# ESTUDIO DE CLUSTERS EN PINGÜINOS

*Gabriela Puerta Bello*

***Resumen*— En este proyecto se aplicaron técnicas de aprendizaje no supervisado para realizar análisis de clustering en un conjunto de datos relacionados con características físicas de pingüinos. El proceso involucró la carga y preparación de los datos, la identificación y tratamiento de los valores nulos o faltantes, también en la selección de características relevantes y en la estandarización de los datos. Posterior a esto, se implementó el método del codo más conocido como (Elbow Method) para determinar el número óptimo de clusters utilizando el algoritmo K-Means. Los resultados obtenidos nos muestran una división adecuada de los datos en tres clusters distintos, haciendo asi que se comprenda mejor las similitudes y diferencias entre las distintas especies o poblaciones de pinguinos.***

## I. Introducción

EL analisis de clustering es una técnica fundamental en el campo del aprendizaje no supervisado por que permite agrupar datos o caracteristicas similares en grupos o tambien llamados clusters. Los pingüinos, son aves que habitan en regiones frías del hemisferio sur, presentan variaciones significativas en sus características físicas según su especie y su hábitat. El estudio de estas variaciones no solo es crucial para la biología y ecología, sino que también ofrece un interesante caso de estudio para la aplicación de técnicas de clustering en el aprendizaje no supervisado. En el contexto en el que estamos se abordó un conjunto de datos que contiene información sobre características de pingüinos como la longitud y profundidad de su pico, este puede variar significativamente entre diferentes especies y sexos, tambien puede ser un indicativo de estos mismos. La longitud de las aletas, la masa corporal que puede variar ampliamente entre especies y su sexo clasificado en Male y Female. A pesar de conocer las características físicas medidas, la dificultad radica en cómo agrupar efectivamente a los pingüinos en clusters significativos basados en esta información, estableciendo así nuestro objetivo principal el cual fue explorar comportamientos similares en los datos y así determinar cuantos grupos naturales de pingüinos pueden existir basandonos en sus características físicas.

Una visualización de una pequeña parte de los datos:

**TABLA I**
*CARACTERISTICAS DE LOS PINGÜINOS*

| LONGITUD DEL PICO | PROFUNDIDAD DEL PICO | MASA CORPORAL | LONGITUD DE ALETAS | SEXO |
|---|---|---|---|---|
| 39.1 | 18.7 | 3750.0 | 181.0 | MACHO |
| 39.5 | 17.4 | 3880.0 | 186.0 | HEMBRA |
| 40.3 | 18.0 | 3250.0 | 195.0 | HEMBRA |
| NAN | NAN | NAN | NAN | NAN |

## II. Modelo

### A. Carga y Preparación de Datos.

En primer lugar se cargaron los datos desde un archivo con formato CSV y se realizó una inspección inicial para identificar valores faltantes y que tipos de datos hay en cada columna. En la tabla 1 podemos apreciar que hay columnas con datos Nulos. Estos valores faltantes se eliminaron utilizando el metodo **"dropna()"** para asegurar la integridad del conjunto de datos. Tambien podemos observar que hay una variable etiquetada como "sex", teniendo información codificada como cadenas de texto. Se realiazo una transformación numerica asignando dos valores, 0 para "Macho" y 1 para "Hembra" mediante el uso de la función **"map()".**

### B. Selección y Estandarización de Caracteristicas.

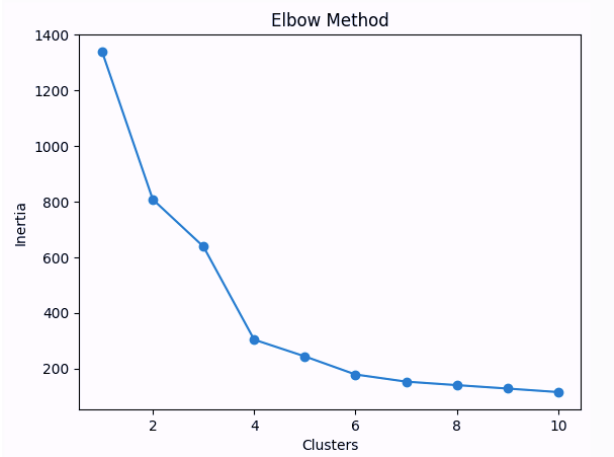
Se seleccionaron las caracteristicas relevantes para el análisis las cuales son: la longitud y profundidad del pico, la longitud de las aletas y la masa corporal de los pinguinos. Estas variables se almacenaron en una matriz con variable denominada X.

Una vez hecho esto, se estandarizaron las caracteristicas utilizando **"StandardScaler()"** de scikit-learn. Esto permitio

escalar todas las variables a una misma escala y evitar que aquellas con valores mas grandes dominen el proceso.

