Universidad del Valle de Guatemala
Inteligencia Artificial - CC3045
Catedrático: Luis Alberto Suriano Saravia

Integrantes:
JULIO ROBERTO HERRERA SABAN
JUAN PABLO PINEDA MELENDEZ
DIEGO ANDRES CRESPO MEALLA

Laboratorio 3 - K-Means y Mixture-Models

Link al repositorio:
https://github.com/jurhs2000/ai-lab3

Link al video:
https://youtu.be/ygW5peVLq0I

Análisis

Primero damos un vistazo a los datos que contiene el *dataset* a trabajar.

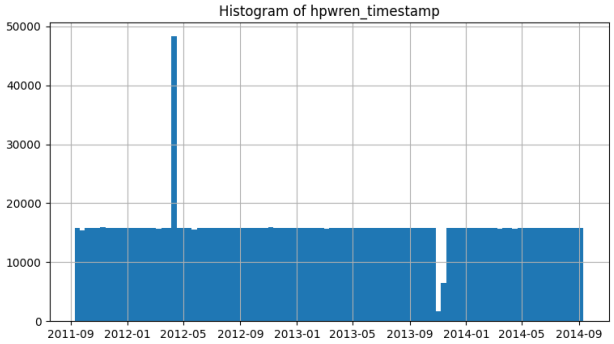
	rowID	hpwren_timestamp	air_pressure	air_temp	avg_wind_direction	avg_wind_speed	max_wind_direction	max_wind_speed	min_wind_direction	min_wind_speed	rain_accumulation	rain_duration	relative_humidity
0	0	2011-09-10 00:00:49	912.3	64.76	97.0	1.2	106.0	1.6	85.0	1.0	NaN	NaN	60.5
1	1	2011-09-10 00:01:49	912.3	63.86	161.0	0.8	215.0	1.5	43.0	0.2	0.0	0.0	39.9
2	2	2011-09-10 00:02:49	912.3	64.22	77.0	0.7	143.0	1.2	324.0	0.3	0.0	0.0	43.0
3	3	2011-09-10 00:03:49	912.3	64.40	89.0	1.2	112.0	1.6	12.0	0.7	0.0	0.0	49.5
4	4	2011-09-10 00:04:49	912.3	64.40	185.0	0.4	260.0	1.0	100.0	0.1	0.0	0.0	58.8

Así como a las medidas estadísticas principales de estas variables.

