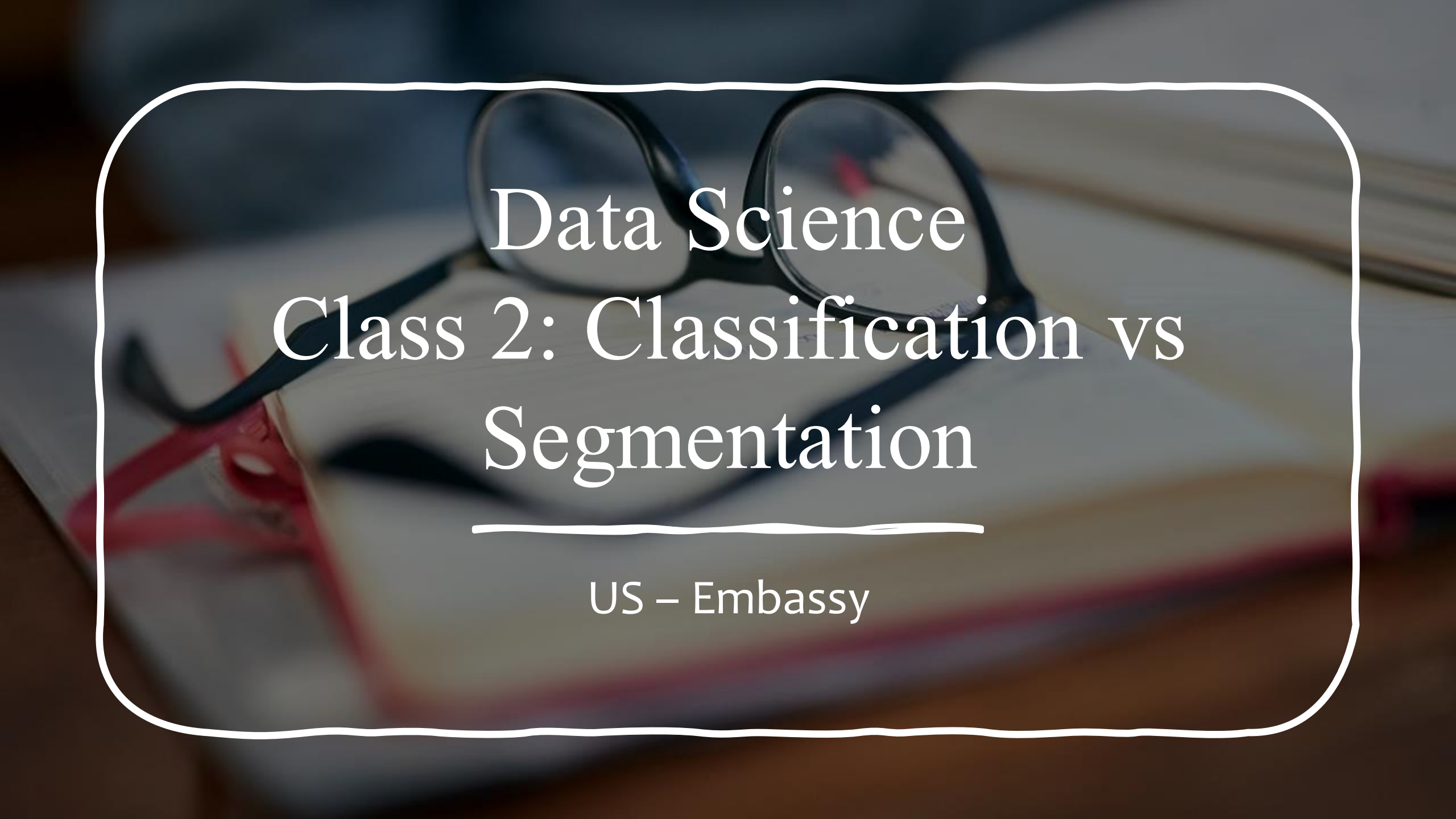
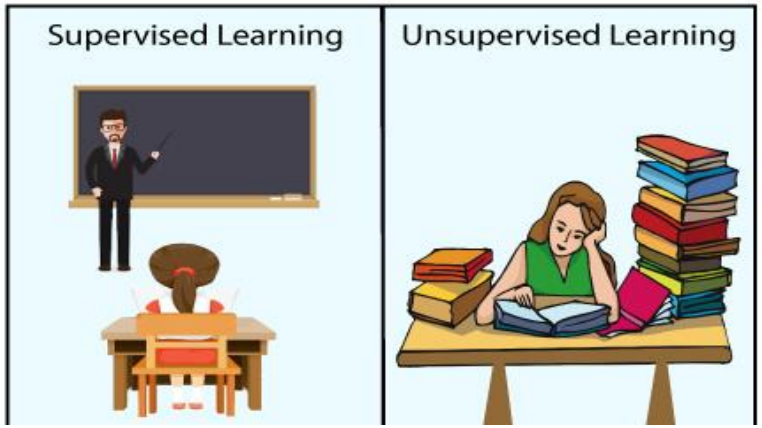
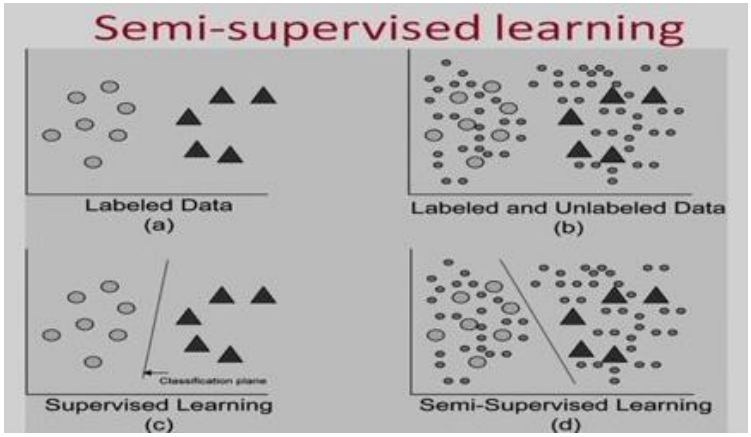


