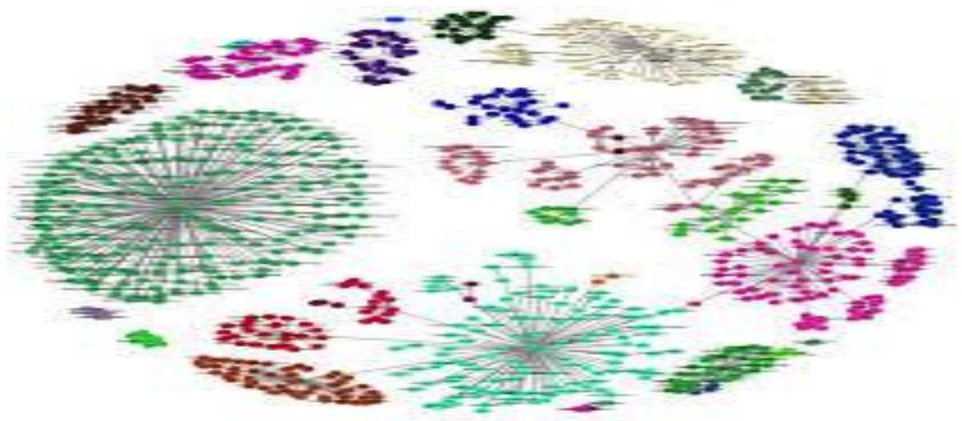
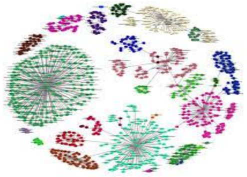


# Key ML Algorithms- KNN

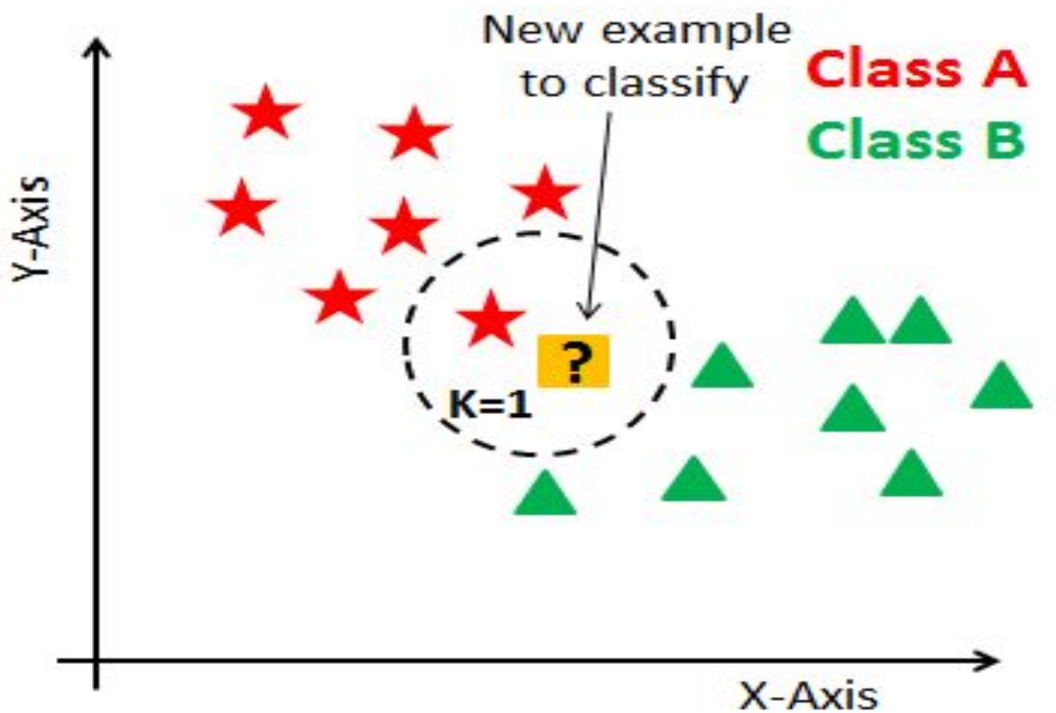
# Agenda



- K-nearest neighbors
- KNN Classifier
- KNN Regressor
- Dimensionality Reduction
- Principal Component Analysis
- Singular Value Decomposition

# What is KNN?

- K Nearest Neighbor is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.
- It is mostly used to classifies a data point based on how its neighbors are classified.

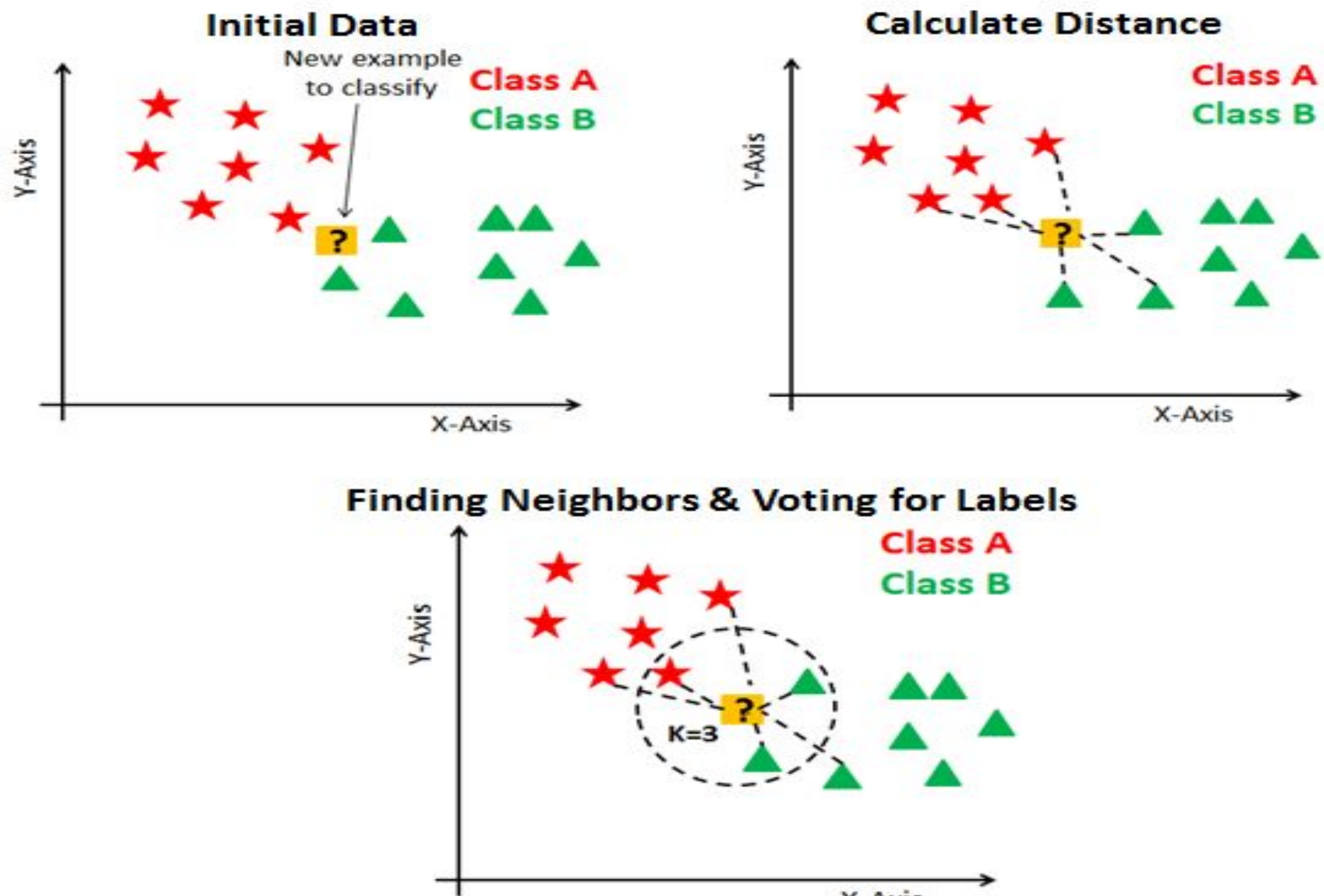


# KNN Steps

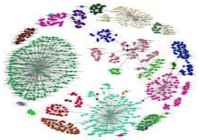
INTERNSHIPSTUDIO

KNN has the following basic steps:

