ALGORITMOS POPULARES DE AGRUPACIÓN NO SUPERVISADA

YULEIDIS MESA

CONJUNTO DE DATOS

	STG	SCG	STR	LPR	PEG	UNS
count	403.000000	403.000000	403.000000	403.000000	403.000000	403.000000
mean	0.353141	0.355940	0.457655	0.431342	0.456360	1.684864
std	0.212018	0.215531	0.246684	0.257545	0.266775	0.986195
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.200000	0.200000	0.265000	0.250000	0.250000	1.000000
50%	0.300000	0.300000	0.440000	0.330000	0.400000	2.000000
75%	0.480000	0.510000	0.680000	0.650000	0.660000	3.000000
max	0.990000	0.900000	0.950000	0.990000	0.990000	3.000000

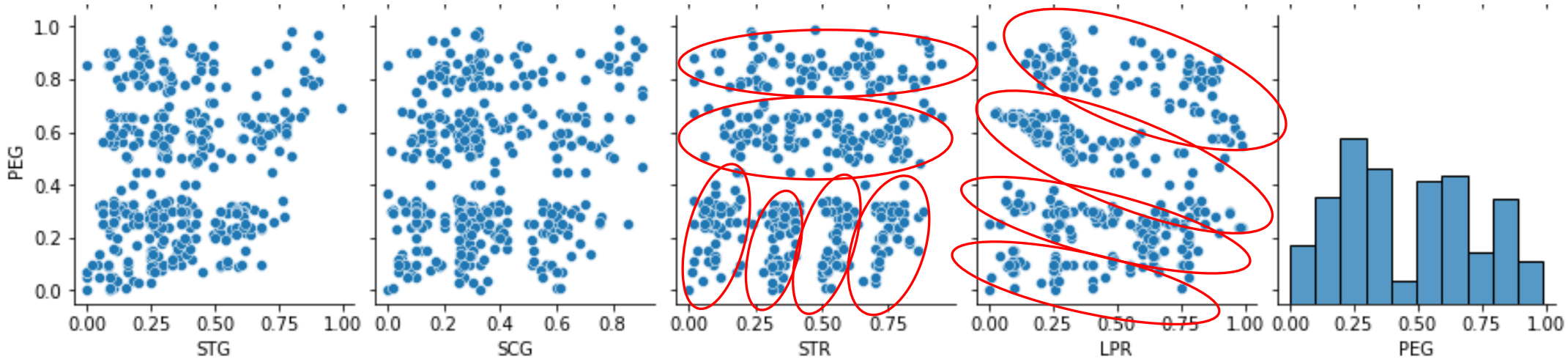
La base de datos contiene un conjunto de datos reales sobre el estado de los conocimientos de los estudiantes sobre el tema de las máquinas eléctricas de corriente continua.

La descripción muestra que en total tenemos 403 observaciones, además que los valores de las variables se encuentran estandarizados de cero a uno, de tal forma que el valor mínimo que toman es 0 y el valor máximo esta al rededor del 0,9.

	0	1
Muy bajo	0	50
Bajo	1	129
Medio	2	122
Alto	3	102

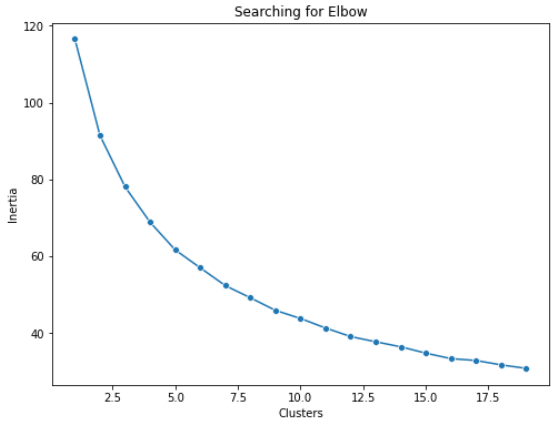
Asimismo, el autor realizó una categorización del nivel de conocimiento de los estudiantes que va desde Muy bajo a Alto. Así el grupo de Muy bajo (0) tiene 50 observaciones, el grupo con nivel Bajo de conocimiento (1) tiene 129, 122 en el nivel Medio (2) y finalmente 102 en el nivel Alto (3). Lo cual será útil para posteriormente evaluar la efectividad de los métodos.

