Conjunto de datos para probar algoritmos (Clustering)

MILDRED CARO ÁLVAREZ

Se trata de un conjunto de datos reales sobre el estado de los conocimientos de los estudiantes sobre el tema de las máquinas eléctricas de corriente continua

Las características de la base de datos son:

•STG (El grado de tiempo de estudio para las materias del objeto de la meta),
•SCG (El grado de repetición del usuario para las materias del objeto de la meta)
•STR (El grado de tiempo de estudio del usuario para los objetos relacionados con el objeto de la meta)
•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)
•UNS (El nivel de conocimiento del usuario)

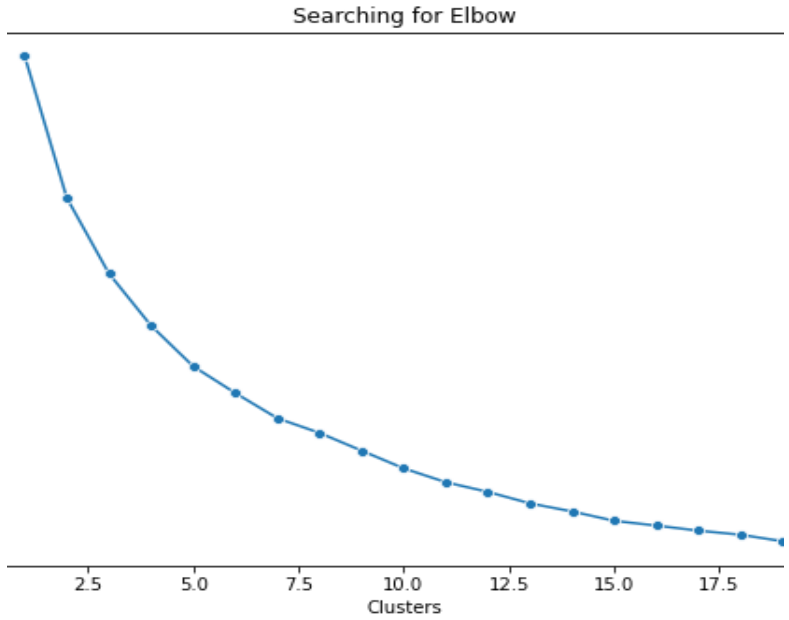
```
data["UNS"].value_counts()
```

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

De acuerdo a la recategorización realizada y a la realizada por el investigador, los 403 estudiantes se han agrupado en 4 niveles para la variable Nivel de Conocimiento del Usuario. Es así como 50 individuos se ubican en nivel "Muy Bajo" (0), 129 en "Nivel Bajo" (1), 122 en "Nivel Medio" y 102 estudiantes en "Nivel Alto".

Aplicaremos 4 algoritmos o metologías de Clustering para el conjunto de datos:
1.K-Means
2.Clustering jerárquico
3.DBSCAN
4.Mean Shift

