DATA ANALYTICS

Data Mining vs Data Analytics

Criteria	Data Mining	Data Analytics
Focus	Discovering patterns and relationships in large datasets	Analysing data to draw insights and make informed decisions
Purpose	Extracting useful information from data	Optimizing business processes, improving customer experiences, developing data-driven strategies
Techniques	Machine learning, statistics, database systems	Statistical analysis, data visualization, data cleaning and processing
Output	Patterns, relationships, trends, anomalies	Insights, recommendations, optimized processes

Correlation

- Correlation is a statistical measure that describes the strength and direction of the relationship between two or more variables.

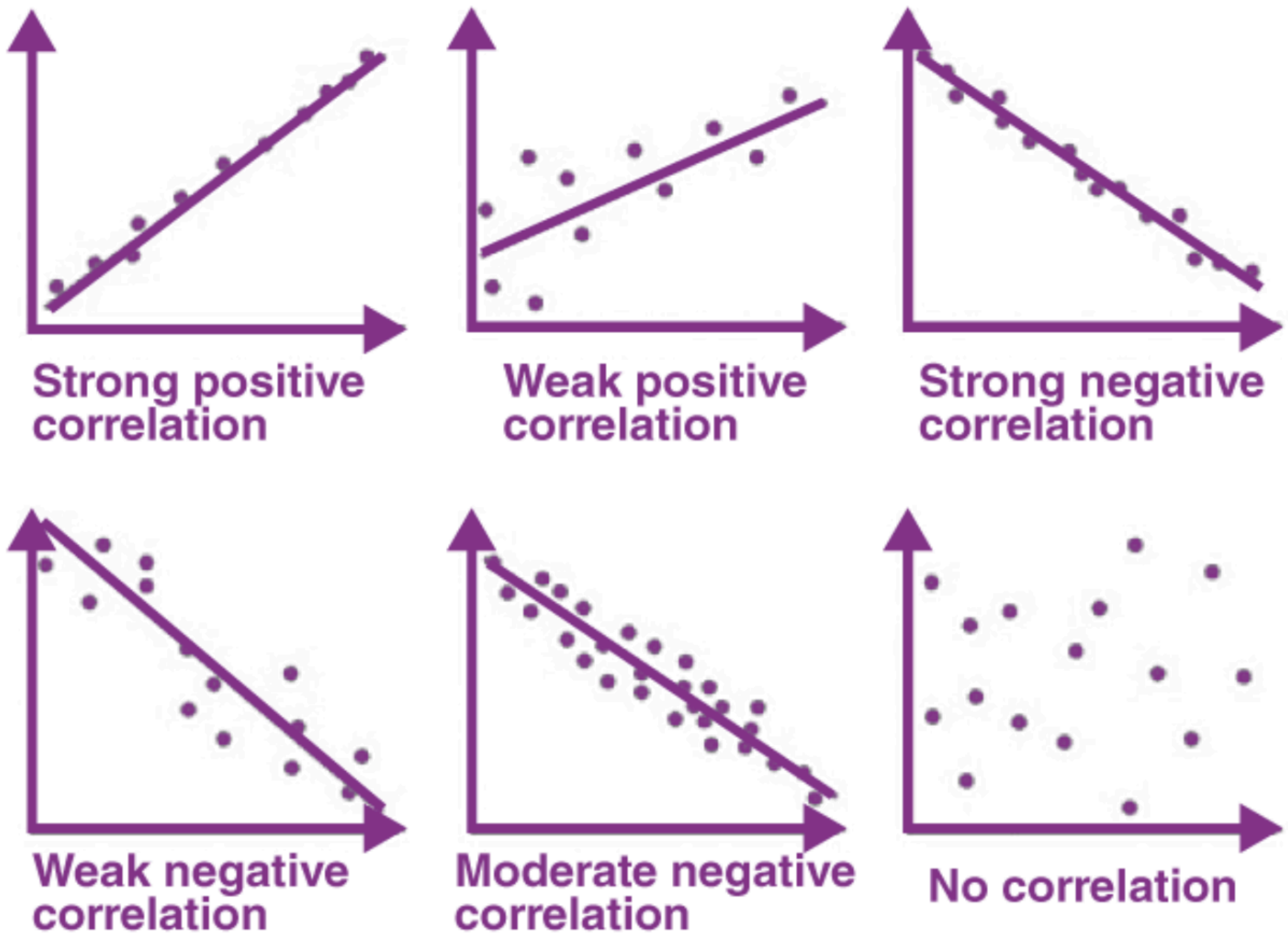
- It is commonly used to analyze data and to identify patterns and relationships between variables. Correlation can be expressed as a value between -1 and 1, where a value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation

