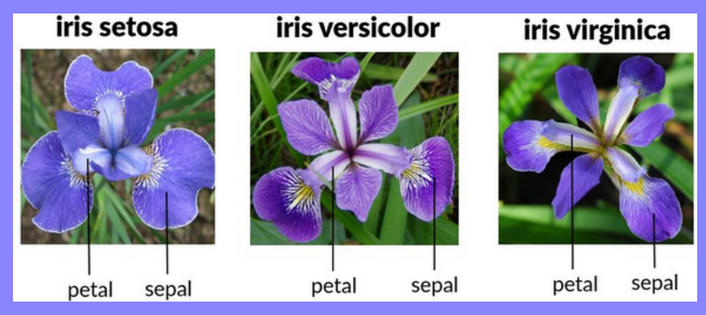
UNSUPERVISED MACHINE LEARNING WITH IRIS DATASET

K-Means Clustering

by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1uipfdSnZYJrdK4kztmKuFOIZhvaVdRIt?usp=sharing



Three classes of IRIS dataset (Mijwil & Abttan, 2021)

INTRODUCTION

IRIS DATASET

Iris dataset is a classic dataset in machine learning and statistics. It classifies three types of iris flowers based on sepal and petal measurements and was originally analyzed by R.A. Fisher in the 1930s (Marsland, 2014).

UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning finds patterns and relationships in data without labeled outcomes. It identifies hidden structures, trends, or groups, making it useful when labeled data is unavailable. This approach helps uncover insights and can complement supervised learning to enhance models. Example of unsupervised machine learning methods:

- Clustering
- Dimensionality reduction
- Anomaly detection
- Association rule-mining

(Sarkar et al., 2017).

RESEARCH PURPOSE

Evaluate the performance of the K-Means clustering model on the Iris dataset.

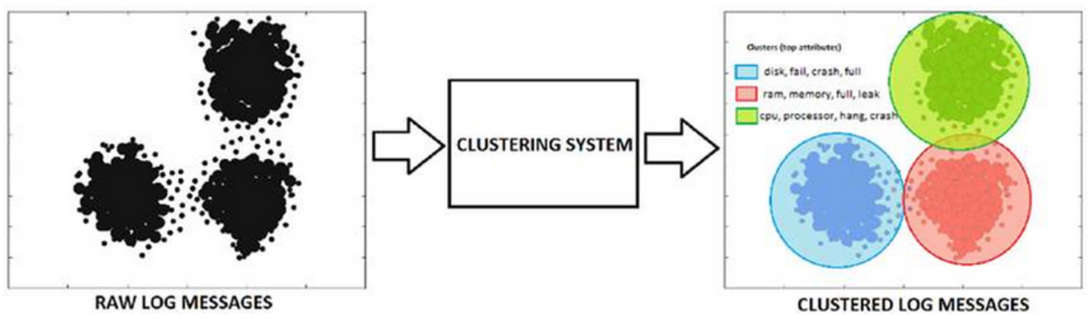
LITERATURE REVIEW

CLUSTERING

Clustering is grouping similar data points into different clusters based on their features. It identifies patterns and relationships in the data without prior labels or supervision.
Types of Clustering Methods:

- Centroid-based (e.g., K-Means, K-Medoids)
- Hierarchical (e.g., Agglomerative, Divisive)
- Distribution-based (e.g., Gaussian Mixture Models)
- Density-based (e.g., DBSCAN, OPTICS)

(Sarkar et al., 2017).



An overview of the working process of Clustering (Sarkar et al., 2017)

K-Means Clustering

K-Means clustering works by grouping data into k clusters (k is chosen by the user).

The algorithm works as follows:

1. k cluster "center" points are created at random locations.
2. For each observation:
   a. The distance between each observation and the k center points is calculated.
   b. The observation is assigned to the cluster of the nearest center point.
3. The center points are moved to the means (i.e., centers) of their respective clusters.
4. Steps 2 and 3 are repeated until no observation changes in cluster membership

K-Means assumes that clusters are roughly circular and of similar size. The key measure used is the distance between each data point and the centroids
(Albon, 2018).

MODEL EVALUATION

Sillhoutte Score

Measures how well data points fit within their assigned cluster compared to other clusters.
The score ranges from -1 to 1:

- Close to 1 → The data point is well-clustered.
- Close to 0 → The data point is between two clusters.
- Close to -1 → The data point may belong to a different cluster.

(Rousseeuw, 1987).