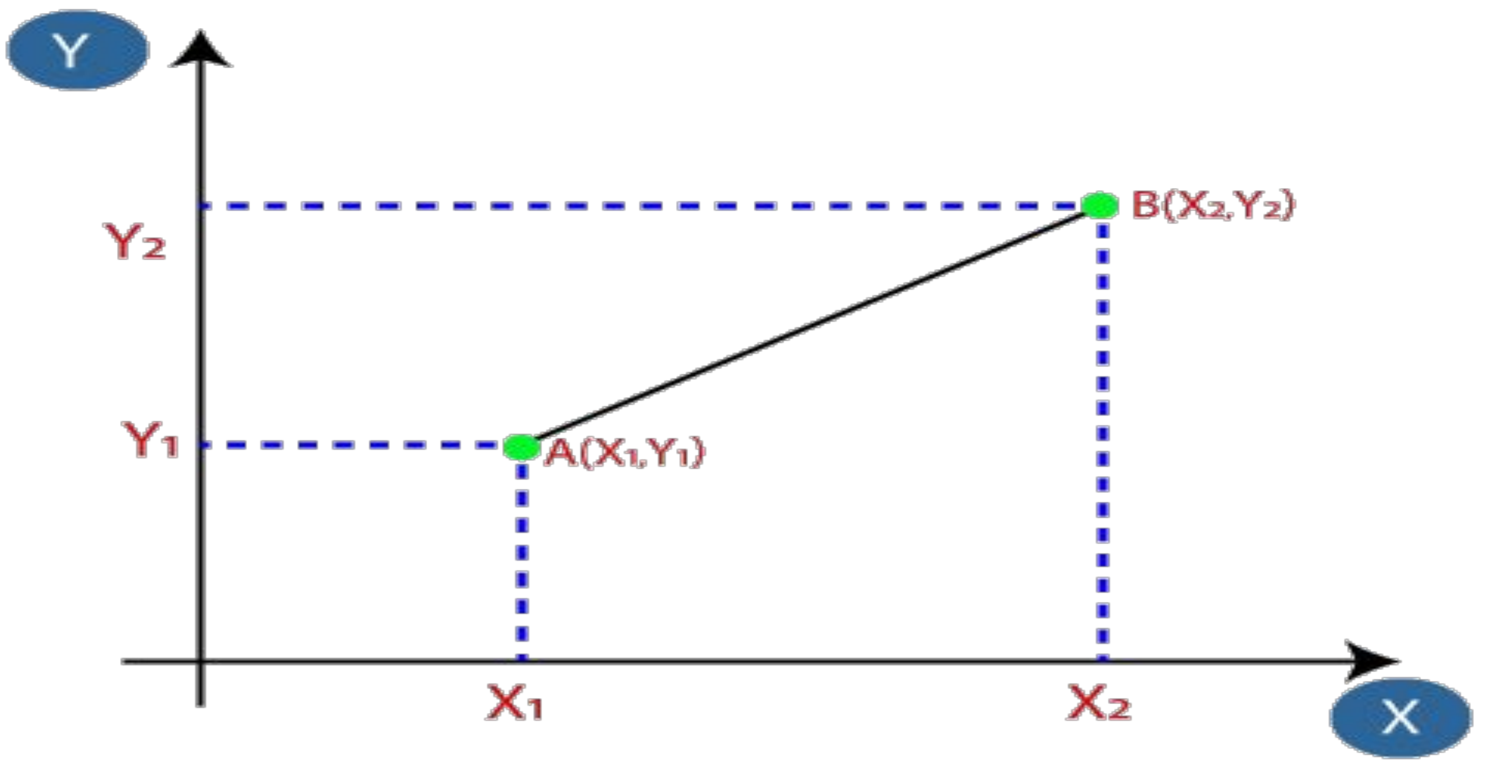
1. Calculate distance
2. Find closest neighbors
3. Vote for labels



# Euclidean Distance

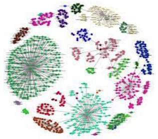


Euclidean Distance represents the shortest distance between two points.




Euclidean Distance between $A_1$ and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

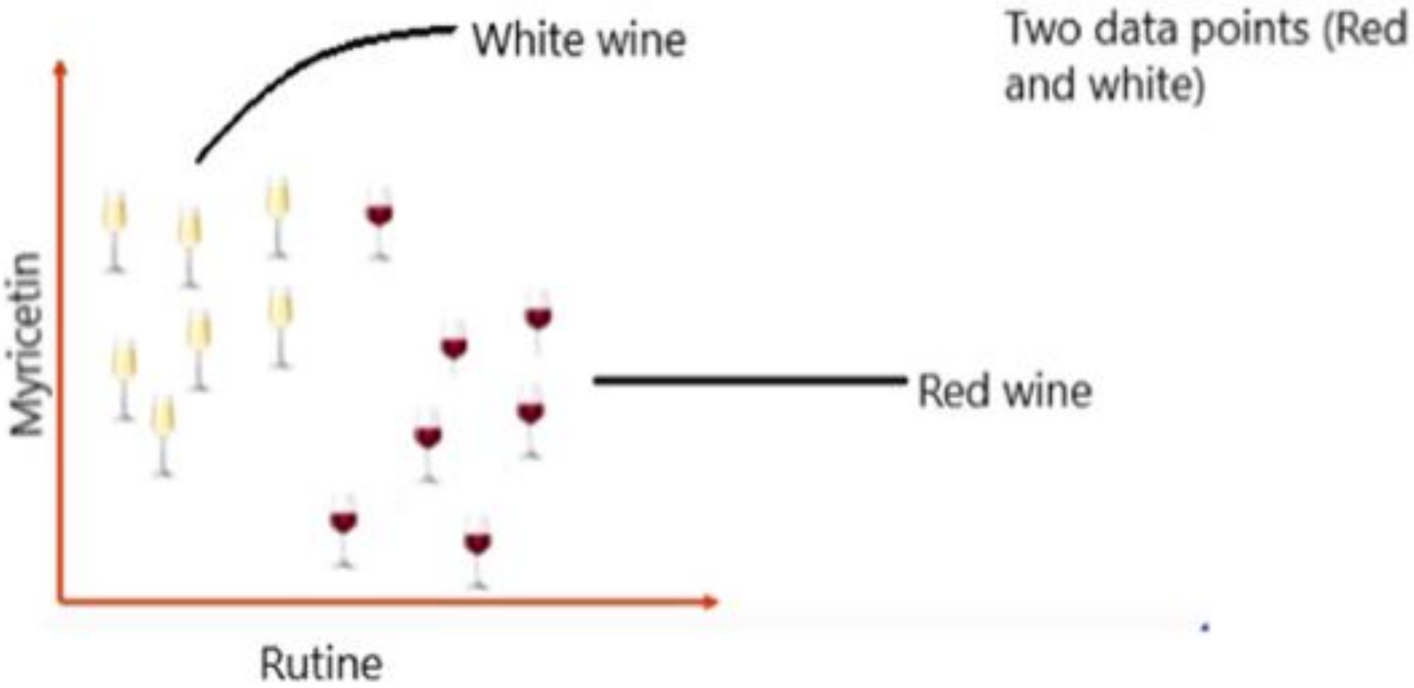
# Choosing the right value for K



- To get the right K, you should run the KNN algorithm several times with different values of K and select the one that has the least number of errors.
- As your value of K increases, your prediction becomes more stable due to the majority of voters.
- When you start receiving an increasing number of errors, you should know you are pushing your K too far.