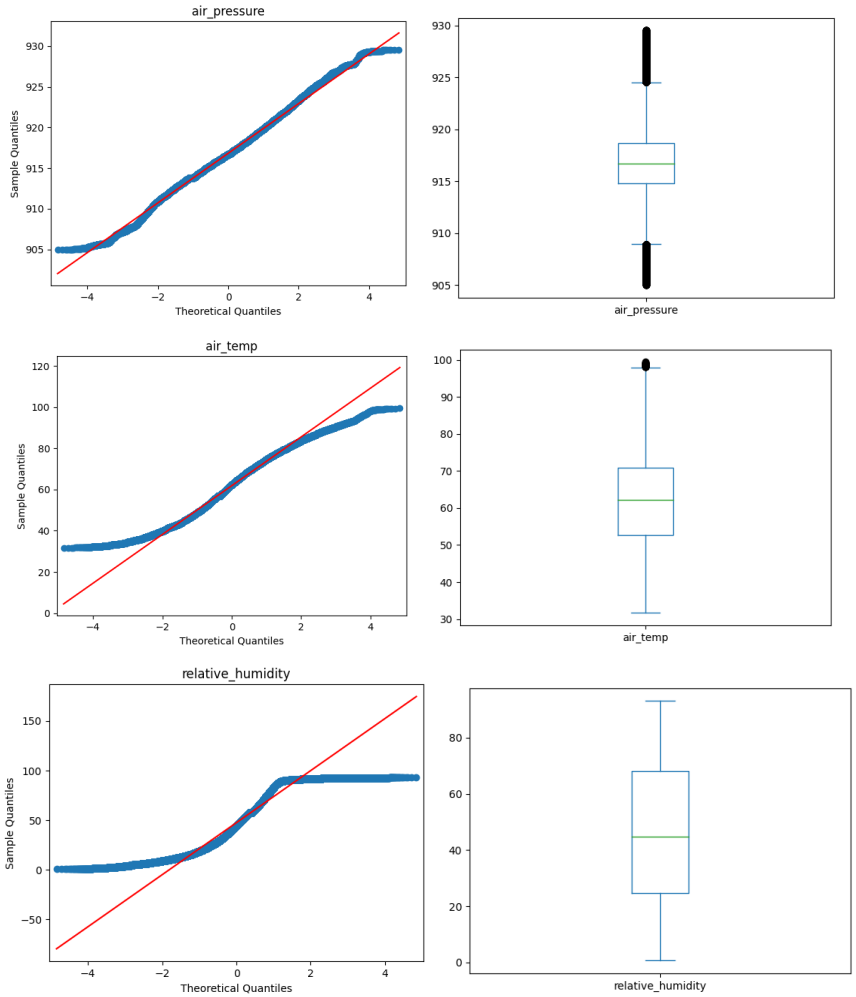
	rowID	hpwren_timestamp	air_pressure	air_temp	avg_wind_direction	avg_wind_speed	...	max_wind_speed	min_wind_direction	min_wind_speed	rain_accumulation	rain_duration	relative_humidity
count	1.587257e+06	1587257	1.587257e+06	1.587257e+06	1.586824e+06	1.586824e+06	...	1.586824e+06	1.586824e+06	1.586824e+06	1.587256e+06	1.587256e+06	1.587257e+06
unique	NaN	1554478	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	2012-04-05 15:12:02	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	32523	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
mean	7.936280e+05	NaN	9.168301e+02	6.185144e+01	1.619654e+02	2.774272e+00	...	3.399813e+00	1.668264e+02	2.133130e+00	1.854836e-03	5.361460e-01	4.760837e+01
std	4.582018e+05	NaN	3.051593e+00	1.183362e+01	9.520812e+01	2.060758e+00	...	2.423167e+00	9.746275e+01	1.745345e+00	9.609716e-01	8.114766e+01	2.621454e+01
min	0.000000e+00	NaN	9.050000e+02	3.164000e+01	0.000000e+00	0.000000e+00	...	1.000000e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7.000000e-01
25%	3.968140e+05	NaN	9.148000e+02	5.270000e+01	6.200000e+01	1.300000e+00	...	1.600000e+00	7.700000e+01	8.000000e-01	0.000000e+00	0.000000e+00	2.470000e+01
50%	7.936280e+05	NaN	9.167000e+02	6.224000e+01	1.820000e+02	2.200000e+00	...	2.700000e+00	1.800000e+02	1.600000e+00	0.000000e+00	0.000000e+00	4.470000e+01
75%	1.190442e+06	NaN	9.187000e+02	7.088000e+01	2.170000e+02	3.800000e+00	...	4.600000e+00	2.120000e+02	3.000000e+00	0.000000e+00	0.000000e+00	6.800000e+01
max	1.587256e+06	NaN	9.295000e+02	9.950000e+01	3.590000e+02	3.230000e+01	...	3.600000e+01	3.590000e+02	3.200000e+01	6.550100e+02	6.330500e+04	9.300000e+01

El siguiente paso y previo al *clustering* es analizar cada una de las variables, para conocer su distribución y el rango de sus valores, para saber si normalizar los datos o no, lo cual ayuda en gran medida a la formación de los *clusters*, esto lo haremos por medio de las gráficas QQ. También buscamos eliminar los *outliers* que pueden afectar a la creación de clusters, estos los veremos con las gráficas de caja y bigotes. El *clustering* trabaja a partir de distancias, así que es conveniente realizar estas operaciones.

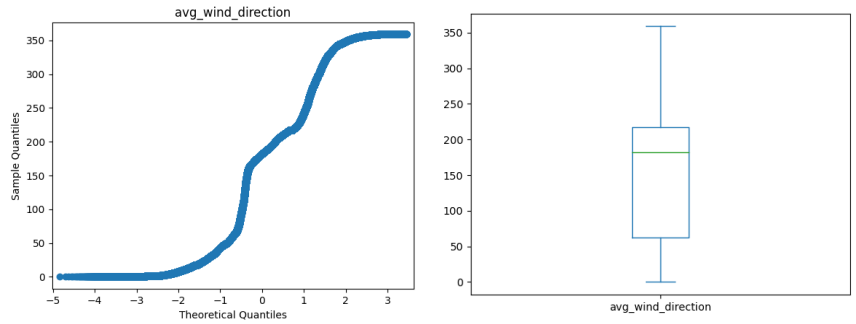
El caso de la variable rowID se descarta ya que simplemente es el identificador.