Data Science
Class 2: Classification vs Segmentation

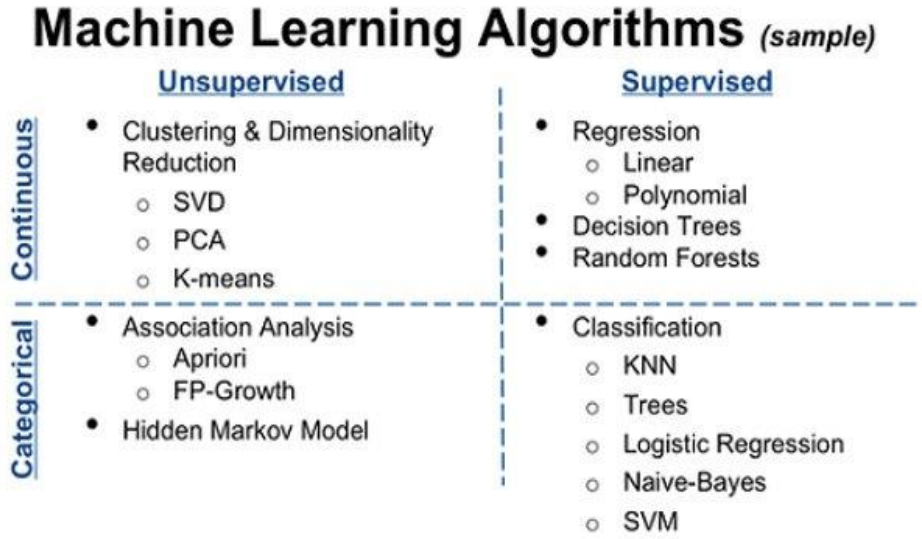
US – Embassy

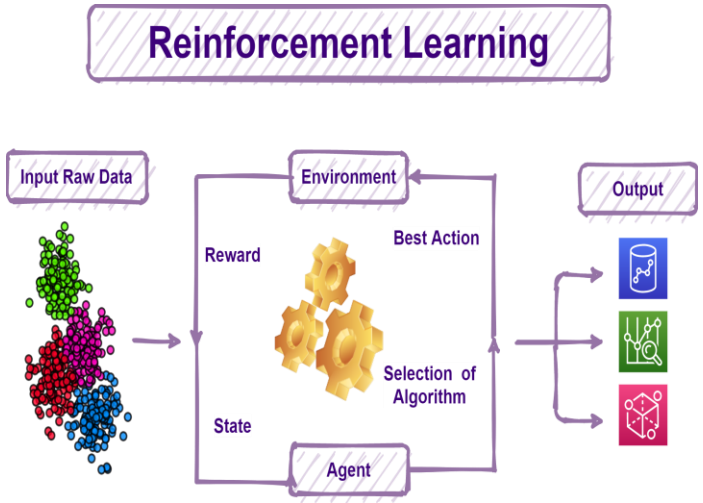
Type of Machine Learning and comparison





We have many type of learning, such as



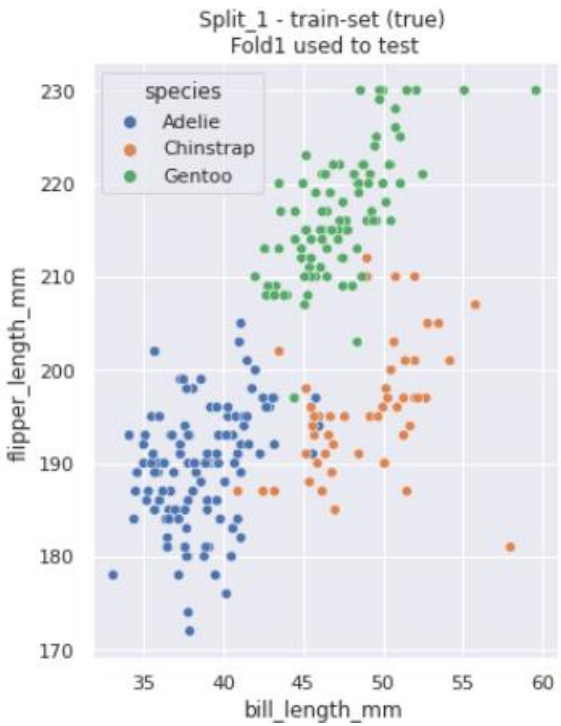


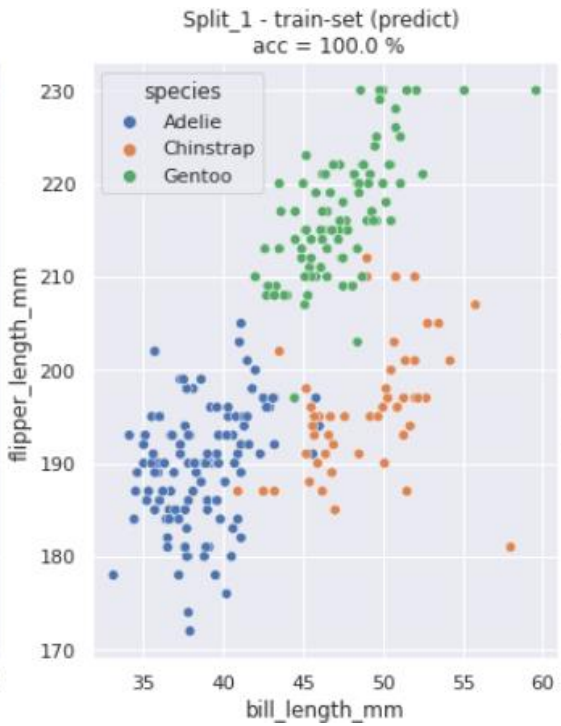
1. Supervised Learning.

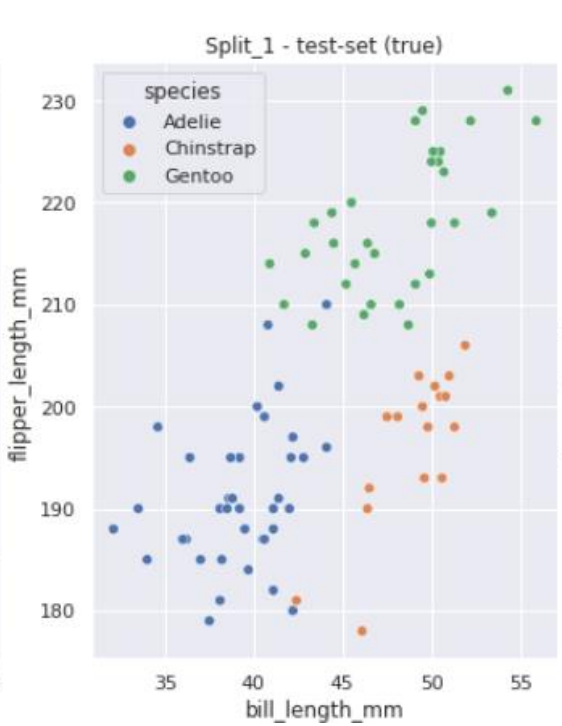
- In the previous lesson, we had studied regression, today we will discuss