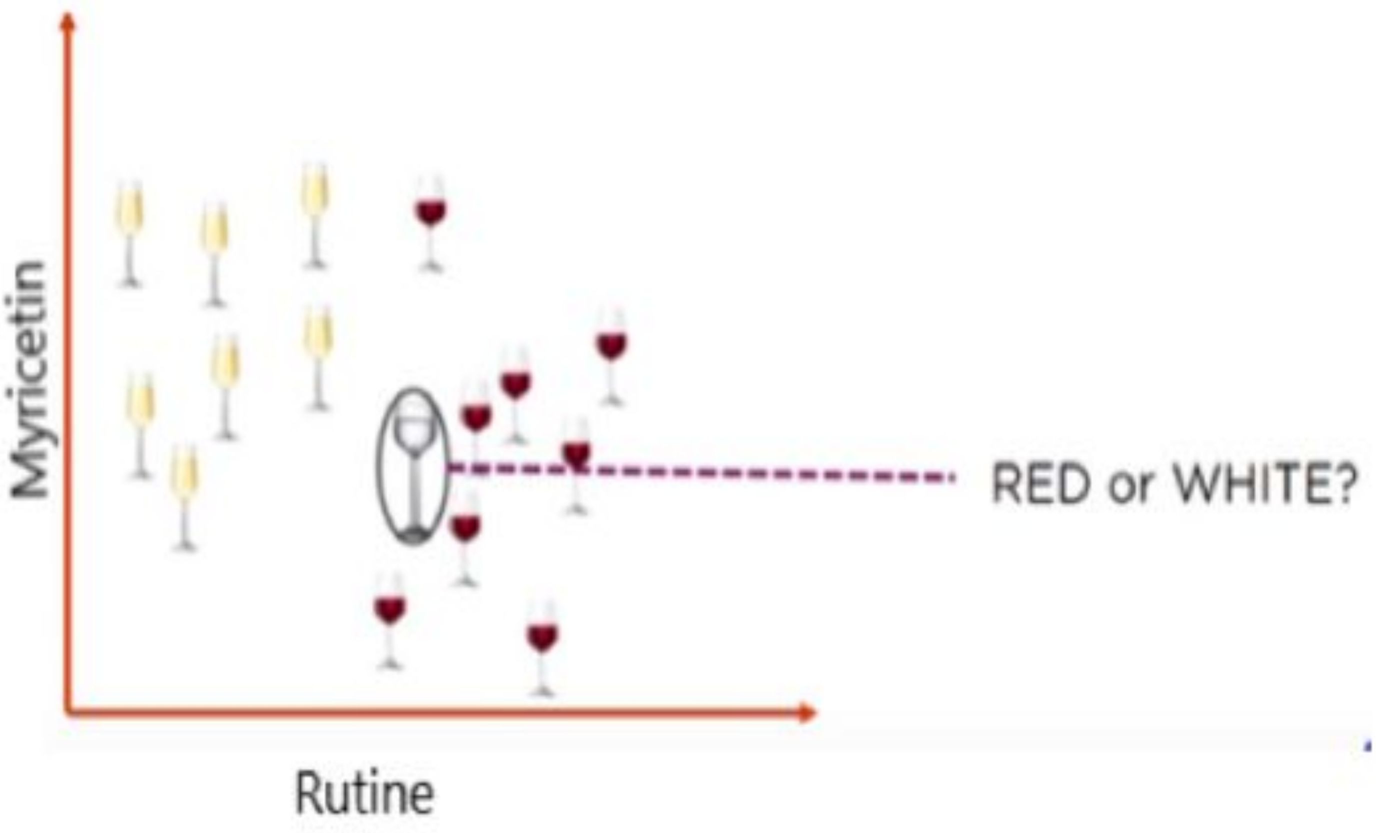
# Example- Red and White wines

- Let’s take below wine example. Two chemical components called Rutine and Myricetin. Consider a measurement with two data points, **Red and White wines**.



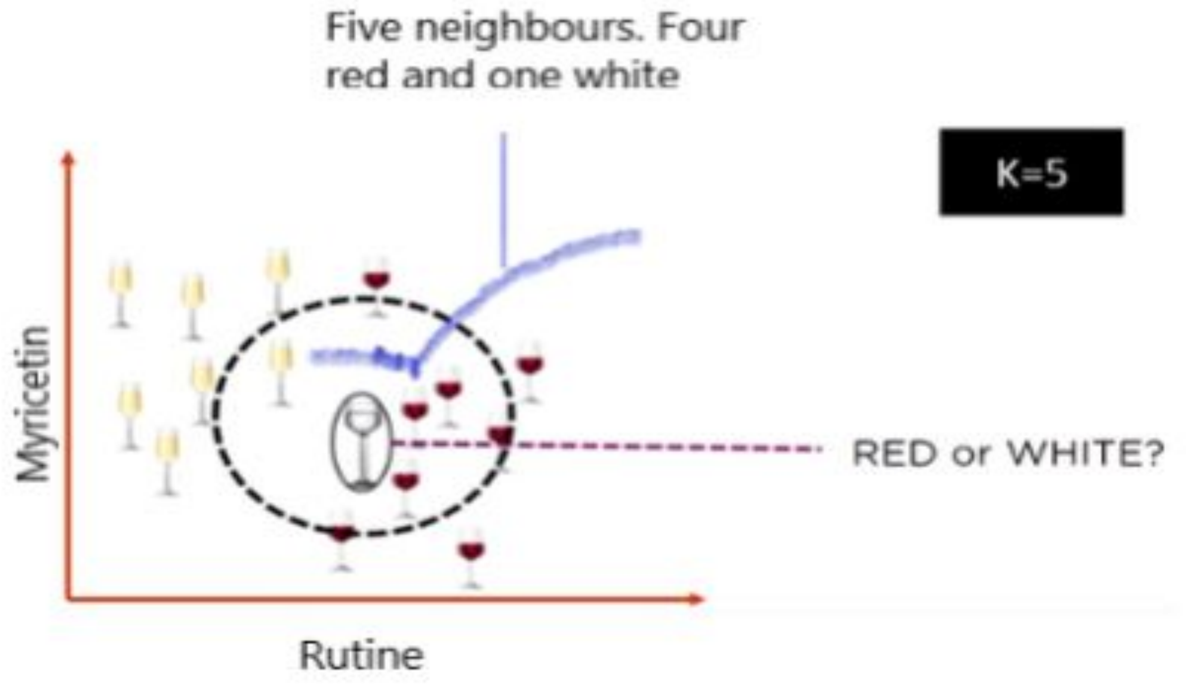
# Example- Red and White wines

- Suppose, if we add a new glass of wine in the dataset. We would like to know whether the new wine is red or white?



# Example- Red and White wines

- So, we need to find out what the neighbours are in this case. Let's say k = 5 and the new data point is classified by the majority of votes from its five neighbours.




- The new point would be classified as red since four out of five neighbours are red.