GRAFICO DE CORRELACIÓN



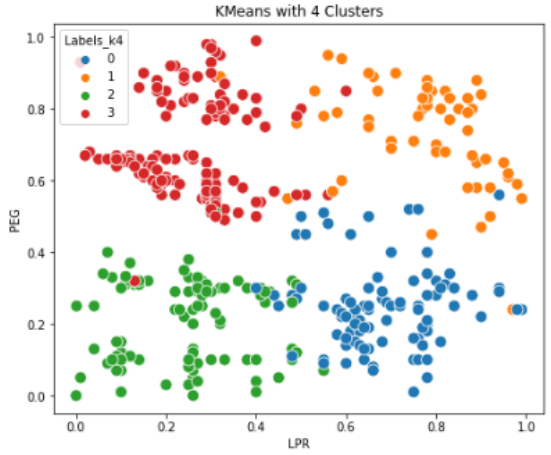
La grafica muestra la correlación entre cada una de las 6 variables, y vemos que para algunos casos no se aprecia segmentación entre grupos, sin embargo, al estudiar la relación entre las variables PEG y LPR, la grafica muestra como se dividen las observaciones en 4 grupos fundamentales, asimismo, la relación de las variables PEG y STR donde se observa la división de los datos en 6 grupos.

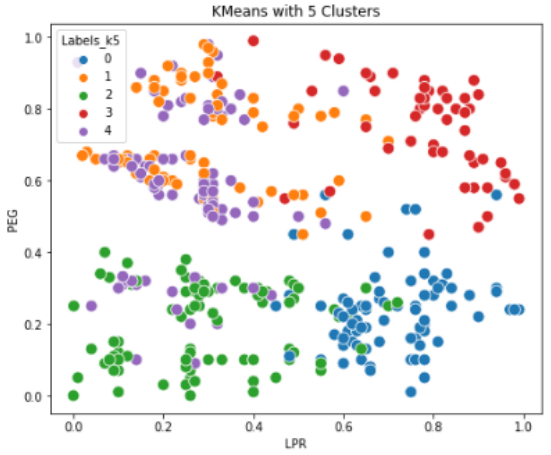
APLICANDO K-MEANS

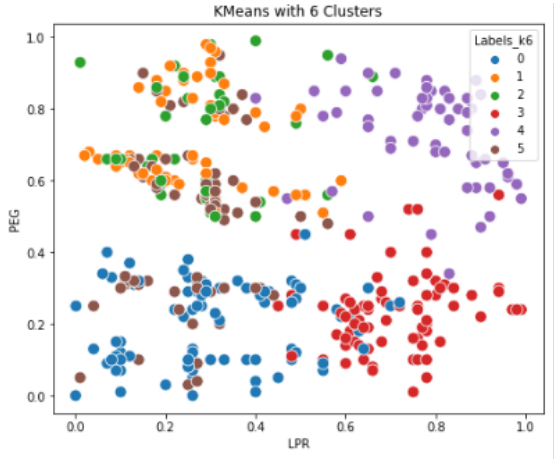


El método del codo nos dice que debemos seleccionar el cluster cuando hay un cambio significativo en la inercia. Como se observa en el gráfico, el punto de inflexión se sitúa de 4 a 6. Conforme a la subdivisión en grupos demostrada en la correlación de las variables LPR y PEG, se procederán a aplicar los métodos de Clustering entre ellas.

Los graficos evidencian que el cluster 4 parece ajustarse mejor a los datos que el cluster 5 y el 6.