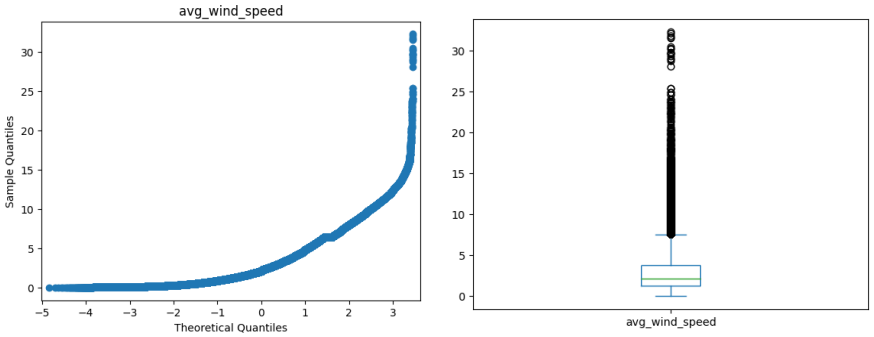


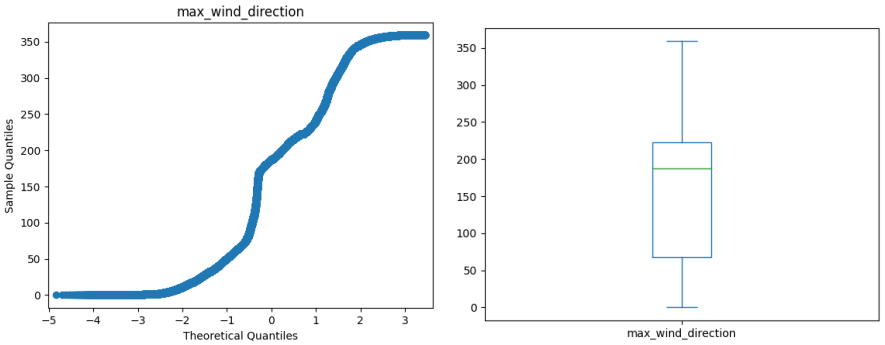
Para la variable hpwren_timestamp se sigue una distribución uniforme a pesar de algunas fechas en las que existen más o menos registros, sin embargo esta no es una variable en la que nos afecte esto.

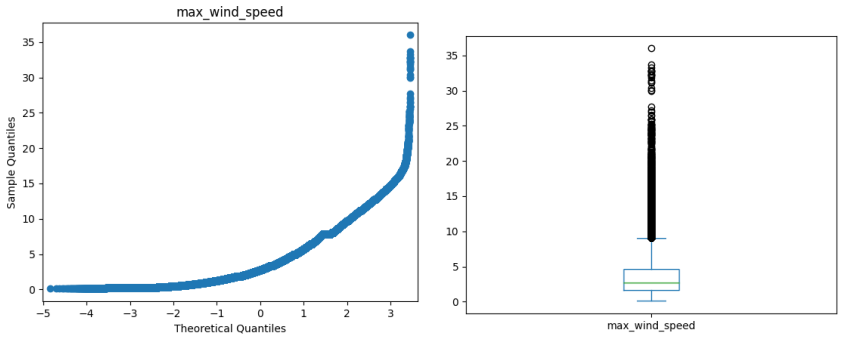


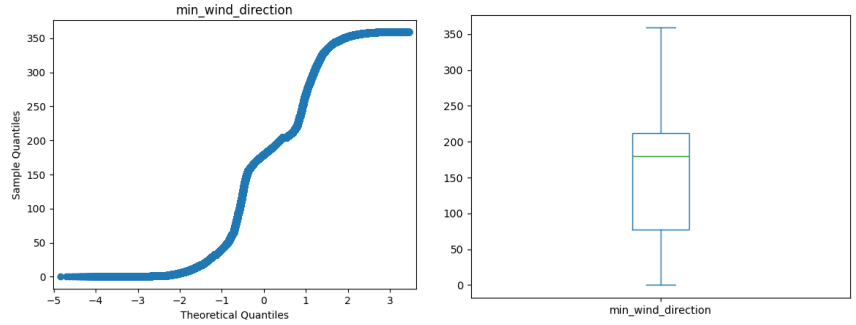
Las variables air_pressure y air_temp y relative_humidity muestran una distribución normal y no se observan datos atípicos que puedan afectar.

