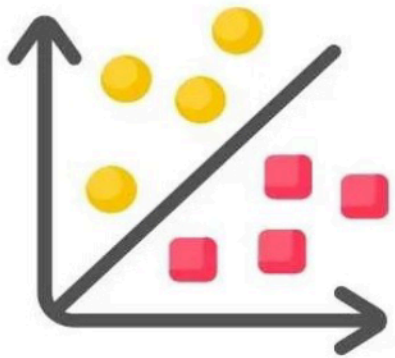
Classification and Regression

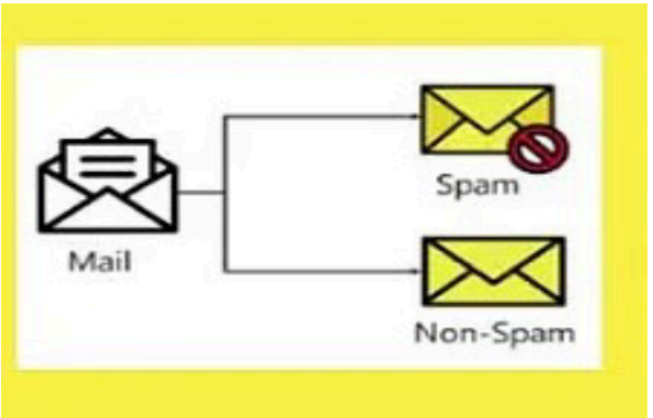
Dr Rani Oomman Panicker

Asst. Prof., CSE, MIT

What is Classification?

- Classification is a **supervised machine learning method** where the model tries to predict the correct label of a given input data.
- Classification algorithms are used to predict/Classify the discrete values
- Eg: hot or cold, male or female, true or false





Classification Types:

- Binary Classification: This classification problem can fall into two classes. (eg: spam or not spam, TB or not TB, 0 or 1)
- Multi-Class Classification: classification can be from more than two classes(eg: wheather: sunny, cloudy, or rainy).

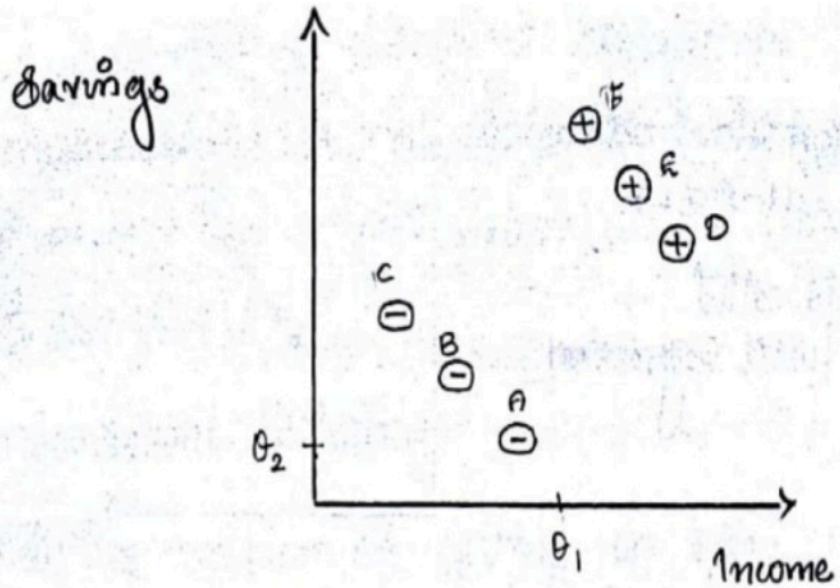
Classification example

Credit Card (Risk associated with a Credit Card)

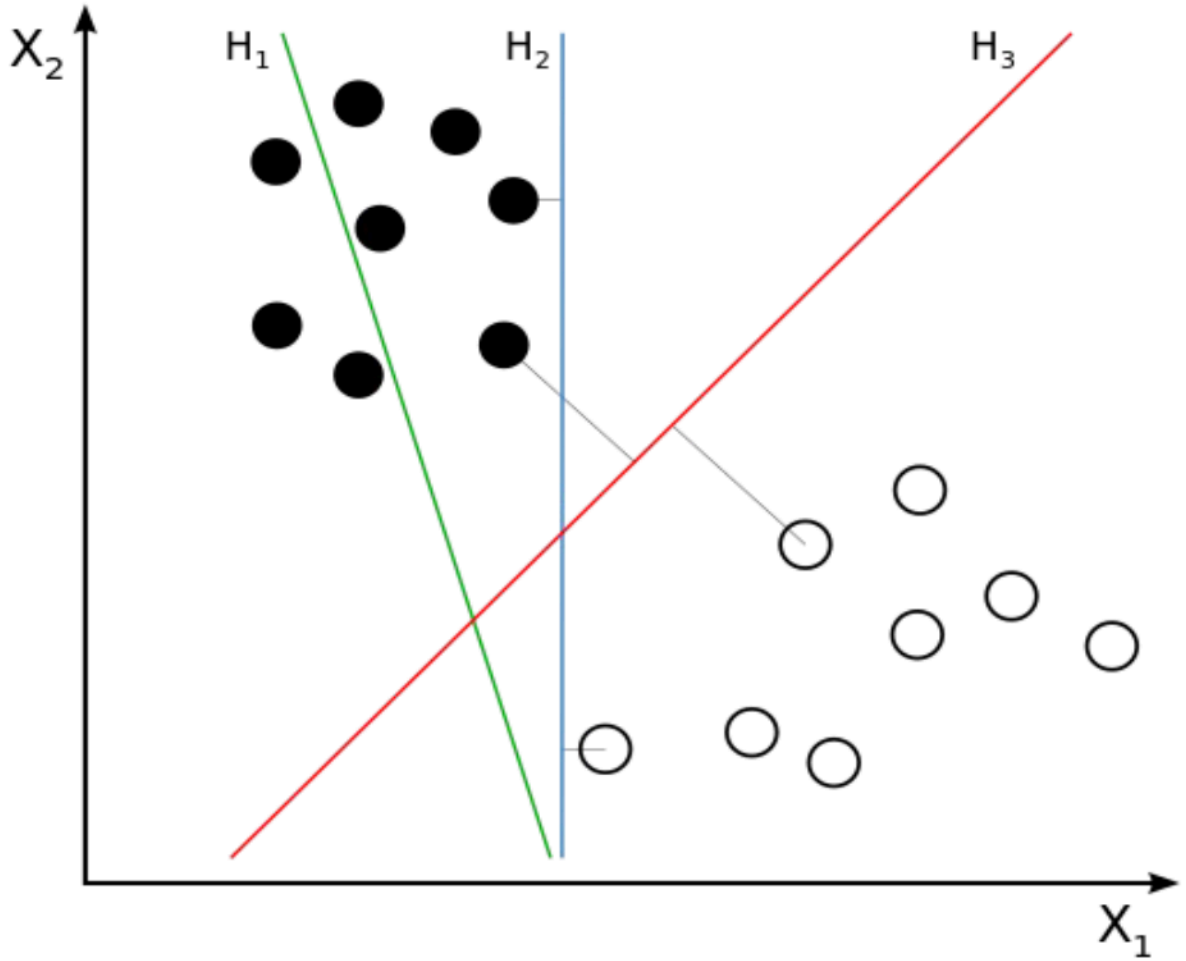
	Income	Savings	Label
A	20000	5000	High_Risk
B	19000	6000	High_Risk
C	18000	7000	High_Risk
D	31,600	11000	Low_Risk
E	30500	12000	Low_Risk
F	30250	13000	Low_Risk

⊕ → Low_Risk

⊖ → High_Risk



Which of these lines, H1, H2, and H3, represents the worst classifier algorithm?



Key Characteristics of Classification:

- 1. Discrete Output: Classification produces a discrete output, which means the predicted values fall into specific categories or labels. For example, it can be used to classify emails as spam or not spam, identify animals in images, or determine the sentiment of a text as positive, negative, or neutral.

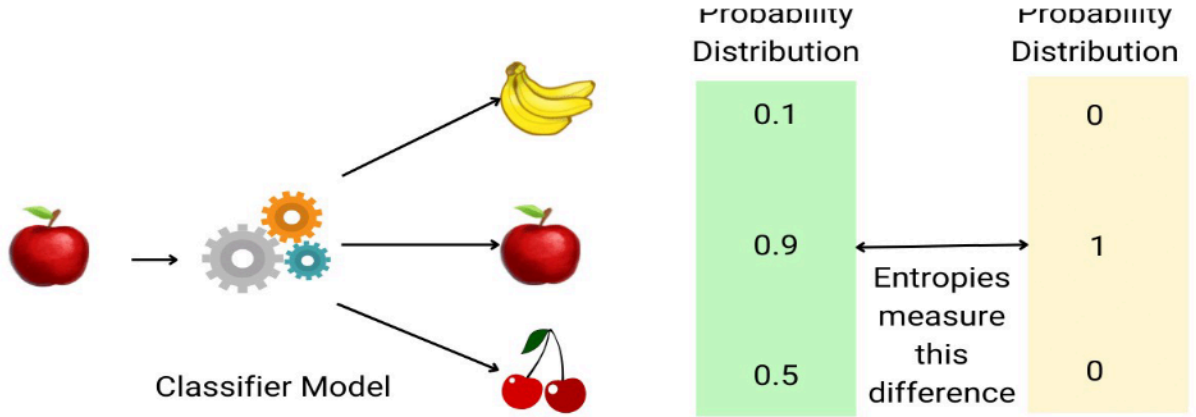
- 2. Supervised Learning: Classification requires *labelled training data*, where each data point is associated with a known class or category. The algorithm learns from this labelled data to make predictions on new, unseen data.

