Agrupación de Datos de Incendios por Clasificación No Supervisada

Prieto, Julian

10/11/2022

Introducción 1

Análisis Exploratorio 2

Conclusiones del análisis exploratorio 3

Clasificación No Supervisada 4

Conclusiones de la Clasificación No Supervisada 5

Futuros Trabajos 6

Apéndices Técnicos 7

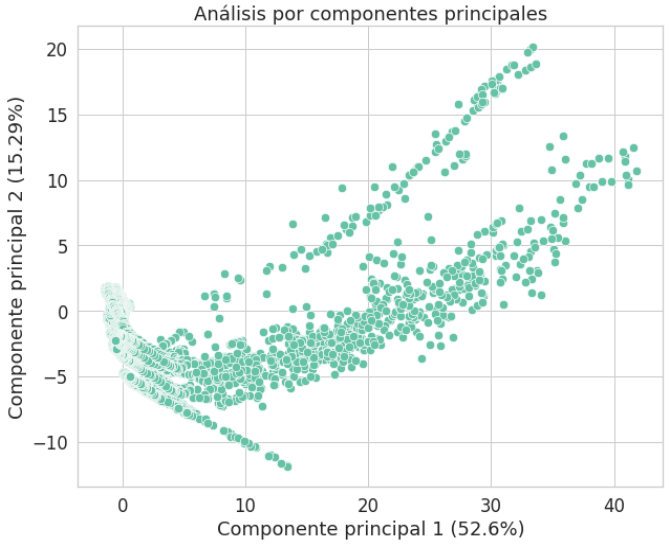
Referencias 8

Introducción

En este informe se detalla la creación de un modelo de agrupamiento con un set de datos con más de 60.000 lecturas de ambientes y fuentes de humo. En cada lectura, se registra la temperatura y la humedad ambiente, los compuestos orgánicos volátiles (medido en partes por mil millones), concentración de CO2, cantidades de hidrógeno molecular bruto y de gas etanol, presión del aire (hPa), tamaño de las partículas y su concentración según su tamaño, el momento exacto en el que se realizó la lectura, y un valor que simplemente representa el conteo de los registros.
Todos estos datos fueron tomados en diferentes entornos así como en interiores y exteriores normales, zonas con fuego de leña o con chimenea de gas, parrillas exteriores de leña, carbón y gas, y exteriores con una alta humedad.

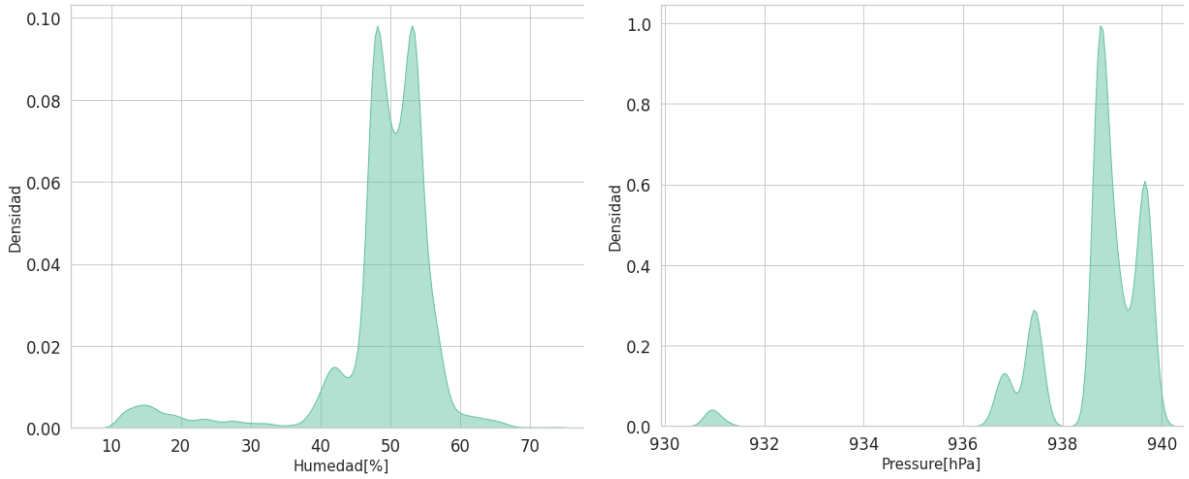
Análisis Exploratorio

Primeramente, se eliminaron las columnas de UTC (tiempo de la prueba) y CNT (conteo de registros) ya que estas no tienen importancia al momento de analizar los datos.
Luego, con la intención de poder graficar información con múltiples dimensiones (+2), se realiza un análisis por componentes principales[1], el cual intenta mostrar con la mayor fidelidad posible los datos del conjunto calculando componentes principales y ordenándolas por mayor varianza, para así poder ver las variables más influyentes en los posibles agrupamientos. Este método permite graficar de la mejor manera posible los datos.
En este caso, el análisis por componentes principales y su respectivo gráfico, logra representar un 68% de la información del set de datos, y como se ve abajo, en este se consiguen observar unas ciertas agrupaciones.