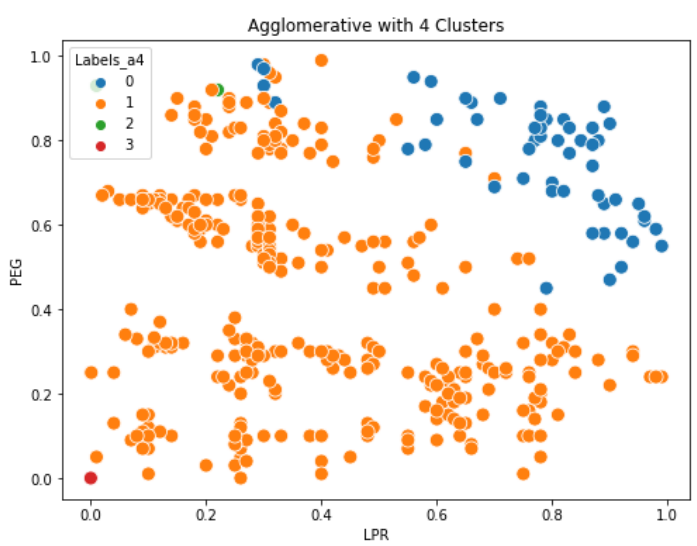


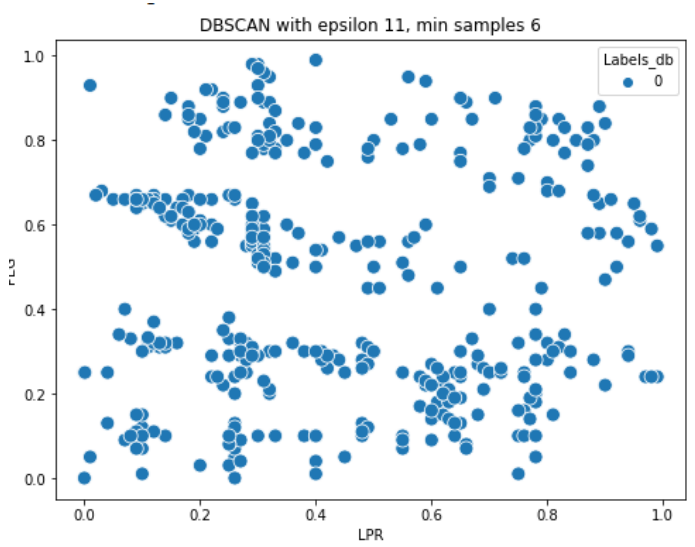
Aplicando k-means con 4 cluster tenemos:

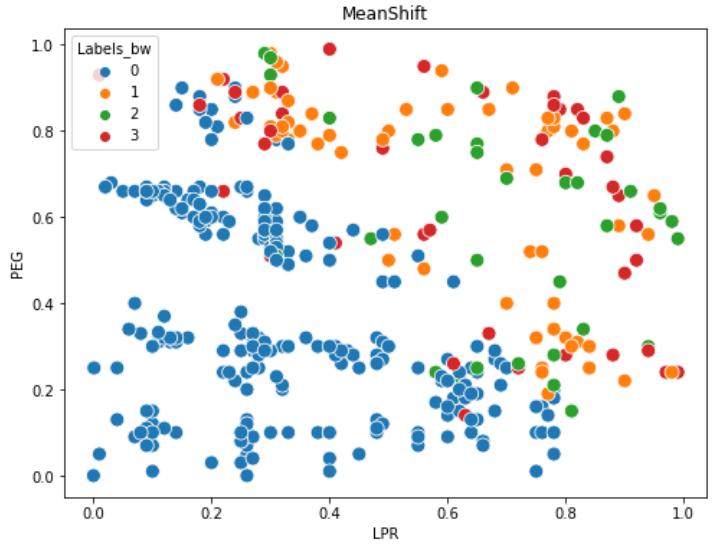
Examen de la materia (LPR) y examen del tema (PEG)

- Etiqueta 0 (Azul): Alto LPR y PEG bajo.
- Etiqueta 1 (Naranja): Alto LPR y PEG alto.
- Etiqueta 2 (Verde): Bajo LPR y PEG bajo.
- Etiqueta 3 (Rojo): Bajo LPR y PEG alto.

APLICANDO OTROS MÉTODOS







Al evaluar la primera gráfica, en comparación con el K-means, podemos evidenciar que el método de **Clustering jerárquico** no se ajusta a los datos, ya que aunque realizar la división en 4 grupos, deja solo dos observaciones en la etiqueta 2 y una en la 3, concentrando el resto de las observaciones en las etiquetas 0 y 1.

El método **DBSCAN** tampoco resulta ser útil para nuestro conjunto de datos, ya que no realiza una subdivisión entre las variables y deja todas observaciones en una única etiqueta.

Finalmente, al probar el algoritmo **Mean-Shift** sobre nuestros datos, vemos que aunque el método realiza un división por grupos en 4 etiquetas, estas se encuentran superpuestas entre si, lo cual no da una buena sensación de ajuste a los datos.

AJUSTE DEL MODELO

Distribución Original

```
1    129
2    122
3    102
0     50
Name: UNS, dtype: int64
```

Distribución K-means

```
3    135
2    113
0     99
1     56
Name: Labels_k4, dtype: int64
```

```
[168] from sklearn.metrics import accuracy_score, mean_absolute_percentage_error
```

```
[177] accuracy_score(data1["UNS_Kmeans"],data1["UNS"])

      0.012406947890818856
```

```
[178] (1 - accuracy_score(data1["UNS_Kmeans"],data1["UNS"]))*100

      98.75930521091811
```

Finalmente, se puede concluir que el modelo K-means con 4 cluster es el que mas se ajusta al conjunto de datos, y aunque la distribución de cada grupo varie con respecto a la distribución originalmente hecha por el autor, la precisión del error es de 0.012, y una precisión de ajuste del modelo es de 98,76.