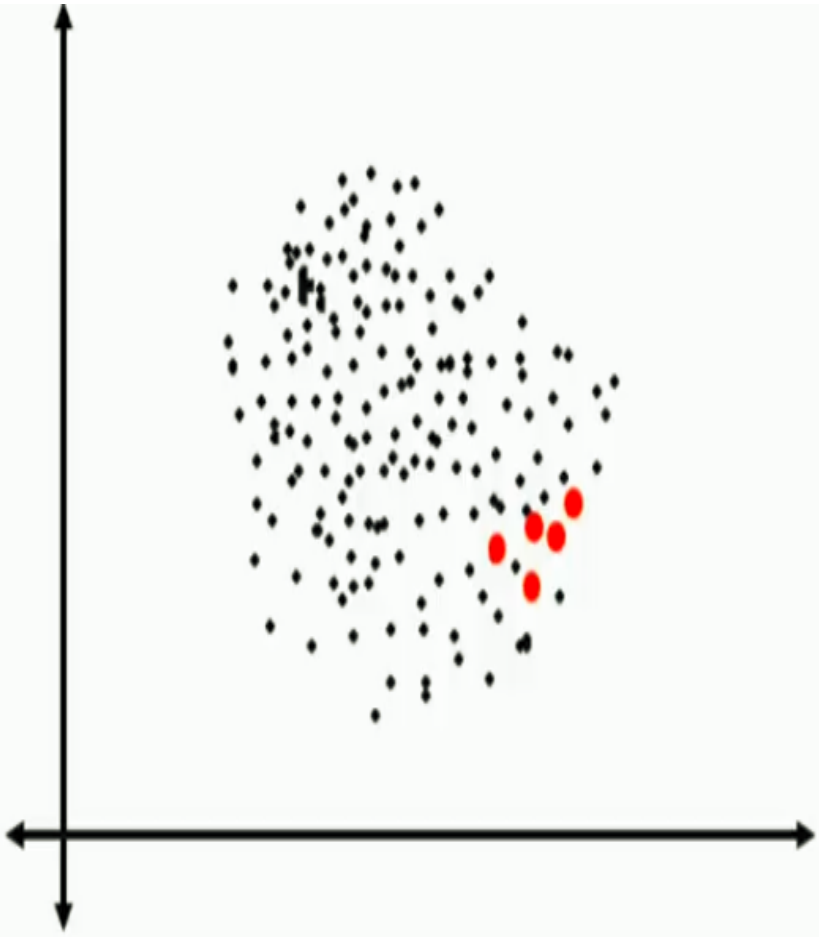
SMOTE - Synthetic Minority Oversampling Technique

IMBALANCED DATA

- Presence of minority class in the dataset

- Challenges related Imbalanced Dataset
  - Biased predictions
  - Misleading accuracy

- Some Examples
  - Credit card frauds
  - Manufacturing defects
  - Rare diseases diagnosis
  - Natural disasters
  - Enrolment to premier institutes



Two Class Classification

No-Fraud → 99.5%

Fraud → 0.5%

RE-SAMPLE

- Balance the classes by Increasing minority or decreasing majority

- Random Under-Sampling
  - Randomly remove majority class observations
  - Helps balance the dataset
  - Discarded observations could have important information
  - May lead to bias

Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

Reduce normal to 90
Fraudulent = 10 or 10%

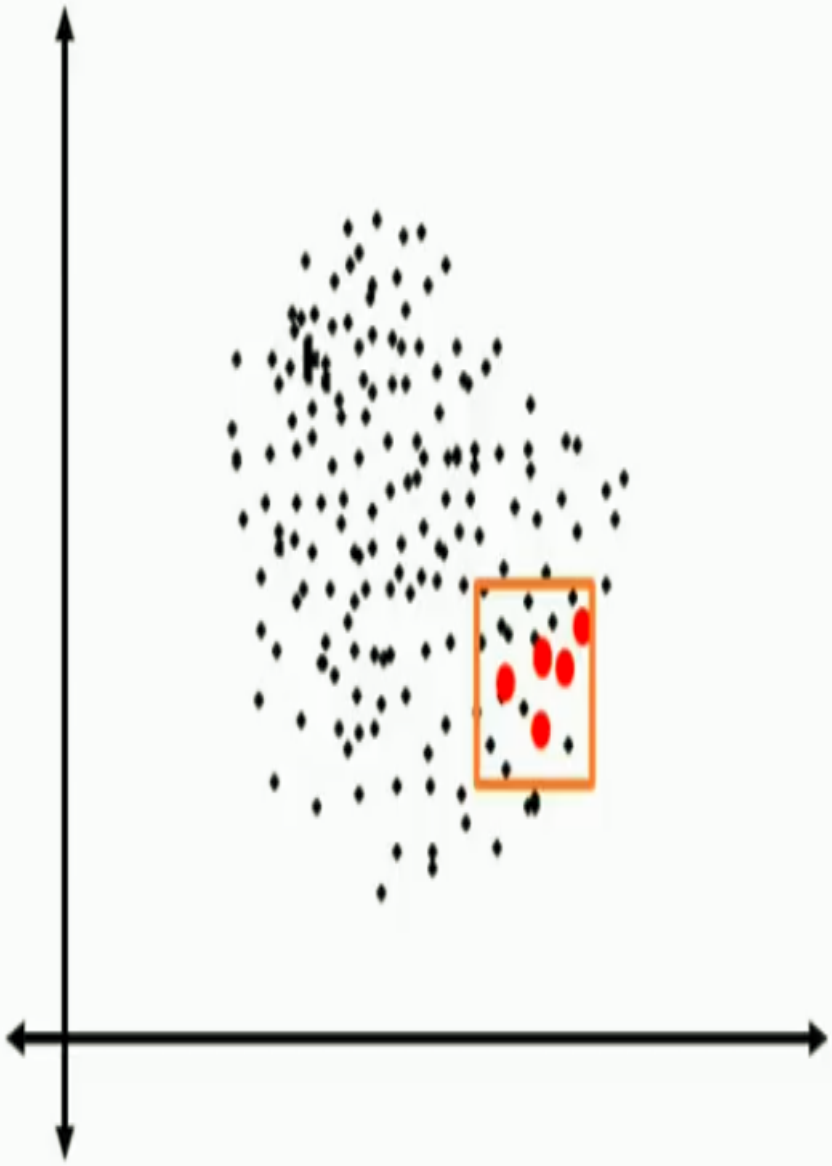
- Random Over-Sampling
  - Randomly add more minority observations by replication
  - No information loss
  - Prone to overfitting due to copying same information

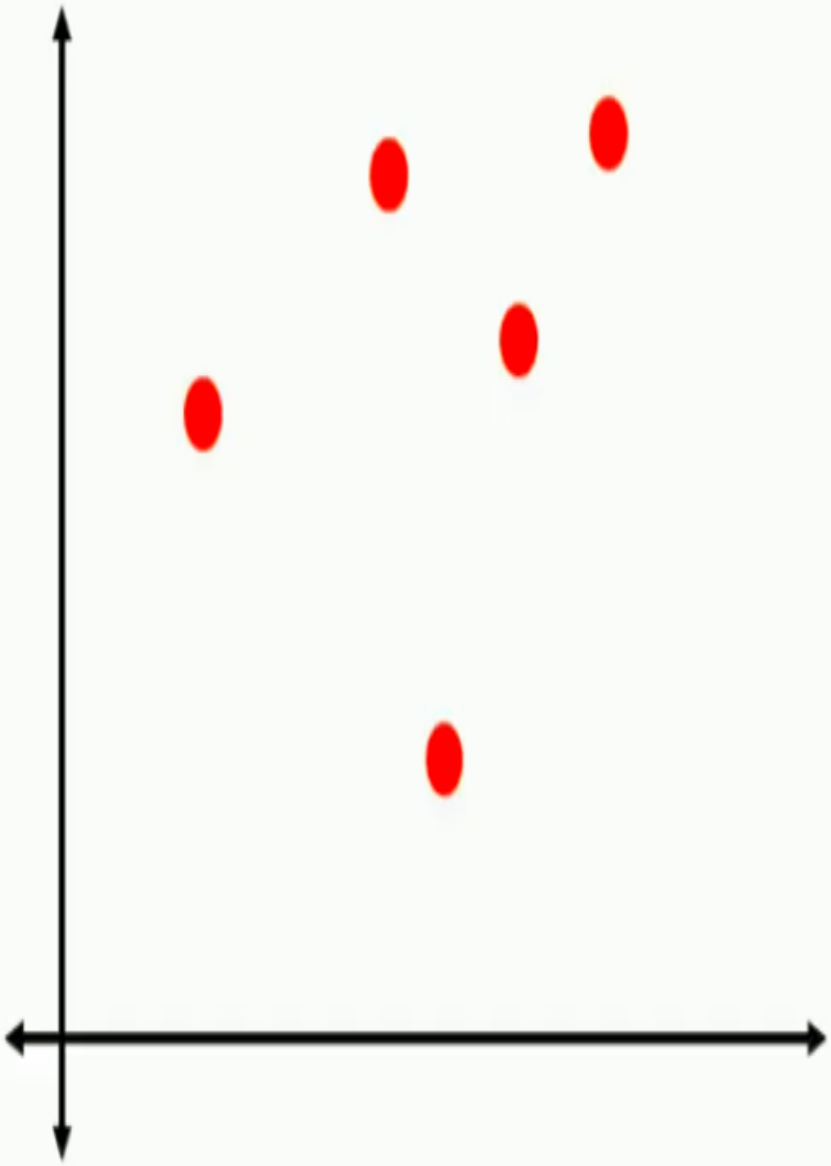
Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