Siguiendo con este análisis, es posible detectar qué variables son las más influyentes en cada componente principal. Observando los datos del análisis, se puede ver que en la primera componente principal, las variables más influyentes para la agrupación son tanto el tamaño de las partículas como también su concentración en el ambiente. Y con respecto a la segunda componente principal, las dos variables más importantes son la humedad del ambiente y la presión del aire.

Teniendo esto en cuenta, es posible graficar las densidades de dos de estas variables, lo que nos permite visualizar de manera sencilla de qué forma se distribuyen los valores de ambas:



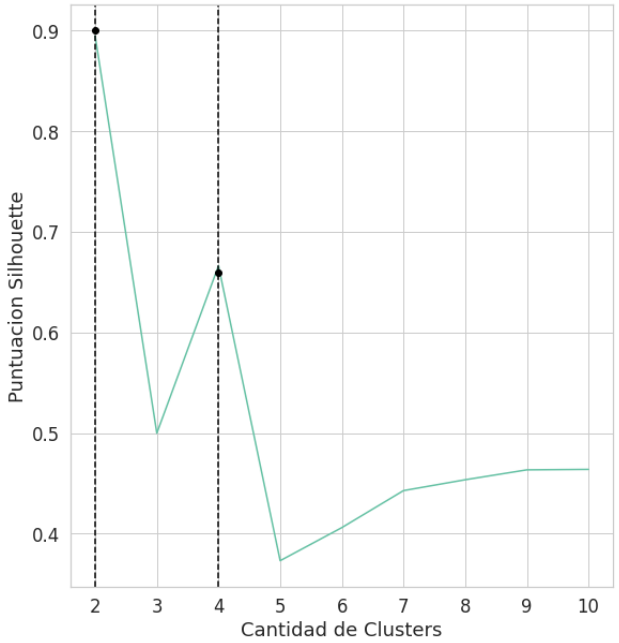
Observando, se puede ver como el porcentaje de humedad del ambiente mayoritariamente varía entre el 45% y el 55%, y los valores de presión entre 938 hPa y 940 hPa.

Conclusiones del análisis exploratorio

El gráfico obtenido gracias al análisis por componentes principales nos permite detectar una cierta agrupación de los datos. Además, es posible reconocer las variables más influyentes a la hora de agrupar, las cuales, son las mediciones de las partículas de materia en el aire como también las mediciones de humedad del ambiente y la presión del aire.

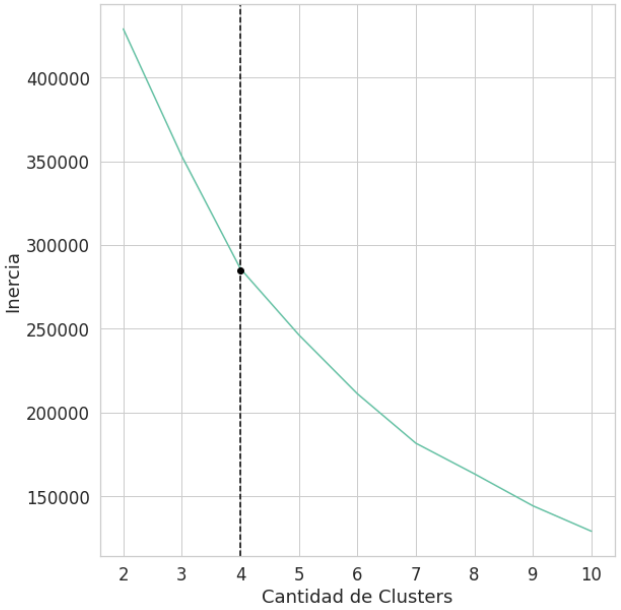
Clasificación No Supervisada

Con la creación de un modelo de clasificación no supervisada, se logra agrupar la información en una cantidad dada de grupos o ***clusters*** según ciertas características elegidas por el modelo.



Debido a que la cantidad de clusters es dependiente a lo que se busque con el análisis, la sección comienza presentando un gráfico donde se muestra la **puntuación de Silhouette**[2] dada según el número de clusters.

