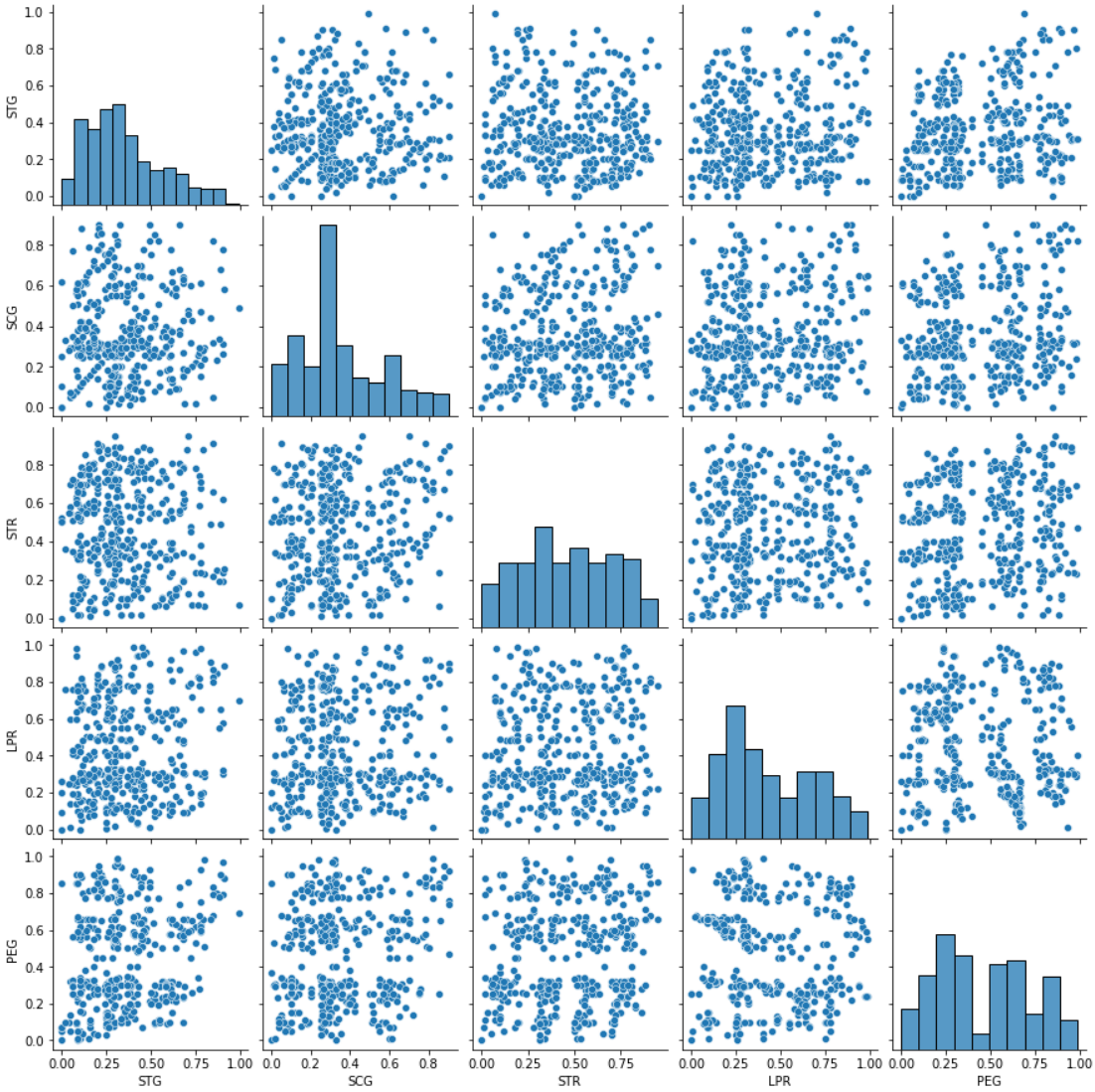
DATA

Se extrae de la tesis Doctoral estudiada un conjunto datos reales sobre el estado de los conocimientos de los estudiantes sobre el tema de las máquinas eléctricas de corriente continua. En este se encuentran las siguientes variables:

- 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)