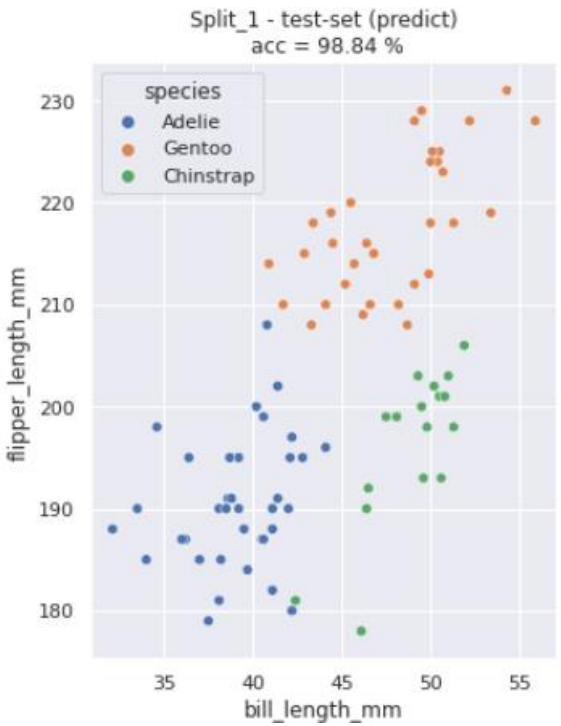
Classification

- The image below is an example taken from [my-results]

