Proyecto de Análisis K-Means Clustering con Excel

Introducción

K-Means es un modelo de clasificación muy utilizado como una introducción a modelos más complejos de Machine Learning, su funcionamiento depende asignar de forma aleatoria K elementos de nuestra muestra siendo nuestro numero de clusters o grupos, los cuales utilizaremos para realizar la clasificación, por medio del cálculo de las medias (centroides) de cada cluster se obtiene una aproximación cada vez más precisa por medio de múltiples iteraciones, hasta que los centroides no tengan cambios entre las iteraciones.

Al ser un modelo sencillo es factible realizarlo por medio de Excel, la cual es una herramienta de hoja de cálculo que además de realizar cálculos sencillos, nos permite procesar y visualizar datos de forma sencilla, permitiéndonos ver cada paso del modelo.

Base de Datos

Para nuestro proyecto de análisis exploratorio utilizaremos una base de datos que mide el peso y el tamaño de 9 especies de peces. La base de datos ha sido adaptada para facilitar este tipo de análisis por lo cual no se debe considerar que se está midiendo la capacidad del modelo, si no que estamos realizando un estudio del modelo.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

En un primer análisis de la base de datos podemos notar que la proporción de cada especie es de aproximadamente 11%, ademas de notar que no existen valores nulos, nuestra base de datos está bien equilibrada y tiene una variedad razonable de datos.

A graph with colorful dots

Description automatically generated

Desarrollo

Como se menciona al inicio, el propósito de este proyecto es estudiar el modelo K-Means por medio de Excel, se debe aclarar que aunque es posible llegar al punto en el que los centroides dejan de moverse, no es necesario completar todas las iteraciones necesarias para estudiar el modelo, en este proyecto nos limitamos a 3 iteraciones para comprobar el funcionamiento del modelo.

Durante el desarrollo experimental no hubo mayores problemas, aunque la base de datos es algo extensa, no hubo problemas con Excel para manejar las múltiples iteraciones del modelo y el procesamiento de datos necesario para calcular los centroides.

A graph with colored spots

Description automatically generated

Conclusión

Aunque Excel es una herramienta muy versátil que permite realizar múltiples modelos, por la naturaleza de este modelo es necesario realizar múltiples procedimientos paso a paso, debido a que no es posible automatizar las iteraciones si no se hace uso de programación para lograrlo.

El modelo es muy eficiente, aunque no es perfecto y para nuestro proyecto no logro la clasificación esperada, sería necesario realizar más modificaciones, ajustar mejor los datos y probar múltiples variables para cada cluster, para lograr mejores resultados de clasificación.

# References

David Langer. (n.d.). *K-means Cluster Analysis With Excel - A Tutorial.* Retrieved from DaveOnData: https://youtu.be/YuxwkchSAW4?si=6vRzlQtrIWdp2VbB

Lo, T.-w. (n.d.). *Fish species sampling weight and height data.* Retrieved from Kaggle: https://www.kaggle.com/datasets/taweilo/fish-species-sampling-weight-and-height-data/data