

## MAESTRIA ESTADISTICA APLICADA Y CIENCIA DE DATOS

ESTUDIANTE: John Fredy Bello Cordero

TEMA: Algoritmo de clustering

Fecha: 26/05/2022

### 1 Punto

La DATA Se toman de la tesis Doctoral cuyos datos se refieren sobre el estado de los conocimientos de los estudiantes sobre el tema de las máquinas eléctricas de corriente continua.

Las variables se encuentran clasificada de la siguiente manera:

- 1- STG (El grado de tiempo de estudio para las materias del objeto de la meta).
- 2- SCG (El grado de repetición del usuario para las materias del objeto de la meta).
- 3- STR (El grado de tiempo de estudio del usuario para los objetos relacionados con el objeto de la meta).
- 4- LPR (El rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta).
- 5- PEG (El rendimiento en el examen del usuario para los objetos de la meta) • UNS (El nivel de conocimiento del usuario).

### 2 Punto

Se determinar en número de el número de datos que es 403 con representación principal además de la media, desviación estándar, así como valor mínimo y valor máximo.

```
data.describe()
```

	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

### 3 Punto

Se clasificaron los datos en variables con nombre específico en 4 columnas

```
] pd.DataFrame(np.unique(data["UNS"],re
```

	muy bajo	bajo	medio	alto
0	0	1	2	3
1	50	129	122	102

### 4 punto

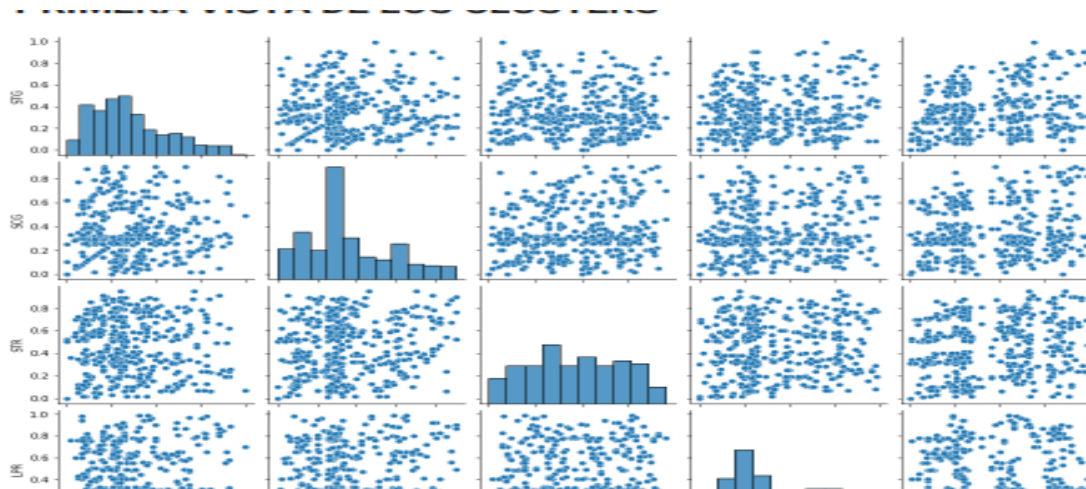
Se delimita a 4 variables con eliminación de UNS

```
[ 29] data0=data[["STG","SCG","STR","LPR","PEG"]]  
data0.head()
```

	STG	SCG	STR	LPR	PEG
0	0.00	0.10	0.50	0.26	0.05
1	0.05	0.05	0.55	0.60	0.14
2	0.08	0.18	0.63	0.60	0.85
3	0.20	0.20	0.68	0.67	0.85
4	0.22	0.22	0.90	0.30	0.90