- Correlation can be visualised using a scatter plot, which shows the relationship between two variables on a graph.

- If the points on the scatter plot form a line or a curve, this suggests that there is a relationship between the variables.

- The slope of the line or curve indicates the direction of the relationship, while the tightness of the points around the line or curve indicates the strength of the relationship.

Correlation



Correlation vs Regression

Criteria	Correlation	Regression
Definition	Describes the strength and direction of the relationship between two or more variables	Estimates the relationship between a dependent variable and one or more independent variables
Purpose	Identifying patterns and relationships between variables	Predicting the value of a dependent variable based on the value of one or more independent variables
Directionality	Can be positive, negative, or zero	Dependent variable always has a positive relationship with independent variables
Output	Correlation coefficient and scatter plot	Regression equation and predicted values
Types	Pearson, Spearman, Kendall	Simple linear, multiple linear, logistic, polynomial, and others

Correlation Coefficient

Subject	Age x	Glucose Level y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Correlation Coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Correlation Coefficient

Subject	Age x	Glucose Level y	xy	x2	y2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Correlation Coefficient

From our table:

Σx = 247

Σy = 486

Σxy = 20,485

Σx2 = 11,409

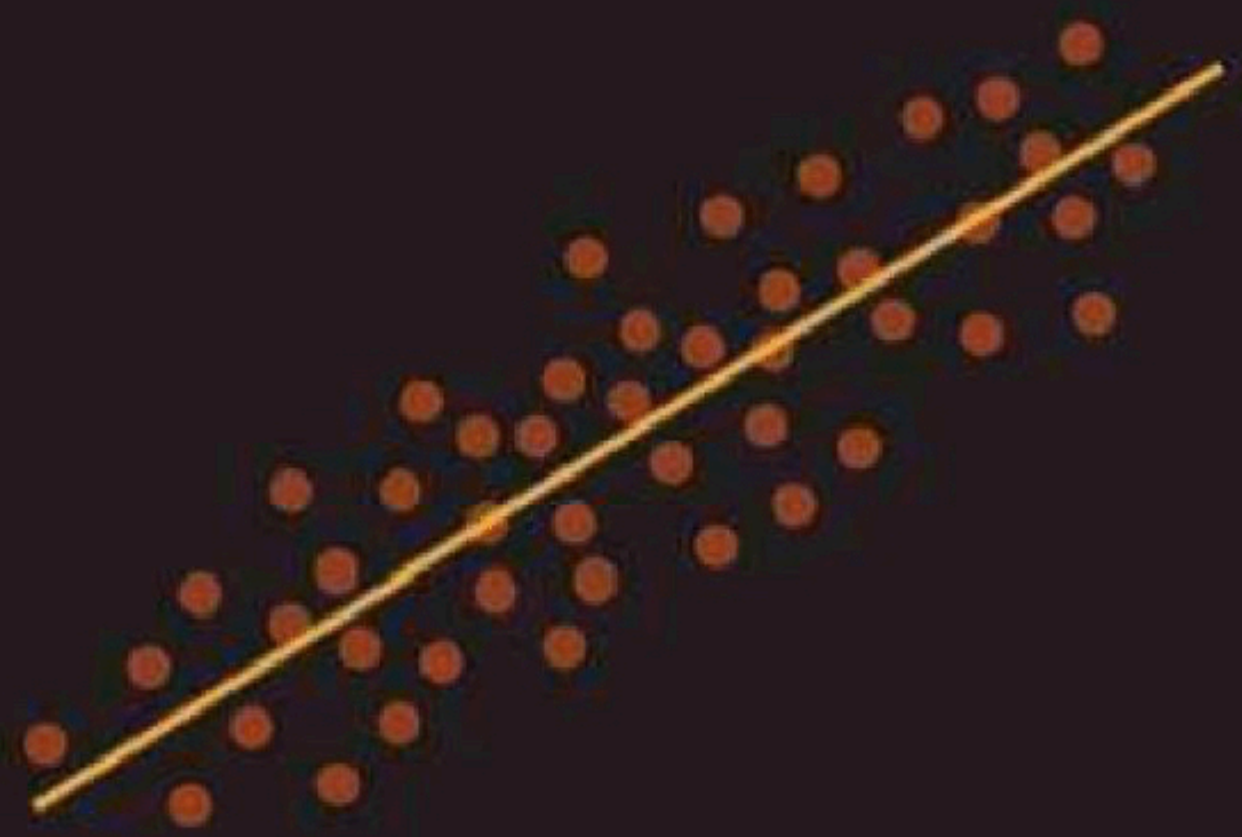
Σy2 = 40,022

n is the sample size, in our case = 6

The correlation coefficient =

6(20,485) – (247 × 486) / [√[[6(11,409) – (2472)] × [6(40,022) – 4862]]]

= 0.5298



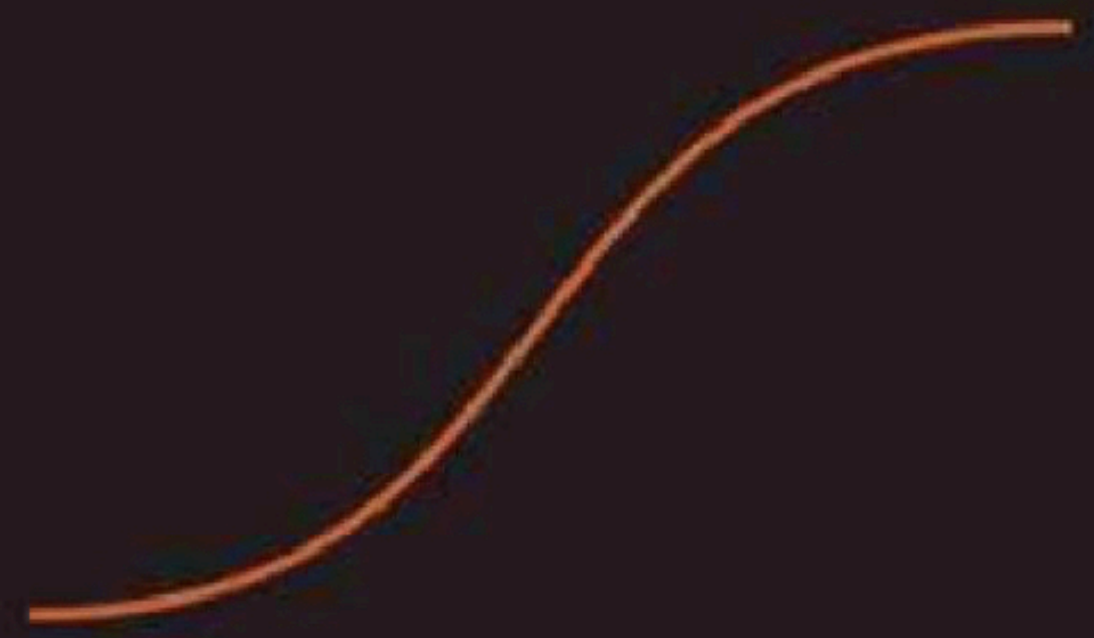
Linear Regression

GOOD

- Simple to implement and efficient to train.
- Overfitting can be reduced by regularization.
- Performs well when the dataset is linearly separable.

