DOUGLAS COLLEGE

FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

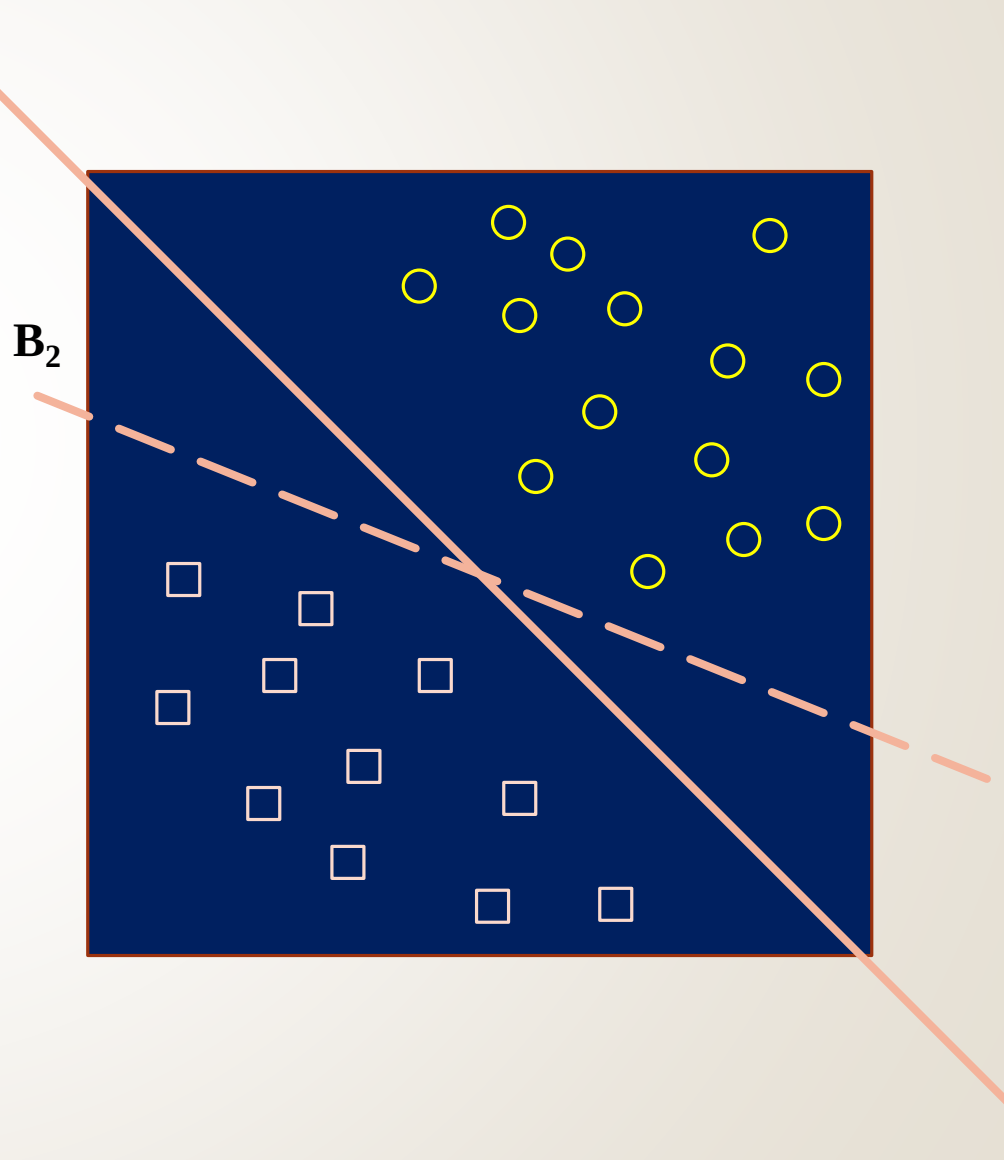
CLASSIFICATION – PART 2

SVM, KNN, NAÏVE BAYES

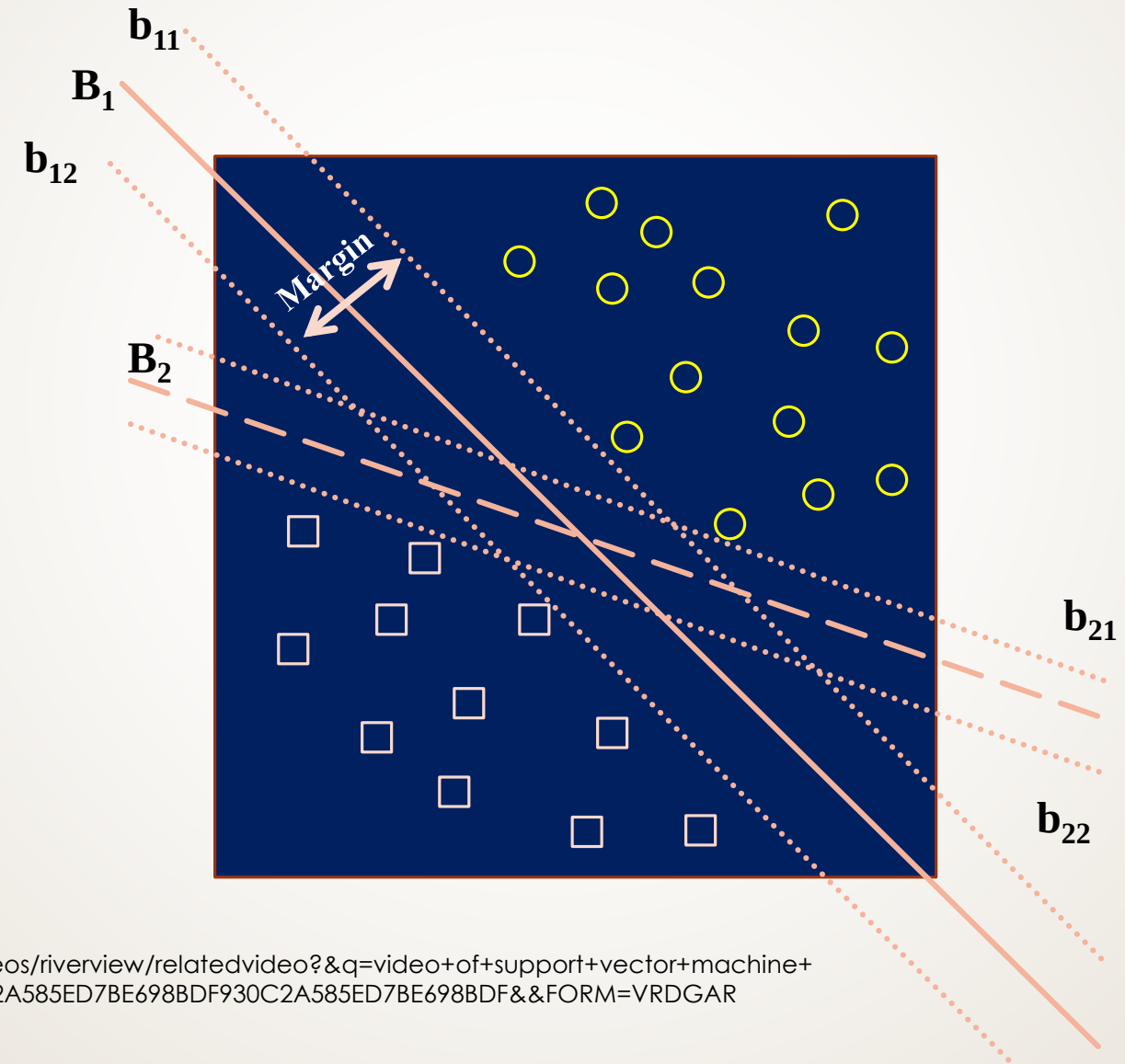
FATEMEH AHMADI

SVM: Support Vector Machine

- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- This best decision boundary is called a **hyperplane**.
- SVM chooses the extreme points/vectors that help in creating the hyperplane.
- These extreme cases are called as **support vectors**, and hence algorithm is termed as **Support Vector Machine**.
- Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