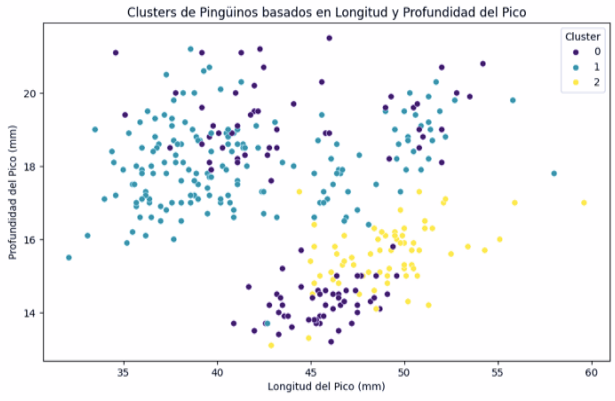
### *C. Determinación del número òptimo de Clusters.*

Para determinar el número óptimo de clusters, se implementó el metodo del codo, mejor conocido como (Elbow Method). Este método consiste en ejecutar el algoritmo K-Means con diferentes valores de k (numero de clusters) y calcular la suma de las distancias cuadradas de las muestras a sus clusters más cercanos. El punto de inflexión en la curva de inertia indica el número óptimo de clusters. En este caso, el gráfico mostro un codo en k = 3, lo que significa que tres clusters es el número óptimo para este conjunto de datos.




### *D. Implementación K-Means*

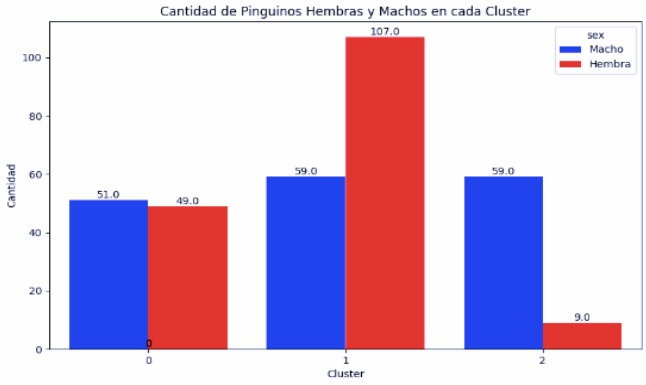
Una vez determinado el número óptimo de clusters como (k = 3), se aplica el algoritmo K-Means al conjunto de datos estandarizado. Este conjunto de datos incluye diversas caracteristicas, entre las que se destacan la longitud y profundidad de sus picos. Estas caracteristicas son claves pues varían significativamente entre especies de pinguinos. Estudios científicos han demostrado que estas medidas pueden ser utilizadas para identificar y a su vez clasificar diferentes especies. Las variaciones de la longitud y profundidad del pico estan relacionadas con los hábitos alimenticios y el nicho ecológico de estos. En conclusión, esto permitió dividir los datos en tres clusters distintos. Cada pingüino perteneciente al conjunto de datos fue asignado a uno de los tres clusters. Los clusters representan grupos de pinguinos con características físicas similares, como se puede visualiar en la siguiente grafica:




### *E. Visualizaciones.*

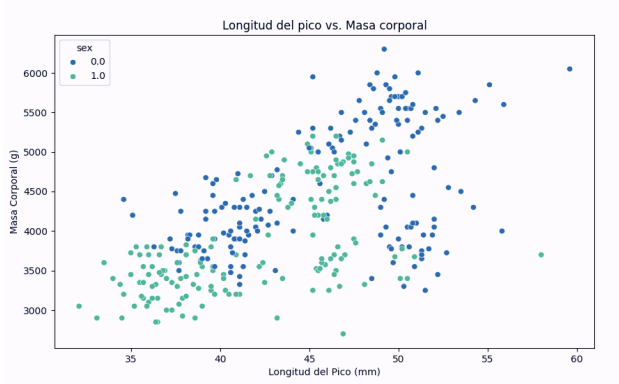
Ademas del gráfico que muestra la distribución de los clusters basados en longitud y profundidad del pico, hemos generado otras visualizaciones que ofrecen una comprensión más completa de la segmentación de los datos y las características de los pinguinos en cada cluster.

- Cantidad de Hembras y Machos en cada Cluster:



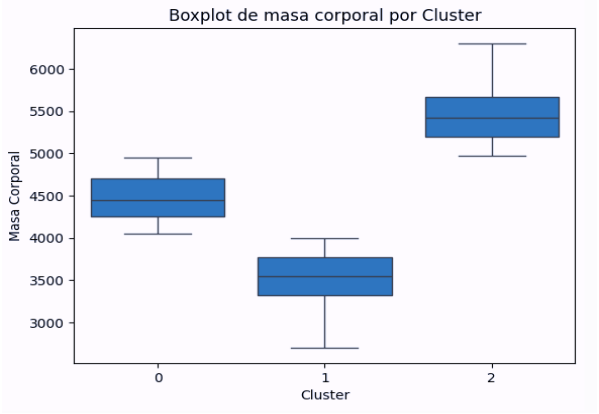

Éste gráfico de barras muestra la cantidad de pingüinos hembras y machos en cada Cluster, se puede observar que el cluster 1 tiene principalmente hembras, el cluster 0 tiene una mezcla mas equilibrada de ambos sexos y el cluster 2 tiene tiene una mayoria de machos.