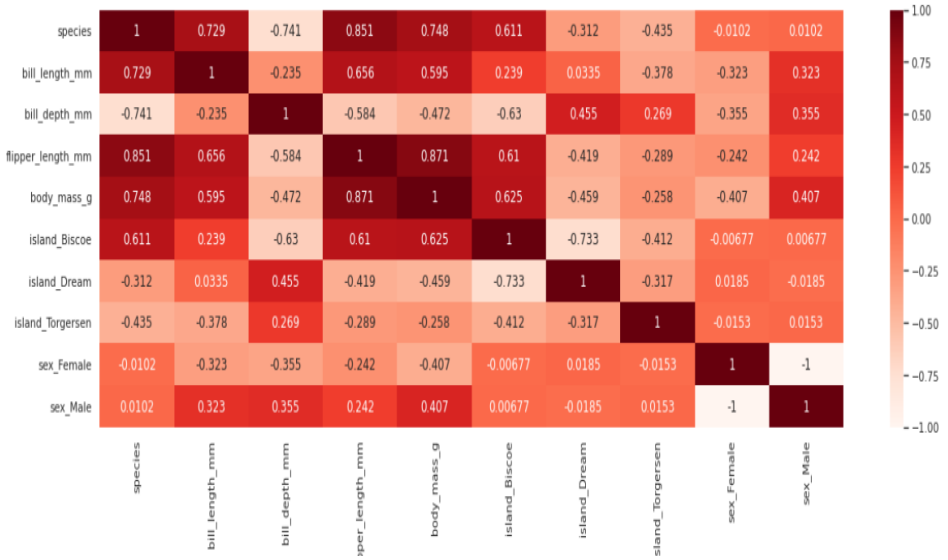


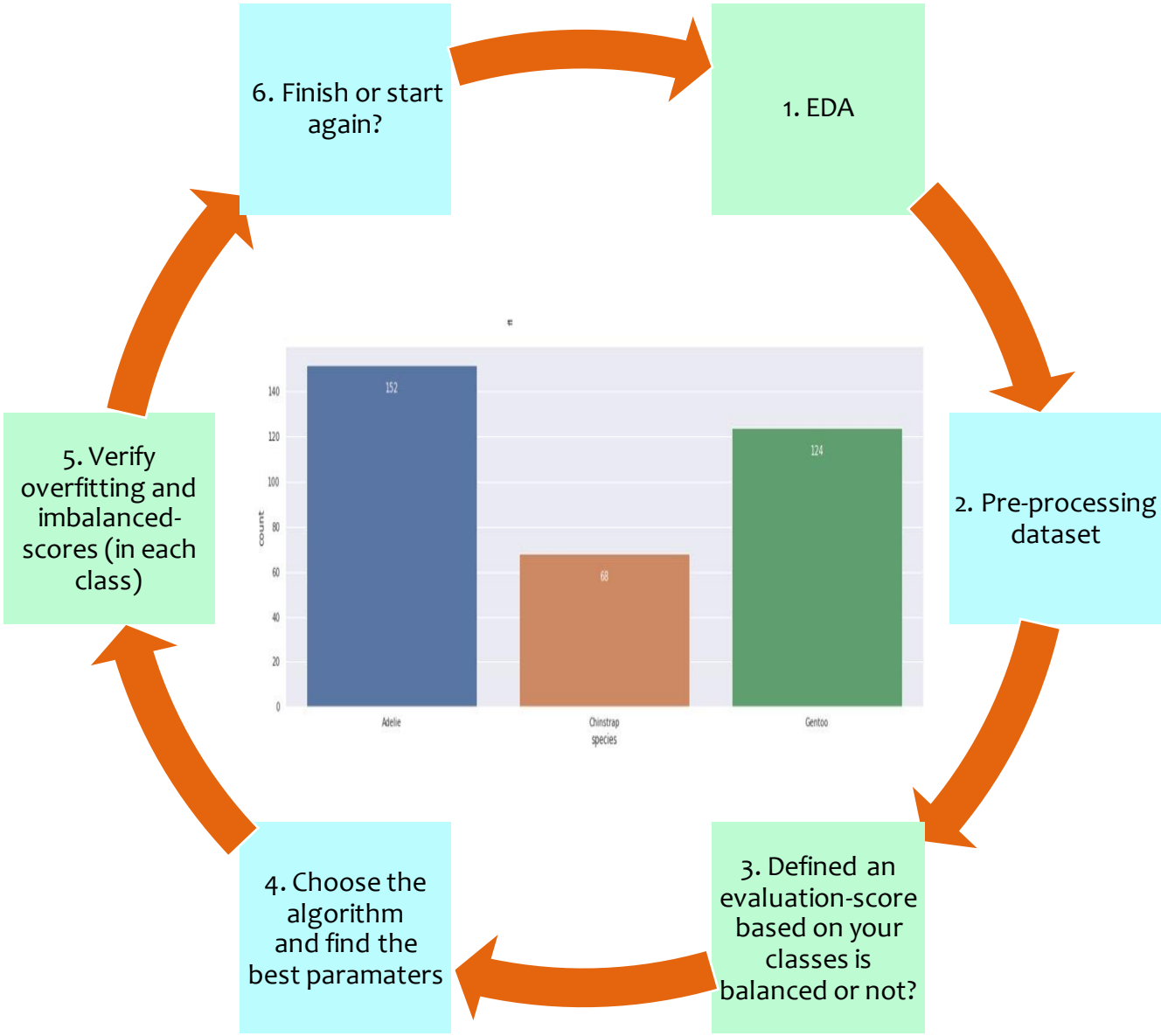






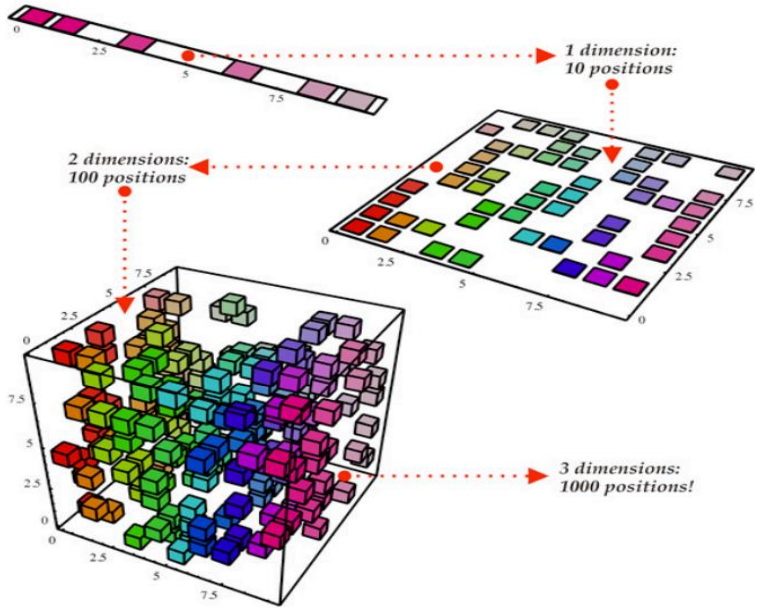
Process in classification





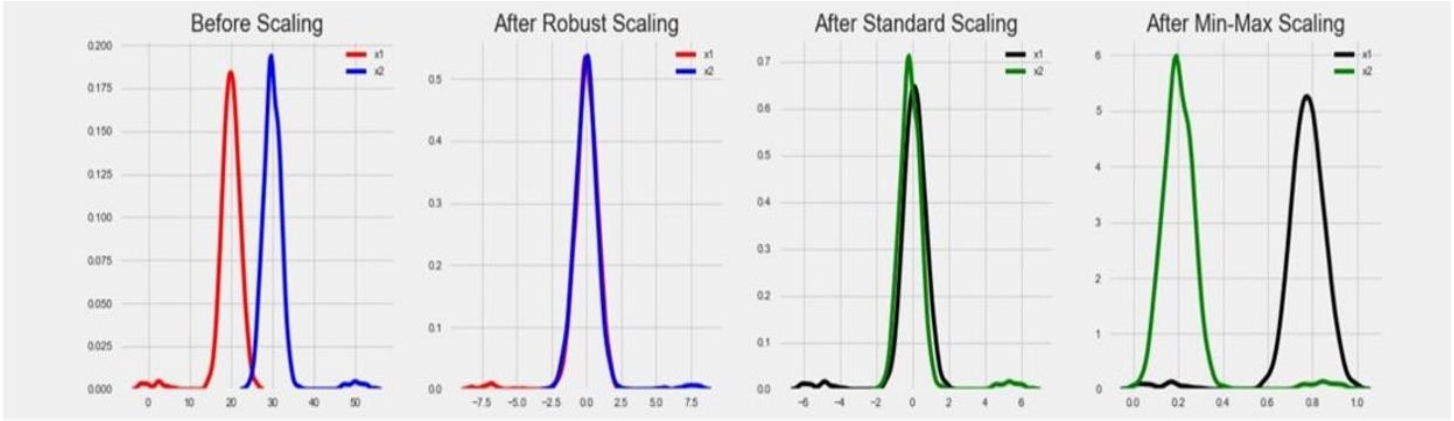
EDA vs Pre-processing

- Visualization the data-relationship
- Dummy-Encoding at category-features
- Impute with missing values
- Apply StandardScaler / MinMax Scaler or not?
- Apply PCA or not?



Human-Readable → Machine-Readable

Pet	Cat	Dog	Turtle	Fish
Cat	1	0	0	0
Dog	0	1	0	0
Turtle	0	0	1	0
Fish	0	0	0	1
Cat	1	0	0	0



Robust Scaler

$$\frac{x_i - Q_1(\boldsymbol{x})}{Q_3(\boldsymbol{x}) - Q_1(\boldsymbol{x})}$$

Standard Scaler

$$\frac{x_i - \text{mean}(\boldsymbol{x})}{\text{stdev}(\boldsymbol{x})}$$

MinMax Scaler

$$\frac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$$

Scores and algorithms

ML Algorithms

- - 1. Logistic Regression
- - 2. Naïve Bayes
- - 3. SGD
- - 4. K-Nearest Neighbours
- - 5. Decision Trees
- - 6. Random Forest
- - 7. Extra Trees
- - 8. SVM
- - 9. AdaBoost Classifier
- - 10. XGBoost Classifier

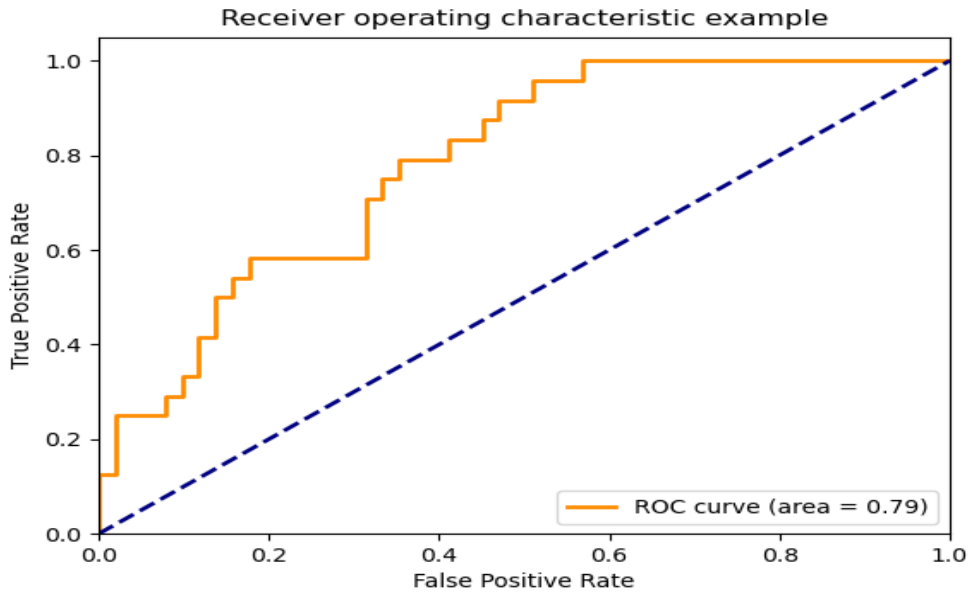
DL models

- 1. ANN
- 2. CNN
- 3. Transfer Learning
  - ResNet
  - Efficience Net
  - VGG16
  - VGG19
  - UNet
  - Bottleneck

Confusion matrix

Actual Class \ Predicted Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

ROC-AUC



2. Unsupervised Learning

In this sections, we only focus on PCA and KMeans-segmentations.
First of all, we will focus on PCA

Principal Component Analysis (PCA)

$$\max_{\delta \mathbf{X}} \mathrm{Var}(\delta \mathbf{X}) = \max_{\delta \mathbf{X}} \delta^T \Sigma \delta$$

