Universidad del Valle de Guatemala
Minería de Datos
Douglas de León Molina
Carné 18037

Laboratorio 2 – Clustering

Análisis exploratorio

El dataset está conformado por 150 registros de 3 variaciones de la flor Iris (Iris Versicolor, Iris Setosa e Iris Virginica), 50 para cada una de las respectivas clases. Contiene las columnas “sepal length” en cm, “sepal width” en cm, “petal length” en cm, “petal width” en cm y “clases” como única variable categórica con los tipos variación de la flor.

Sumado a esto, se provee la siguiente tabla de estadísticas del dataset:

	Min	Max	Mean	SD	Correlation
Sepal length	4.3	7.9	5.84	0.83	0.7826
Sepal width	2	4.4	3.05	0.43	-0.4194
Petal length	1	6.9	3.76	1.76	0.949
Petal width	0.1	2.5	1.2	0.76	0.9565

Hipótesis u objetivo

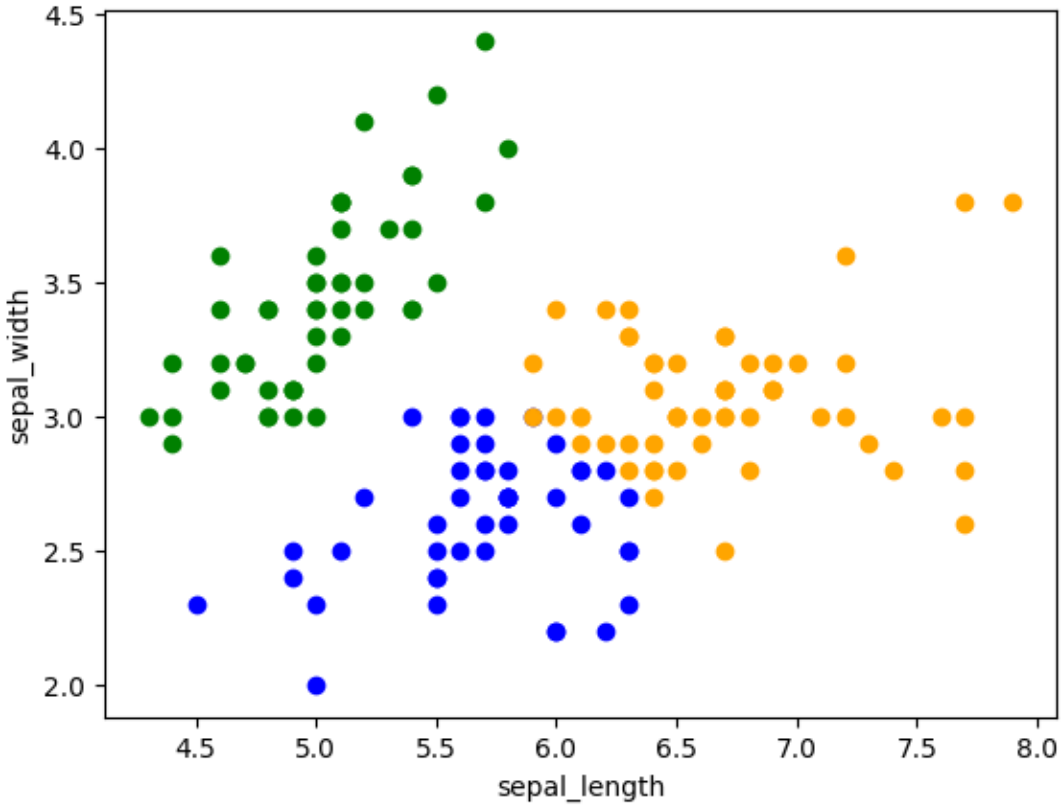
El objetivo es crear un modelo de segmentación con el algoritmo K-Means para las variaciones de la flor Iris con el dataset *iris.csv* y comparar los resultados contra los valores reales del dataset.

Solución y exploración

Para crear el modelo se tomó en cuenta las estadísticas proveídas en el análisis exploratorio en las cuáles se identificó que hay una alta correlación en las variables de petal length y petal width. Se tomó la decisión de eliminar del dataset la variable de petal_width para hacer la segmentación, por la misma razón. Para identificar los centroides, se tomó 3 muestras del dataset y se aplicó el algoritmo y luego se generó una gráfica para ver los segmentos generados (en Resultados).

Resultado

Para observar los resultados se creó la siguiente gráfica en la que se compara el “sepal length” contra el “sepal width”:



Como referencia tenemos la gráfica del dataset original para ver cómo se distribuyen realmente las variaciones de la flor.