Adjusted Rand Index (ARI)

ARI measures the similarity between the clustering results generated by the algorithm and the original labels, taking into account the possibility of random clustering.
ARI values range from -1 to 1:

- 1 → Perfect clustering according to the original labels.
- 0 → Clustering occurs randomly.
- < 0 → Clustering is worse than random clustering.

(Scikit-learn developers, n.d.)

RESEARCH METHODOLOGY

DATA SOURCE → secondary data from the website https://www.kaggle.com/datasets/uciml/iris, accessed on December 17, 2024.

RESEARCH VARIABLES

Independent Variables (Features):
represented by the symbol "X"

- Sepal Length (cm)
- Sepal Width (cm)
- Petal Length (cm)
- Petal Width (cm)

Dependent Variable (for ARI):
represented by the symbol "y"
Species (Iris-setosa, Iris-versicolor, Iris-virginica)

ANALYSIS STEPS

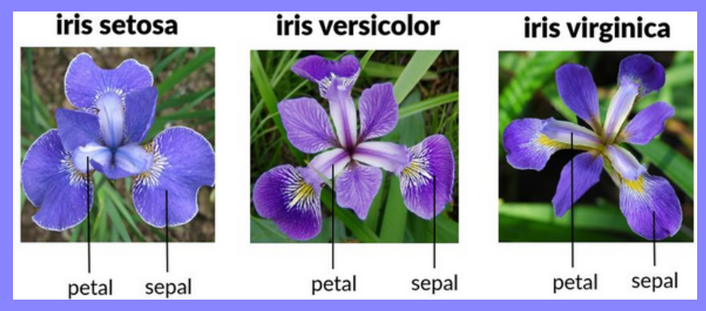
1. Identify the problem and input the dataset.
2. Exploratory Data Analysis.
3. Train the models and evaluate their performance.
4. Summarize the results and draw conclusions.

UNSUPERVISED MACHINE LEARNING WITH IRIS DATASET

K-Means Clustering

by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1uipfdSnZYJrdK4kztmKuFOIZhvaVdRIt?usp=sharing



Three classes of IRIS dataset (Mijwil & Abttan, 2021)

DISCUSSION

1. INPUT THE DATASET

change all column names into lowercase

```
df = pd.read_csv('/content/iris.csv')
```

```
df.columns = df.columns.str.lower().str.replace(' ', '_')
```

```
df.head()
```

	id	sepallengthcm	sepalwidthcm	petallengthcm	petalwidthcm	species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

```
# Divide the dataset into X and y
X = df[['sepallengthcm', 'sepalwidthcm', 'petallengthcm', 'petalwidthcm']]
y = df['species']
```

- X is for k-means clustering
- y is for model evaluation on Adjusted Rand Index (ARI)

2. EXPLORATORY DATA ANALYSIS

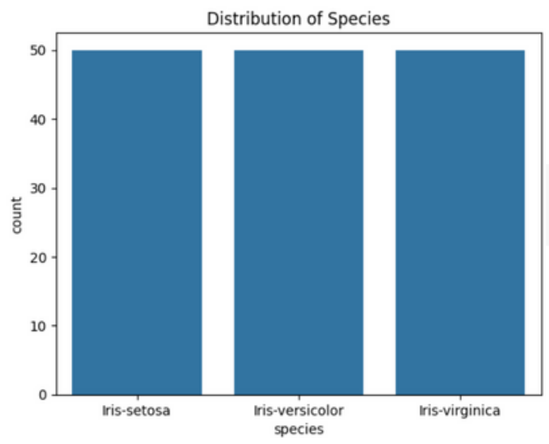
Clean, explore, and visualize the data

```
df.isnull().sum()
# To see the missing values
```

