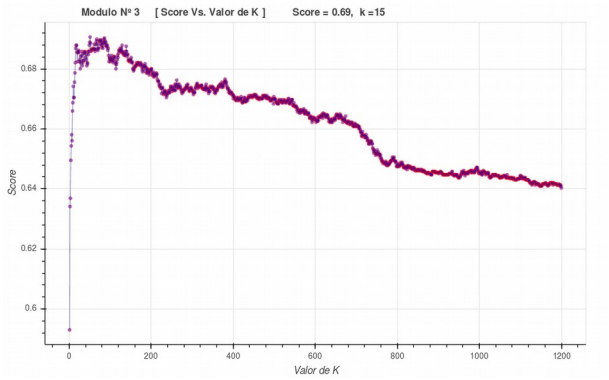

Figura 13. Score vs K

En la Fig.13 se observa que con un valor de K= 15 se obtiene un score de 0,69, lo cual indicaría que es el mejor **K** para predecir.

No resulta necesario tomar una muestra aleatoria para comprobar la efectividad del algoritmo en la clasificación porque es notorio que, en comparación con los casos anteriores, el score obtenido es menor. Este criterio se toma en base a lo comentado anteriormente sobre el objetivo de este trabajo.

Dada esta baja puntuación para esta última experiencia, se procedió a agregar la presión como una dimensión más para los datos con el objetivo de comprobar si los datos se pueden diferenciar con mayor claridad, es decir, en vez de usar un plano para los datos (humedad – temperatura), se utiliza un tercer eje de forma que los datos queden diferenciados. Al hacer esto

y ejecutar nuevamente el algoritmo, se tiene el siguiente gráfico de Score vs K (Fig.14)

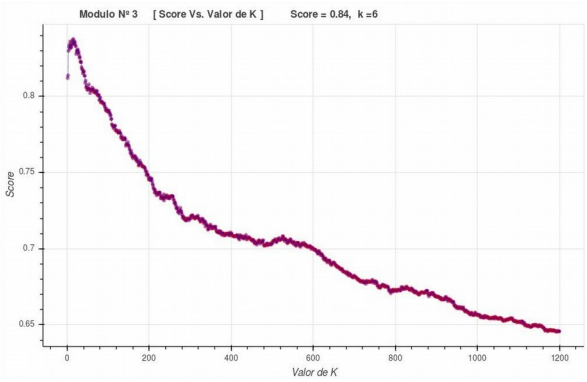

Figura 14. Score vs K

Al observar la Fig.14, se puede ver que el efecto de agregar una tercera variable a los datos, mejora la puntuación del algoritmo.

A continuación se comentan las conclusiones obtenidas en base a las experiencias realizadas.

5. Conclusiones

En base a las experiencias realizadas, se pudo observar que en la primera, la cual consiste en tomar dos grupos de valores, el score es de 0,975 tomando un K = 8, mientras que, para el caso de tomar tres grupos de muestras, el score es de 0,837 tomando un K = 22 y para la última experiencia el score es de 0,69 con valor de K = 15. Esto demuestra que separar los datos en una mayor cantidad de grupos deviene en una menor puntuación (score) del algoritmo. Una posible explicación se puede dar desde el punto de vista de la cercanía que tienen los valores que componen a los grupos, con respecto a la nueva muestra, es decir, en el caso de haber menos grupos, los valores pertenecientes a estos que se encuentren cercanos a la nueva muestra pueden ser de uno u otro grupo.

Por otro lado, también es posible ver que al aumentar la cantidad de grupos, los datos comienzan a tener similitud entre sí (Fig.12), lo cual conduce a pensar que con solo dos parámetros (temperatura y humedad) no resulta suficiente para clasificar una clase de otra, es decir, resulta necesario agregar una o más variables. En el algoritmo de KNN utilizado, esto se puede lograr agregando una variable más de forma que el plano mostrado en las distintas figuras de humedad-temperatura, se transforme en un volumen. Esta variable a sumar puede ser la presión del ambiente, con lo cual el conjunto de datos quedaría conformado por un volumen cuyas tres dimensiones sean: presión-temperatura-humedad. De esta manera, agregando otra dimensión a los datos se obtendrá una mayor diferencia entre los **K** vecinos lo cual deviene en una mejora en cuanto a la puntuación (Fig.14). En otras palabras, cuanto mayor sea el número de variables que caractericen a los datos, el algoritmo de KNN obtiene mejores resultados [10].

