

Mentoría Churn Prediction - Aprendizaje No Supervisado

Input:

Archivo input editado del TP N°2, luego de la “curación” de datos.

Entregable:

- Se puede ir desarrollando cada punto en la misma notebook donde se escriba el código.
- Se debe subir el entregable a un repositorio GitHub o enviar el link a un Google Colab.
- Tener en cuenta que si bien pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.
- Luego de cada análisis es importante poder obtener una conclusión de lo observado.

Se analizarán los datos de los usuarios de la plataforma de streaming de música para identificar grupos de clientes “similares”, aplicando modelos de clústering.

Tal vez la empresa pueda aplicar distintas estrategias de marketing para estos diferentes grupos identificados.

1. Preparación de los datos para aplicar modelos de clústering.

Evaluación visual e intuitiva de a dos variables por vez: Realizar visualizaciones de cómo se comportan las variables presentes en el dataset mediante el gráfico “pairplot”.

- a. ¿Es posible en alguna de ellas visualizar “grupos” separados? ¿Qué variables presentan dicha separación?
- b. ¿Se puede visualizar algún impacto de la variable target (Churn) en los posibles grupos?

- c. Si evalúa que es necesario, seleccione sólo algunas variables con las cuales se realizarán modelos de clustering.

2. Creación de modelos de clustering: K-means

- a. Buscamos el número de clusters óptimo observando la inercia de cada uno.
- b. Correr K-means utilizando como parámetro el número de clusters obtenido en el punto anterior.
- c. Analizar la cantidad de casos en cada clúster.
- d. Grafique de a dos variables incluyendo el label obtenido en el clustering. ¿Es posible determinar alguna separación de grupos?

Documentación:

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

3. Creación de modelos de clustering: Algoritmo Jerárquico

- a. Realizar las aglomeraciones utilizando el algoritmo "AgglomerativeClustering".
- b. Analizar la cantidad de casos en cada clúster.
- e. Grafique de a dos variables incluyendo el label obtenido en el clustering. ¿Es posible determinar alguna separación de grupos?
- f. Como conclusión ¿cuál algoritmo de clustering parece haber funcionado mejor para el set de datos trabajado?

Documentación:

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

Deadline de entrega: 09/09/2022.