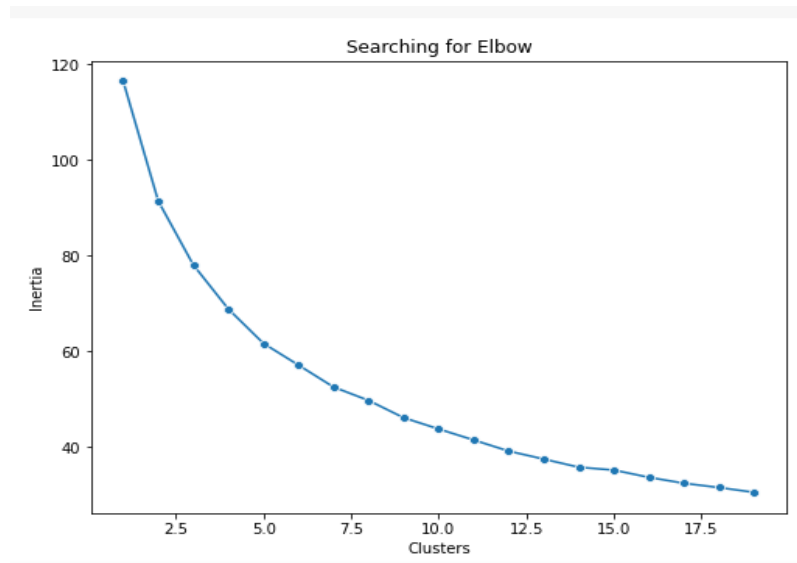
### 5 punto

Se toman las variables LPR y PEG de esta forma se establece una relación entre cada 2 variables a partir de un gráfico de dispersión dando la siguiente imagen.



## 6 punto

A partir del método del codo, se buscó el punto de inflexión entre si hay cambios de inercia mediante la gráfica, es decir buscamos el punto de inflexión entre 4 y 6 por cambios en la inercia representada en la gráfica.



## 7 punto

Se procede a realizar algoritmo de K-MEANS y a partir del análisis de 2 variables las cuales de discriminan a continuación

- ❖ LPR (El rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta).
- ❖ PEG (El rendimiento en el examen del usuario para los objetos de la meta).

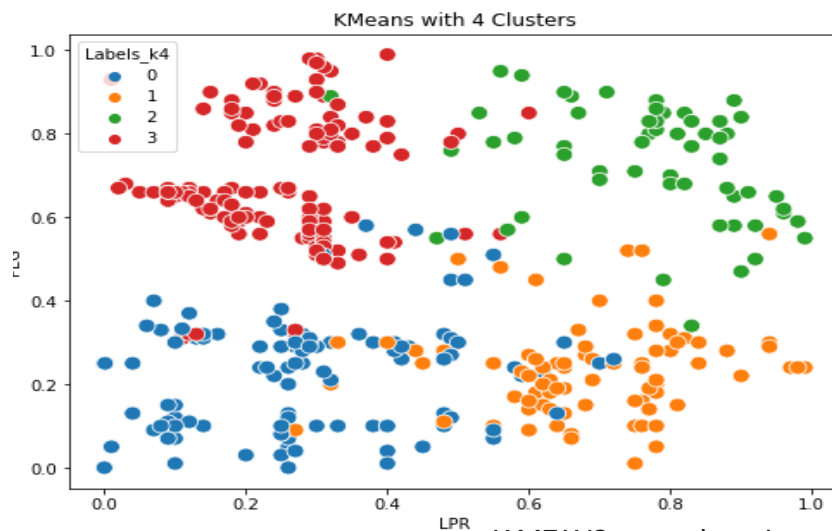
Se obtienen 4 cluster que se mencionan a continuación

1 Cluster 0: Bajo rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Alto rendimiento en el examen del usuario para los objetos de la meta

2 Cluster 1: Alto rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Alto rendimiento en el examen del usuario para los objetos de la meta

3 Cluster 2: Bajo rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Bajo rendimiento en el examen del usuario para los objetos de la meta

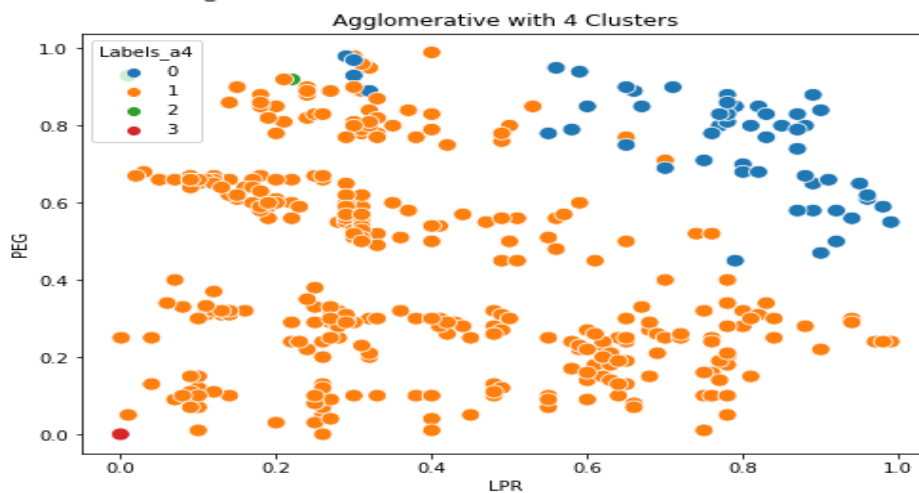
4 Cluster 3: Alto rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Bajo rendimiento en el examen del usuario para los objetos de la meta