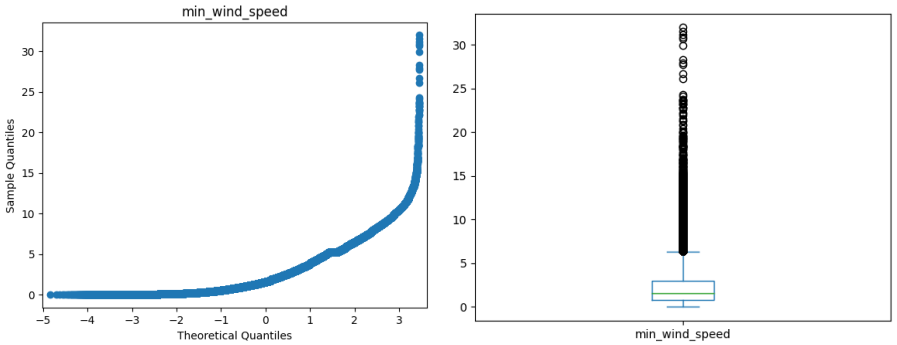

avg_wind_direction
Sample Quantiles
Theoretical Quantiles
avg_wind_direction


avg_wind_speed
Sample Quantiles
Theoretical Quantiles
avg_wind_speed

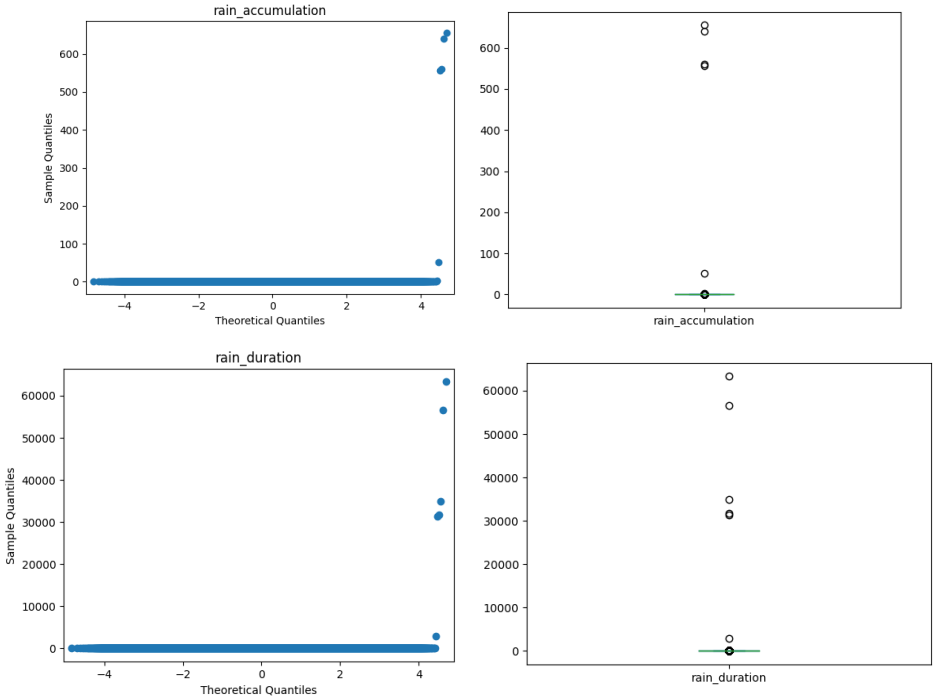

max_wind_direction
Sample Quantiles
Theoretical Quantiles
max_wind_direction


max_wind_speed
Sample Quantiles
Theoretical Quantiles
max_wind_speed

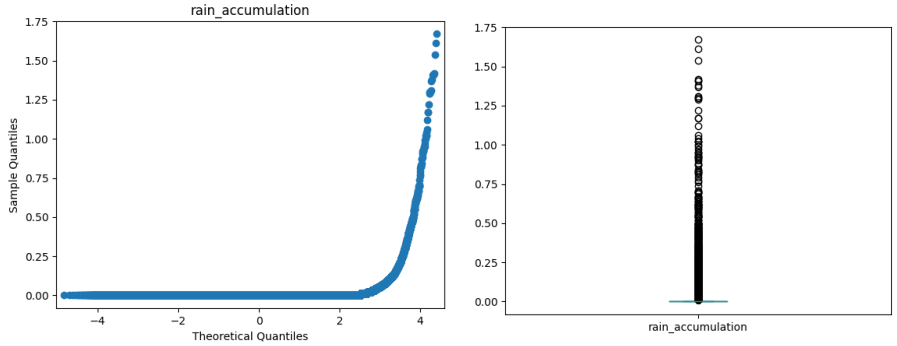

min_wind_direction
Sample Quantiles
Theoretical Quantiles
min_wind_direction

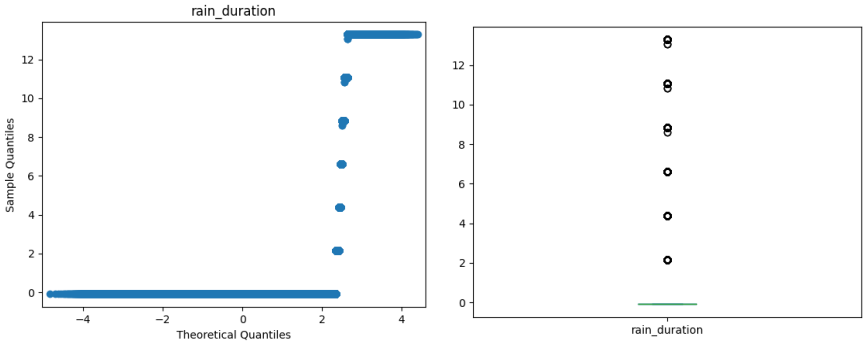


Las variables avg_wind_direction, avg_wind_speed, max_wind_direction, max_wind_speed, min_wind_direction y min_wind_speed no muestran una distribución normal, sin embargo, los cuantiles sí muestran una distribución uniforme y en cuanto a los datos atípicos, sí existen datos atípicos pero según lo que representa la variable, no los consideraremos como *outliers*.