	0
id	0
sepallengthcm	0
sepalwidthcm	0
petallengthcm	0
petalwidthcm	0
species	0

No missing value

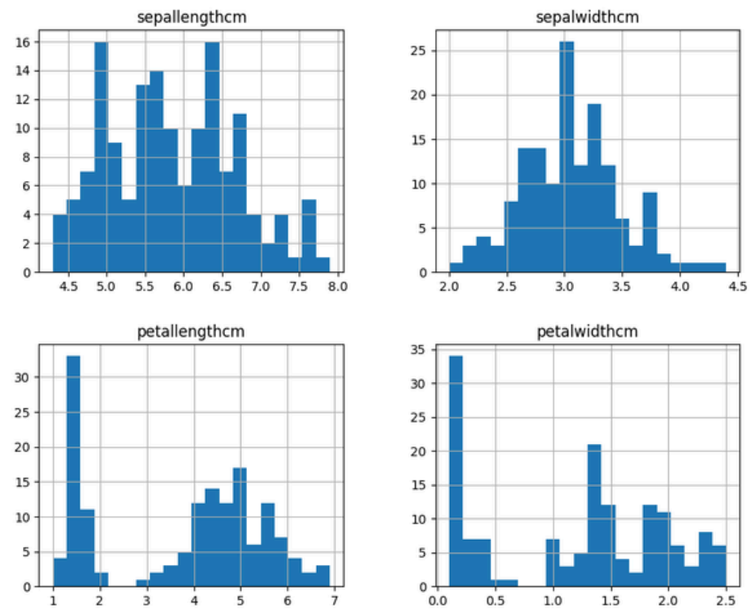
```
# Species distribution
sns.countplot(x='species', data=df)
plt.title('Distribution of Species')
plt.show()
```



the amount of data between species is evenly distributed

```
# Species distribution
print(df['species'].value_counts())
```

```
species
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: count, dtype: int64
```

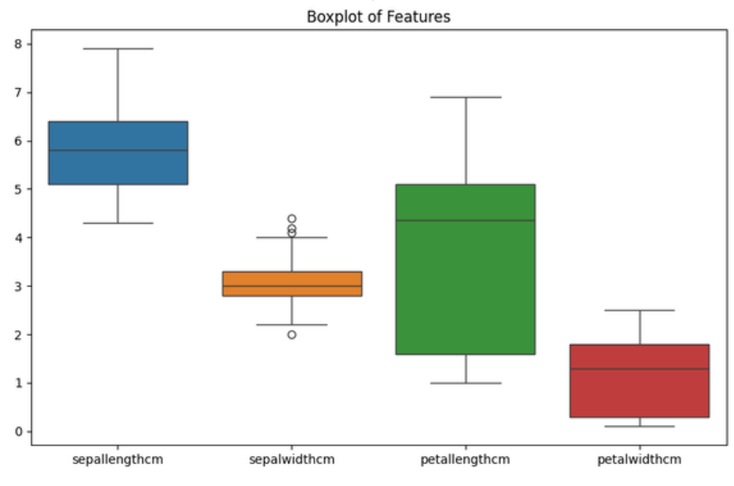


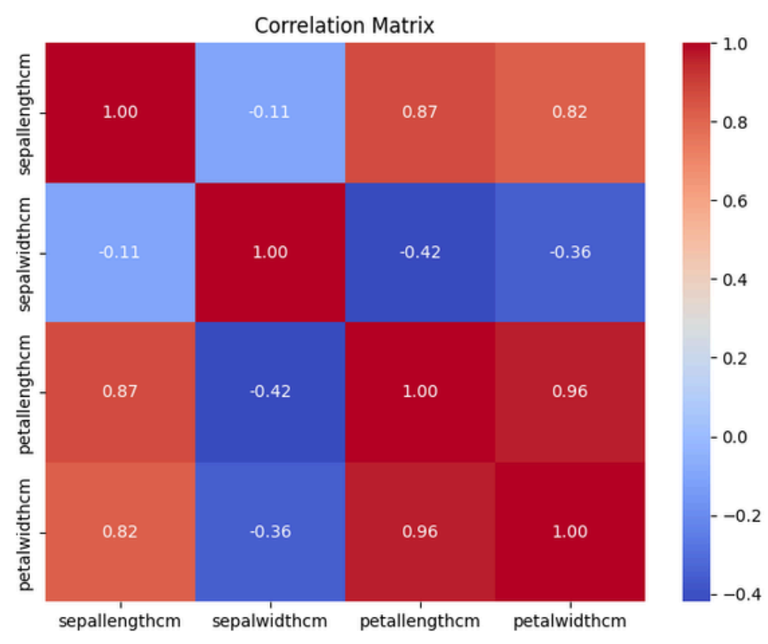
Data Distribution & Outlier Detection

```
X.describe()
```

	sepallengthcm	sepalwidthcm	petallengthcm	petalwidthcm
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Outliers can be ignored