En él se marcan los 2 puntos elegidos de cantidad de clusters que se graficaron en el trabajo. Esto se determina observando los máximos locales del gráfico, los cuales según este método, son los más sugeridos para la utilización.

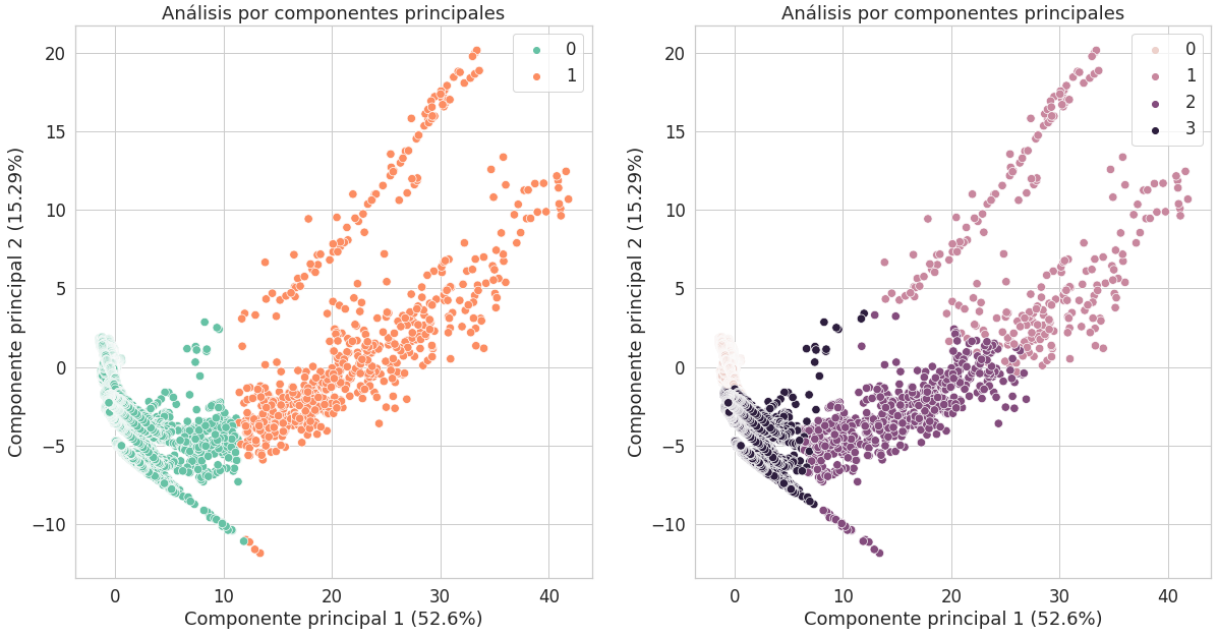


Por otro lado, se analizó también usando el **método Elbow**[3], el cual nos deja con que la cantidad óptima de clusters que se deben utilizar para este conjunto es de 4.

Por los resultados obtenidos por el método Silhouette y el método Elbow, para este trabajo se agruparán los datos en 2 y 4 clusters.

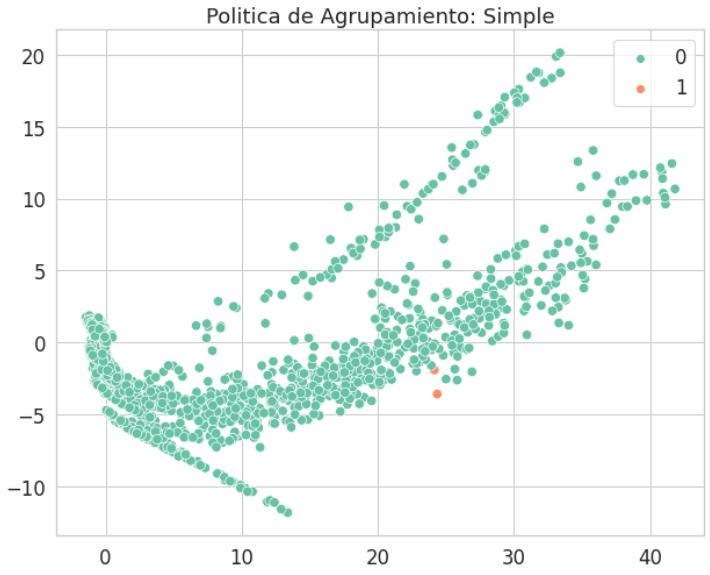
Con el objetivo de luego poder comparar todos los gráficos obtenidos por los métodos de clustering, se presentarán los casos más relevantes de cada método para que luego sea más sencillo elegir el más adecuado.