BAD

- Assumes that the data is independent which is rare in real life.
- Prone to noise and overfitting.
- Sensitive to outliers.



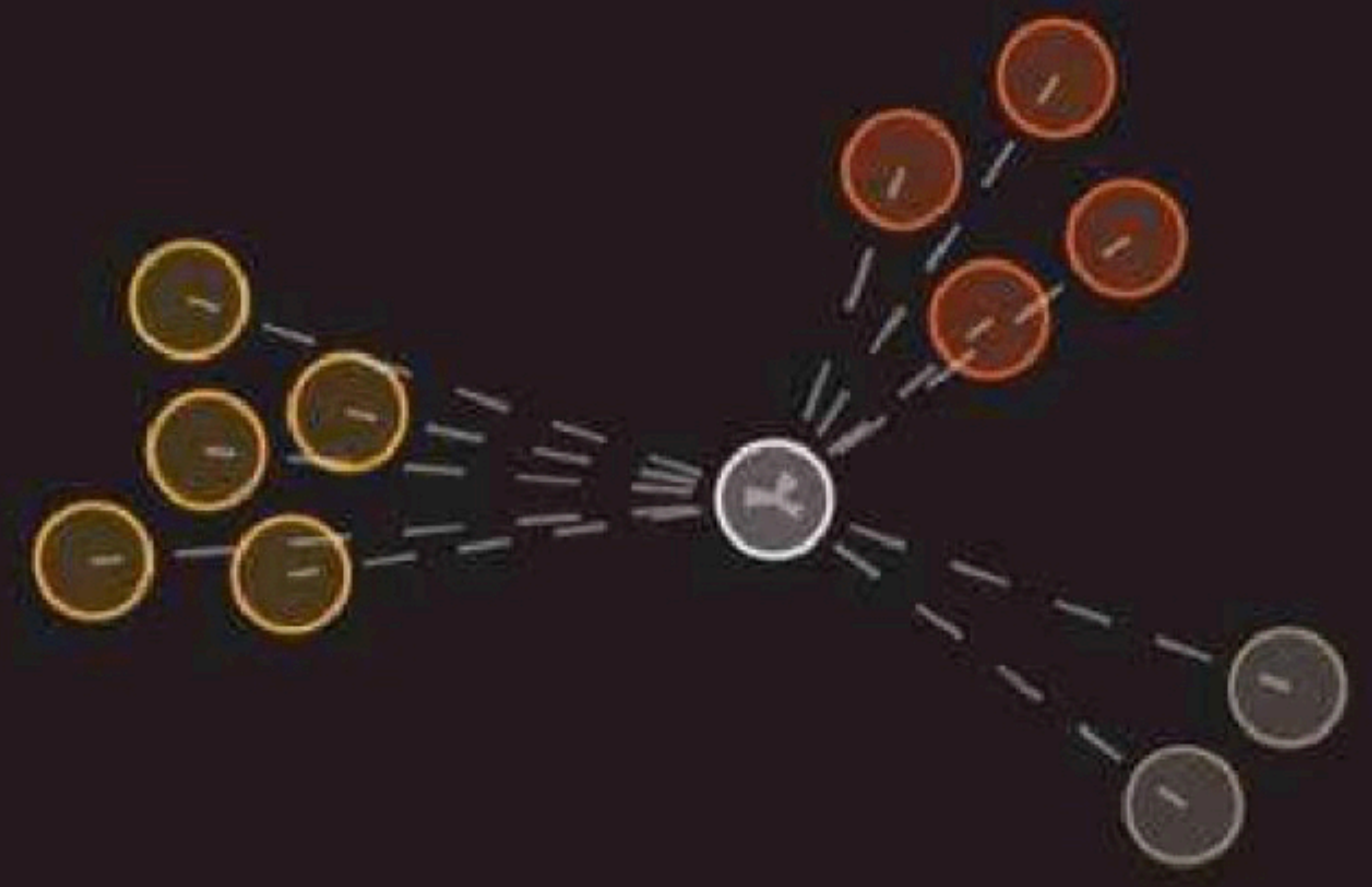
Logistic Regression

GOOD

- Less prone to over-fitting but it can overfit in high dimensional datasets.
- Efficient when the dataset has features that are linearly separable.
- Easy to implement and efficient to train.