Método K- Means

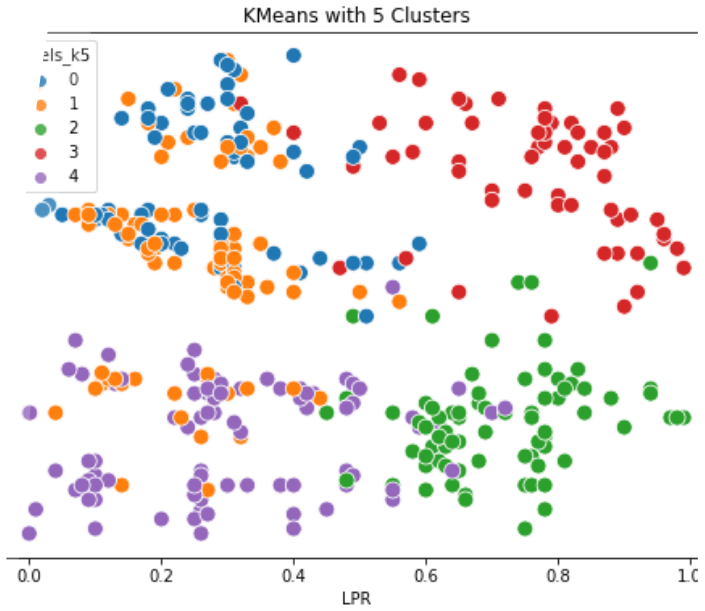


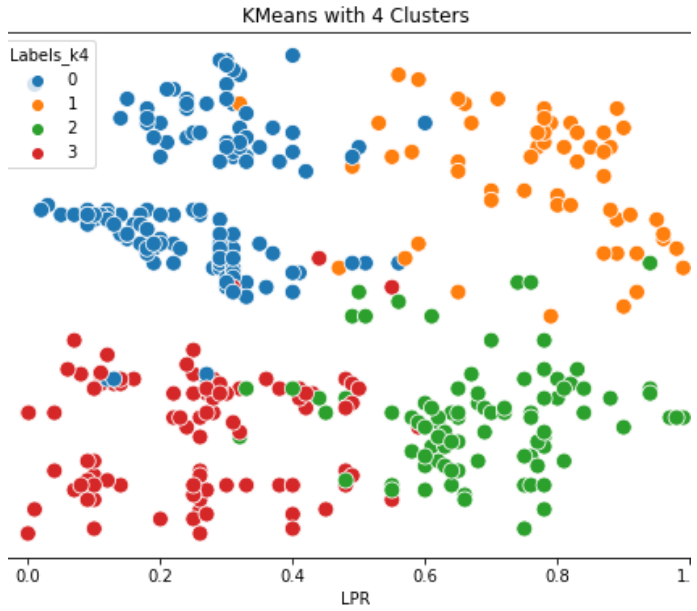
Conforme a la gráfica de codo, podemos inferir que el conjunto de datos se puede agrupar en aproximadamente 5 categorías.

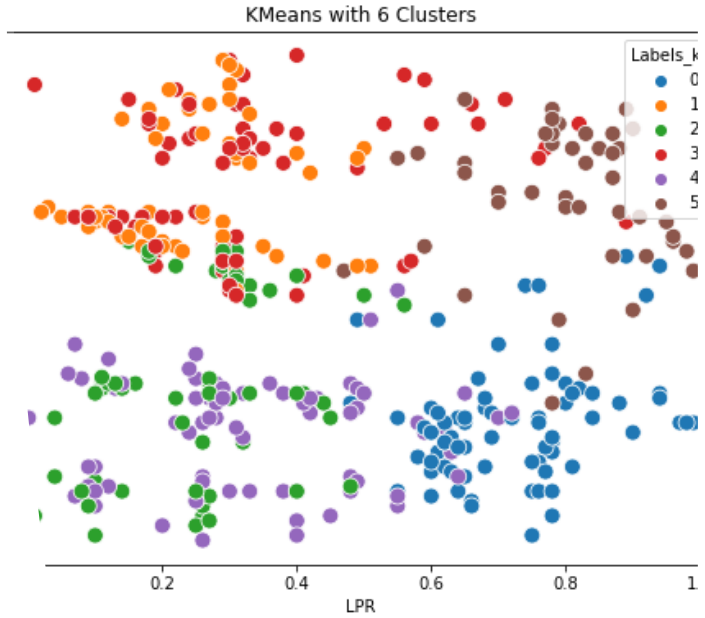
```
(1-accuracy_score(data1["UNS_Kmeans"],data1["UNS"]))*100
```

```
89.57816377171216
```

Este método se ajusta en aproximadamente un 89,5% para categorizar el conjunto de datos.