- 3. Common Algorithms: Several classification algorithms are available, including *logistic regression, decision trees, support vector machines (SVM), k-nearest neighbours (K-NN), and various deep learning techniques like neural networks.*

- 4. Evaluation Metrics: Classification models are typically evaluated using metrics like *accuracy, precision, recall, F1-score, and the confusion matrix*. These metrics help assess the model's ability to correctly classify data points into the appropriate classes.

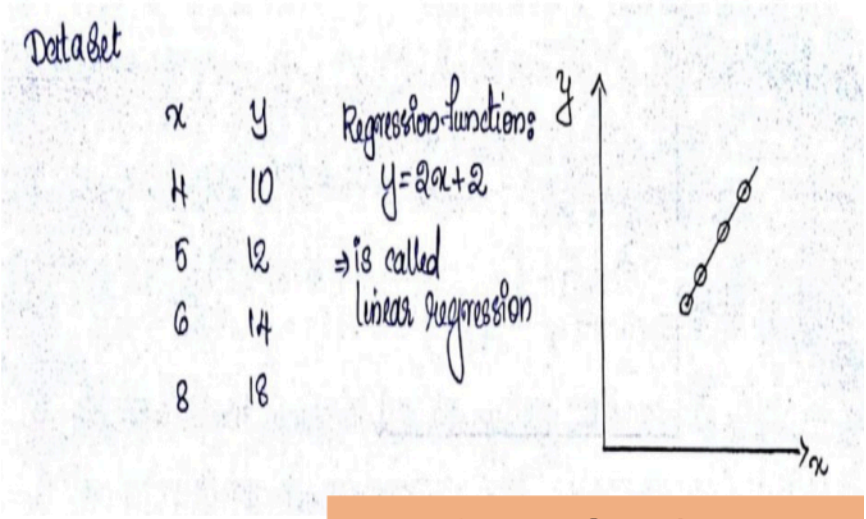
For example, suppose there are three class labels, **[Apple, Banana, Cherry]**. But the problem is that machines don't have the sense to understand these labels. That's why we need to convert these labels into a machine-readable format. For the above example, we can define **Apple = [1,0,0], Banana = [0,1,0], and Cherry = [0,0,1].** Once the machine learns from these labelled training datasets, it will give probabilities of different classes on the test dataset like this: **[P(Apple), P(Banana), P(Cherry)].**

These predicted probabilities can be from one type of probability distribution function (PDF), and the actual (true) labelled dataset can be from another probability distribution function (PDF). If the predicted distribution function follows the exact distribution function, the model is learning accurately. **Note:** These PDF functions are continuous. As a similarity between classification and regression, if the predicted PDF follows the actual PDF, we can say the model learns the trends.



Regression

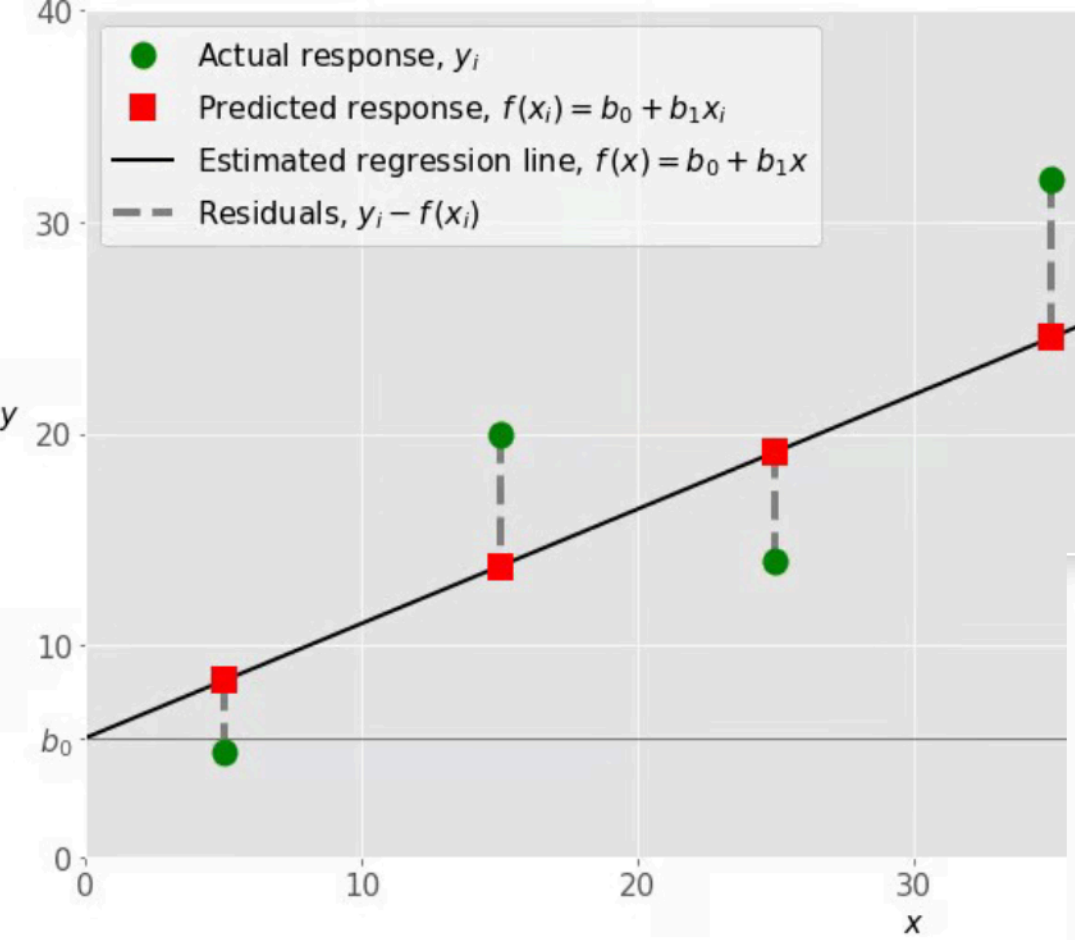
What is Regression?



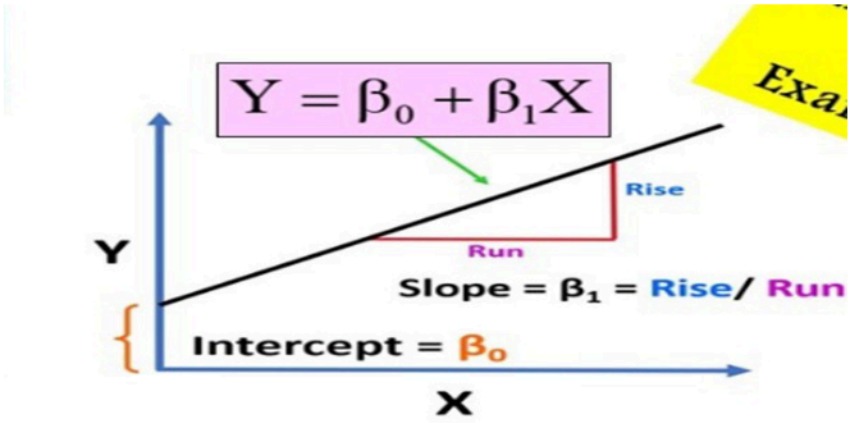
- is a supervised learning task
- regression aims to predict continuous numerical values.
- In regression, the ***output is a real number*** *that can fall within a range*.

Experience → Salary
Height → weight

- Regression models are used to ***find relationships between input features and the target variable,*** allowing for the prediction of numeric outcomes.
- Regression is a predictive modelling technique that models the relationship between a dependent variable and one or more independent variables.
- Regression analysis aims to estimate the dependent variable's value based on the independent variables' importance.



Regression Line or line of best fit



Cost Function

We want the Regression line to resemble the dataset as closely as possible. In other words, we want the line to be as close to actual data points as possible. It can be achieved by minimizing the vertical distance between the actual data point and the fitted line. I calculate the vertical distance between each data point and the line. This distance is called the **residual**.

So, in a regression model, we try to minimize the residuals by finding the line of best fit. The residuals are represented by the vertical dotted lines from actual data points to the line. We can try to minimize the sum of the residuals, but then a large positive residual would cancel out a large negative residual. For this reason, we minimize the sum of the squares of the residuals.