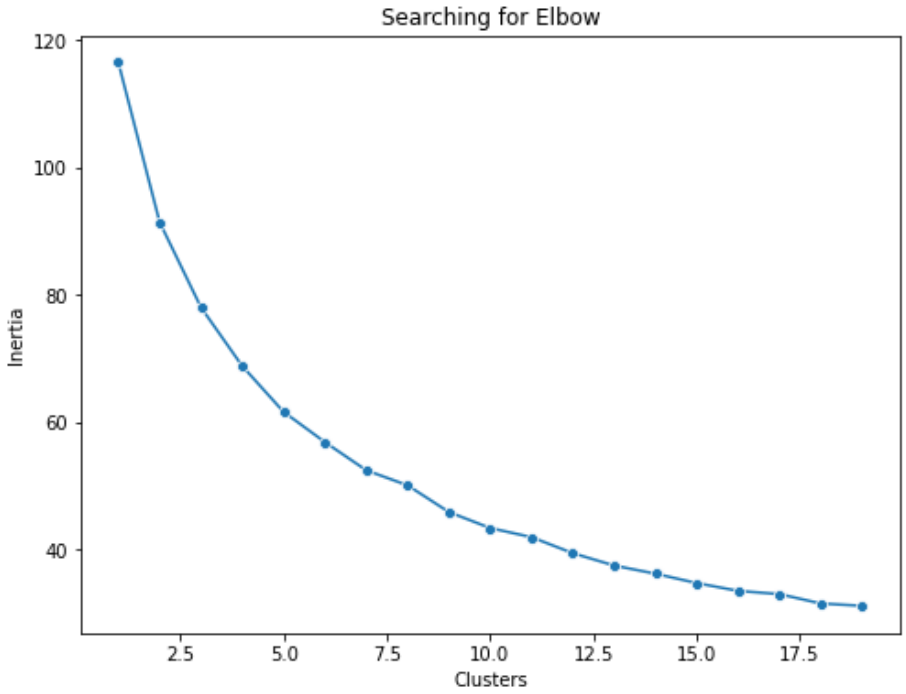
PRIMERA VISTA DE LOS CLUSTERS



Se compara mediante gráfico de dispersión la relación existente entre cada par de variables para determinar a una primera vista la presencia de clusters. Se decide estudiar la pareja que representa las variables LPR y PEG

¿CUANTOS CLUSTERS?

Determinamos mediante el método del codo si hay cambios de inercia mediante la gráfica, es decir buscamos el punto de inflexión, el cual se puede apreciar que está entre 4 y 6 clusters para posteriormente aplicar los algoritmos de Machin Learnig en este caso de problemas estudiados.

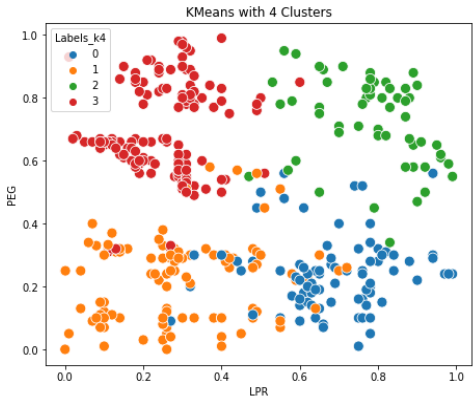


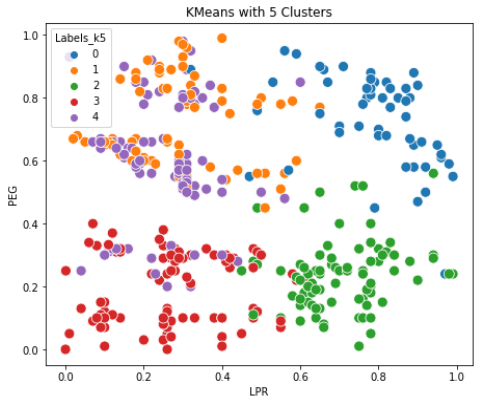
APLICANDO CLUSTER KMEANS

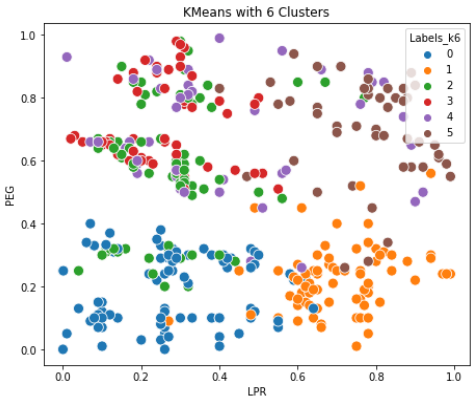
Al parecer por lo observado en los gráficos con 4, 5 y 6 clusters la mejor opción para KMEANS sería tomar 4 clusters. Para las dos variables analizadas; 1. LPR (El rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta) y 2.PEG (El rendimiento en el examen del usuario para los objetos de la meta), tenemos el sigueinte análisis:

- **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
- **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