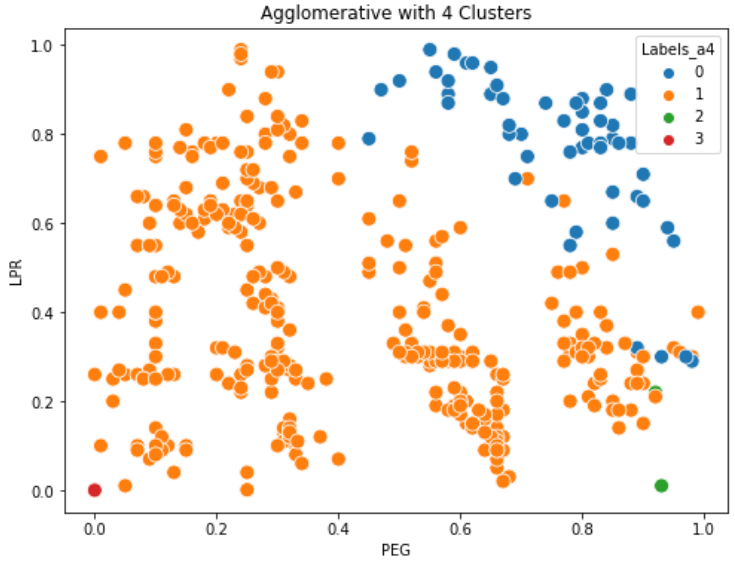


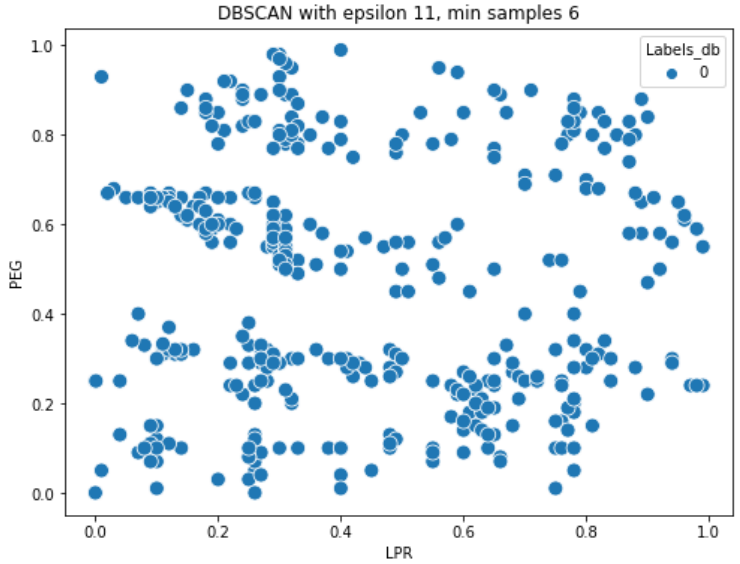


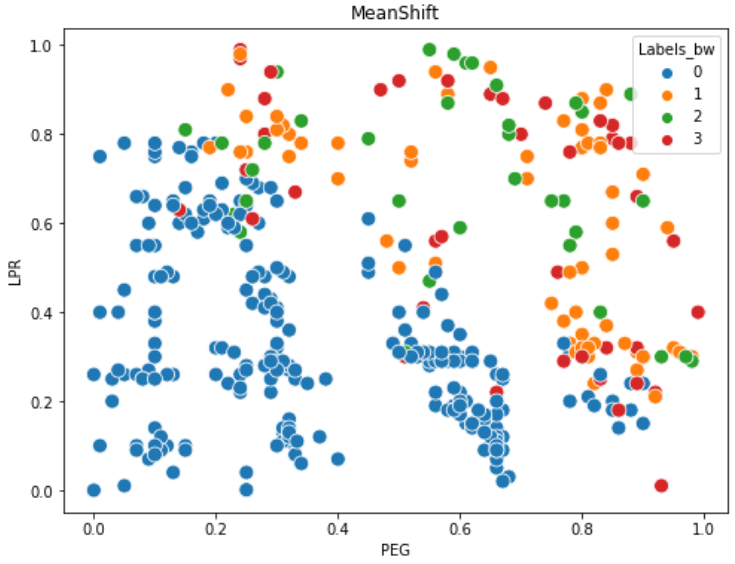
Agrupando las observaciones en 4 categorías, se evidencian mejor la diferencias entre los grupos, con algunas excepciones en algunos datos. Sin embargo, son mucho más notorios estos grupos en comparación con las pruebas de los 5 y 6 clústers.

En ese orden de ideas, las categorías son:

- Clúster 0: LPR bajo y PEG alto
- Clúster 1: LPR alto y PEG alto
- Clúster 2: LPR alto y PEG bajo
- Clúster 3: LPR bajo y PEG bajo







Con los métodos DBSCAN y Mean Shift, al igual que con el Aglomerative, no es muy evidente una agrupación de los datos en categorías, por lo que para este caso puntual, resulta conveniente el uso del algortimo K-Means con 4 clústers.