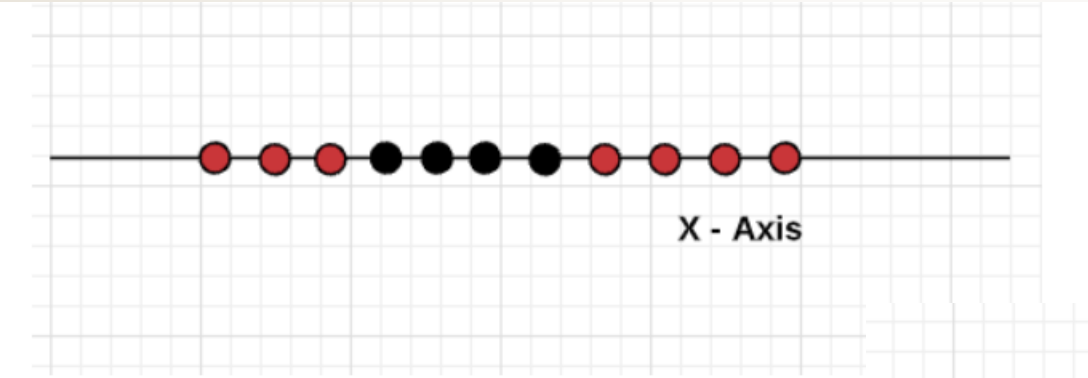


SVM: Support Vector Machine



https://www.bing.com/videos/riverview/relatedvideo?&q=video+of+support+vector+machine+how+it+works&&mid=930C2A585ED7BE698BDF930C2A585ED7BE698BDF&&FORM=VRDGAR