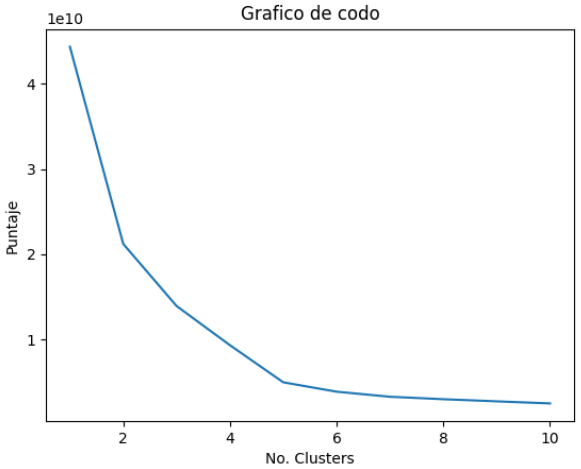
Por último las variables rain_accumulation y rain_duartion sí muestra valores atípicos que consideramos outliers, por lo que procedemos a removerlos. También se observa que la variable rain_duartion representa valores con rangos muy grandes en comparación con las otras, por lo que procedemos a normalizar sus valores. Estos son los resultados con dichas operaciones de limpieza aplicadas.





Determinación de cantidad de clusters

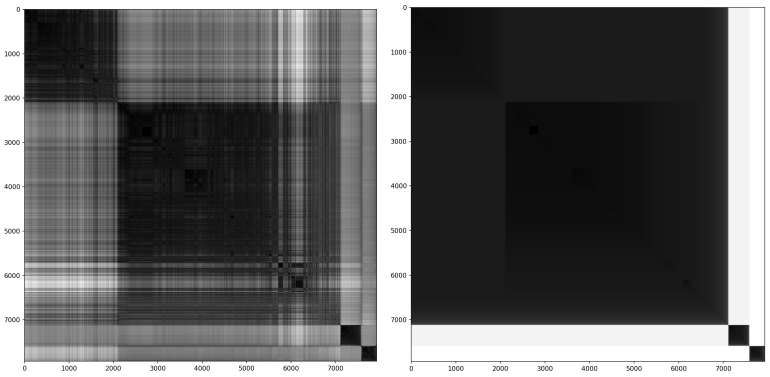
Como principal método para determinar la cantidad óptima de *clusters* utilizamos el método de codo, para ello seleccionamos las variables cuantitativas que en este caso son todas excepto hpwren_timestamp y también descartamos rowID, por lo que no es necesario convertir ninguna a categórica.