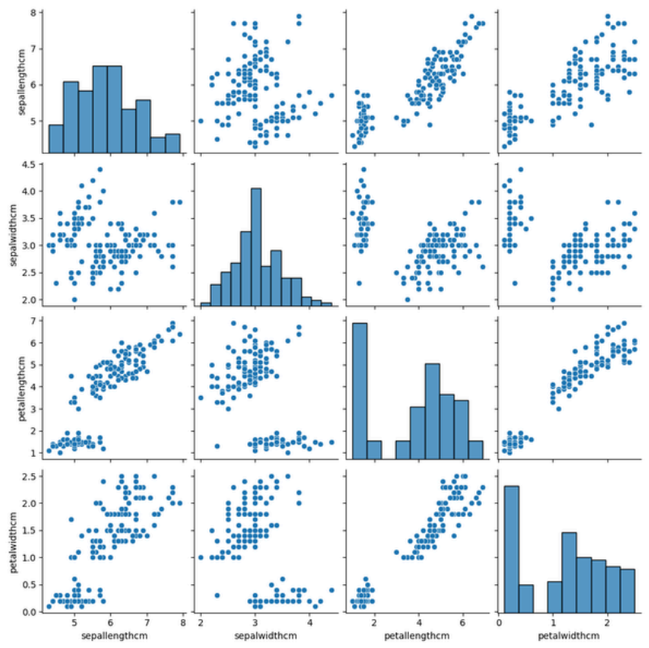
Correlation between features

Correlation Matrix shows the strength of the relationship between features.

- a number close to 1 or -1 → have strong correlation
- 0 → don't have a correlation

(Sarkar et al., 2017)

Features that have strong correlation
- petallengthcm & petalwidthcm (0.96)
- sepallengthcm & petallengthcm (0.87)
- sepallengthcm & petalwidthcm (0.82)

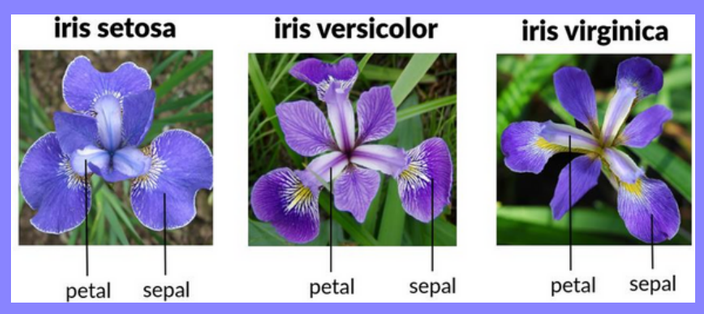


UNSUPERVISED MACHINE LEARNING WITH IRIS DATASET

K-Means Clustering

by: Raniah Mufidah Admayana

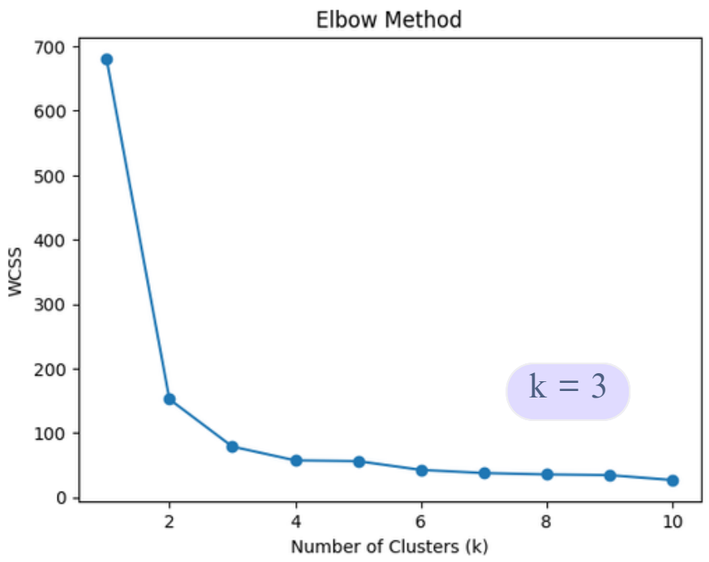
Code in Google Colab = https://colab.research.google.com/drive/1uipfdSnZYJrdK4kztmKuFOIZhvaVdRIt?usp=sharing



Three classes of IRIS dataset (Mijwil & Abttan, 2021)

3. K-MEANS CLUSTERING

1. Search the Best Number of Cluster (k)



- The elbow method is a technique used to find the best k in K-Means clustering by measuring WCSS (Within-Cluster Sum-of-Squares), which represents the variance within each cluster.
- A lower WCSS means better clustering.
- As k increases, WCSS decreases, but at some point, the drop slows down. This turning point, called the "elbow," is considered the optimal k value (Cui, 2020).