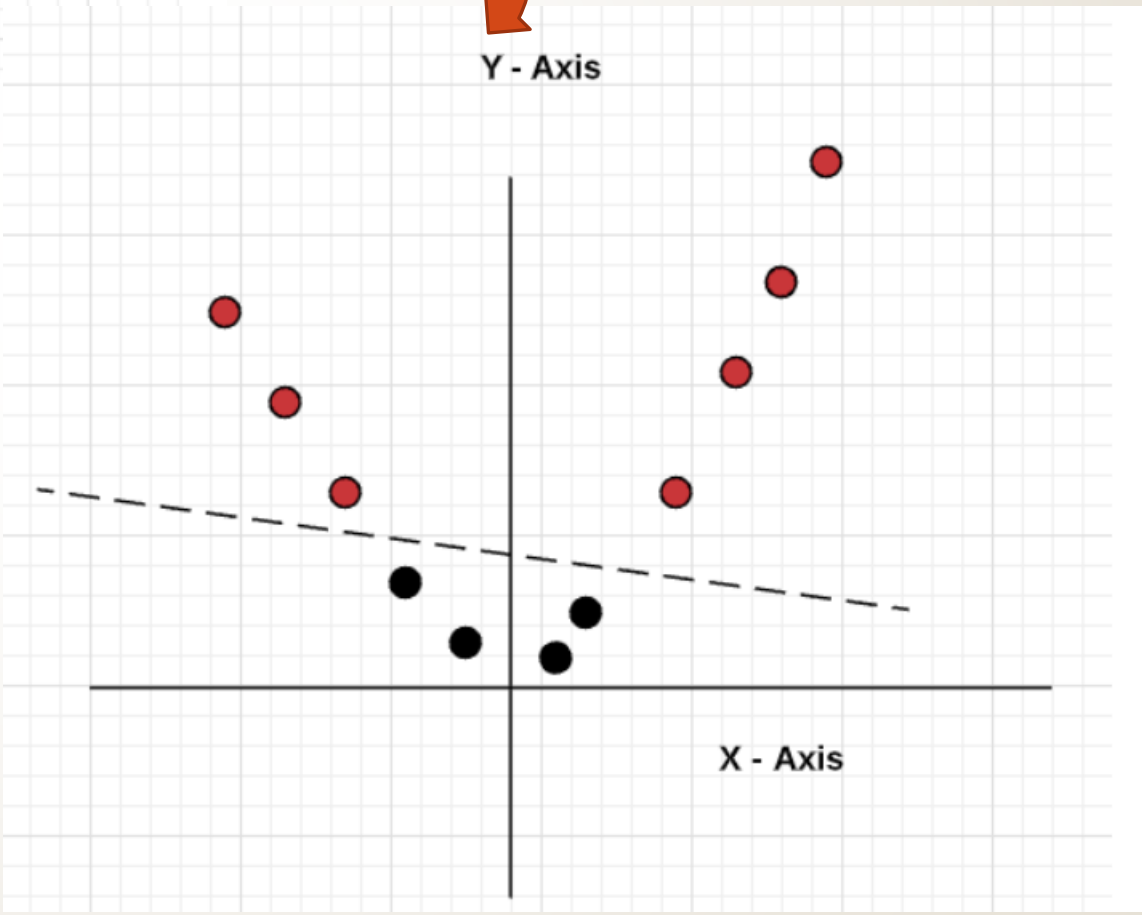
Using Kernel Functions



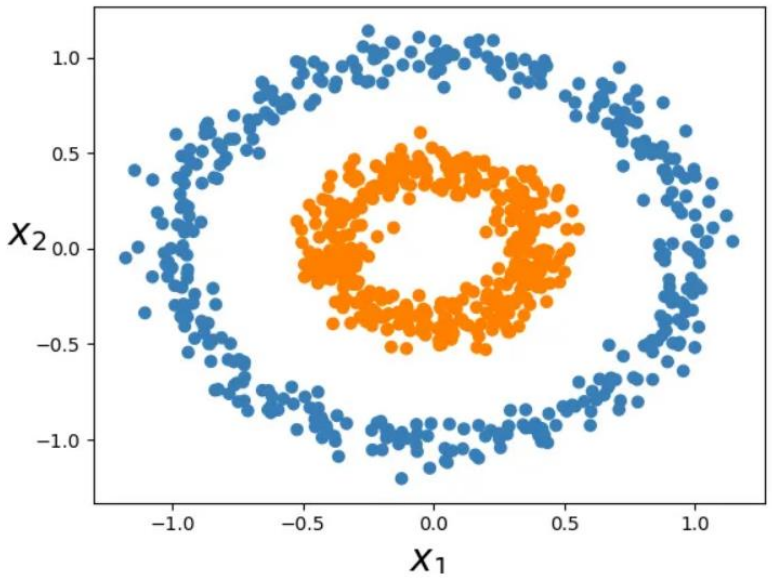
Adding another dimension to the data by using the function for example:

Y = X^2 (X-squared)

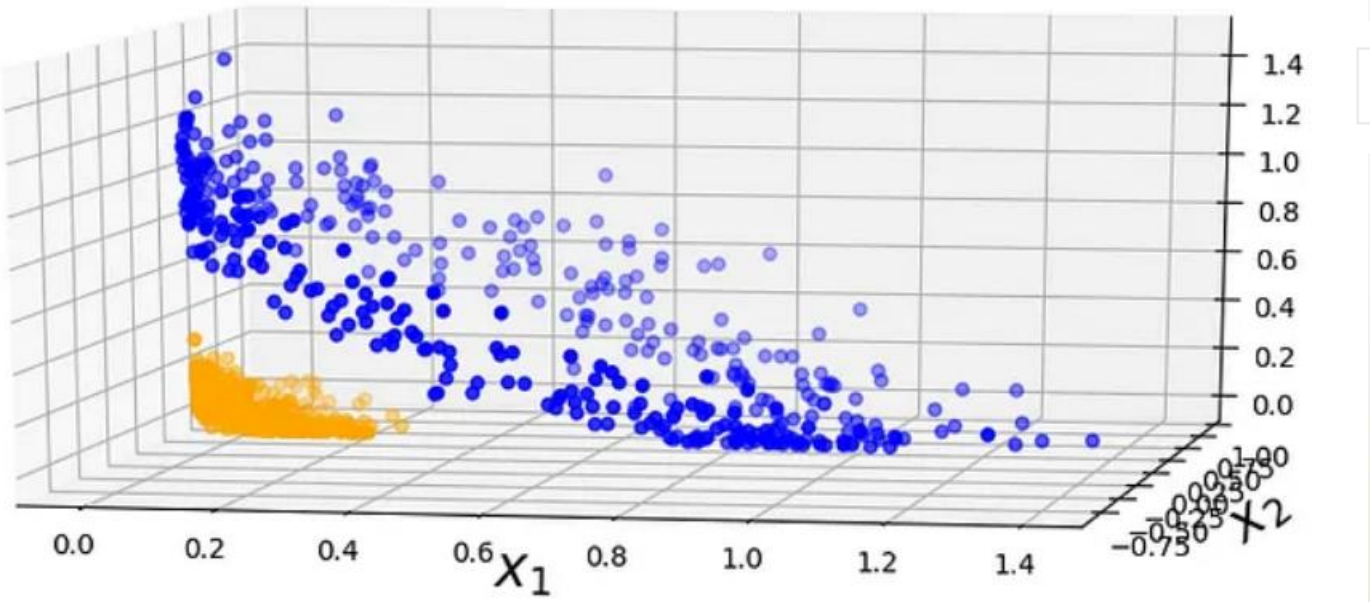
Now the data points can be linearly separable.