La gráfica de codo nos indica que la cantidad óptima de *clusters* es 5.

También podemos aplicar otros métodos, en este caso se busca reafirmar lo ya indicado por la gráfica de codo con el método Hopkins y las gráficas VAT e iVAT, aunque esto a menor medida ya que su procesamiento toma mucho tiempo así que solo se usa una muy pequeña porción de los datos, 25% para obtener el puntaje de Hopkins y 0.05% para las gráficas VAT, las cuales al representar un porcentaje tan pequeño, no muestran dan indicios claros sobre la cantidad óptima de *clusters*.

Hopkins score: `0.06161414333517891`. Esto nos indica que nuestra *data* no está uniformemente distribuida. Un valor por encima a 0.3 se considera alto, indica que la *data* está uniformemente distribuida y que el *clustering* no puede ser muy útil para el *dataset*.

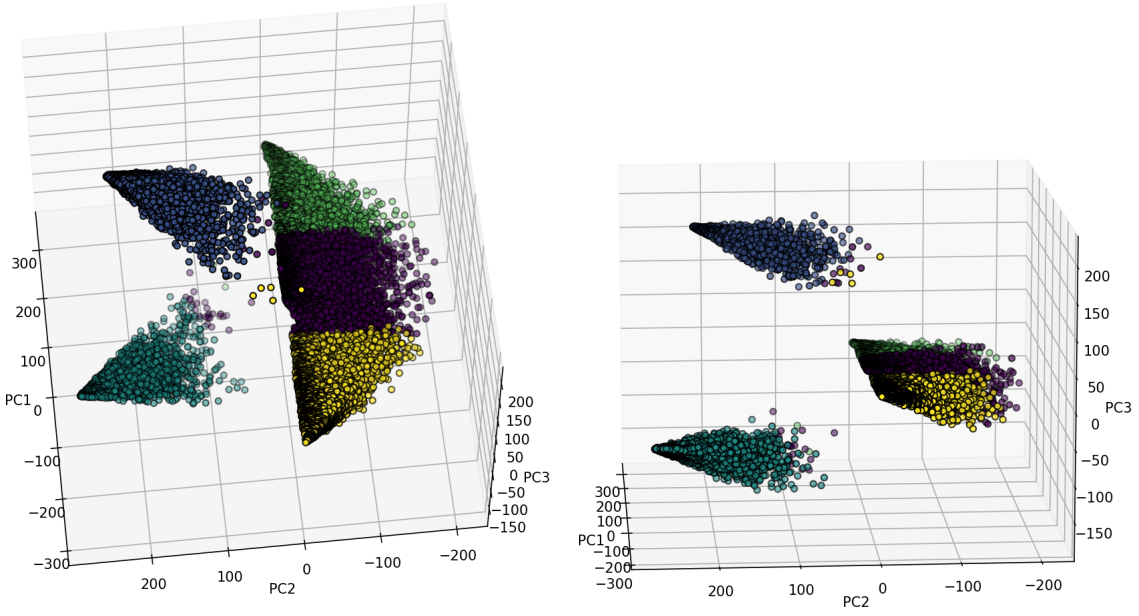


Como se esperó, las gráficas VAT e iVAT no nos muestran indicios de la cantidad de *clusters*, por lo que el valor a usar será el dado por la gráfica de codo, 5 *clusters*.