# More Example

INTERNSHIPSTUDIO

## K-NN IN ACTION

- Consider the following data:
  A={weight,color}
  G={Apple(A), Banana(B)}

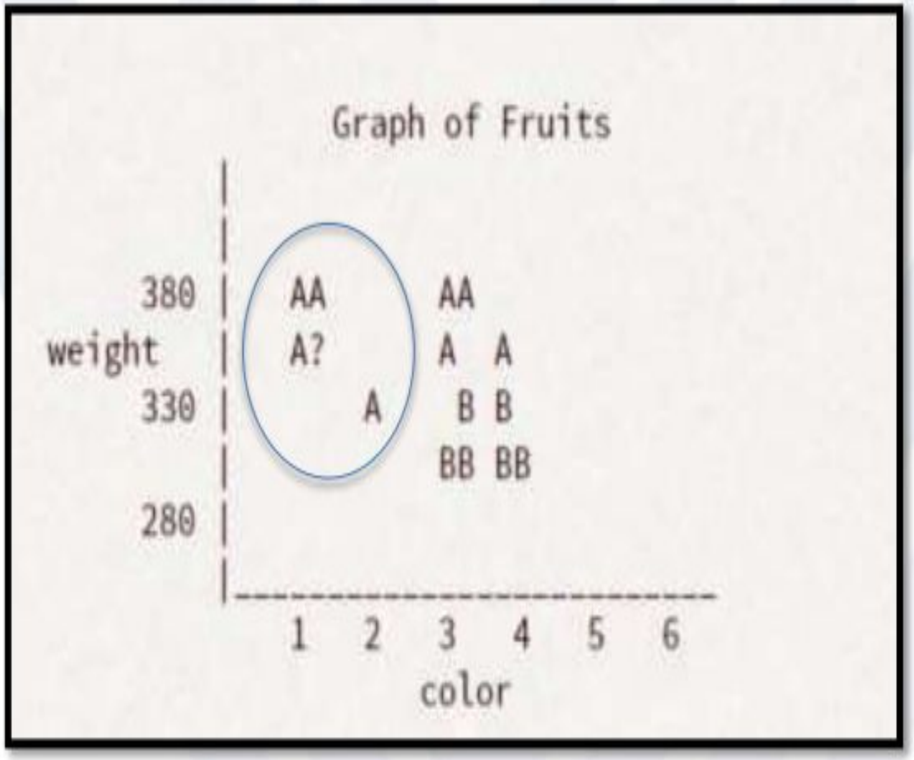
- We need to predict the type of a fruit with:
  weight = 378
  color = red

| weight (g) | color | Type of fruit |
|---|---|---|
| 303 | 3 | Banana |
| 370 | 1 | Apple |
| 298 | 3 | Banana |
| 277 | 3 | Banana |
| 377 | 4 | Apple |
| 299 | 3 | Banana |
| 382 | 1 | Apple |
| 374 | 4 | Apple |
| 303 | 4 | Banana |
| 309 | 3 | Banana |
| 359 | 1 | Apple |
| 366 | 1 | Apple |
| 311 | 3 | Banana |
| 302 | 3 | Banana |
| 373 | 4 | Apple |
| 305 | 3 | Banana |
| 371 | 3 | Apple |

# More Example

## PLOTTING

- Using K=3,
  Our result will be,



# KNN Advantages & Disadvantages

## Advantages :-

- Quick calculation time
- Simple algorithm – to interpret
- Versatile – useful for regression and classification
- High accuracy – you do not need to compare with better-supervised learning models
- No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.

## Disadvantages :-

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory – need to store all of the training data
- Given that it stores all of the training, it can be computationally expensive

# K Nearest Neighbor

## K-NN VARIATIONS

- Weighted K-NN: Takes the weights associated with each attribute. This can give priority among attributes.

Ex: For the data,

Weight: $w(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\lambda|\mathbf{x} - \mathbf{x}_i|_2^2\right)$

Probability: $\Pr(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} w(\mathbf{x}, \mathbf{x}_i)\delta(y, y_i)}{\sum_{i=1}^{n} w(\mathbf{x}, \mathbf{x}_i)}$

Where,

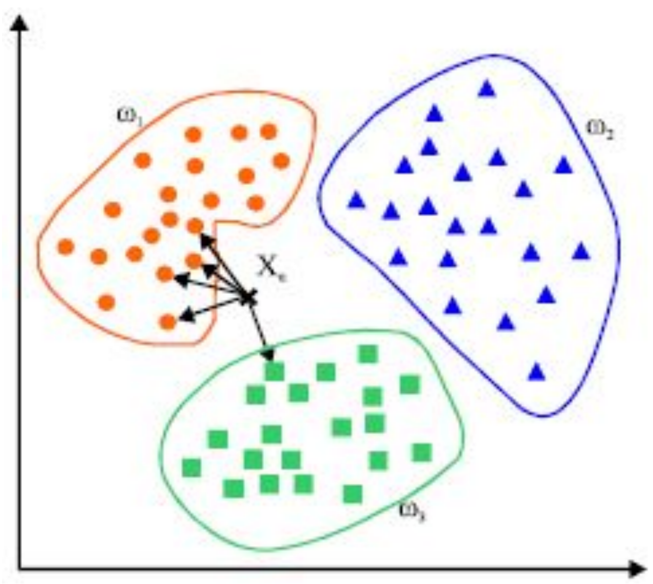
$$\delta(y, y_i) = \begin{cases} 1 & y = y_i \\ 0 & y \neq y_i \end{cases}$$

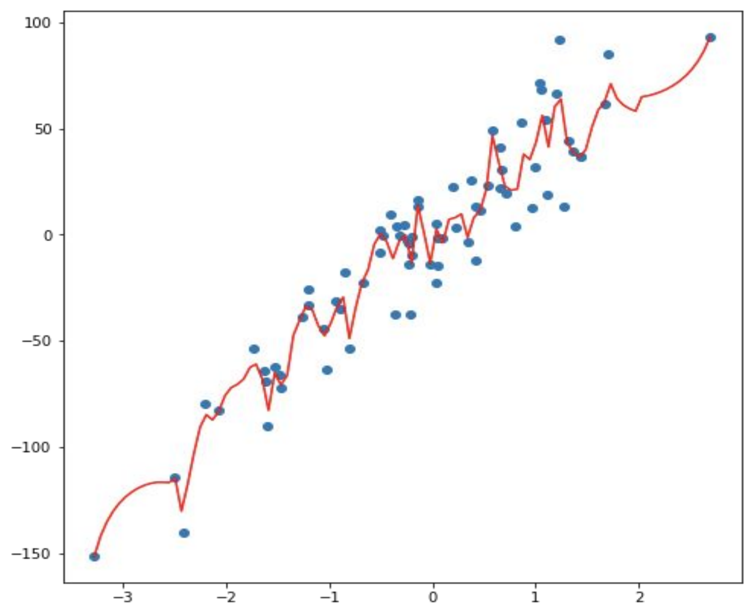
| | $d(x_i, x)$ | $w_i$ |
|---|---|---|
| $x_1$ | 2 | 0.5 |
| $x_2$ | 2 | 0.5 |
| $x_3$ | 2 | 0.5 |
| $x_4$ | 2 | 0.5 |
| $x_5$ | 0.7 | 1/0.7 |
| $x_6$ | 0.8 | 1/0.8 |

Above is the resulting dataset

# KNN Classifier vs KNN Regressor

The key differences are:
1. KNN regression tries to predict the value of the output variable by using a local average.
2. KNN classification attempts to predict the class to which the output variable belong by computing the local probability.