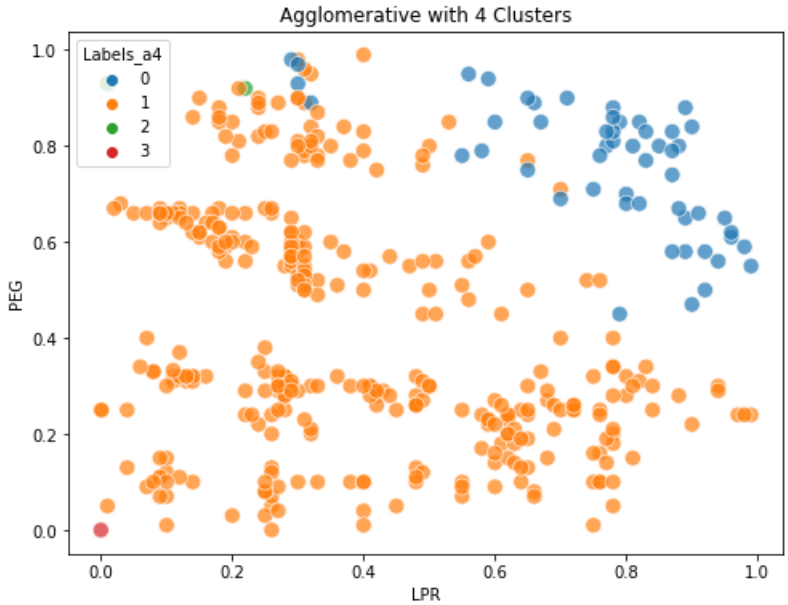
- **Cluster 2:** Bajo rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Bajo rendimiento en el examen del usuario para los objetos de la meta
- **Cluster 3:** Alto rendimiento en el examen del usuario para los objetos relacionados con el objeto de la meta, Bajo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







APLICANDO CLUSTER JERÁRQUICO



Al aplicar el algoritmo de cluster jerárquico con 4 clusters podemos concluir a simple vista para este número el KMEANS agrupa de mejor forma los datos.

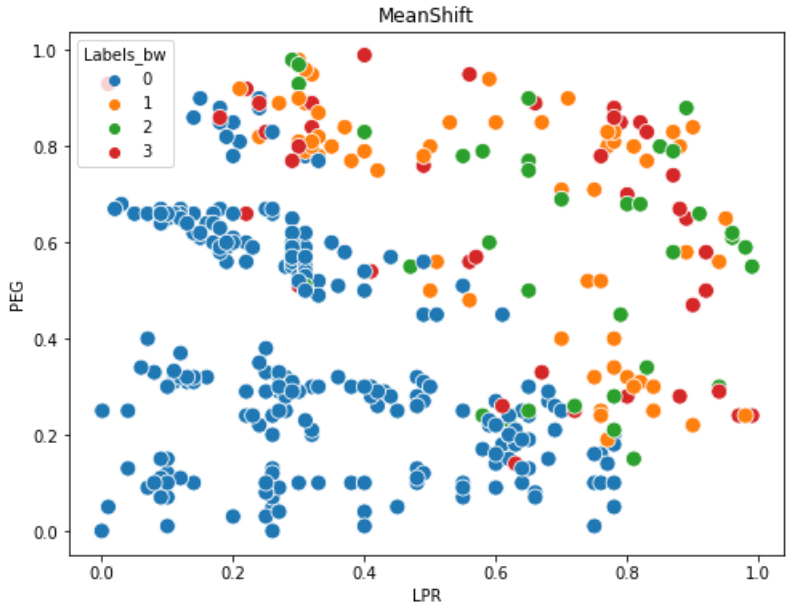
APLICANDO DBSCAN

Al utilizar la herramienta o algoritmo DBSCAN podemos observar que definitivamente para este conjunto de datos no es para nada adecuado al momento de generar clusters, incluso variando los parámetros Epsilon y Puntos Mínimos Epsilon el algoritmo no es adecuado.

Lo más probable es que DBSCAN no funcione por que la cantidad de datos es muy pequeña para que el aplique su estructura. Lo ideal es que se tengan más datos o información.

APLICANDO MEANSHIFT

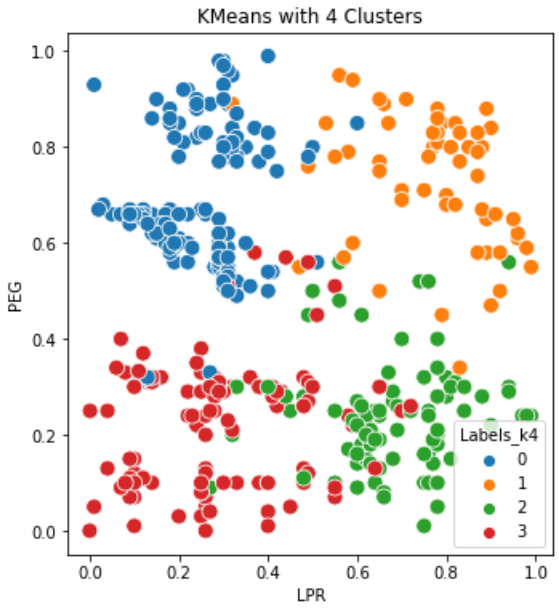
Para el algoritmo MEANSHIT se verifica el quantile que nos pueda arrojar un arreglo de datos en 4 clusters, para este caso sería el quantile=0.0819. Notamos que para este tipo de datos este algoritmo tampoco es muy adecuado ya que existen mucho cruce de datos entre los distintos clusters.