- k = 3 because it marks the point where the decrease in WCSS starts to slow down.
- Additionally, k = 3 is chosen to create three clusters, representing the three Iris species in the dataset.

2. Cluster the Data

```
kmeans = KMeans(n_clusters=3, random_state=1)
kmeans.fit(X)

cluster_labels = kmeans.labels_
print(cluster_labels)
```

```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 2 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 0 2 2 2 2
 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2
 2 0]
```

```
# Add cluster results to original DataFrame `df`
df['cluster'] = cluster_labels
# Add cluster results into DataFrame
X['cluster'] = cluster_labels
```

```
ct = pd.crosstab(df['species'], df['cluster'])
print(ct)
```

```
cluster           0   1   2
species
Iris-setosa       0  50   0
Iris-versicolor  47   0   3
Iris-virginica   14   0  36
```

So from this k-means clustering:
- Cluster 0 = iris-versicolor
- Cluster 1 = iris-setosa
- Cluster 2 = iris-virginica

This code tell us that:
- All 50 Iris-setosa flowers were assigned to Cluster 0
- 47 Iris-versicolor flowers were in Cluster 1, and 3 were in Cluster 2
- 36 Iris-virginica flowers were in Cluster 2, and 14 were in Cluster 1

```
silhouette_avg = silhouette_score(X, cluster_labels)
print(f'Silhouette Score: {silhouette_avg:.2f}')
```

```
Silhouette Score: 0.68
```

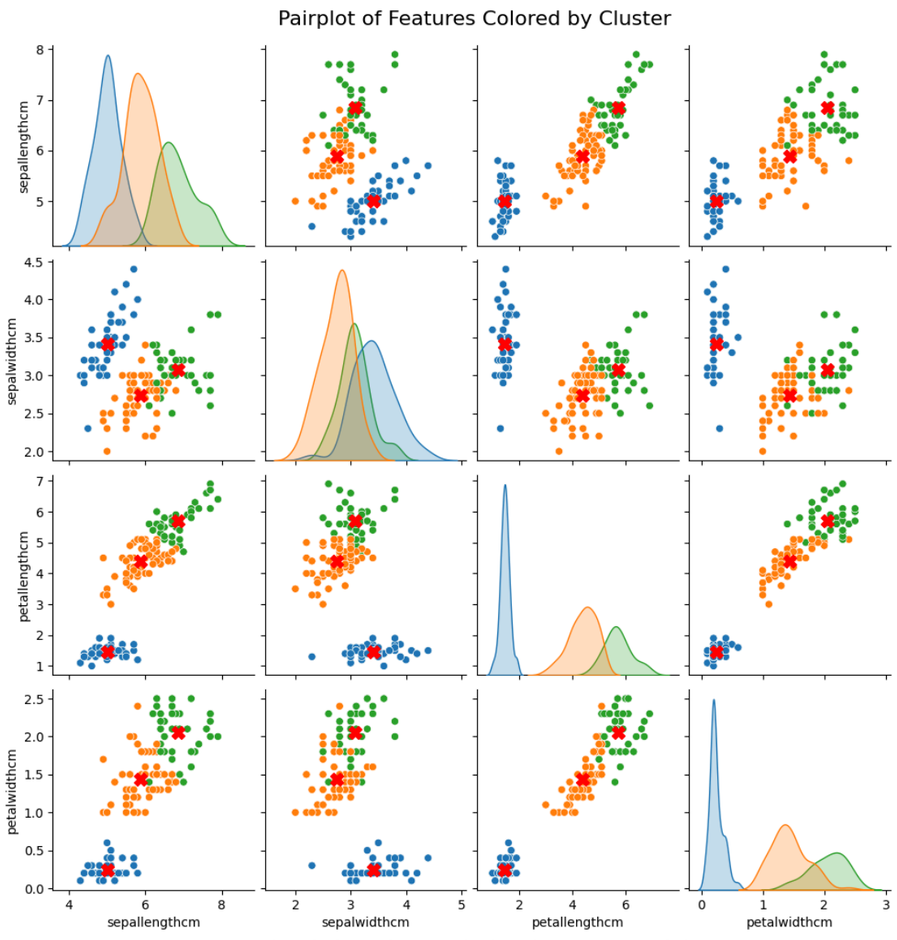
Silhouette Score = 0.68

```
from sklearn.metrics import adjusted_rand_score
ari = adjusted_rand_score(y, df['cluster'])
print(f'Adjusted Rand Index: {ari:.2f}')
```