Using Kernel Functions



$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$



$$z = x^2 + y^2$$

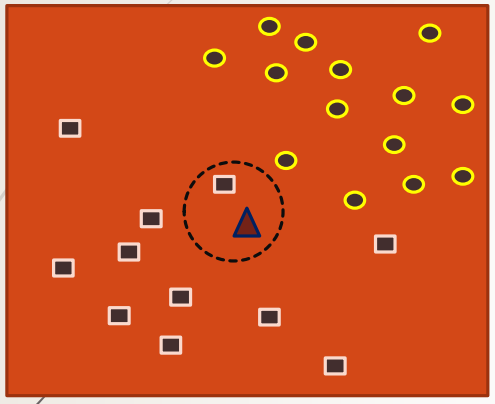
KNN: K-Nearest Neighbors

- **K-Nearest Neighbors (KNN)** is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

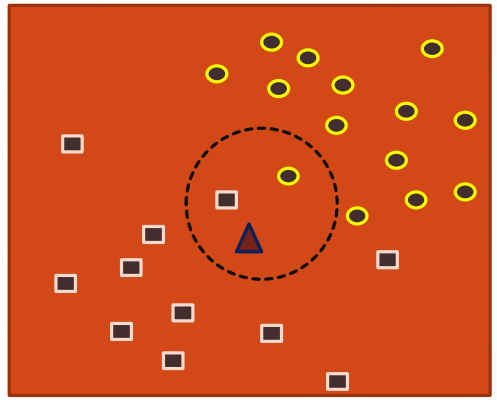
- KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. For this purpose, we use below distance metrics:

  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance

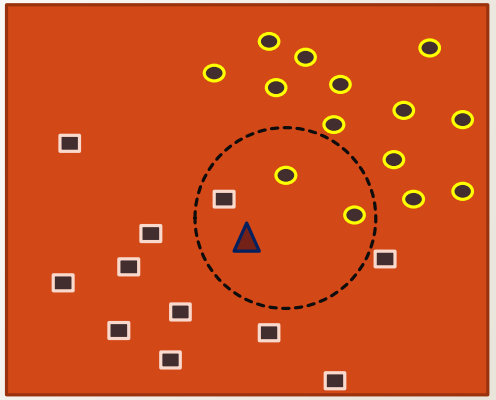
KNN: K-Nearest Neighbors



1- Nearest Neighborhood Class =



2- Nearest Neighborhood Class?



3- Nearest Neighborhood Class =

Naïve Bayes

For discrete features of Refund and Marital Status:

T_ID	Refund	Marital Status	Income	Cheat
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

1- For discrete features

$$P(A_i|C_k) = \frac{|Aik|}{N_{ck}}$$

2- For continues features

$$P(A_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{ik}}} e^{-\frac{(A_i-\mu_{ik})^2}{2\sigma^2_{ik}}}$$

$$P(A_i \mid C_k) = P(\text{Refund} = \text{Yes} \mid \text{Yes}) = \frac{P((Refund=Yes) \cap (Cheat=Yes))}{P(Cheat=Yes)} = \frac{0}{3/10} = 0$$

$$P(A_i \mid C_k) = P(\text{Marital Status} = \text{Married} \mid \text{No}) = \frac{4/10}{7/10} = 4/7$$

Naïve Bayes

For continues feature of Income:

$P(A_i \mid C_k)$= P(Income = 120 | No) = ?

$$\text{Mean} = \mu_{ik} = \frac{125+100+70+120+60+220+75}{7} = 110$$

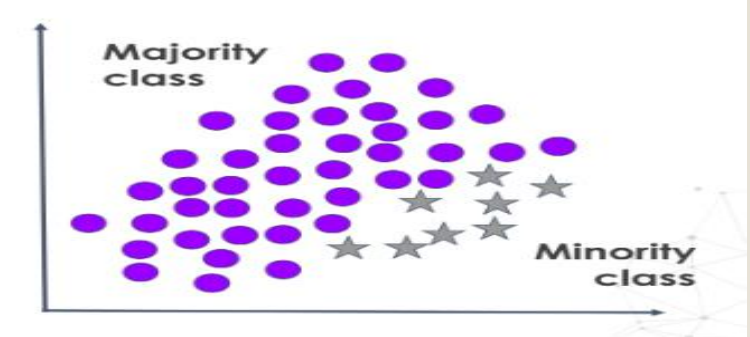
$$\text{Variance} = \sigma^2_{ik} = \frac{(125-110)^2+(100-110)^2+(70-110)^2+(120-110)^2+(60-110)^2+(220-110)^2+(75-110)^2}{6} = 2975$$