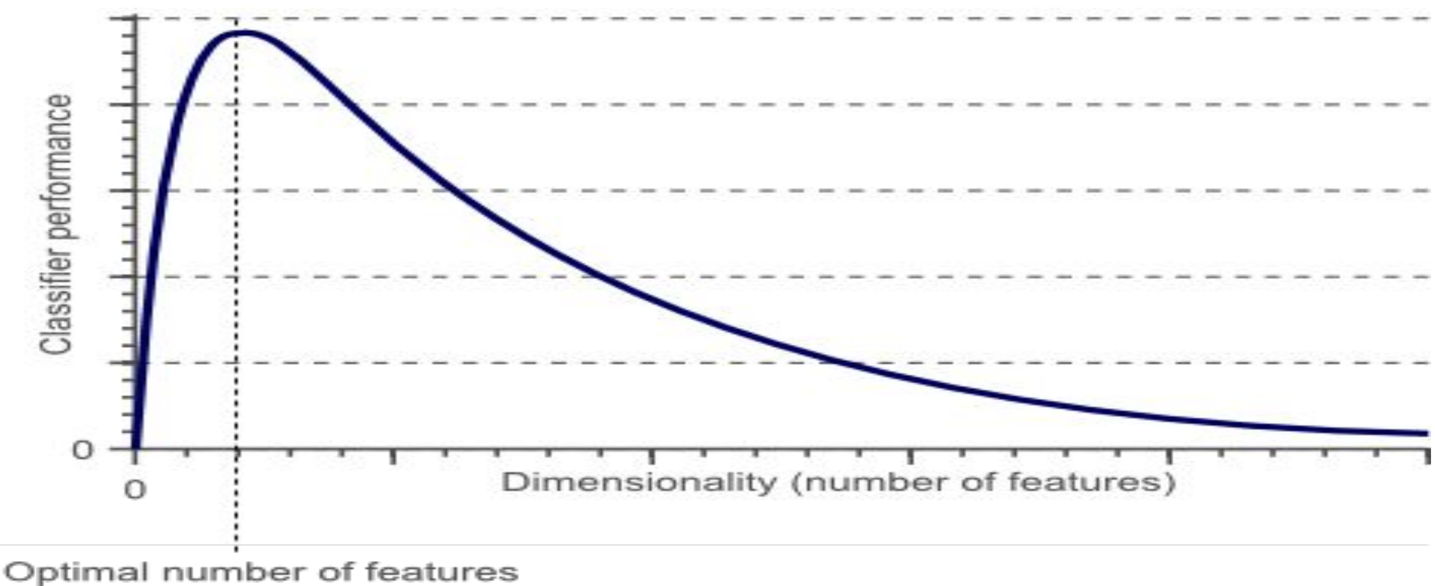



- Q.1 What is KNN?
- Q.2 What is the need of KNN?
- Q.3 What is the origin of KNN?
- Q.4 Can we use KNN in classification problems?

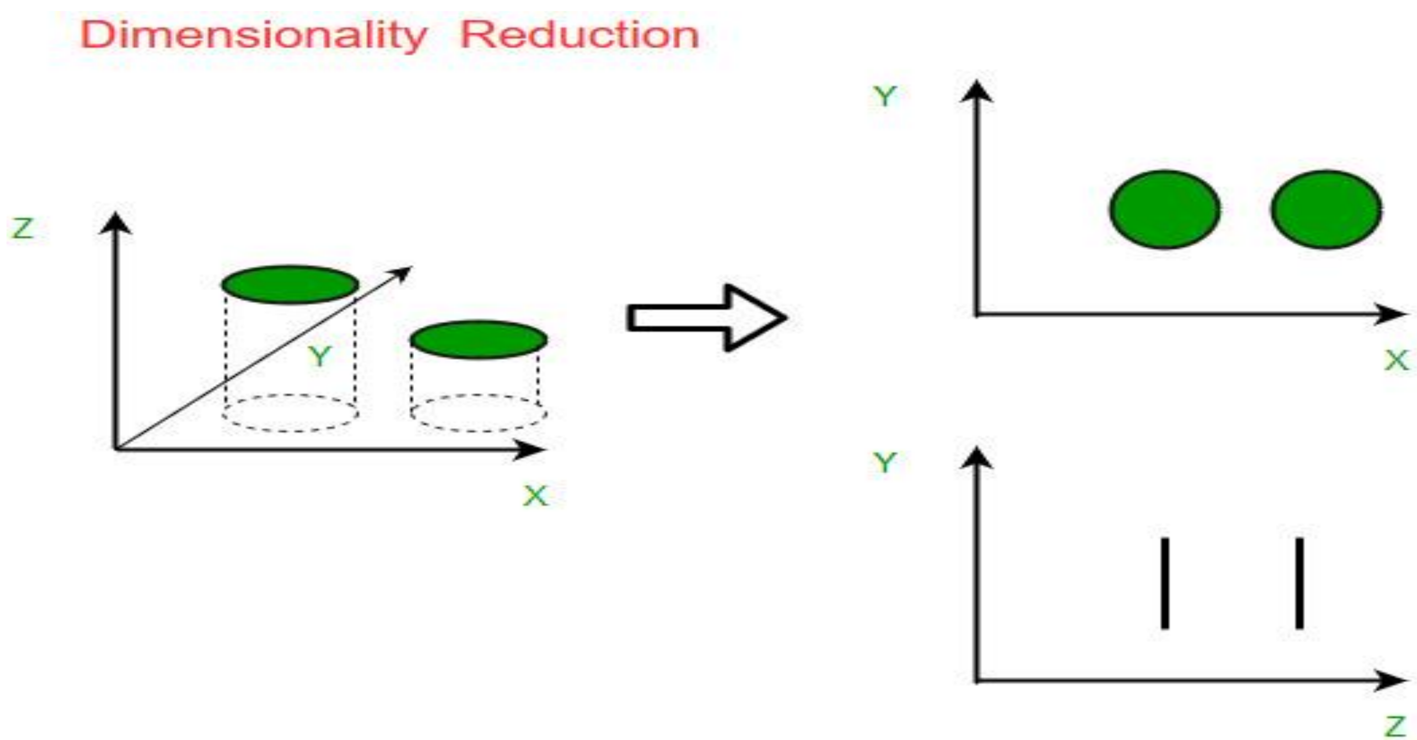
# Dimensionality Reduction

- We are generating a tremendous amount of data daily.  As data generation and collection keeps increasing, visualizing it and drawing inferences becomes more and more challenging.
- As the number of features increases, the model becomes more complex. The more the number of features, the more the chances of overfitting. A machine learning model that is trained on a large number of features, gets increasingly dependent on the data it was trained on and in turn overfitted, resulting in poor performance on real data.



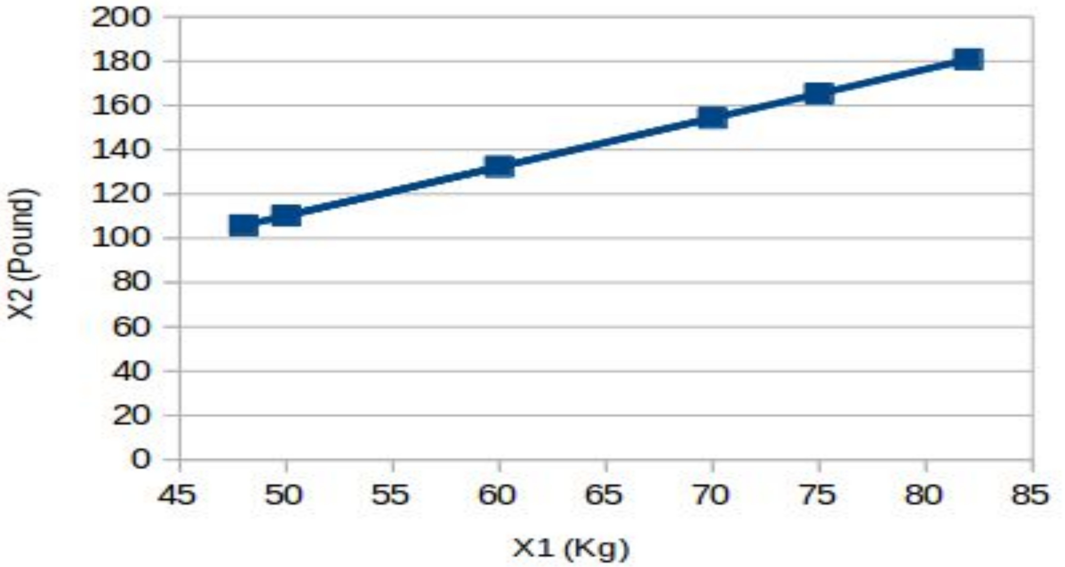
# Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- It can be divided into feature selection and feature extraction.