Mathematically, we denote actual data points by yi and predicted data points by ŷi. So, the residual for a data point i would be given as

```
di = yi -  ŷi
```

The sum of the squares of the residuals is given as:

```
D = Σ di**2       for all data points
```

This is the **Cost function**. It denotes the total error present in the model, which is the sum of the

Examples of simple linear regression

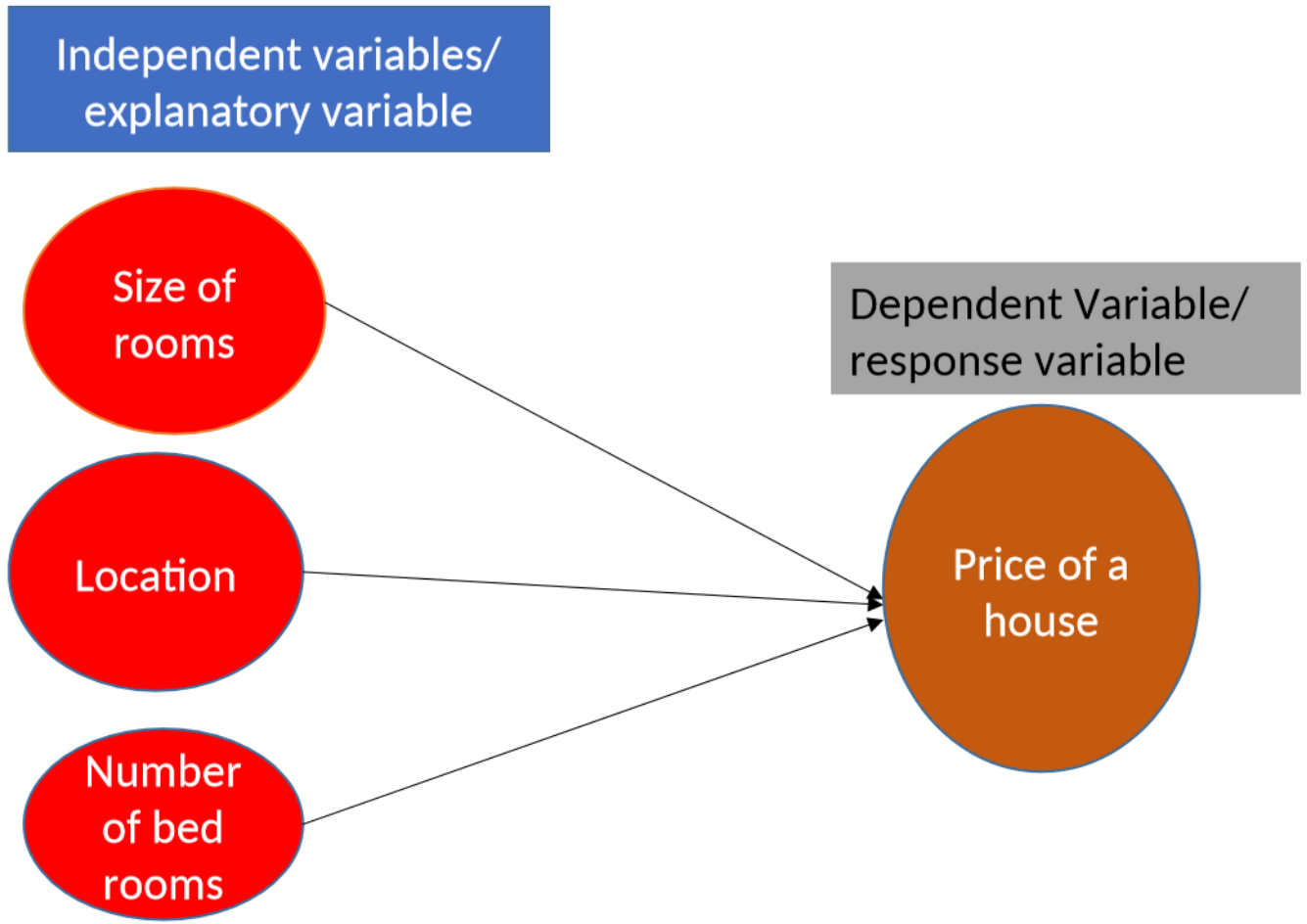
Comparing Dependent and Independent Variables

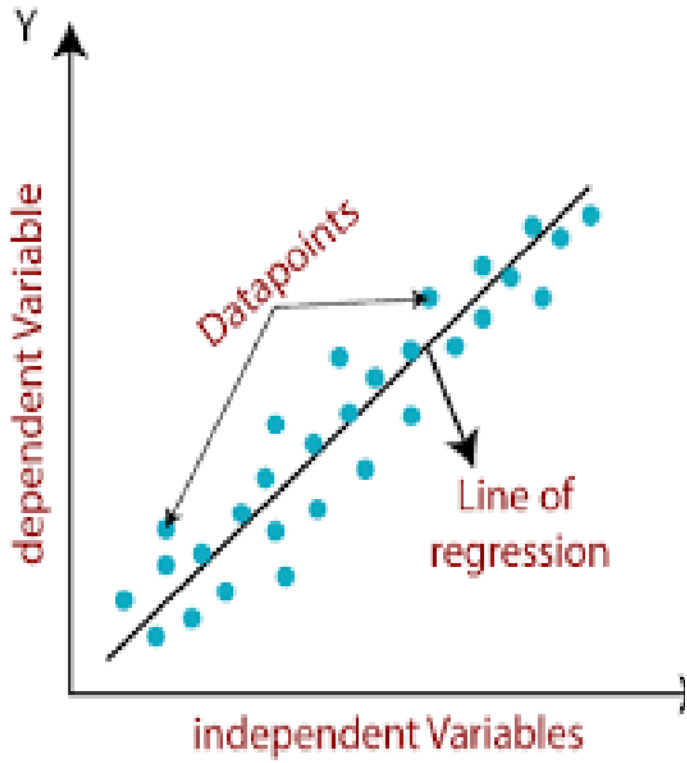
Research Topic	Independent Variable	Dependent Variable
Plants grow faster in warmer temperatures.	Temperature	Plant Growth
To what extent does traffic affect a person's mood?	Traffic	Mood
People walk slower after drinking coffee.	Drinking Coffee	Walking Speed

Multiple linear Regression example

location, size, and number of bedrooms ⟹ price of a house.

Brand, year ,engine capacity, mileage ⟹ price of a car





Key Characteristics of Regression:

- 1. Continuous Output: Regression produces a continuous output, which means the predicted values can be any real number within a certain range.

  eg:- predicting house prices, stock prices, or a patient's blood pressure.

- 2. Supervised Learning: Similar to classification, regression requires labelled training data. The algorithm learns from this data to establish relationships between the input features and the continuous target variable.

- 3. Common Algorithms: Common regression algorithms include linear regression, decision tree regression, random forests, and regression neural networks. These algorithms are chosen based on the specific characteristics of the data and the problem.

- 4. Evaluation Metrics: Regression models are evaluated using metrics such as mean squared error (MSE), mean absolute error (MAE), R-squared (coefficient of

Classification Use cases:

- 1. Spam Detection: Classify emails as spam or not spam based on their content and characteristics.
- 2. Image Classification: Identify objects or patterns in images, such as classifying images of animals or recognising handwritten digits. (cat or dog)
- 3. Sentiment Analysis: Determine the sentiment of textual data, such as product reviews, social media posts, or comments, as **positive, negative, or neutral**.
- 4. Disease Diagnosis: Classify medical conditions based on patient data and diagnostic tests.

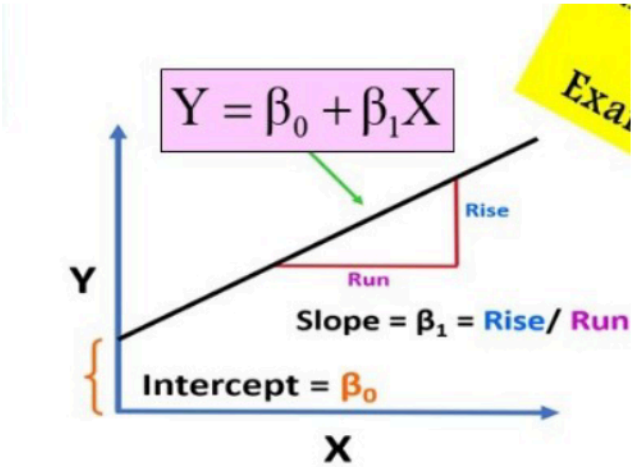
  (cancer or not cancer)

Regression Use cases:

- 1. House Price Prediction: Predict the price of a house based on features like location, size, and number of bedrooms.
- 2. Stock Price Forecasting: Use historical stock data to forecast future stock prices and trends.
- 3. Weather Forecasting: Predict temperature, rainfall, or other meteorological variables for a specific location and time.
- 4. Demand Forecasting: Estimate future demand for products or services based on historical sales data.

Types of Regression

1. Simple linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Logistic Regression
5. Ridge Regression
6. Lasso Regression



1. Simple Linear Regression

Simple Linear Regression, "*models relationship between two variables in data*"

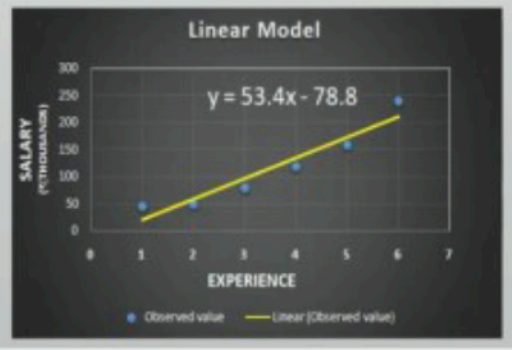
Experience	Salary (₹Thousands)
2	50
4	100
6	250
8	300
9	440
10	700

Observed Data

Simple Linear Regression

Model Construction

Linear Model



New Data

Experience	Salary (₹Thousands)
7	?