Para empezar, se realiza el clustering con el método **K-Means**[4]. Agrupando los datos en 2 y en 4 clusters, se consigue los siguientes dos gráficos:

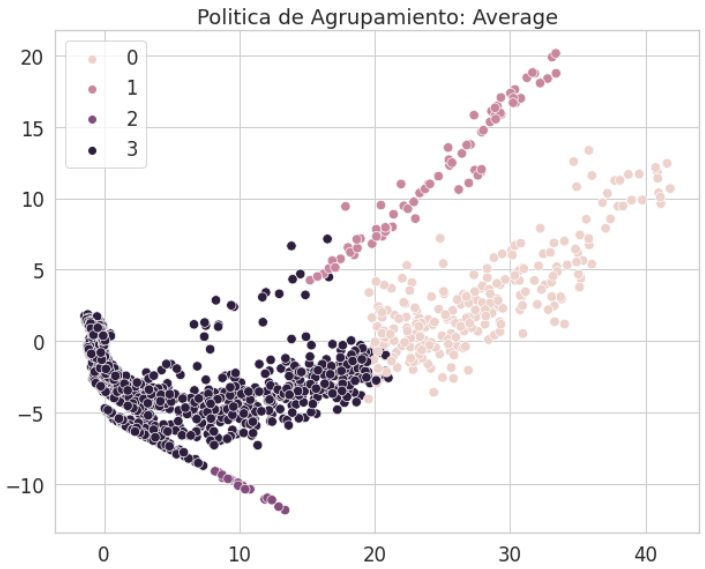


En el primer gráfico, se puede observar una agrupación bastante marcada que según los datos está dada por el tamaño y concentración de las partículas en el aire. En el segundo gráfico, se dan resultados similares, ya que las agrupaciones también están dadas por las características de las partículas.

Luego, se continúa utilizando el método de clustering **Aglomerativo**[5] con 2 y 4 clusters, pero en este caso, se varía en la política de agrupamiento utilizada. Esto quiere decir que se modifica que distancia es usada al formar los clusters.Los gráfico más interesantes que difieren de los previamente obtenidos son los siguientes:

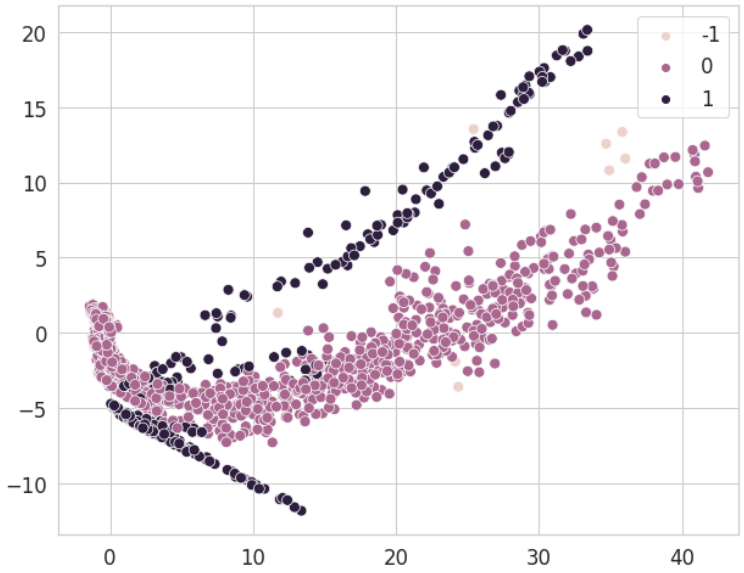


En este caso, se agrupó en dos clusters usando la política de agrupamiento “Simple”. Con lo que se llegó a agrupar la gran mayoría de los datos en un solo cluster, dejando 2 datos dentro de otro grupo distinto identificandolos como valores atípicos.



En este otro, se usó 4 clusters y la política de agrupamiento "Average". Como se ve, se obtuvo un agrupamiento distinto al obtenido con K-Means, separando de diferentes maneras. Según las medias de cada clusters, se identifica que las diferencias entre ellos residen mayormente en la cantidad total de compuestos orgánicos volátiles en el ambiente

Por último, se utilizó el método DBSCAN[6]. Este difiere bastante de los nombrados anteriormente, ya que la cantidad de cluster es elegida por el método y no por nosotros. Luego de varias pruebas, se obtuvo un gráfico muy diferente al resto, que podría ser de gran ayuda:



En él, se identifican dos clusters (0 y 1), y un grupo denominado como -1 donde se colocan los supuestos outliers o valores atípicos. Este gráfico resulta interesante porque da cierta idea de profundidad, dejando notar como el cluster violeta está "por encima" o "más cerca" que el otro grupo.
Lo curioso de esta agrupación, es que analizando las medias de cada variable en los clusters 0 y 1, se puede observar como estas son muy similares casi idénticas. Por este motivo es difícil determinar el motivo por el cual estos se diferencian.