Donde se evidencia que el algoritmo K-MEANS parece la mejor opción

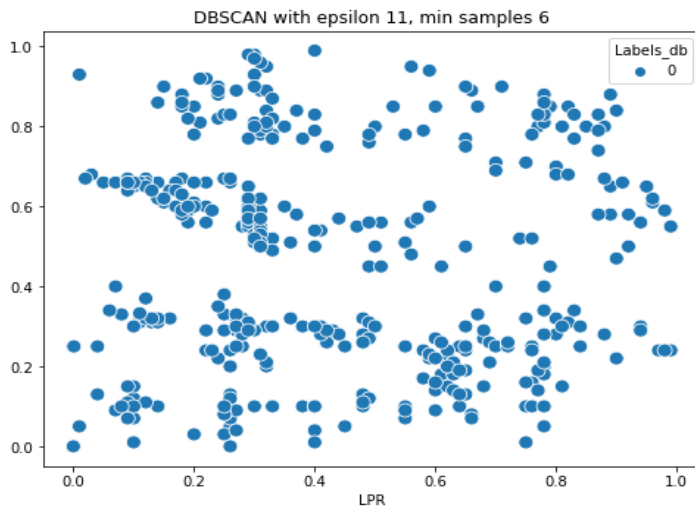
## 8 punto

Se aplica el algoritmo de CLUSTERING JERARQUICO que en comparación con K-MEANS Agglomerative Clustering no parece funcionar



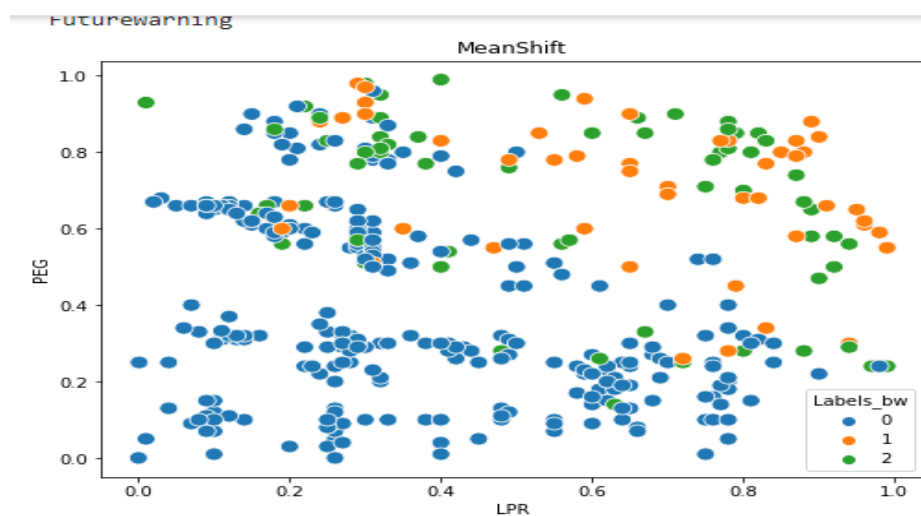
## 9 Punto

Se procede a la realización del algoritmo de DBSCAN que En comparación con K-MEANS DBSCAN no parece funcionar a pesar de modificaciones en las interacciones.



## 10 Punto

Se procede a la realización de Algoritmo Mean Shift que en comparación con K-MEANS Mean Shift no parece funcionar aun con modificaciones es las interacciones



## CONCLUSIONES:

LA MEJOR OPCION FUE K-MEANS DONDE SE EVIDENCIA 4 CLUSTERS EN RELACION CON 4 METODOS  
ADEMAS SE DETERMINO DOS INDICADORES EL PRMER CON UNA PRECISION DEL 94% Y UN  
PORCENTAJE DE ACERCAMIENTO DEL 93.