Model Usage: Prediction

Experience	Salary (₹Thousands)
7	295

Predicted Value

Steps of Simple LR:

1. Identify dependent and independent variables
2. Understand the linear relationship
3. Find slope and intercept
4. Find the best fit line (linear model) or regression line

$$Y = \beta 0 + \beta 1 * X$$

STEP1: *Dependent & Independent Variables*

- **Data** - *collection of recorded observations.*
- **Variable** –*describes the observation.*

1. **Dependent Variable (*Response Variable or Outcome Variable*)**
2. **Independent Variable (*Explanatory Variable*)**

Example Task : Predicting sales of ***ice cream*** based on ***Temperature***

Data (D)

Temp. (° C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

Variables (Columns)

Instances (Rows)

Dependent Variable : Ice Cream Sales

Independent Variable : Temperature

STEP 2: Linear Relationship *(Positive vs Negative)*

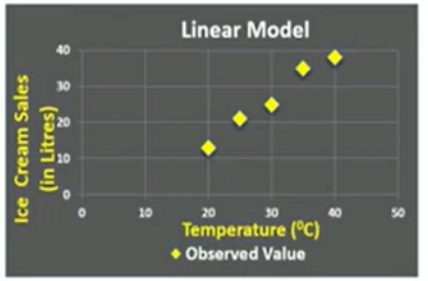
Example 1 :Ice Cream Sales Prediction

Positive Linear Relation

Independent Variable (X-axis)

Temp. (° C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

Dependent Variable (Y-axis)



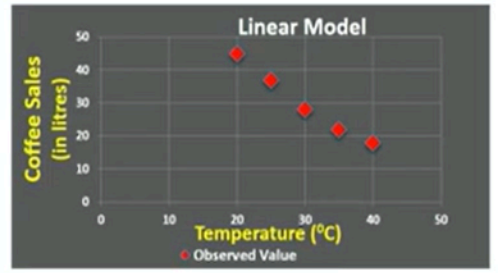
Example 2: Coffee Sales Prediction

Negative Linear Relation

Independent Variable (X-axis)

Temp. (° C)	Coffee Sales (in litres)
20	45
25	37
30	28
35	22
40	18

Dependent Variable (Y-axis)



STEP 3: Find the best fit line (Linear Model)

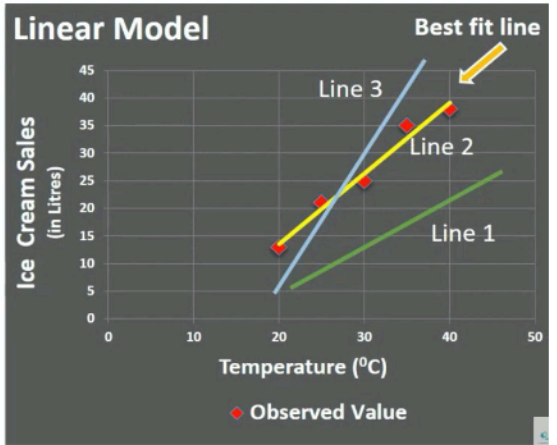
- Simple Linear Regression works with 2-variables in Data
- Best fit line is a straight line that best fits the observed data point
- Best fit Line (Regression line) is the constructed linear Model.

Temp. (° C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

Simple Linear Regression

1. Identifies a Positive Linear Relation
2. Finds the line that best fit the observed data

Example: Line 2 best fit the data

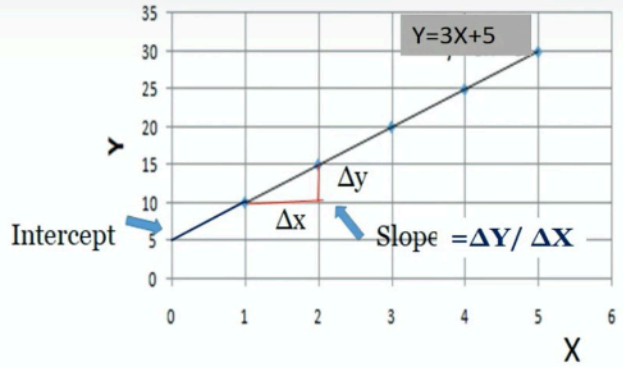


Quick recap on Line Equation

- Two variables => **2D graph**
- Line is represented as,

Y=a (X) + b
Y =3X+5

X-independent variable
Y- Dependent Variable
a- slope , b- intercept



Slope measure rate of change in Y when X changes

Intercept a point where the line cuts the Y-axis

To Draw a Line in 2D we have to find slope and intercept

Find Slope and Intercept for a line ?

Least Square regression method for finding regression (best fit) line

Finding slope and intercept for a straight line
Y=a (X) + b

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - a(\sum x)}{n}$$

This formula is also derived using **Covariance, Correlation Coefficient, Standard deviation & Variance** of X and Y

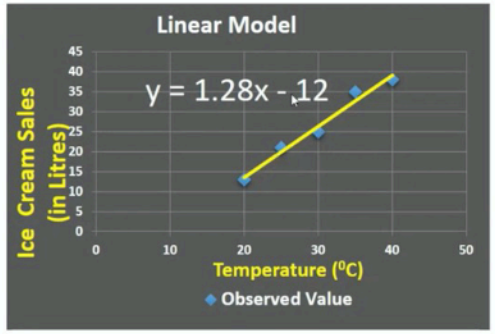
a-slope; b- intercept; X- Independent Variable ; Y-Dependent Variable; n number of data instances

Example :

Ice Cream Sales Prediction

Y =1.28 (X) -12

Temp. (° C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

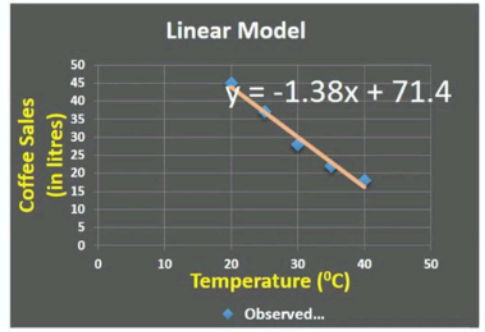


Ice cream sales= **1.28** (Temperature)**-12**

Coffee Sales Prediction

Y=-1.38 X+71.4

Temp. (° C)	Coffee Sales (in litres)
20	45
25	37
30	28
35	22
40	18



Coffee sales= **-1.38** (Temperature)**+71.4**

Simple Linear Regression (without error component)

$Y=aX+b$

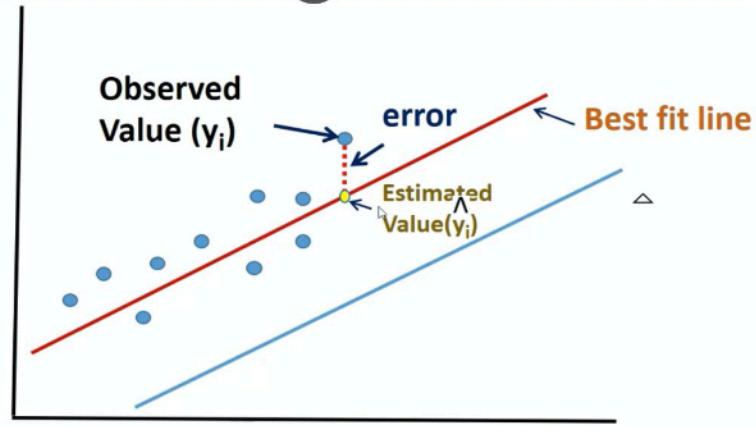
(or)

Simple Linear Regression (with error component $\mathcal{E}$)

$Y= \beta_o X+\beta_1 +\mathcal{E}$

Model Evaluation?

Best fit line is selected based on the least **residual error**



Sum of Squares Error (SSE) or Sum of Squares Residual Error

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output — Coefficients — Input — Error

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

Sample problem of LR

An Ice cream shop owner wants to **predict the sales of ice cream** based on temperature. A data sample of his transaction is given in Table 1. Find a regression line that can do the prediction.

Table 1

Temp. (^{0} C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

What will be the sales of ice cream when temperature =28 degree Celsius?

Answer

Step 1: For each data instances or point (x, y) calculate x^2 and xy

Step 2: Find the sums : Σx, Σy, Σx2 and Σxy

Temp. (^{0} C)	Ice Cream Sales (in litres)
20	13
25	21
30	25
35	35
40	38

X (Temp)	Y (Ice Cream Sales)	x^2	xy
20	13	400	260
25	21	625	525
30	25	900	750
35	35	1225	1225
40	38	1600	1520
Σx =150	**Σy=132**	**Σx^2=4750**	**Σxy=4280**

Step 3: **Calculate Slope (a):**

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

=(4280*5-(150*132))/((5*4750)-(150)^2)

=1.28

Step 4: **Calculate intercept (b):**

$$b = \frac{\sum y - a(\sum x)}{n}$$

= (132- (1.28) (150))/5 **=-12**

Slope Intercept

$\hat{y}$ =1.28 (X) -12

Ice cream Sales =1.28 (Temperature)-12

What will be the sale of ice cream if temperature is 28°c ?