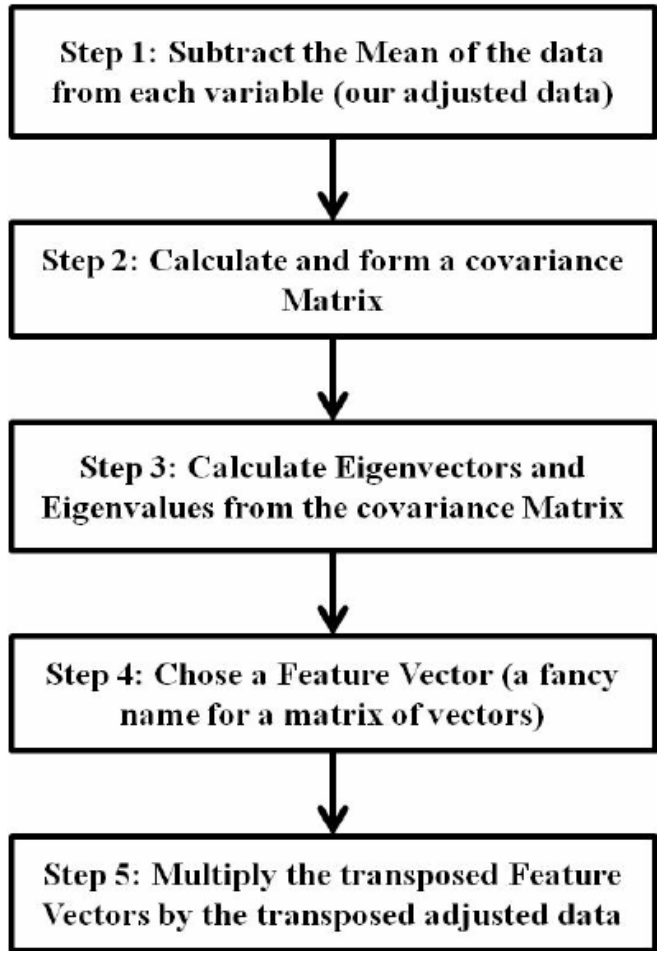
$$\begin{cases} (P) & : & \delta = \underset{\delta}{\mathrm{argmax}}(\delta^T \Sigma \delta) \\ \text{s.t} & : & \| \delta \|^2 = 1 \end{cases}$$

$$\mathbf{Y} = \Gamma^T (\mathbf{X} - \mu_{\mathbf{X}}).$$

$$\mathbb{E}\mathbf{Y} = \mathbf{0} \text{ and } \mathrm{Var}\mathbf{Y} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_d \end{pmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d.$$

References

PCA_sklearn_python



Segmentation and applications

Market segmentation

- 1. Behavioral
- 2. Demographic
- 3. Psychographic
- 4. Geographic

Others.

- Transactional segmentation
- Technographic segmentation
- etc.

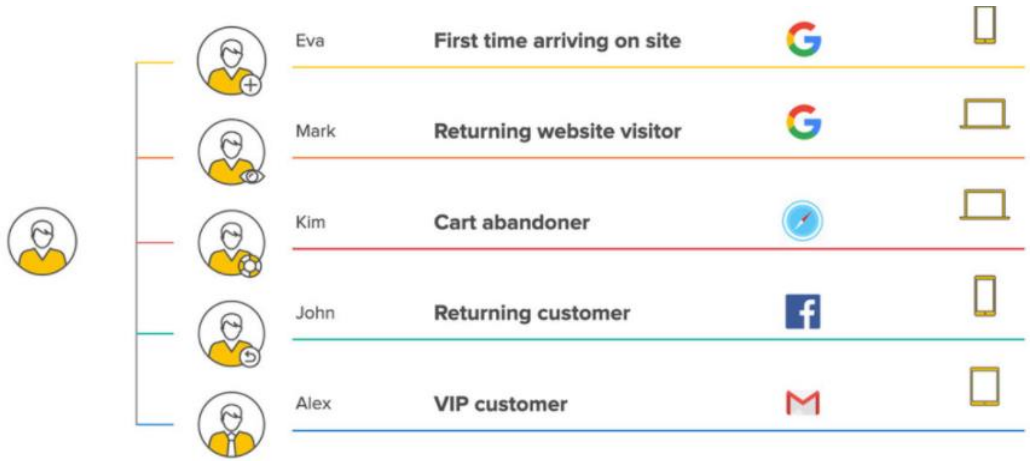
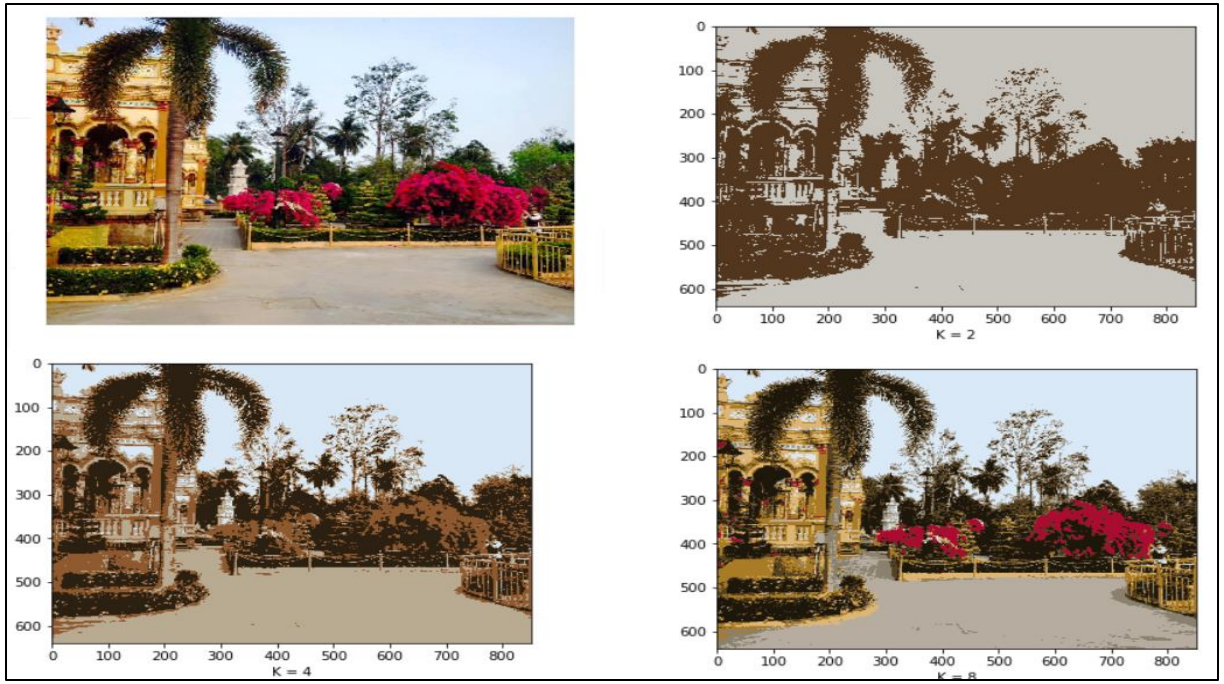


Image segmentation

(compressed image)

References:

1. Market segmentation
2. Image segmentation



Motivation & benifits

Do you know why we focused on market-segmentation (then verify by using classification)?
By segmenting your market you'll be able to understand your customer's needs better and how you can fulfill these better than your competition.