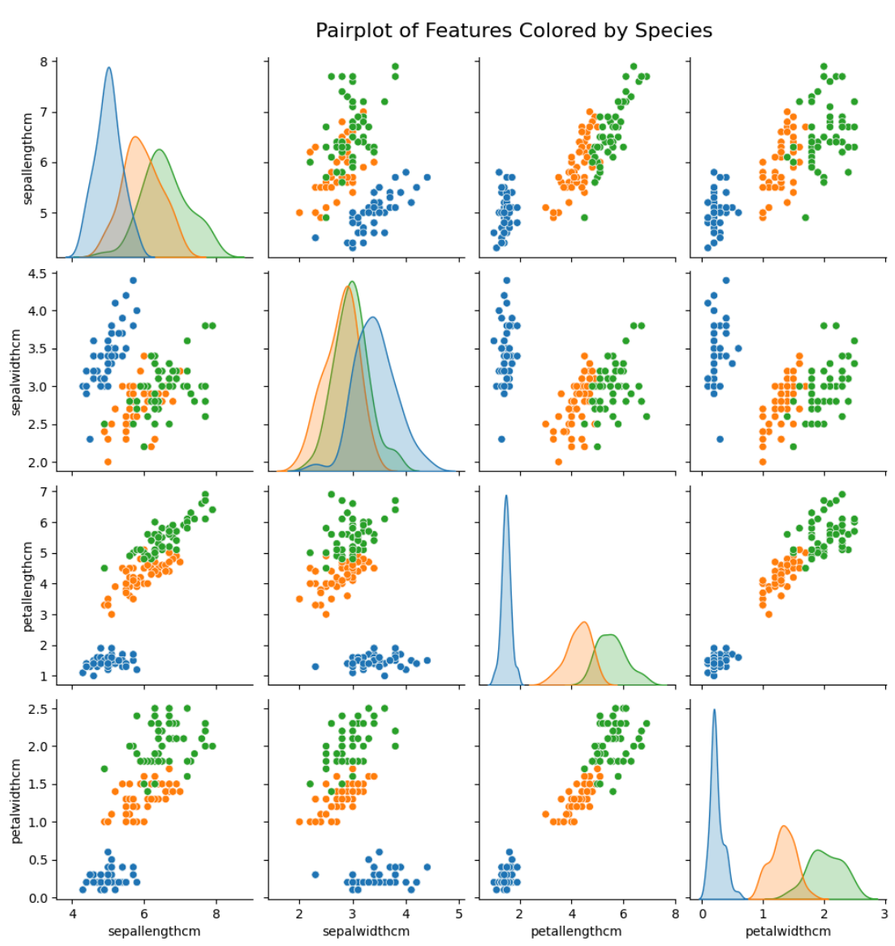
```
Adjusted Rand Index: 0.72
```

Adjusted Rand Index = 0.72

3. Visualization - Pairplot of Features Colored by Cluster vs Species



the 'x' red mark is the centroid



CONCLUSION

The K-Means clustering model applied to the Iris dataset achieved:
- Silhouette Score: 0.68, indicating well-defined clusters.
- Adjusted Rand Index (ARI): 0.72, showing a strong match with the true labels.

CODE IN COLAB

Google Colab
google.com
https://colab.research.google.com/drive/1uipfdSnZYJrdK4kztmKuFOIZhvaVdRIt?usp=sharing

REFERENCE

Albon, C. (2018). Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning.
Cui, M. (2020). Introduction to the K-Means clustering algorithm based on the Elbow Method. Accounting, Auditing and Finance, 1, 5–8. https://doi.org/10.23977/accaf.2020.010102
Marsland, S. (2014). Machine learning: An Algorithmic Perspective, Second Edition. CRC Press.
Mijwil, M. M., & Abttan, R. A. (2021). Utilizing the genetic algorithm to pruning the C4.5 decision tree algorithm. Asian Journal of Applied Sciences, 9(1). https://doi.org/10.24203/ajas.v9i1.6503
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Sarkar, D., Bali, R., & Sharma, T. (2017). Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems. Apress.
Scikit-learn developers. (n.d.). *sklearn.metrics.adjusted_rand_score*. Scikit-learn. Retrieved February 21, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html