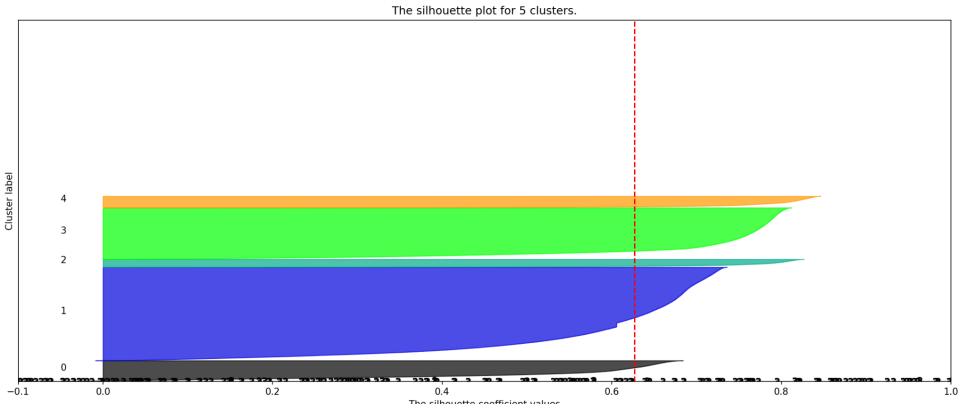
Clustering

K-Means

Ya que el modelo es creado con más de 3 variables, para poder visualizar los *clusters* de una manera que podamos comprender la clasificación, es necesario reducir la dimensionalidad de los resultados, es decir pasar las *features* a un números que podamos visualizar, para ello usamos el método de Principal Component Analysis (PCA) para comprimir nuestras *features* a 3 dimensiones.

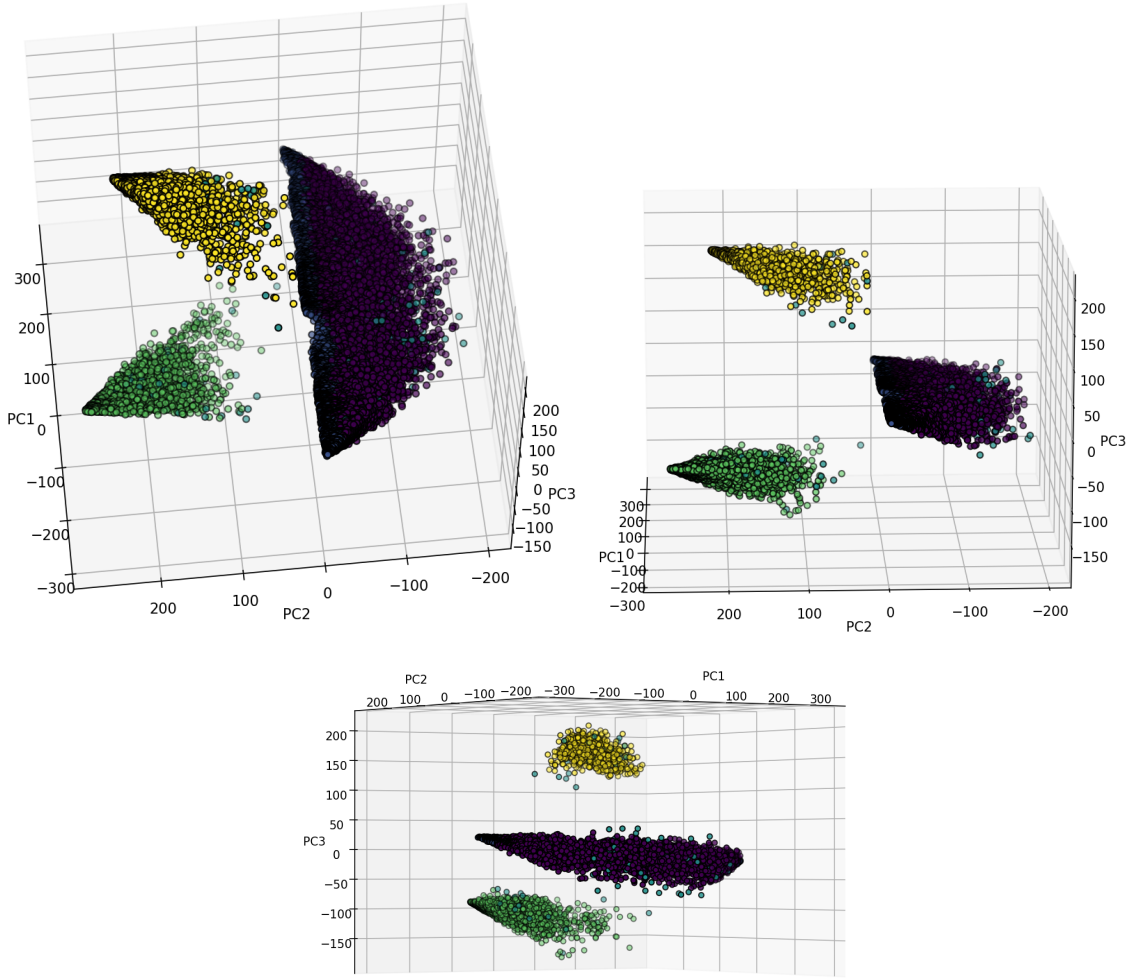


Y por último determinamos la eficiencia del clustering por medio del método de la silueta, tanto el promedio de esta como de su visualización por medio de la gráfica, en este caso se obtuvo un valor de `0.6269168516366032`, el cual al ser más cercano a 1 que a -1 nos indica que es un buen agrupamiento.