1) More effective marketing

2) More efficient spending

3) Higher quality leads

4) Identifying niche markets

5) Improved customer retention

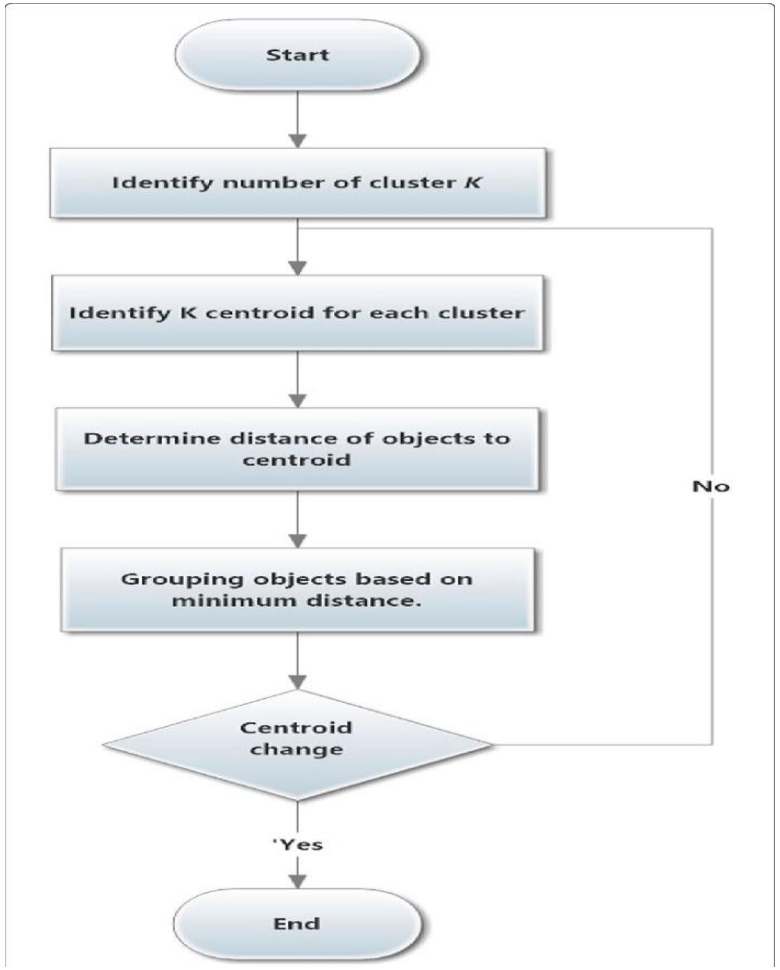
6) Differentiating your brand

7) More focus

Some useful clustering-algorithms

- K-Means : reference
- Mini batch K-Means : reference
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering : reference, code
- Spectral Clustering
- Gaussian Mixture Model
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
- etc.

K-Means clustering algorithm



How to find the centers and the label vectors

- **1) Find Y (label), for fixed M (center)** Assume that we have found the `centers`, then determine the `label vectors` to minimize the loss function

$$\mathbf{y}_i^* = \underset{\mathbf{y}_i \in \mathbf{Y}}{\operatorname{argmin}} \sum_{j=1}^{K} y_{ij} \| x_i - m_j \|^2 ,$$

by the condition on the `label vector`, we obtain

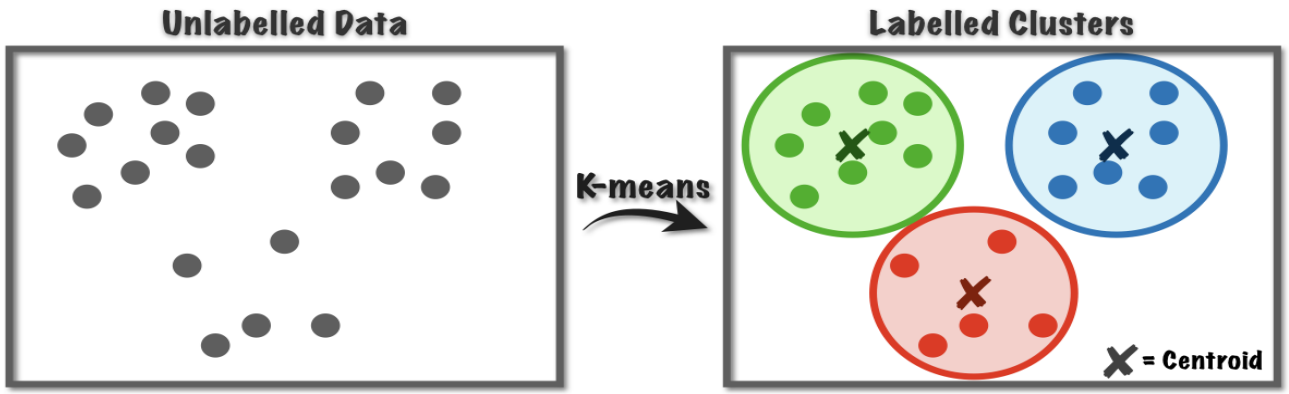
$$j^* = \underset{1 \le j \le K}{\operatorname{argmin}} \| x_i - m_j \|^2$$

- **2) Find M, for fixed Y** Now, suppose we found the label vectors, then

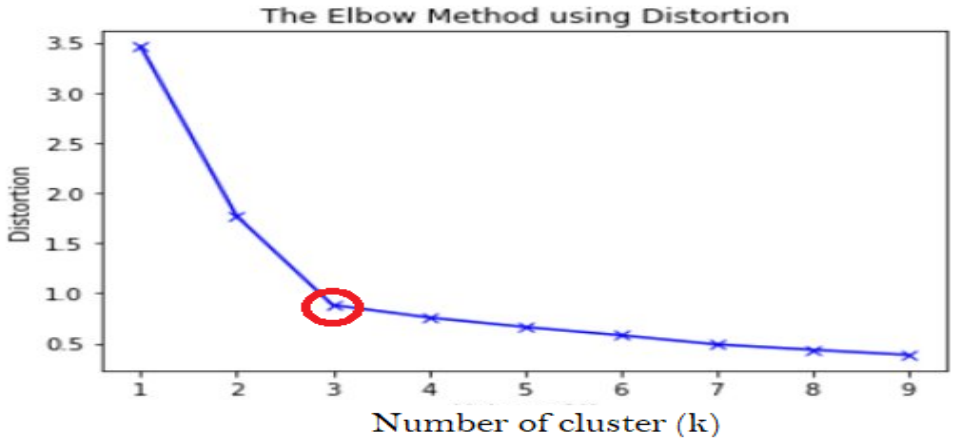
$$\mathbf{m}_j^* = \underset{\mathbf{m}_j \in \mathbf{M}}{\operatorname{argmin}} \sum_{i=1}^{N} y_{ij} \| x_i - m_j \|^2$$