$$P(A_i \mid C_k) = P(\text{Income} = 120 \mid \text{No}) = \frac{1}{\sqrt{2\pi\sigma^2_{ik}}} e^{-\frac{(A_i-\mu_{ik})^2}{2\sigma^2_{ik}}}$$

$$= \frac{1}{\sqrt{2*3.14*2975}} e^{-\frac{(120-110)^2}{2*2975}}$$

$$= 0.0072$$

Imbalanced Datasets

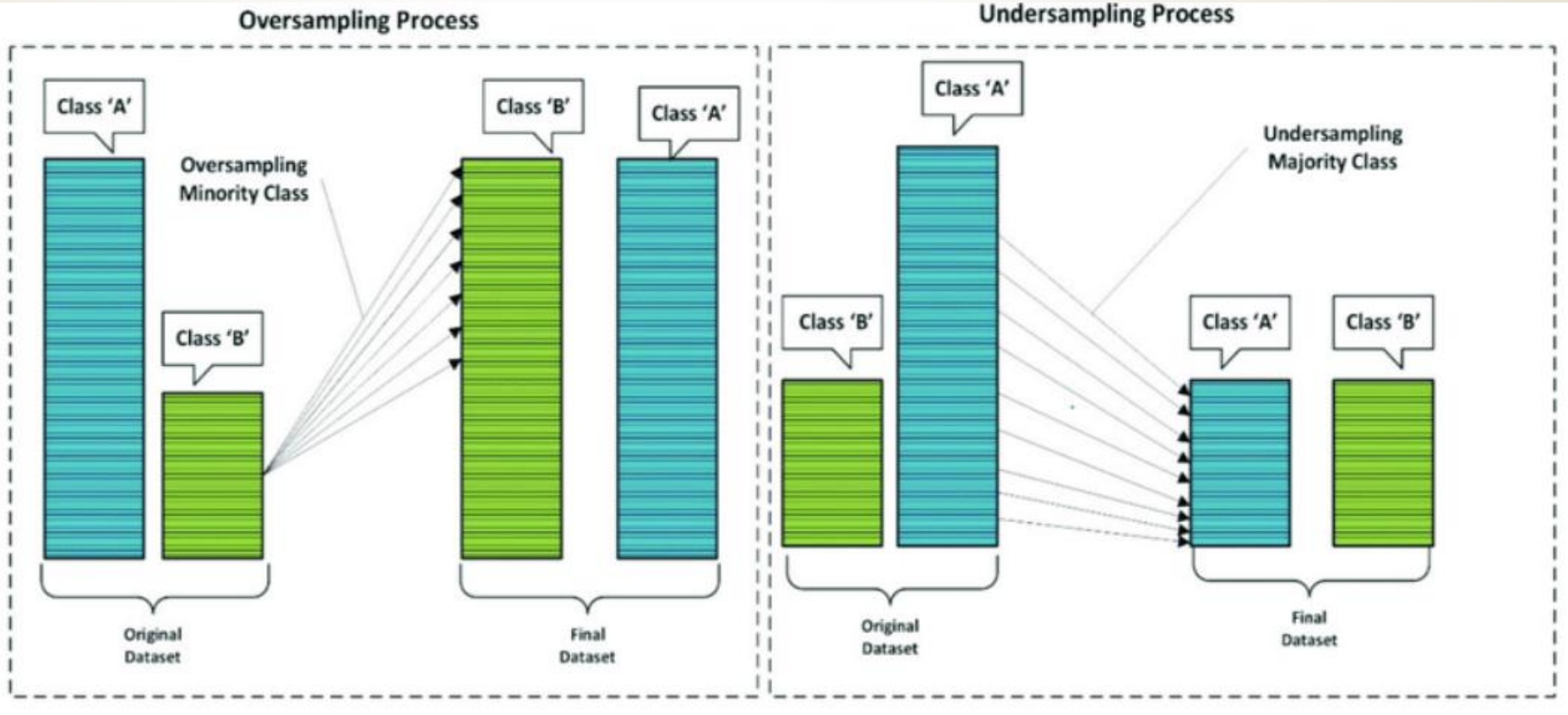


- In imbalanced datasets, one class is significantly more represented than the other(s). In other words, imbalanced datasets have disproportionate numbers of observations in each category of the target variable, with one or more classes being extremely under-represented. This could make it difficult for machine-learning algorithms to learn how to discriminate between them.