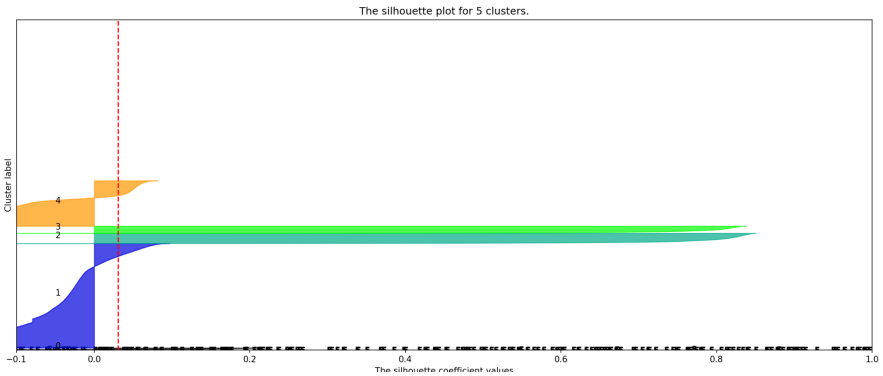


Mixture-Models

Los resultados del modelo Mixture-Models son similares a los de K-Means, sin embargo no se distingue esa separación clara entre los 3 *clusters* que se posicionan al centro, llegando a parecer que solamente creó 3 *clusters*.



Y en cuanto al rendimiento del modelo Mixture-Models por medio del método de la silueta, se obtuvo un valor de `0.030831404553935265` lo cual indica que no es ni un buen ni un mal *clustering*, sin embargo comparando con los resultados de K-Means, resultó en un agrupamiento poco eficiente.



Preguntas

- ¿Qué diferencias resultan de los clusters dados en K-Means y Mixture-Models para el dataset utilizado?

K-Means realizó una agrupación en los 5 clusters que se plantearon inicialmente dando un resultado rápido y efectivo, en cambio, Mixture-Models no realizó la agrupación en los 5 clusters, solamente realizó 3 clusters. Creemos que esto se dio debido a que Mixture-Models junto 3 clusters dentro de 1 solo, esto pudo haber causado que el puntaje de la silueta sea tan bajo, 0.308.

- ¿Por qué cree que se dan estas diferencias?

Principalmente se debe a los algoritmos que utiliza cada modelo y que están implementados internamente en la librería, donde K-Means va agrupando según distancias y Mixture-Models por estadística va asignando cada dato a los *clusters*. También hay que considerar que K-Means se adaptó bien a lo indicado por la gráfica de codo ya que la gráfica de codo es generada utilizando K-Means.

- ¿Ha variado la selección del número de clusters entre K-Means y Mixture-models? ¿Por qué?

Sí, K-means ha dividido en 5 *clusters* como se le indicó mientras Mixture-Models aunque se le indicaron 5, dividió solamente en 3 *clusters* principales. Esto se debe a que K-Means va agrupando a los *clusters* a base de los centroides, por lo que si se especifica un número de *clusters* es común que de un resultado de agrupación a cada uno de estos centroides, en cambio Mixture-Models es un modelo probabilístico que utiliza el algoritmo Expectation-Maximization (EM) que va agrupando los datos según su similitud.

- ¿En qué casos usaría K-means? ¿Por qué?

K-means lo utilizaremos en los casos donde buscamos optimizar la rapidez, facilidad y efectividad. Utilizaremos este modelo cuando se posee un análisis claro y comprensión de las variables así como en este laboratorio, donde se tenía bastante confiabilidad de la división en 5 *clusters*.

- ¿En qué casos usaría Mixture-Models? ¿ Por qué?

Mixture-Models lo utilizaremos en los casos donde buscamos optimizar la efectividad y complejidad , ya que , es más tardado pero brinda resultados más precisos, esto nos permite utilizarlo cuando no conocemos a fondo los *clusters* que debemos generar, dejando así que el modelo utilice el algoritmo de *Expectation-Maximization* para ir reduciendo a los *clusters*.