^

Y=1.28 (X) -12

= 1.28 (28)-12 = 23.84 litres

Linear Regression problem sample Qn

Solved Examples

Question: Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x$ = 20	$\sum y$ = 25	$\sum x^2$ = 120	$\sum xy$ = 144

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{4\times 144 - 20\times 25}{4\times 120 - 400}$$

$$b = 0.95$$

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{25\times 120 - 20\times 144}{4(120) - 400}$$

$$a = 1.5$$

Linear regression is given by:

$y = a + bx$

$y = 1.5 + 0.95\ x$

The equation is Y on X, where the value of Y changes with a variation in the value of X. The equation is X on Y, where the change in X variable depends upon the Y variable's deviation.

REGRESSION

① Find the equation of regression lines for the following data.

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Regression Equation of X on Y

$$X - \bar{x} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = \frac{N\Sigma xy - \Sigma x . \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

Regression Equation of Y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$b_{yx} = \frac{N\Sigma xy - \Sigma x . \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X = 28$	$\sum Y = 77$	$\sum X^2 = 140$	$\sum Y^2 = 875$	$\sum XY = 334$

Table 9.7

$$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4,$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$$

Regression coefficient of X on Y

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

(i) Regression equation of X on Y

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

$$X - 4 = 0.929(Y - 11)$$

$$X - 4 = 0.929Y - 10.219$$

$\therefore$ The regression equation X on Y is $X = 0.929Y - 6.219$

(ii) Regression coefficient *of* Y on X

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{7(334) - (28)(77)}{7(140) - (28)^2}$$

$$= \frac{2338 - 2156}{980 - 784}$$

$$= \frac{182}{196}$$

$\therefore \quad b_{yx} = 0.929$

(iii) Regression equation of Y on X

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$Y - 11 = 0.929\,(X - 4)$$

$$Y = 0.929X - 3.716 + 11$$

$$= 0.929X + 7.284$$

The regression equation of Y on X is $Y = 0.929X + 7.284$

Example 9.16

For 5 pairs of observations the following results are obtained $\sum X=15$, $\sum Y=25$, $\sum X2=55$, $\sum Y2=135$, $\sum XY=83$ Find the equation of the lines of regression and estimate the value of X on the first line when $Y=12$ and value of Y on the second line if $X=8$.

Here $N=5$, $\overline{X}=\frac{\Sigma X}{N}=\frac{15}{5}=3$, $\overline{Y}=\frac{\Sigma Y}{N}=\frac{25}{5}=5$

and the regression coefficient

$$b_{xy}=\frac{N\Sigma XY-\Sigma X\Sigma Y}{N\Sigma Y^2-(\Sigma Y)^2}=\frac{5(83)-(15)(25)}{5(135)-(25)^2}=0.8$$

The regression line of X on Y is

$$X-\overline{X}=b_{xy}(Y-\overline{Y})$$

$$X-3=0.8(Y-5)$$

$$X=0.8\,Y-1$$

When $Y=12$, the value of X is estimated as

$$X=0.8\,(12)-1=8.6$$

The regression coefficient

$$b_{yx}=\frac{N\Sigma XY-\Sigma X\Sigma Y}{N\Sigma X^2-(\Sigma X)^2}$$

$$=\frac{5(83)-(15)(25)}{5(55)-(15)^2}=0.8$$

Thus $b_{yx}=0.8$ then the regression line Y on X is

$$Y-\overline{Y}=b_{yx}(X-\overline{X})$$

$$Y-5=0.8(X-3)$$

$$Y=0.8X+2.6$$

When $X=8$ the value of Y is estimated as

$$Y=0.8(8)+2.6$$

$$Y=9$$

$Y-5=0.8(X-3)$

$=0.8X+2.6$

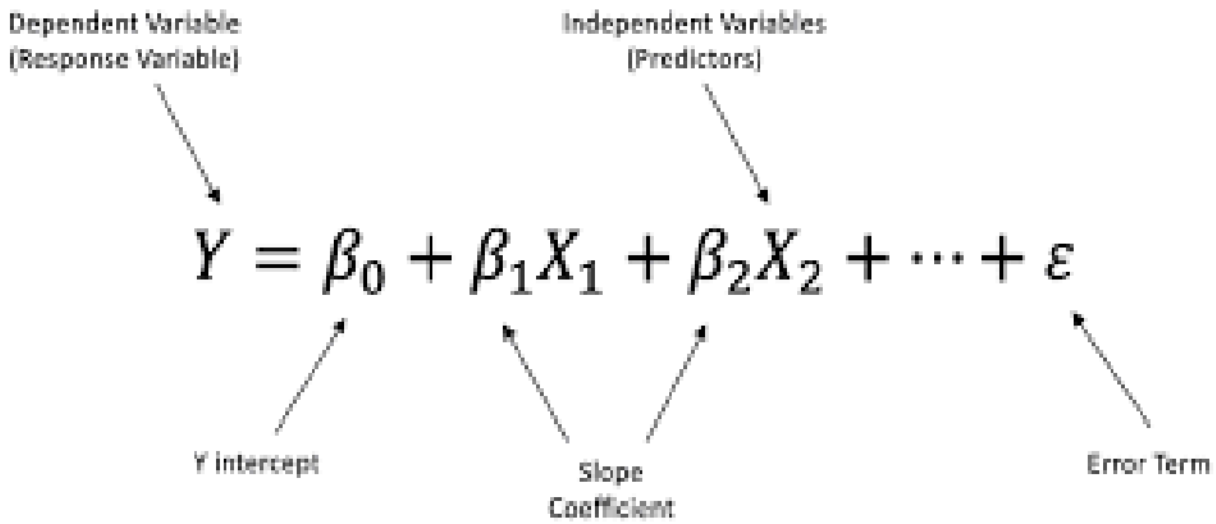
When $X=8$ the value of Y is estimated as

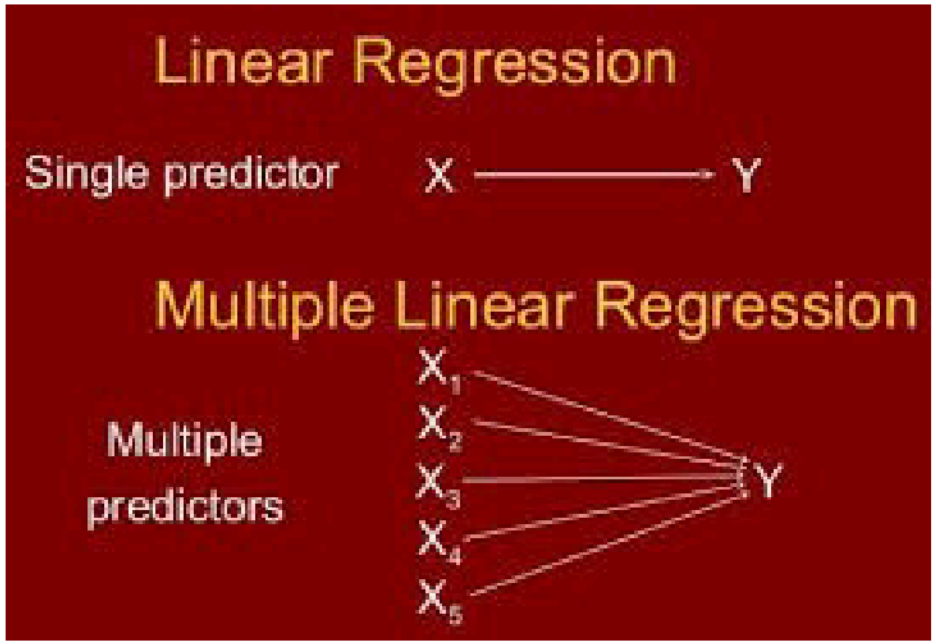
$=0.8(8)+2.6$

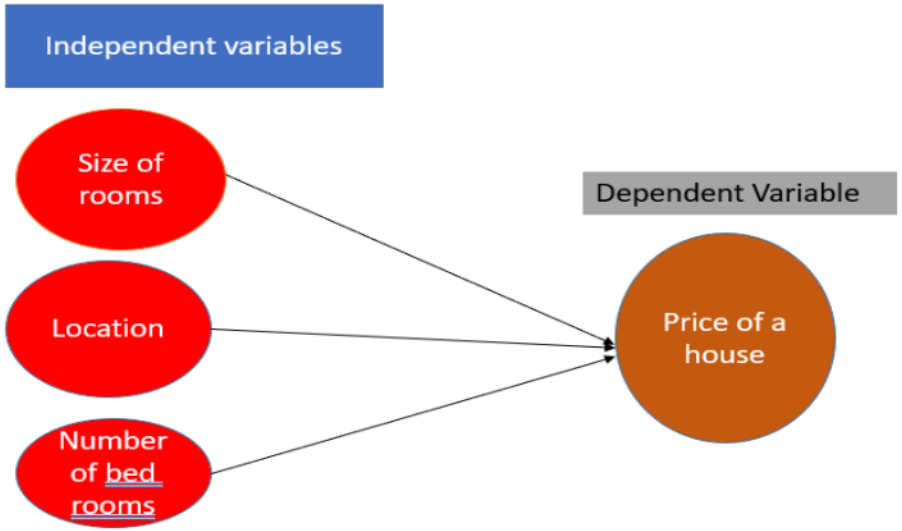
$=9$

2. Multiple Linear Regression

Multiple linear regression is a model for predicting the value of **one dependent variable** based on **two or more independent variables.**