Conclusiones de la Clasificación No Supervisada

Luego del análisis de los 3 métodos utilizados y con la observación de todos los gráficos obtenidos, se puede decir que el método más acertado en el DBSCAN. Esto se puede saber fácilmente al comparar su gráfico con el gráfico realizado con la clasificación original del set de datos[7].

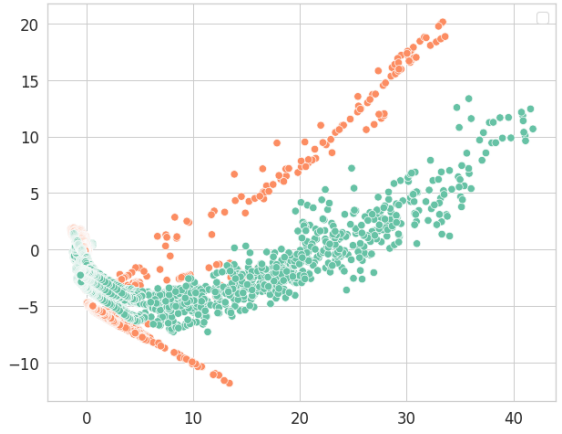
Futuros Trabajos

Para futuros trabajos se podría incluir en el análisis del clustering aglomerativo otras formas de calcular la distancia entre datos que no sea la usada por defecto (Euclidiana) como por ejemplo: “manhattan”, “cosine”, o “precomputed”.

Apéndices Técnicos

[1] **Análisis por Componentes Principales**: El análisis por componentes principales intenta graficar con la mayor representatividad posible la información dada en múltiples variables, separando y ordenando sus componentes principales según su varianza y mostrando únicamente las dos primeras. Además, nos brinda información acerca de qué variables influyen más en cada componente principal.
[2] **Silhouette Score**: Es un método de interpretación y validación dentro del análisis de grupos. Determina que tan bien se ha clasificado una muestra en su respectivo cluster calculando la distancia entre cada agrupación.
[3] **Método Elbow**: Este método utiliza la distancia media de las observaciones a su centroide (centro del cluster). Cuanto más grande es el número de clusters, la varianza intra-cluster tiende a disminuir, y cuando menor sea esta, mejor. Esto debido a que significa que los clusters son más compactos.
[4] **K-Means**: Es un algoritmo de clustering o agrupamiento de información que intenta separar las muestras en n grupos de igual varianza, minimizando lo llamado “*inercia*” (Criterio de suma de cuadrados dentro del grupo). Este método se basa en la distancia Euclidiana para determinar qué muestras entran en los diferentes clusters.
[5] **Aglomerativo**: Algoritmo del mismo tipo que K-Means el cual, a diferencia de este, permite elegir la forma en que se mide la distancia entre cada dato y la métrica usada para calcular dicha distancia. En este método cada observación se asigna a su propio cluster, para luego calcular la similitud o distancia entre cada uno de los clusters y “juntar” los dos que sean más parecidos. Este método es bueno para identificar clusters de poco tamaño.
[6] **DBSCAN**: Algoritmo de clustering de densidad que, según dos valores dados por el usuario (radio y puntos mínimos), decide en cuántos clusters separar la información dada. El radio determina la longitud específica que formará una cierta área, y los puntos mínimos hace

referencia a la cantidad mínima de datos que debe haber en cada área para que esta se transforme en un cluster. Este algoritmo es eficaz en la identificación de grupos en un contexto espacial, y no se ve afectado por el “ruido” de los datos. DBSCAN trabaja con la idea de que si un dato en particular pertenece a un clusters, debería estar cerca de un montón de otros datos en ese mismo cluster.
[7]El gráfico obtenido al mostrar la información con su clasificación original es (Obtenido del tp1):



Referencias

Los links utilizados fueron los siguientes:

2.3. Clustering — scikit-learn 1.1.3 documentation
Segmentación utilizando K-means en Python (machinelearningparatodos.com)
sklearn.metrics.silhouette_score — scikit-learn 1.1.3 documentation
sklearn.cluster.AgglomerativeClustering — scikit-learn 1.1.3 documentation
Clustering Jerárquico en R (rstudio-pubs-static.s3.amazonaws.com)
sklearn.cluster.DBSCAN — scikit-learn 1.1.3 documentation
Introducción al Clustering DBSCAN - StatDeveloper