- Longitud del Pico vs Masa Corporal:



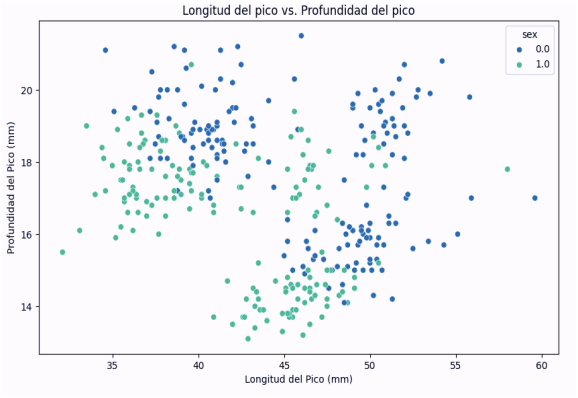
Se puede visualizar en este grafico la relación entre la longitud del pico y la masa corporal de los pingüinos, con los puntos coloreados según el sexo, 0.0 para Macho y 1.0 para Hembra. Existe una correlación positiva entre la longitud del pico y la masa corporal; es decir, a medida que la longitud del pico aumenta, también lo hace la masa corporal.

- Masa Corporal por Cluster:




Este gráfico muestra la distribución de la masa corporal para cada uno de los clusters identificados, podemos notar que el cluster 0 tiene una masa corporal mediana alrededor de 4500 g, con un rango que se extiende aproximadamente de 4000 g a 5000 g, el cluster 1 tiene una masa corporal más baja, con una mediana alrededor de 3500 g y un rango de aproximadamente 3000 g a 4000 g y el cluster 2 presenta la mayor masa corporal, con una mediana cerca de 5500 g y un rango de aproximadamente 5000 g a 6000 g.

- Longitud vs Profundidad del Pico:




En esta gráfica se permite observar el comportamiento de estas variables en función del sexo de los pingüinos. Siendo 0.0 macho y 1.0 hembra, se puede notar que los picos mas cortos y menos profundos los ocupan en mayor medida los puntos verdes, o sea las hembras y los picos más largos y profundos estan más densamente poblados por puntos azules, lo que significa que son mas Machos los que tienden a tener dichos picos de esa forma, haciendo sentido al diformismo sexual de esta especie.

## III. RESULTADOS

Despues de aplicar el algoritmo de clustering y obtener los clusters es fundamental evaluar la calidad y coherencia de los resultados. Esto se puede lograr aplicando evaluaciones y métricas específicas como lo son:

- ***Coeficiente de Silueta:*** Esta es una medida que cuantifica que tan separados están los clusters en comparación con la distancia media entre los puntos en el mismo cluster. En nuestros resultados arroja un coeficiente de silueta para el conjunto de datos sin escalar de 0.57, mientras que para el conjunto de datos escalado es de 0.25. La diferencia entre los coeficientes de silueta nos muestra que el escalamiento de las caracteristicas tienen un impacto significativo.
- ***Índice Calinski-Harabasz:*** Es la relación entre la varianza de un punto de datos comparado con los puntos de otros clusters, frente a la varianza comparada con los puntos de su cluster. Cuanto mayor sera el valor del índice mejor será la calidad del clustering, en nuestro caso se ha obtenido un índice de 604.1009035590978, significando que hemos obtenido clusters bien separados.

El análisis realizado demuestra que el algoritmo K-Means ha segmentado efectivamente los datos en clusters bien definidos y separados. Las métricas de evaluación validan la calidad del producto obtenido, este enfoque permite una mejor comprensión de las similitudes y diferencias entre las distintas especies de pingüinos.

## REFERENCIAS


[1] Evaluación de Modelos de Aprendizaje No Supervisado | AI Planet (formerly DPhi). (s.f.). AI Planet (formerly DPhi). https://aiplanet.com/learn/unsupervised-learning-es/analisis-y-tecnicas-de-clustering/1624/evaluacion-de-modelos-de-aprendizaje-no-supervisado

[2] *Métricas De Evaluación De Modelos En El Aprendizaje Automático*. (s.f.). DataSource.ai. https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

[3] https://www.researchgate.net/publication/371489018_Swarm_based_automatic_clustering_using_nature_inspired_Emperor_Penguins_Colony_algorithmK. Elissa, "Title of paper if known," no puplicado.

[4] Youssefaboelwafa. (2023, 31 de octubre). *Clustering Penguins Species (K-means Clustering)*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/code/youssefaboelwafa/clustering-penguins-species-k-means-clustering

[5] *Clustering Penguins Species*. (s.f.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species