Example problem (Multiple LR):
https://www.statology.org/multiple-linear-regression-by-hand/

Comparison

Parameter	Simple LR	Multiple LR
Definition	Models the relationship between one dependent and one independent variable.	Models the relationship between one dependent and two or more independent variables.
Equation	y = b0 + b1x + e	y = b0 + b1x1 + b2x2 + … + bnxn + e
Complexity	Simpler	More Complex
Risk of Overfitting	Lower	Higher
Multicollinearity Concern	Not applicable, as there's only one predictor.	A primary concern; having correlated predictors can affect the model's accuracy and interpretation.
Applications	Predict salary based on experience	based on location, size, and number of bedrooms predicts the price of a house.
Visualization	2D scatter plot	3D plot

3. Polynomial Regression

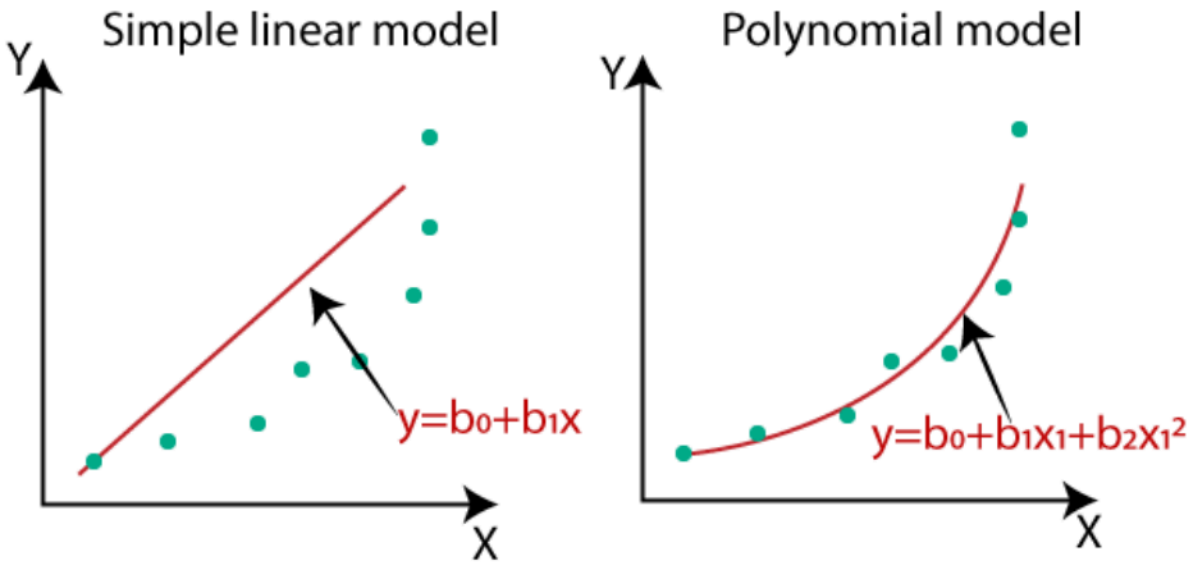
Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as **nth degree polynomial**.

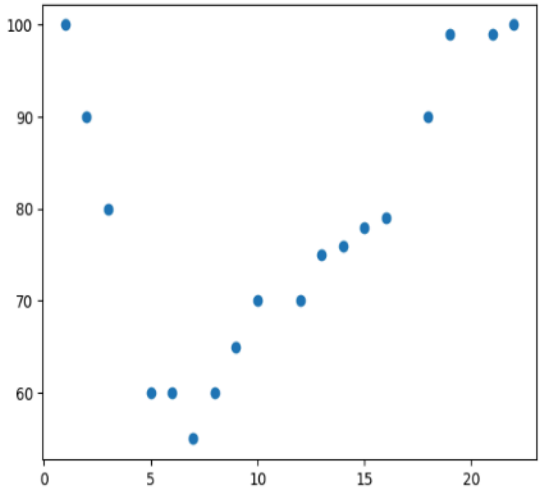
The Polynomial Regression equation is given below:

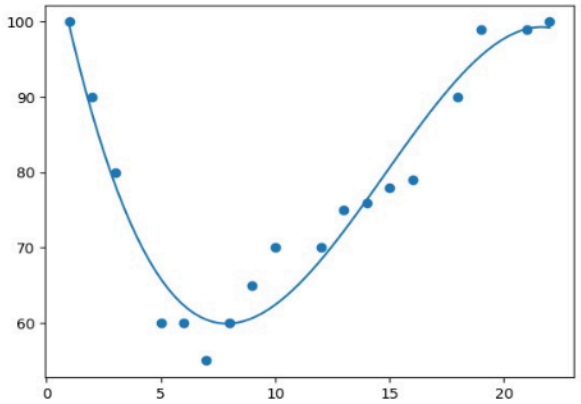
Polynomial Regression

A form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an n^{th} degree polynomial in x

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$







Need for Polynomial Regression:

If we apply a linear model on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, **the error rate will be high, and accuracy will be decreased**.

So for such cases, where **data points are arranged in a non-linear fashion**, we need the Polynomial Regression model.

Simple Linear egression

$$y = b_0 + b_1 x_1$$

Multiple Linear egression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x$$

olynomial Linear egression

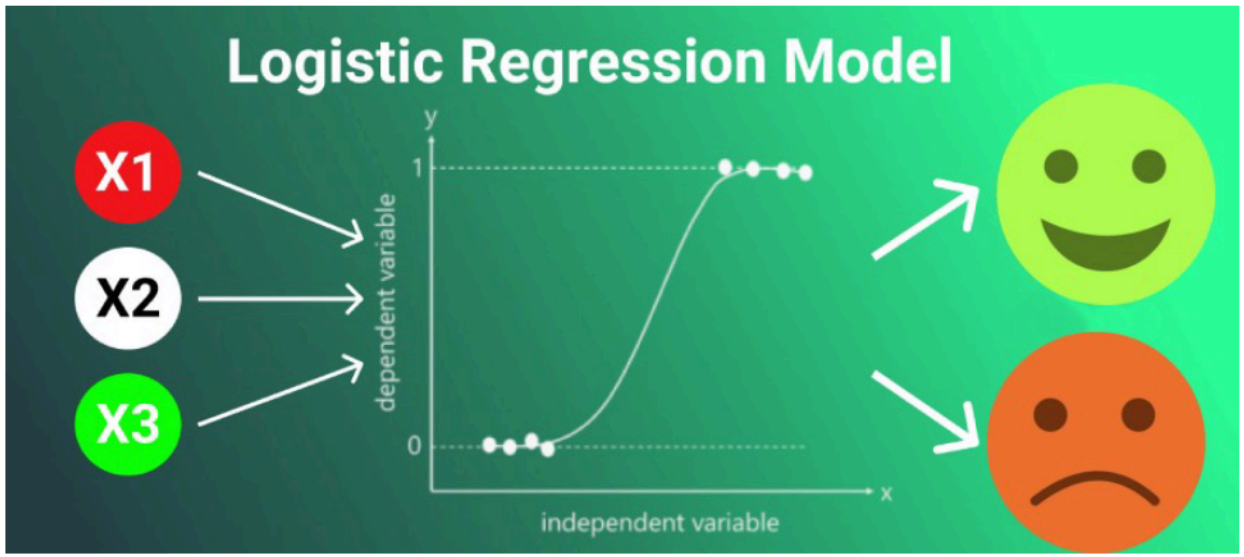
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x$$

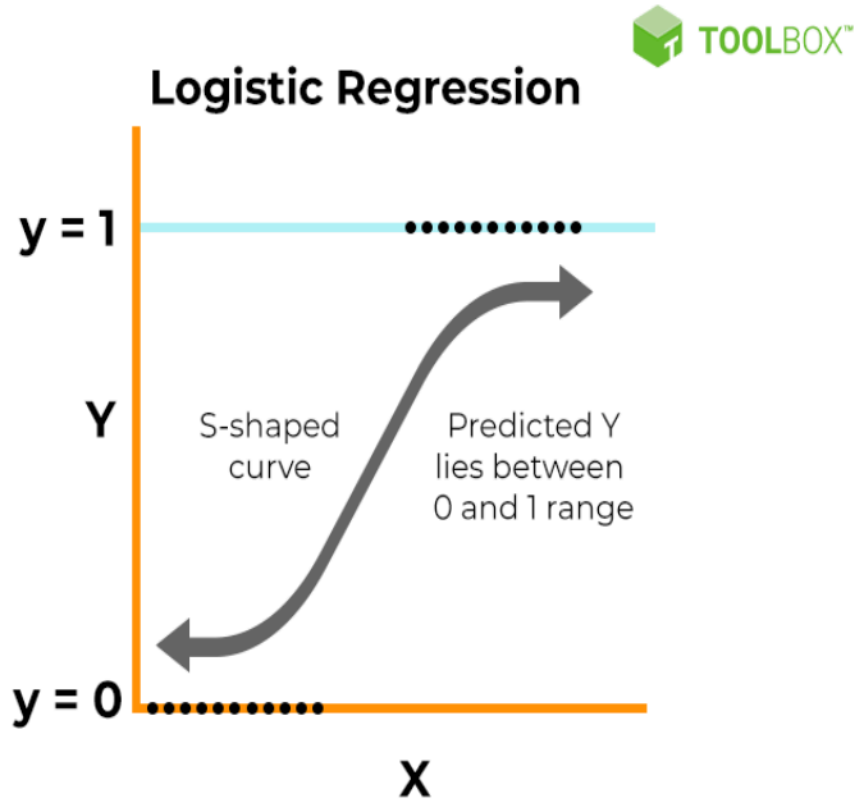
4. Logistic Regression

Logistic regression is a supervised machine learning algorithm that accomplishes **binary classification tasks** by **predicting the probability of an outcome, event, or observation**.