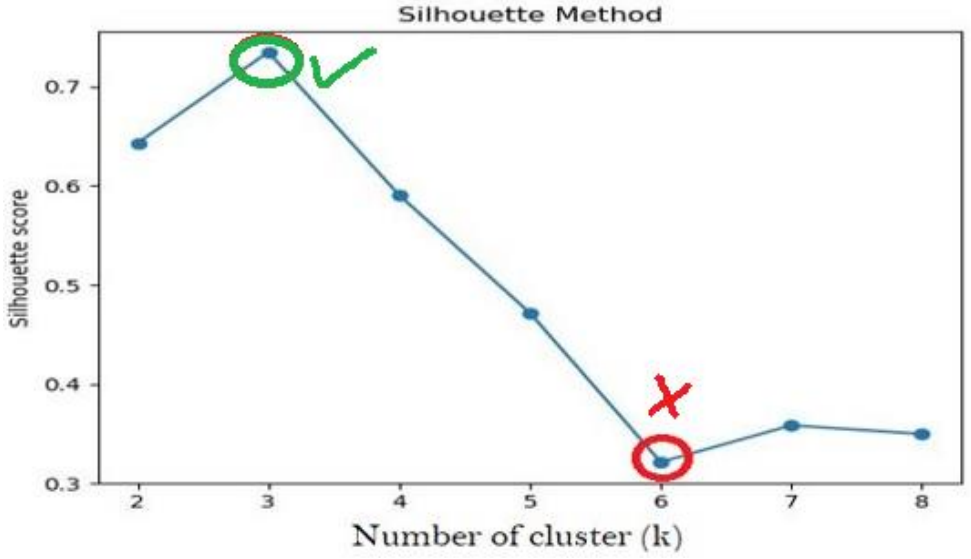
and hence, we get

$$\mathbf{m}_j^* = \frac{\sum_{i=1}^{N} y_{ij} \mathbf{x}_i}{\sum_{i=1}^{N} y_{ij}}$$



Elbow & Silhouette methods





$$L(\mathbf{Y}, \mathbf{M}) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \| x_i - m_j \|^2$$

$\mathbf{y}_k = (y_{k1}, \ldots, y_{kK})$ for $k \in 1, \ldots, K$ is the label of the cluster k^{th}.

$$\sum_{j=1}^{K} y_{ij} = 1, \quad \forall\, i \in 1, \ldots, N.$$

N is the number of observation
K is number of clusters
$X = (X_1, \ldots, X_d)$ is datapoint,
m_j is center j^{th}

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

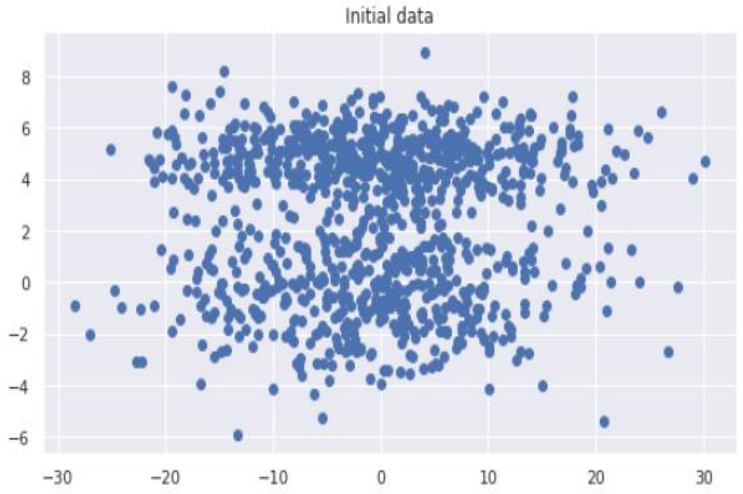
For each data point $i \in C_i$ (data point i in the cluster C_i), let

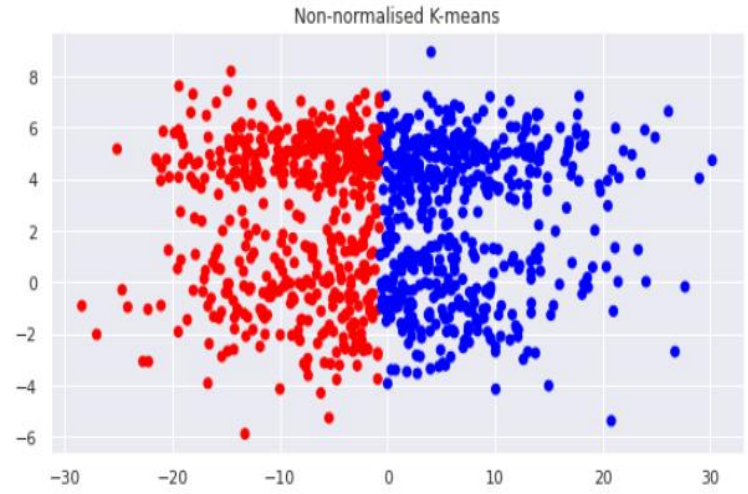
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

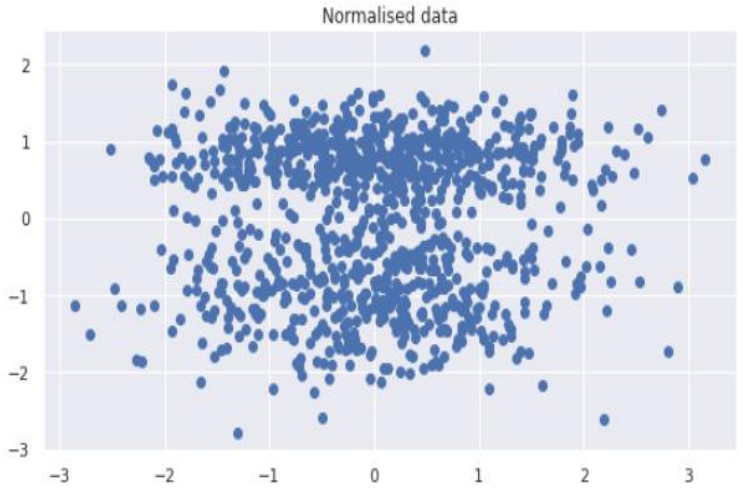
For each data point $i \in C_i$, we now define