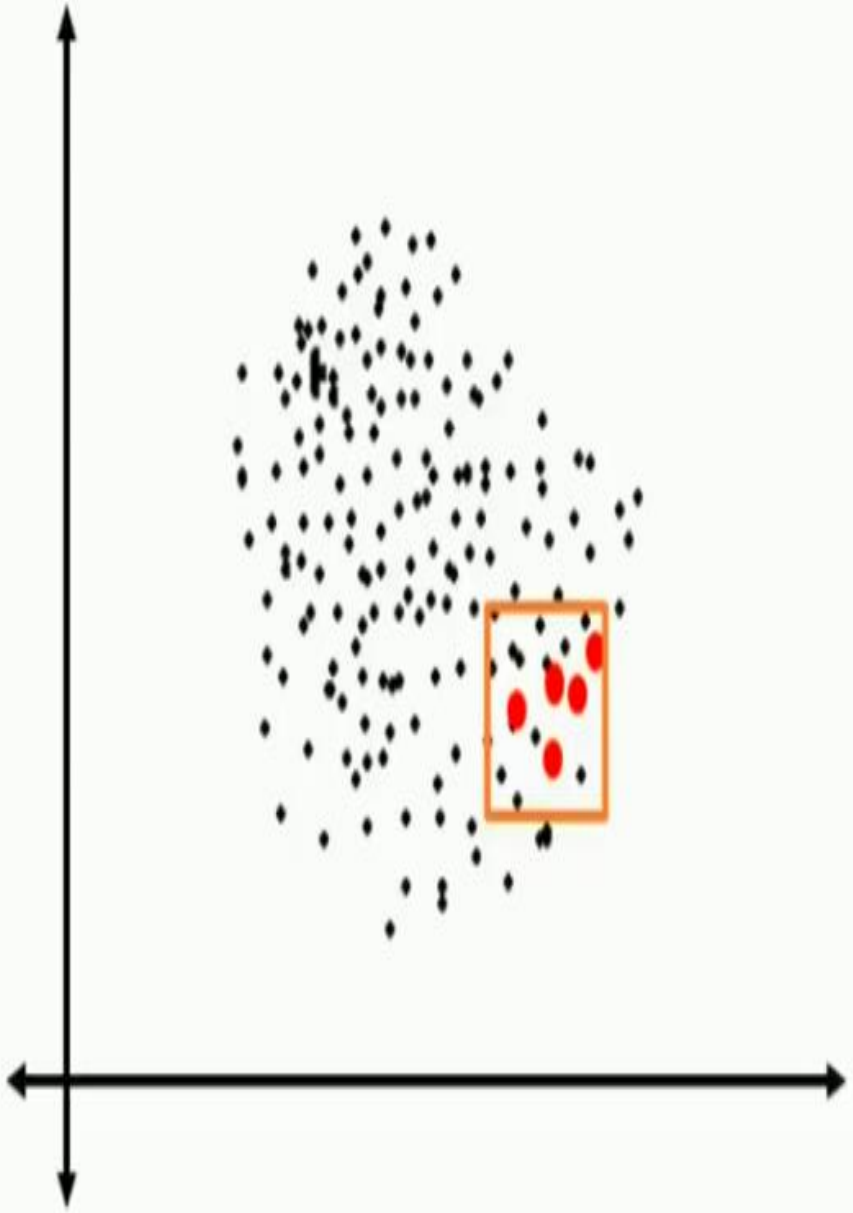
Increase fraudulent by 100
Fraudulent 110 or 10%