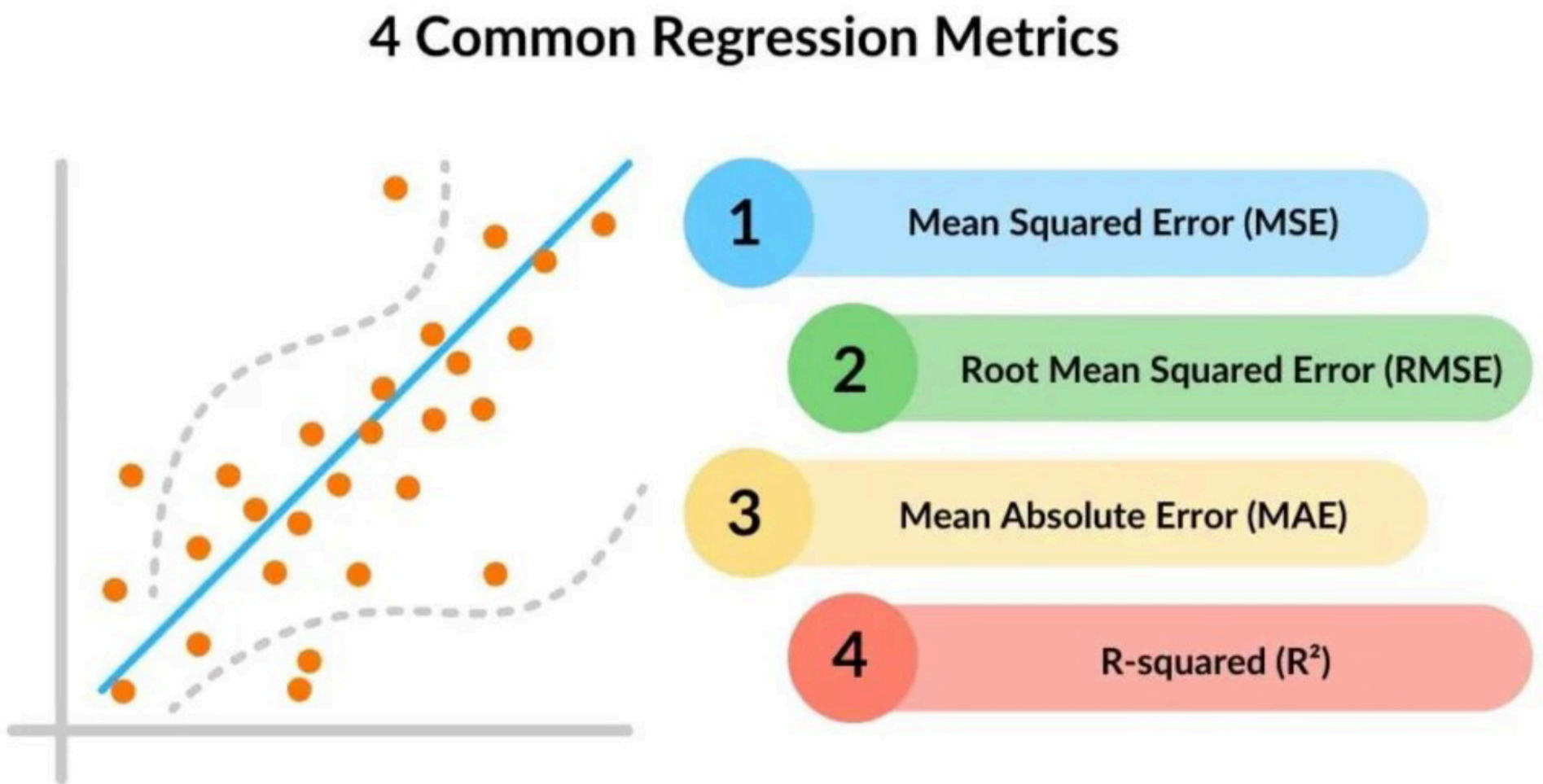
$$f(x) = \frac{1}{1 + e^{-x}}$$

The model delivers a ***binary or dichotomous outcome limited to two possible outcomes***: yes/no, 0/1, or true/false.





The 4 Most Common Performance Metrics for Regression Models



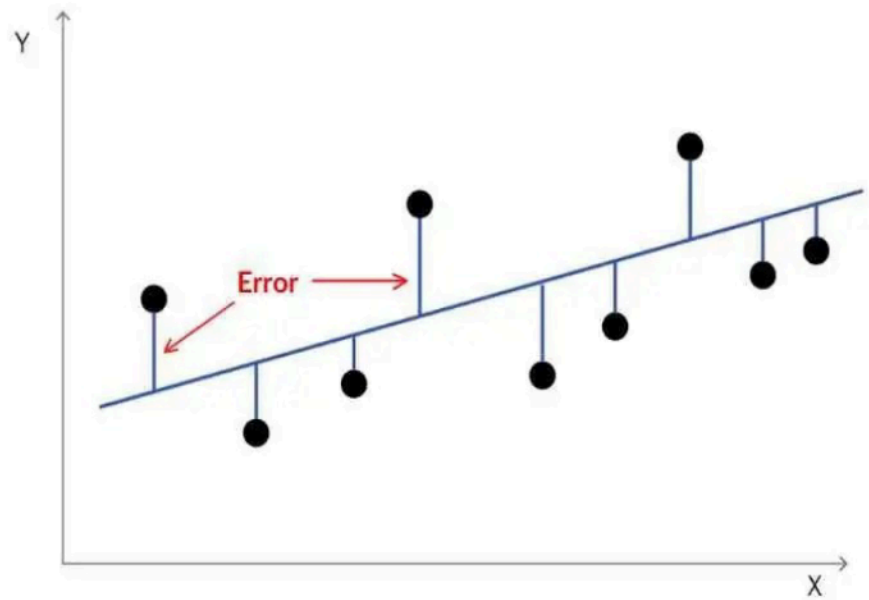
1. Mean Squared Error (MSE)

MSE calculates the average squared difference between predicted and actual values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{Actual}_i - \text{Predicted}_i)^2$$

where *yi* represents the actual value, *y^i* represents the predicted value, and *n* is the number o observations.

MSE measures the average squared error, with higher values indicating more significant discrepancies between predicted and actual values.



$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}}$$

$$n = number\ of\ considered\ points$$

$$actual_i = actual\ value$$

$$predicted_i = the\ predicted\ value$$

Root Mean Squared Error (RMSE)

The square root of the average squared distance between the actual and predicted values.

A lower RMSE value indicates a better model.

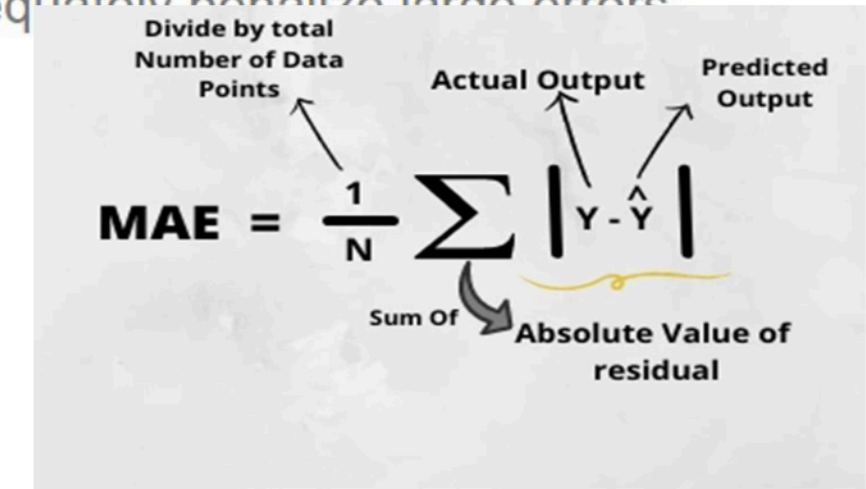
3. Mean Absolute Error (MAE)

MAE computes the average absolute difference between predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\text{Actual}_i - \text{Predicted}_i|$$

It measures the average magnitude of errors, with higher values indicating larger discrepancies between predicted and actual values.