BAD

- Should not be used when the number of observations are lesser than the number of features.
- Assumption of linearity which is rare in practise.
- Can only be used to predict discrete functions.



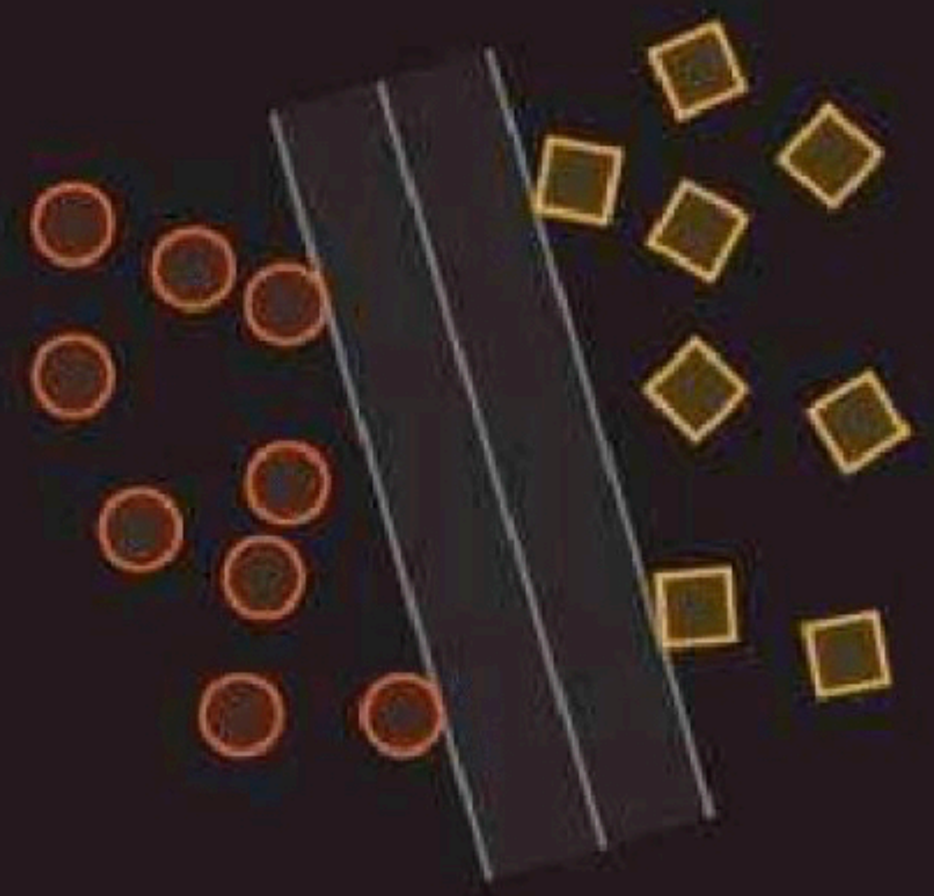
K Nearest Neighbour

GOOD

- Can make predictions without training.
- Time complexity is O(n).
- Can be used for both classification and regression.

BAD

- Does not work well with large dataset.
- Sensitive to noisy data, missing values and outliers.
- Need feature scaling.
- Choose the correct K value.



Support Vector Machine

GOOD
- Good at high dimensional data.
- Can work on small dataset.
- Can solve non-linear problems.

BAD
- Inefficient on large data.
- Requires picking the right kernal.

$$P(y|x) = \frac{P(x \mid y)P(y)}{P(x)}$$

Naive Bayes

GOOD

- Training period is less.
- Better suited for categorical inputs.
- Easy to implement.

BAD

- Assumes that all features are independent which is rarely happening in real life.
- Zero Frequency.
- Estimations can be wrong in some cases.

Regression Example