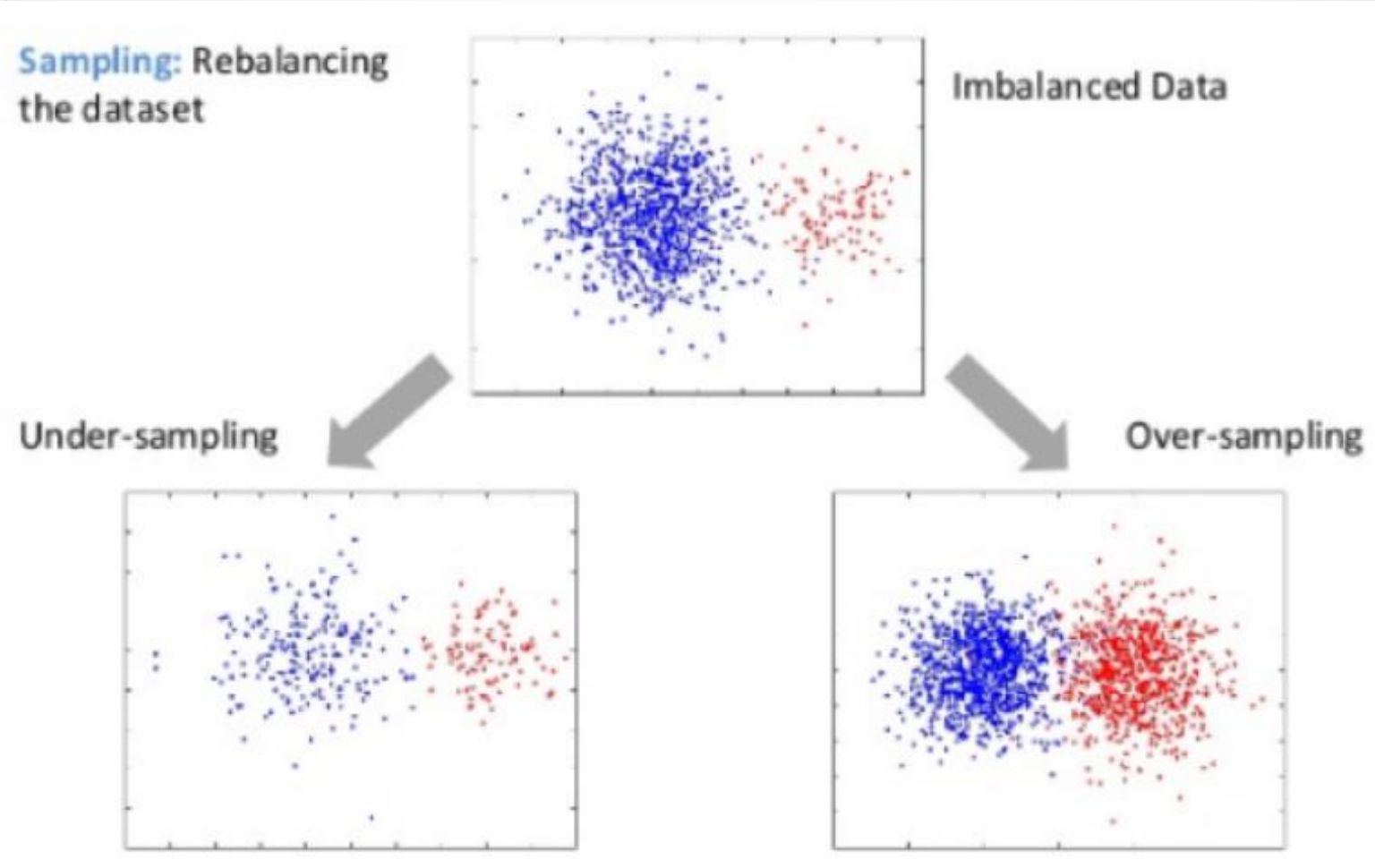
- Imbalanced datasets are common in the real world and often lead to biased predictions and poor overall performance of the machine learning model.

- The degree of imbalance can vary significantly and may be caused by factors like natural unequal distribution or sampling bias in data collection. Understanding the characteristics and differences between binary and multiclass imbalanced data and minority and majority classes will help us address them better.

Imbalanced Datasets



Imbalanced Datasets



Reference

Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 9.