De todo esto, se puede concluir que, para el caso del presente trabajo donde se pretende buscar cuál es la mejor división de los grupos que permita obtener la máxima puntuación del algoritmo con dos dimensiones (temperatura y humedad), resulta útil entonces separar al conjunto de datos en dos grupos. Esto permitirá que, al agregar ventilación al Datacenter, se pueda conocer en qué períodos de meses conviene que la misma aumente.

Teniendo como base este análisis de clasificación, resulta importante como trabajo a futuro, poder integrar un sistema de ventilación que se active según la distribución de los datos separados en dos grupos. Esto permitirá obtener dos cosas: por un lado, disminuir la temperatura del ambiente y, por otro, tener un nuevo conjunto de datos que resulte interesante para analizar y aplicar este algoritmo. Todo esto facilitará generar una dinámica en el estudio de los datos para automatizar la ventilación.

En este trabajo se analizó el algoritmo de KNN que es una técnica de aprendizaje supervisado. Por otro lado, existen algoritmos de clustering con técnicas de aprendizaje no supervisado, como por ejemplo, el algoritmo K-Means, el cual puede agrupar datos similares en base a las características o propiedades que poseen. Resulta interesante, para futuros trabajos, analizar el dataset utilizado bajo la aplicación de este algoritmo. De esta manera, se podrán encontrar diferentes formas de agrupar los datos, en vez de agrupar por meses como se hizo en este trabajo.

Agradecimientos

Este Trabajo se realizó en base al Proyecto de Investigación PROAPI 2017, desde la Carrera de Ingeniería Informática de la Universidad Nacional de Avellaneda (UNDAV), bajo el título de “Mantenimiento de parámetros del ambiente del Laboratorio de Redes y Sistemas de Computación mediante protocolos de IoT”. El Datacenter bajo estudio representa al Laboratorio de Redes de la Carrera Ingeniería Informática de la UNDAV.

Referencias

[1] Abhinav Rathor., Manasi Gyanchandani.: A Review at Machine Learning Algorithms Targeting Big Data Challenges. 2017. ISBN: 978-1-5386-2361-9

[2] Jeff Dyck.: Machine Learning for Engineering. Solido Design Automation. 2018. 978-1-5090-0602-1

[3] Truúcă, M., Soran, A., "The benefits of data center temperature monitoring", Datacenter Department. 2015 INCDTIM. ISBN: 978-6-0673-7040-9

[4] Truúcă, M., Soran, A., Abrudean, M., "Environmental Parameters Control in Datacenter", Datacenter Department. INCDTIM. 2014. ISBN: 978-1-4799-3732-5

[5] Gil Costa, V., Errecalde, Marcelo., Taranilla, María Teresa., "Un Modelo de Detección de Anomalías en una LAN usando K-NN y Técnicas de Computación de Alto Desempeño". 2017. Universidad Nacional de San Luis. VI Workshop de Agentes y Sistemas Inteligentes

[6] Barrionuevo, M., Lopresti, Mariela., Miranda, Natalia., Piccoli, Fabiana., "Learning to detect spam messages", LIDIC. Universidad Nacional de San Luis. XXIII Congreso Argentino de Ciencias de la Computación. 2005. ISBN: 978-950-34-1539-9

[7] Yevhenii Udovychenko, A., Chaikovsky, I., “Ischemic Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps”. Physical and Biomedical Electronics Department National Technical University of Ukraine. 2015 IEEE 35th International Conference on Electronics and Nanotechnology. ISBN: 978-1-4673-6534-5

[8] Eyupoglu, C., "Implementation of Color Face Recognition Using PCA and k-NN Classifier", Department of Computer Engineering Istanbul Commerce University. 2016. ISBN: 978-1-5090-0445-4

[9] Lantz, Brett, "Machine Learning with R", Second Edition. 2015. Packt Publishing. ISBN: 978-1-78439-390-8.

[10] Trunk, G., "A Problem of Dimensionality: A Simple Example", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 1979. ISSN: 1939-3539

[11] Saha, S., Majumdar, A., "Data Centre Temperature Monitoring with ESP8266 Based Wireless Sensor Network and Cloud Based Dashboard with Real Time Alert System", Dept. of Electronics and Communication Engineering, RCC Institute of Information., Former Head of State e-Governance Mission Team (SeMT), Tripura National Institute for Smart Government (NISG), India. ISBN: 978-1-5090-4724-6