# Dimensionality Reduction

- Now consider a case in which we have, say 100 variables (p=100).
- Here we have weights of similar objects in Kg (X1) and Pound (X2). If we use both of these variables, they will convey similar information.



# Dimensionality Reduction

- So, it would make sense to use only one variable. We can convert the data from 2D (X1 and X2) to 1D (Y1) as shown below:
- Similarly, we can reduce p dimensions of the data into a subset of k dimensions (k<<n).

# Components of DR

## Feature selection:

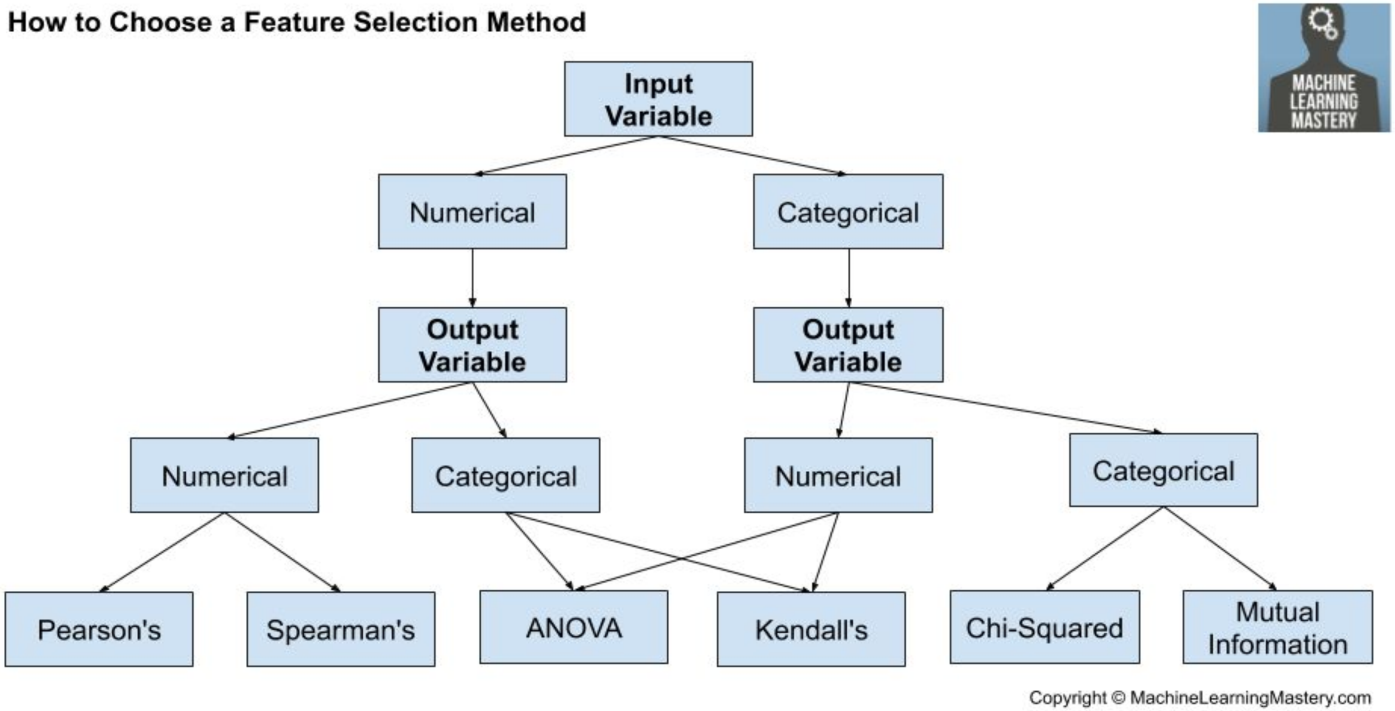
- In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  - Filter
  - Wrapper
  - Embedded

## Feature extraction:

- This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

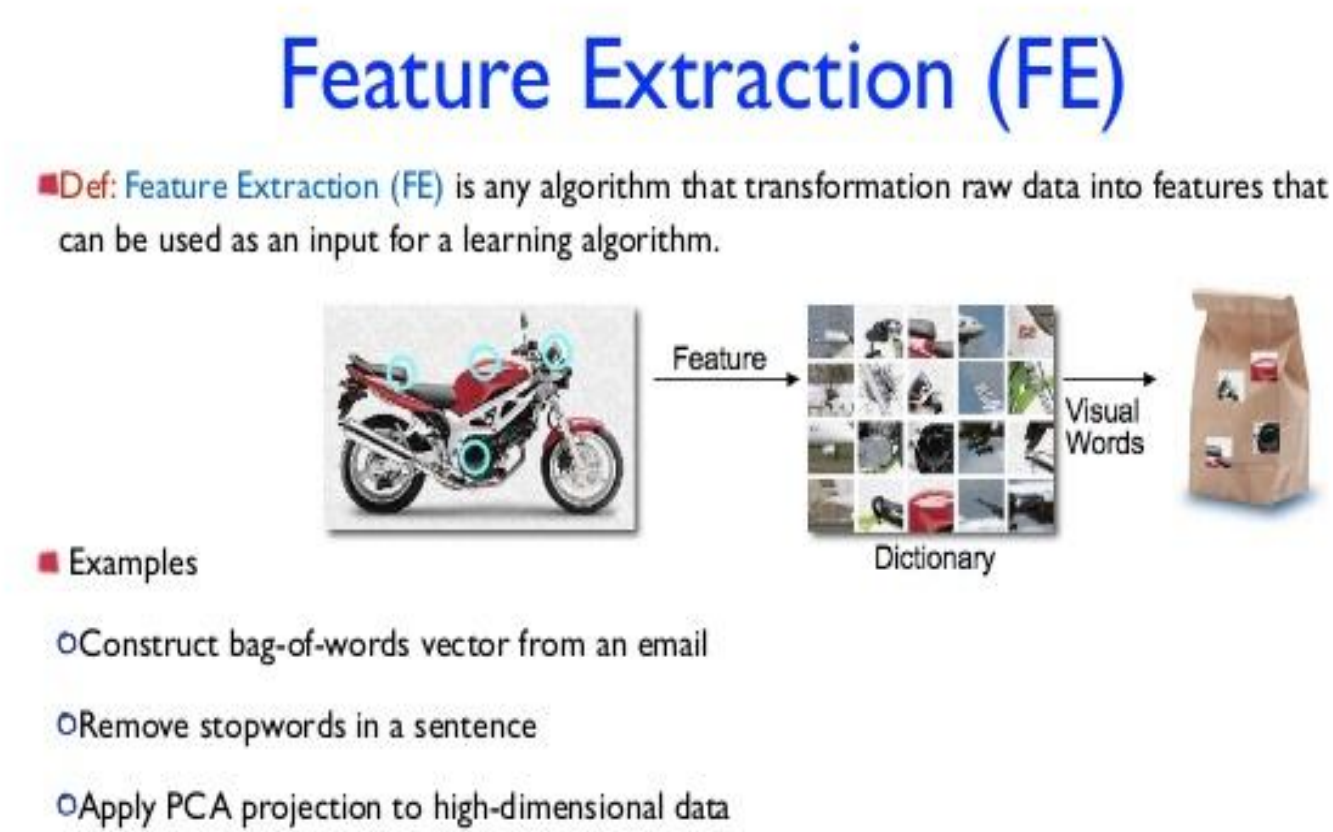
# Feature Selection

- **Feature selection** methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.
- Feature selection is primarily focused on removing non-informative or redundant predictors from the model.



# Feature Extraction

Feature extraction is a type of dimensionality reduction where a large number of pixels of the image are efficiently represented in such a way that interesting parts are captured effectively.