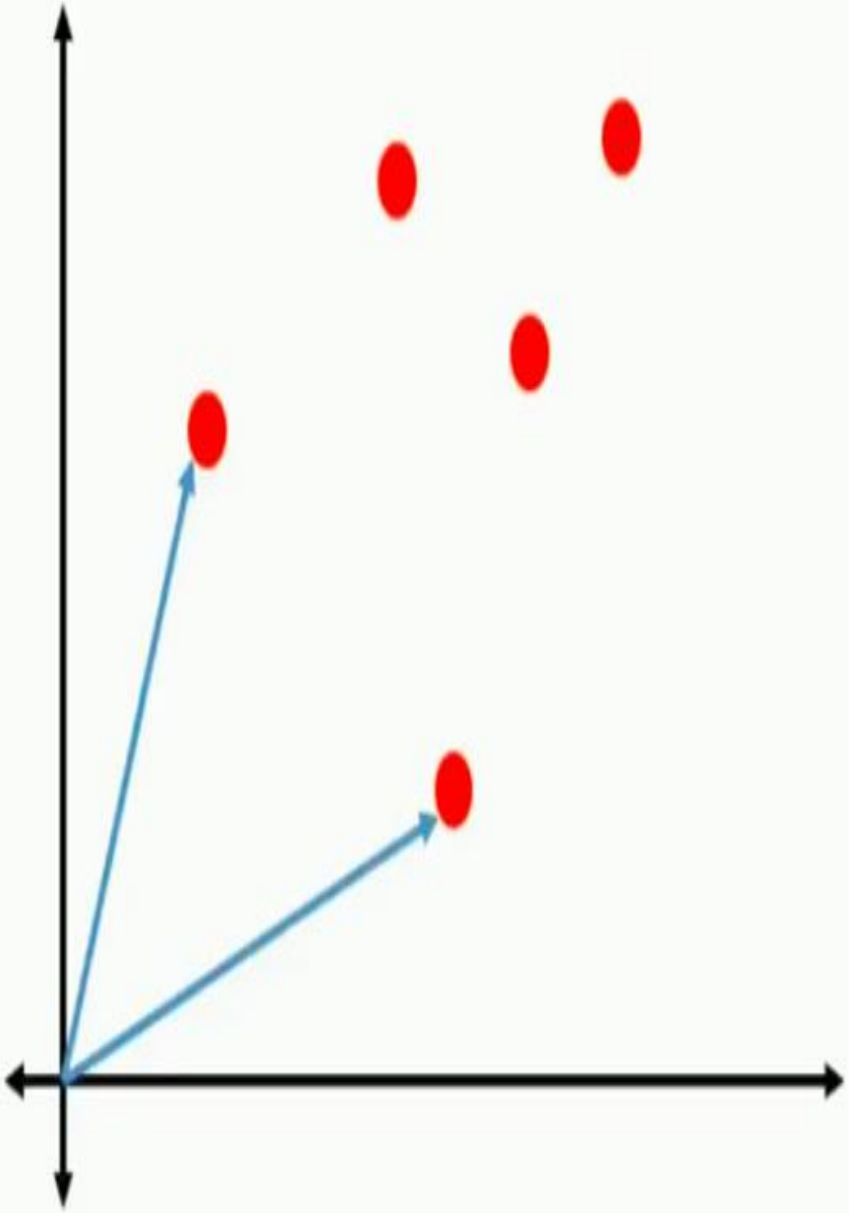
SMOTE

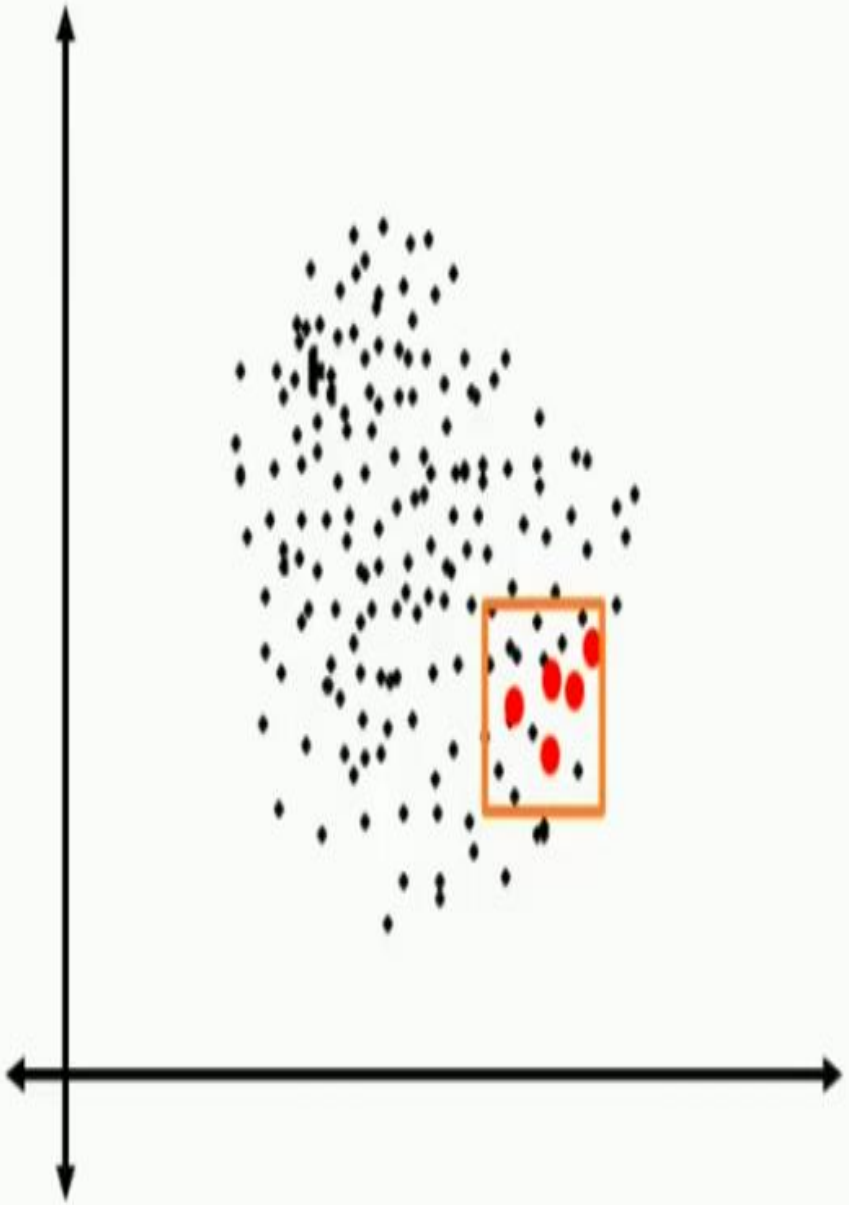
- Synthetic Minority Oversampling Technique
- Creates new “Synthetic” observations
- SMOTE Process
  - Identify the feature vector and its nearest neighbour
  - Take the difference between the two
  - Multiply the difference with a random number between 0 and 1
  - Identify a new point on the line segment by adding the random number to feature vector
  - Repeat the process for identified feature vectors