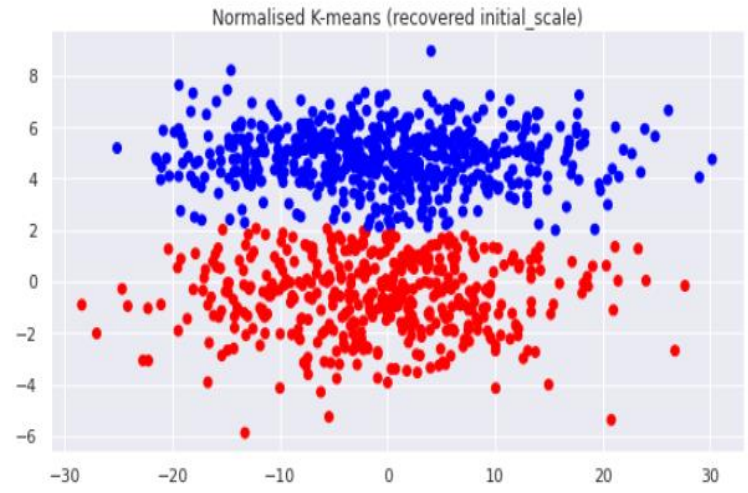
$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

Importance of anomallies-removal and scaling in clustering









Verify by using classification

Clustering is done on unlabelled data returning a label for each datapoint. Classification requires labels. Apply classification algorithms into your clusters-models

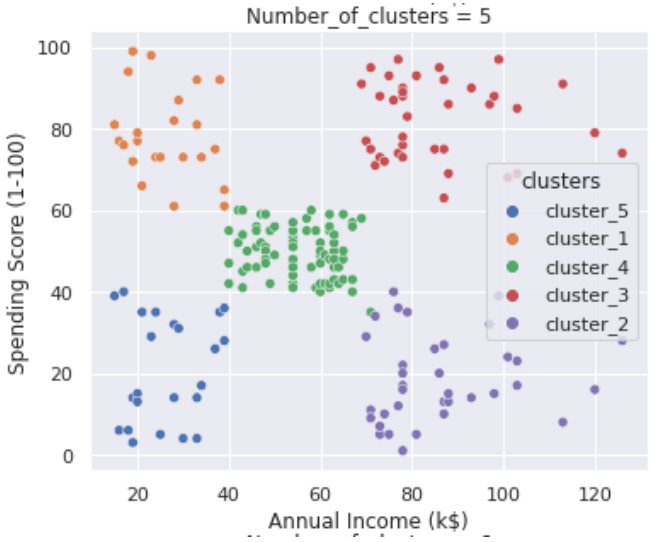
- Initially, check the quantity in each cluster is balanced or not?

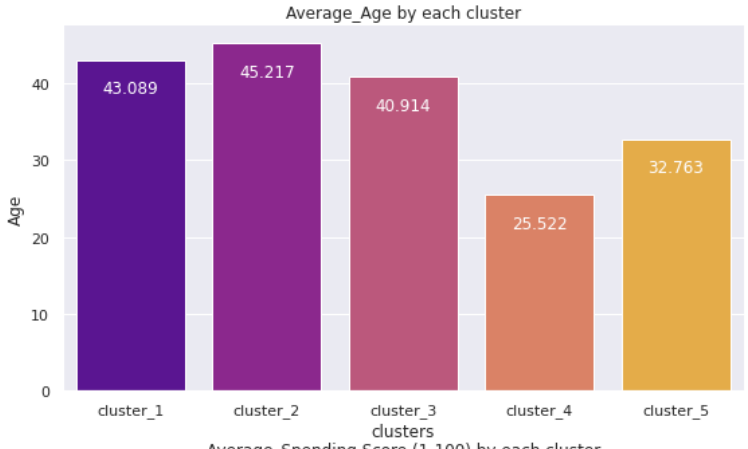
- Then, apply Logistic Regression, SVM, Naive Bayes, XGBoost, Random Forest, etc. into your model.

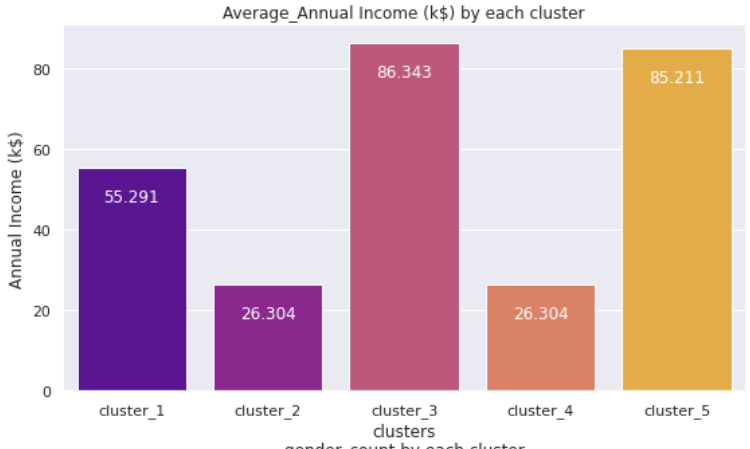
- Evaluate the metrics at each cluster

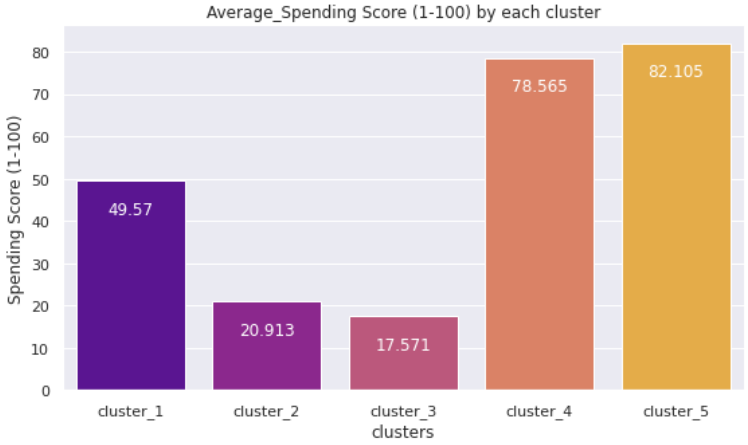
Defined clusters

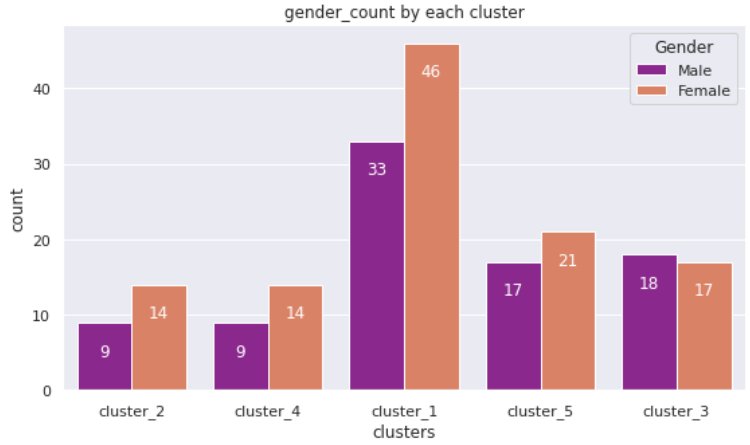
Demographic segmentation











Summary

- Suppervised vs Unsupervised Learning
- Classification and application
- PCA
- Segmentation and its applications
- K-Means clustering algorithm
- Metrics evaluation in classification vs segmentation

Thank for your considerations