# Methods to Perform DR

# Advantages of DR

The fewer features our training data has, the lesser assumptions our model makes and the simpler it will be. But that is not all and dimensionality reduction has a lot more advantages to offer, like..

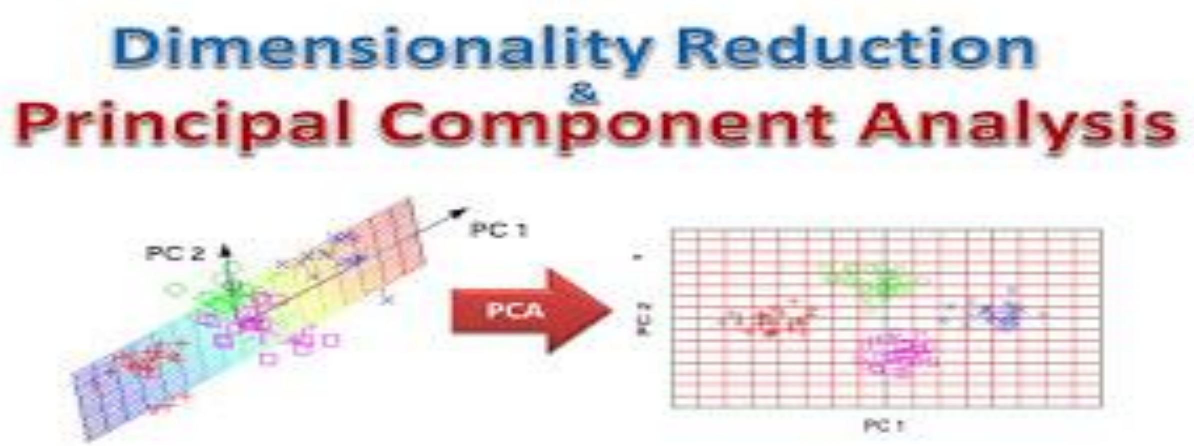
- Dimensionality Reduction helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.
- Dimensionality Reduction helps in data compressing and reducing the storage space required
- It fastens the time required for performing same computations.
- If there present fewer dimensions then it leads to less computing. Also, dimensions can allow usage of algorithms unfit for a large number of dimensions.
- It takes care of multicollinearity that improves the model performance. It removes redundant features. For example, there is no point in storing a value in two different units (meters and inches).

# QUIZ!

INTERNSHIPSTUDIO

- Q.1 What is Dimensionality Reduction?
- Q.2 What is the need of Dimensionality Reduction?
- Q.3 What is Feature Selection?
- Q.4 What is Feature Extraction?

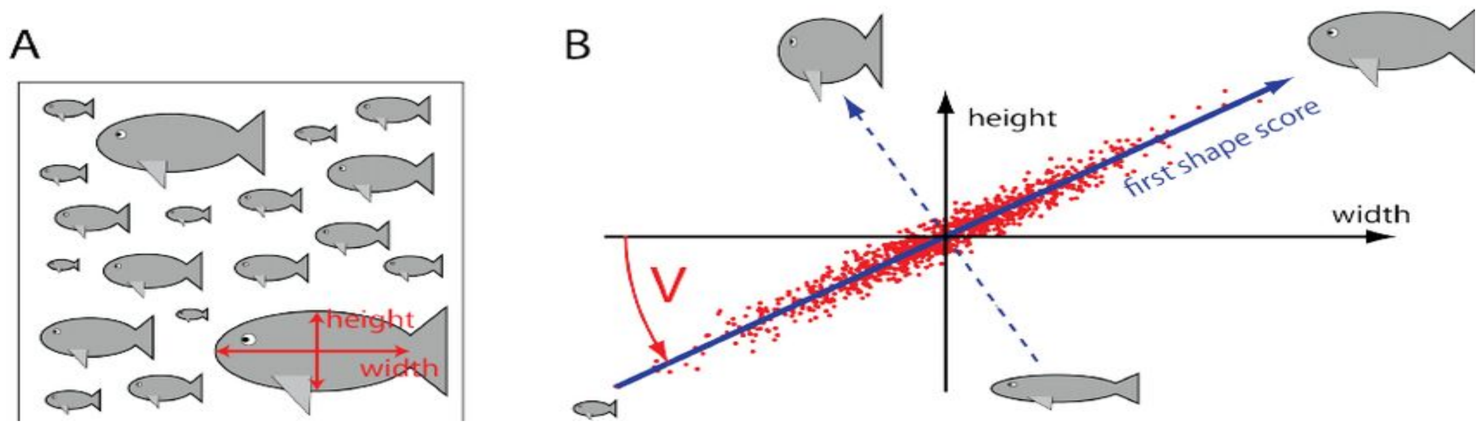
# Principal Component Analysis



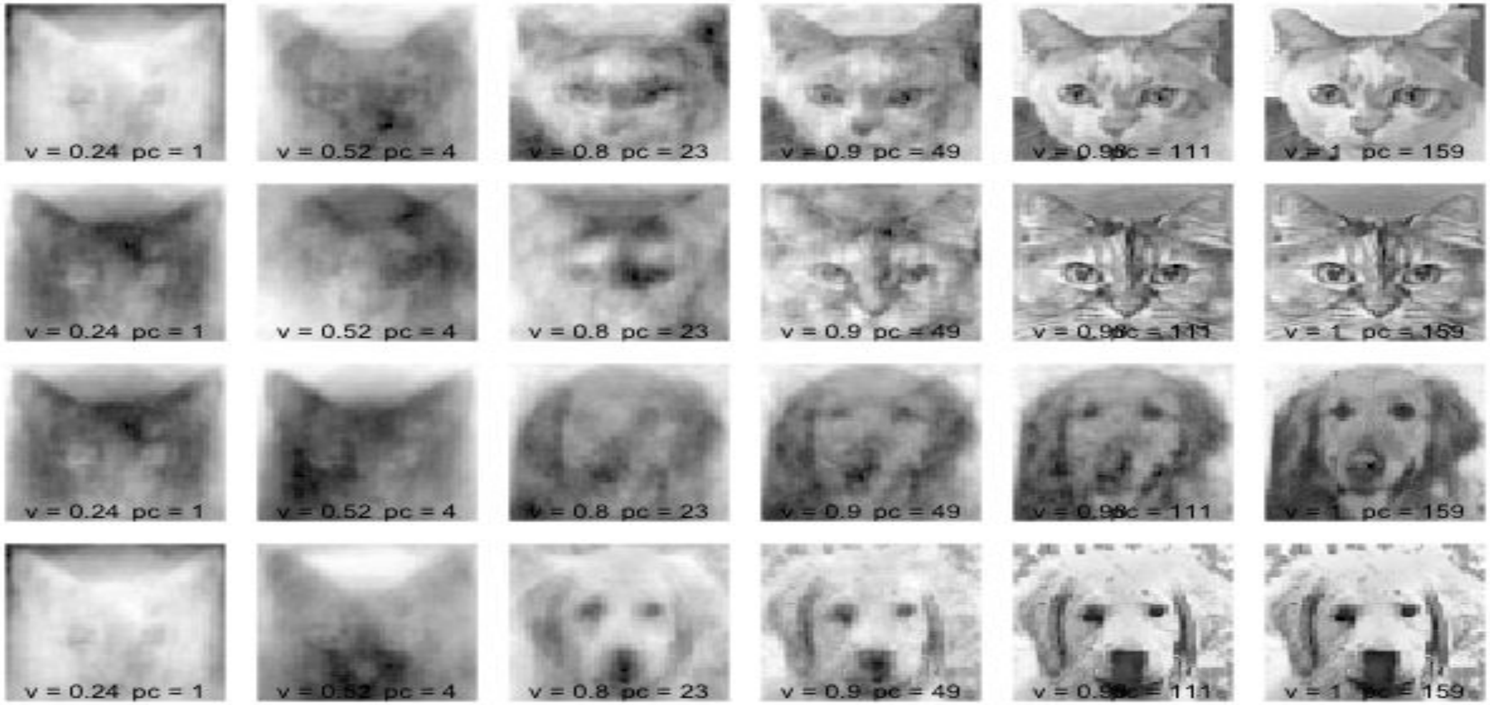

- PCA is a feature extraction method and extracts the most important information.
- This in turn leads to compression because the less important information are discarded. With fewer data points to consider, it becomes simpler to describe and analyse the dataset.
- PCA can be seen a trade-off between faster computation and less memory consumption versus information loss.
- PCA is considered as one of the most useful tools for data analysis.