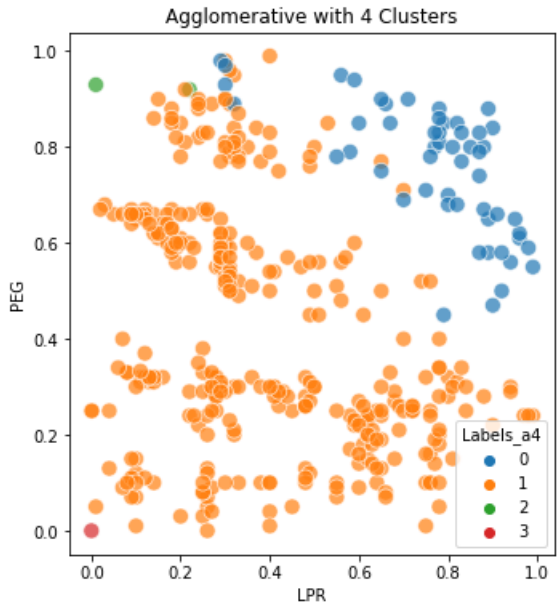


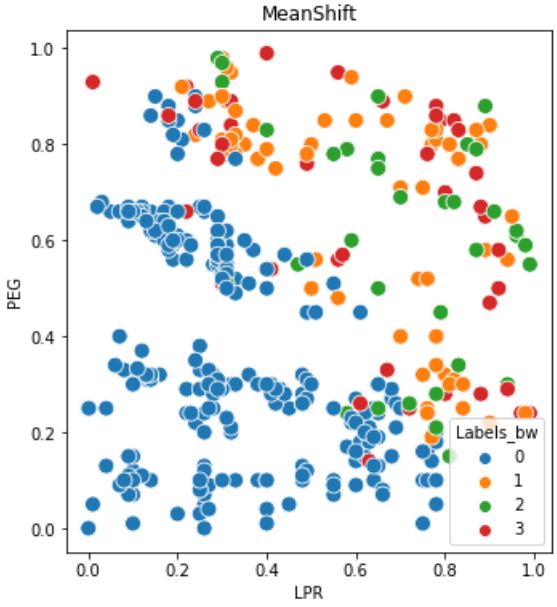
CONCLUSIONES

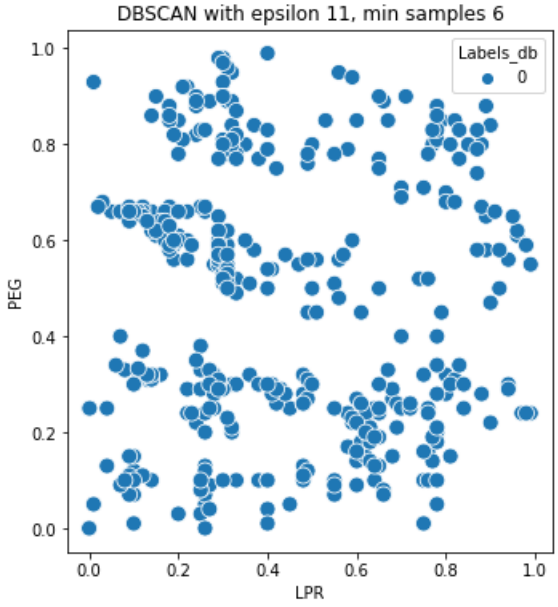
La primera conclusión a la que se llega y de acuerdo con lo expuesto anteriormente es que el mejor método de cluster para estos datos es el arrojado por medio del algoritmo KMEANS, ya que como se muestra en la figura siguiente donde se comparan los 4 métodos con 4 clusters el que menor evidencia traslapes entre puntos de clusters distintos es precisamente KMEANS.

Se establecen 2 indicadores, el primero es el porcentaje de acercamiento del método KMEANS a los datos proporcionados en la investigación en cuanto a niveles, este nos arroja un 93,07 %, ver tabla siguiente. El según indicador es el grado de precisión del algoritmo, el cual, se calcula mediante la resta de 1 menos el cociente de los valores predichos por el algoritmo sobre el total de los datos, al final se multiplica por 100, en este último se obtiene un grado de precisión del 95,54 % muy parecido al primer indicador, Ver tabla anterior.









Indicador	%
% De acercamiento	93.07
Precisión	94.54