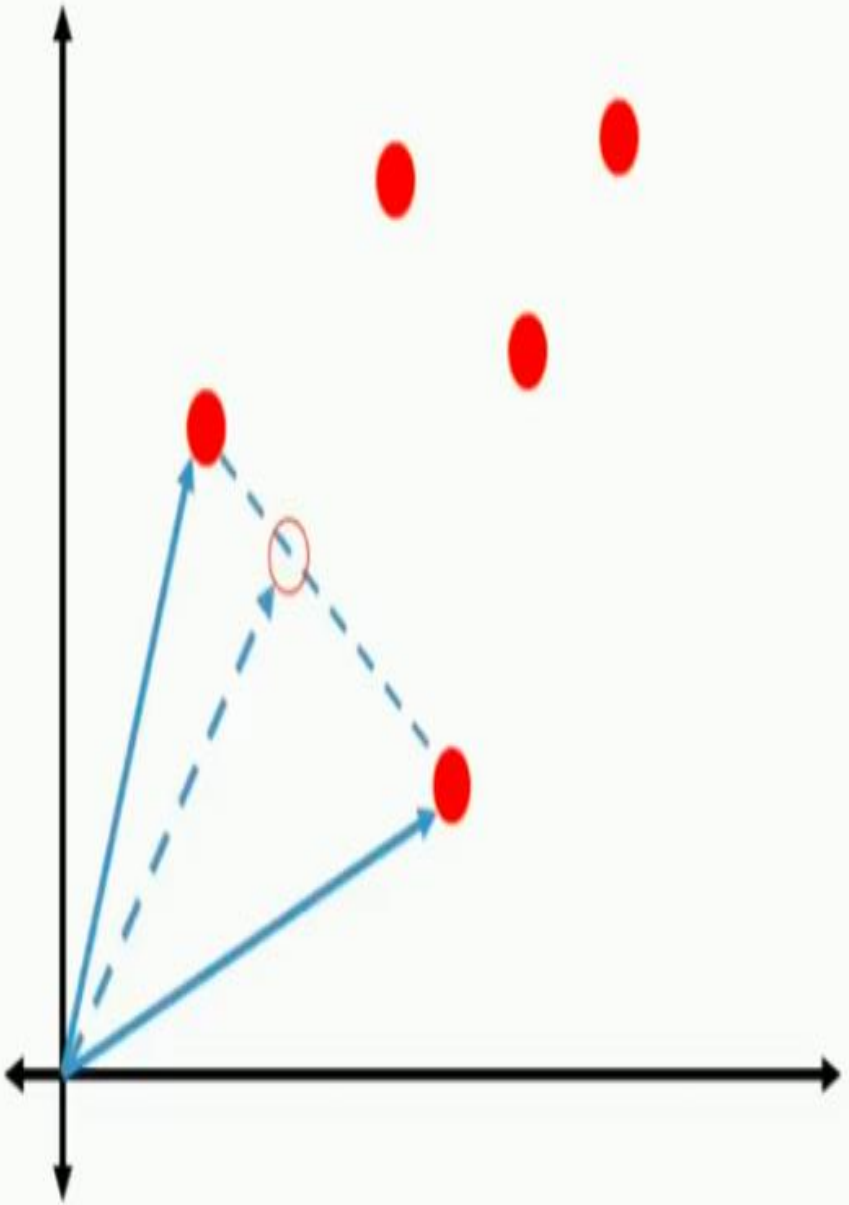


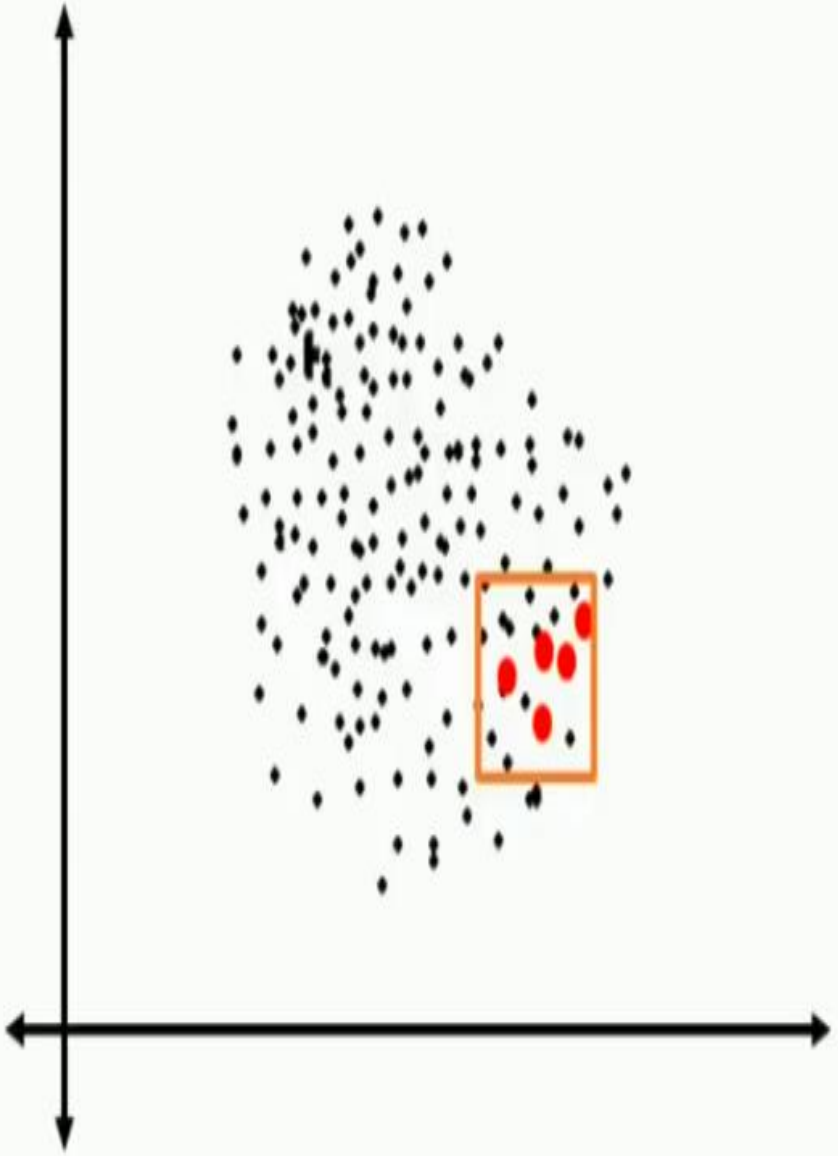