# Example of PCA

- We can describe the shape of a fish with two variables: height and width.
- Given the height, we can probably estimate the width; and vice versa. Thus, we may say that the shape of a fish can be described with a single component.
- This doesn't mean that we simply ignore either height or width. Instead, we transform our two original variables into two orthogonal (independent) components
- The first component (blue line) will explain most of the variation in the data.
- The second component (dotted line) will explain the remaining variation.



# Example- Facial Recognition




- PCA can be applied for facial recognition. For 90% capture variance, only a third of the components had to be retained. This may be sufficient for Machine Learning applications.

# Computation Of PCA

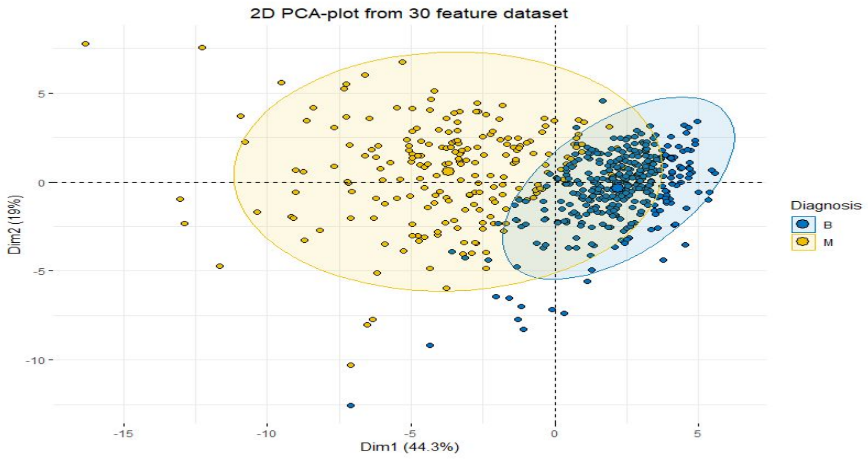
The below steps need to be followed to perform dimensionality reduction using PCA:

- Standardization of the data
- Computing the covariance matrix
- Calculating the eigenvectors and eigenvalues
- Computing the Principal Components
- Reducing the dimensions of the data set

# When should I use PCA?

- Do you want to reduce the number of variables, but aren't able to identify variables to completely remove from consideration?
- Do you want to ensure your variables are independent of one another?
- Are you comfortable making your independent variables less interpretable?

If you answered "yes" to all three questions, then PCA is a good method to use. If you answered "no" to question 3, you **should not** use PCA.



# PCA- Advantages/Disadvantages

INTERNSHIPSTUDIO

## ADVANTAGES

Both objective and subjective attributes can be used.

- It can be done accurately (only) with the help of Statistical software.
- Direct inputs from treatments.
- There is flexibility in naming and using dimensions.
- PCA is useful for finding new, more informative and uncorrelated components.
- It reduces dimensionality by rejecting lower variance components.

# DISADVANTAGES

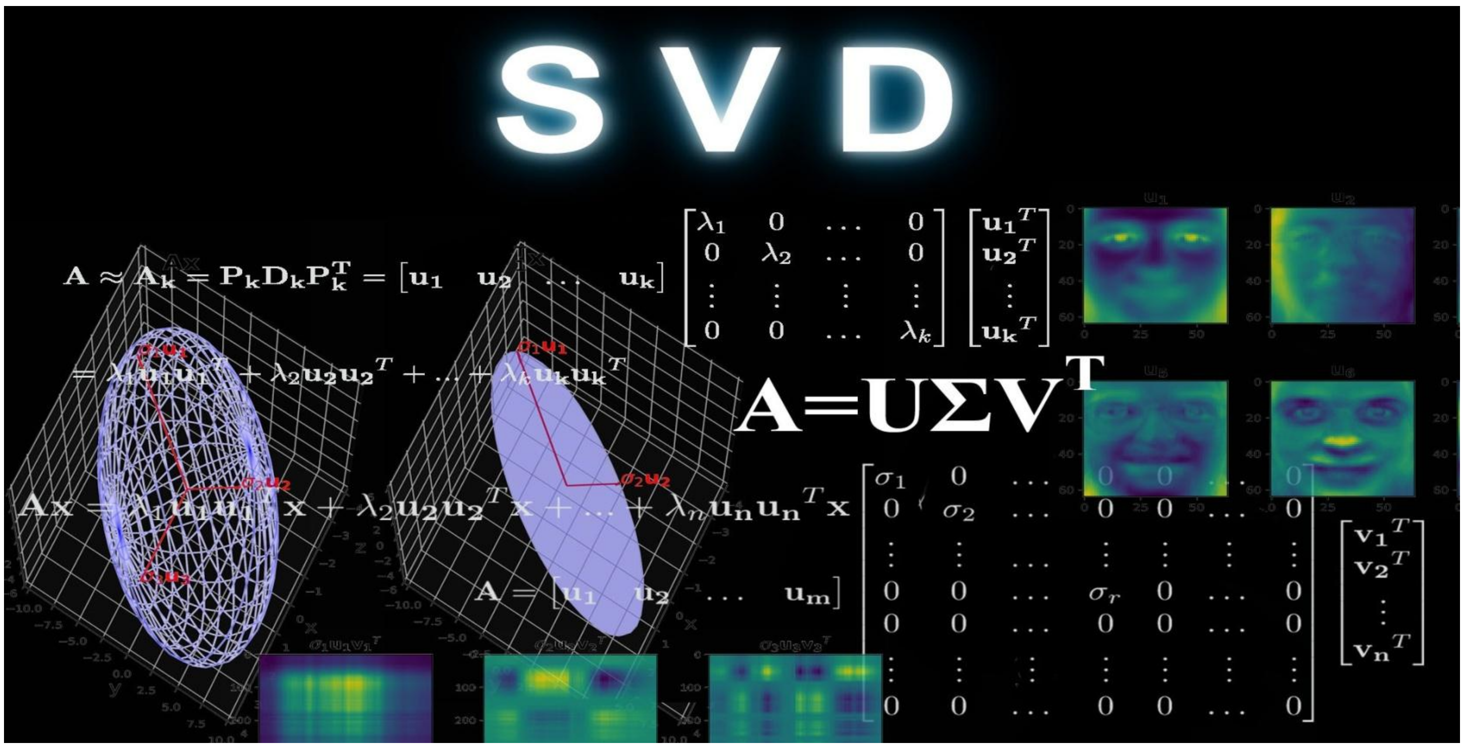
- Usefulness depends on the researchers' ability to develop a complete and accurate set of attributes - If important attributes are missed, precision of the procedure is reduced accordingly.

- Naming of the factors (independent variables) can be difficult - multiple attributes can be highly correlated with no apparent reason.

- If the observed variables are completely unrelated, PCA analysis is unable to produce a meaningful pattern.

# Singular Value Decomposition

- Singular Value Decomposition (SVD) is another dimensionality reduction technique in data science.
- SVD allows us to extract and untangle information.




- Q.1 What are the properties of KNN?
- Q.2 What do you mean by PCA?
- Q.3 Explain Euclidean Distance?
- Q.4 What do you mean by SVD?
- Q.5 What is the role of PCA?

# Thank You

INTERNSHIPSTUDIO