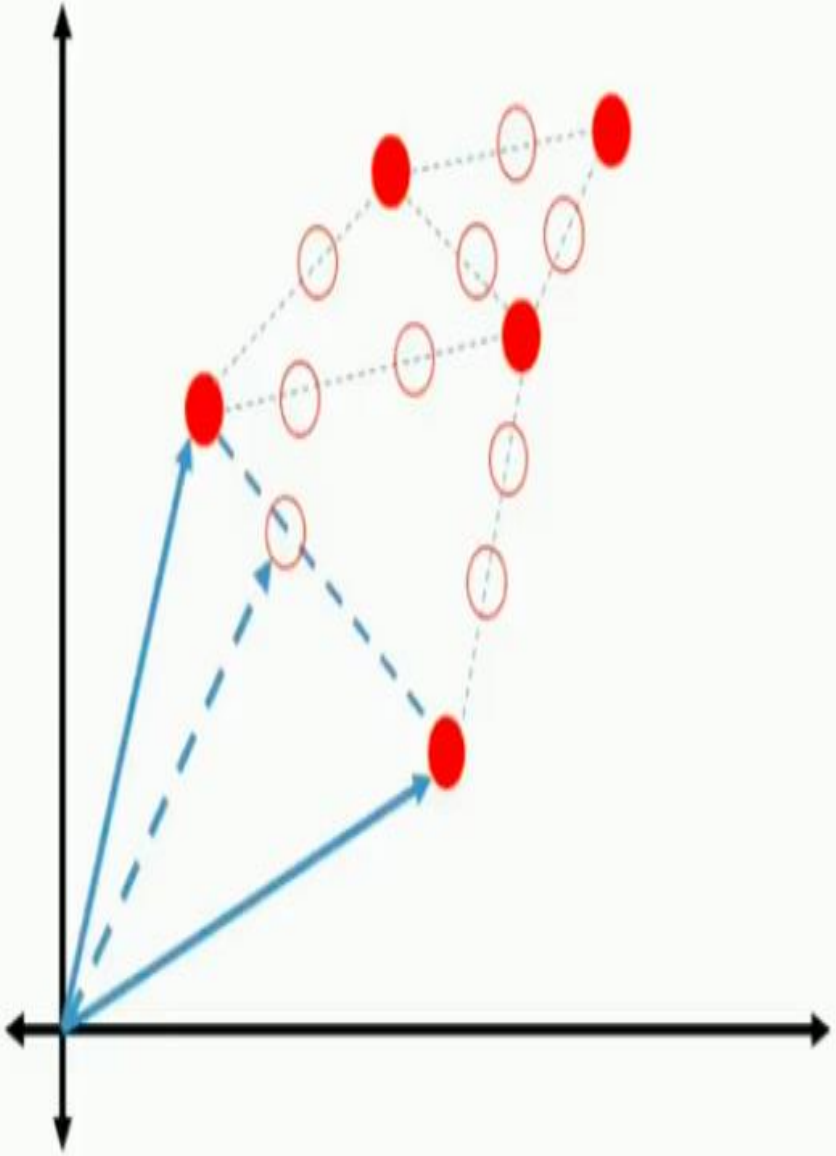












Thanks