MAE is less sensitive to outliers than MSE but may not adequately penalize large errors.



4. R-squared (R^2)

R^2 measures the proportion of variance in the dependent variable explained by the independent variables.

$$R^2 = 1 - \frac{SSR}{SST}$$

where SSR is the sum of squared residuals, and SST is the total sum of squares.

R2 score

A value between 0 and 1 that indicates how well the regression predictions fit the data. An R2 score of 1 (100%) means the predictions are a perfect fit.

https://medium.com/@shuklapratik22/error-calculation-techniques-for-linear-regression-ae436b682f90

A Step-by-Step Implementation of Simple Linear Regression in Google Colab|MachineLearning

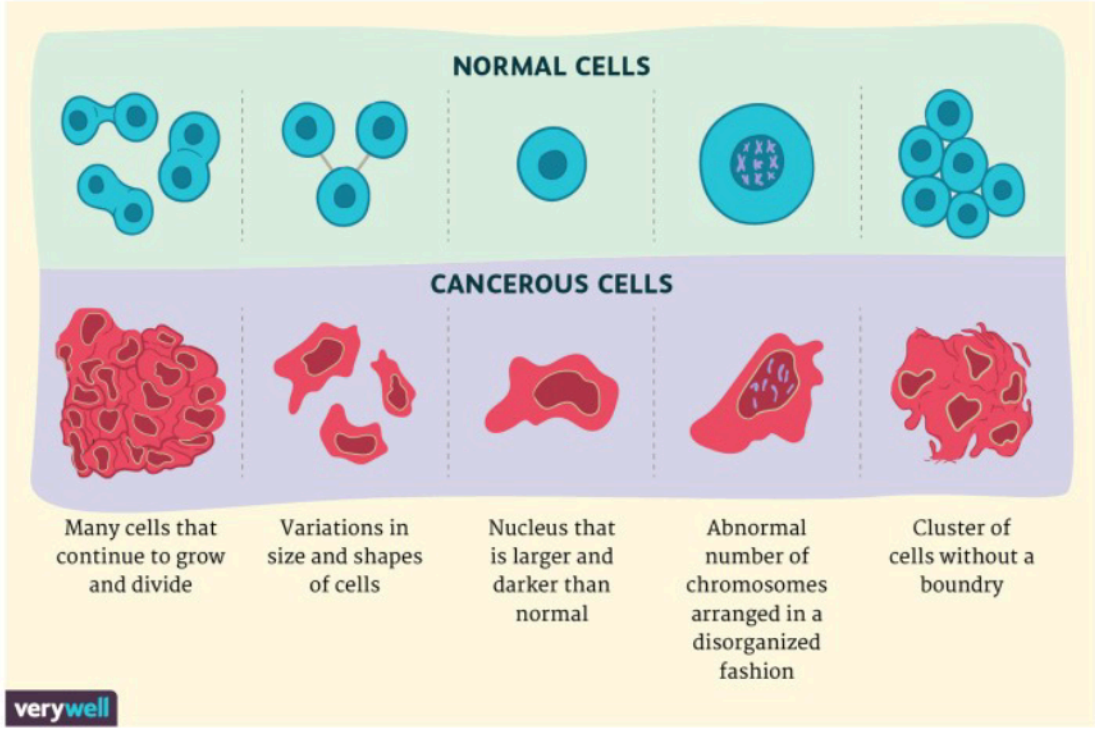
https://www.youtube.com/watch?v=v-cwS5tDeAA

https://learn.saylor.org/mod/page/view.php?id=55340

Metrics for accessing classification accuracy

Evaluation Matrices

Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right.



Normal cells: 90, Cancerous cells :10

Normal cells correctly identified : 90

Cancerous cells correctly identified : 0

Accuracy of the classifier : 90%

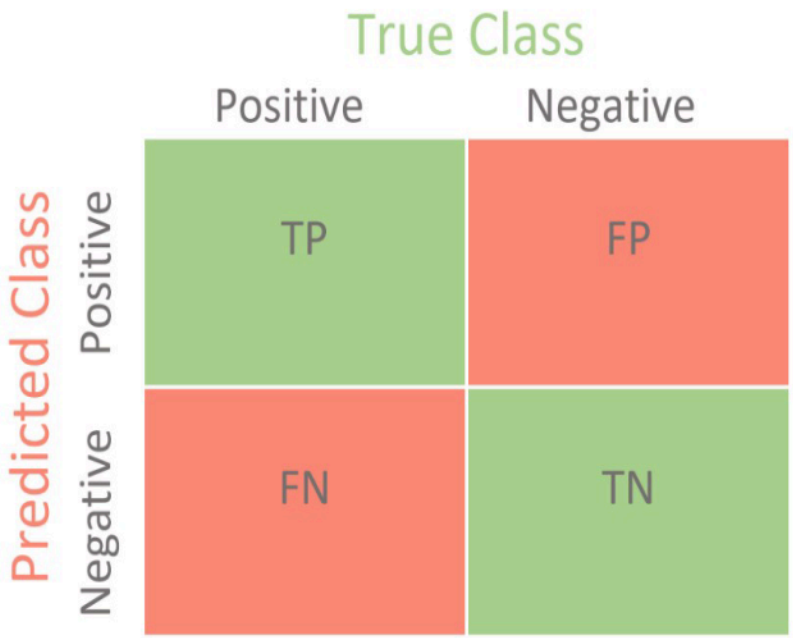
$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Accuracy is a good metric when classes are balanced and the cost of errors is similar across classes.

When Accuracy Might Be Misleading: **Imbalanced Datasets, Misclassification Costs, Multiclass and Multilabel Problems.**

Confusion matrix

TP=True positive
TN=True negative



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- False Positive (FP)
- The actual value was negative, but learning algorithm classified as positive (Also known as a **"Type I error."**)

- False Negative (FN)
- examples are actually positive, but learning algorithm classified as negative (Also known as a **"Type II error."**)

- True Positive (TP) :
- Actual is positive, and is predicted to be positive.

- True Negative (TN) :
- Actual is negative, and is predicted to be negative.

- A confusion matrix is a technique ***for summarizing the performance of a classification algorithm***.

Cat +ve

Dog –ve // cat –Dog Calssification problem

- Cat(+)→ cat(+) TP
- Dog(-)→ Dog (-) TN

- Cat(+)→Dog(-) FN
- Dog(-)→Cat(+) FP

Precision and **Recall** are two key metrics derived from the confusion matrix, but they measure different aspects of a model's performance:

Precision measures the proportion of **correctly predicted positive observations** out of total predicted positives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Precision = \frac{TP}{TP + FP}$$

Interpretation: Precision focuses on how many of the predicted positive results are actually correct.

High Precision: Few false positives.

Use Case: When **false positives** are costly or dangerous, precision becomes more important.

Spam Email Detection: Sending a legitimate email to the spam folder can frustrate users.

Recall (Sensitivity or True Positive Rate)

Definition: **Recall measures** the proportion of actual positive observations that are correctly identified.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Recall = \frac{TP}{TP + FN}$$

Interpretation: Recall focuses on identifying as many positive observations as possible.

High Recall: Few false negatives.

Use Case: When **false negatives** are costly or dangerous, recall becomes more important.

Medical Diagnosis: Missing a disease (false negative) could have life-threatening consequences.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

When you need to balance precision and recall, especially if the cost of false positives and false negatives is roughly equal, you can use the F1 score

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Problem:

```
true_labels =      [1, 1, 0, 0, 1, 1, 0, 1, 1]  # 1 for spam, 0 for non-spam
predicted_labels = [1, 0, 1, 0, 0, 1, 1, 0, 1]
```

```
true_labels =      [1, 1, 0, 0, 1, 1, 0, 1, 1]  # 1 for spam, 0 for non-spam
predicted_labels = [1, 0, 1, 0, 0, 1, 1, 0, 1]

# Calculate true positives (TP), false positives (FP), and false negatives (FN)
n = len(true_labels)
tp = sum(1 for i in range(n) if true_labels[i] == 1 and predicted_labels[i] == 1)
tn = sum(1 for i in range(n) if true_labels[i] == 0 and predicted_labels[i] == 0)
fp = sum(1 for i in range(n) if true_labels[i] == 0 and predicted_labels[i] == 1)
fn = sum(1 for i in range(n) if true_labels[i] == 1 and predicted_labels[i] == 0)

print("True positives (TP):", tp)
print("True negatives (TN):", tn)
print("False positives (FP):", fp)
print("False negatives (FN):", fn)

# Calculate accuracy
accuracy = (tp + tn) / (tp + tn + fp + fn)
print("Accuracy:", accuracy)

# Calculate precision
precision = tp / (tp + fp)
print("Precision:", precision)

# Calculate recall
recall = tp / (tp + fn)
print("Recall:", recall)
```

```
# Calculate F1-score
f1 = 2 * (precision * recall) / (precision + recall)
print("F1-score:", f1)
```

Output:

```
True positives (TP): 3
True negatives (TN): 1
False positives (FP): 2
False negatives (FN): 3
Accuracy: 0.4444444444444444
Precision: 0.6
Recall: 0.5
F1-score: 0.5454545454545454
```

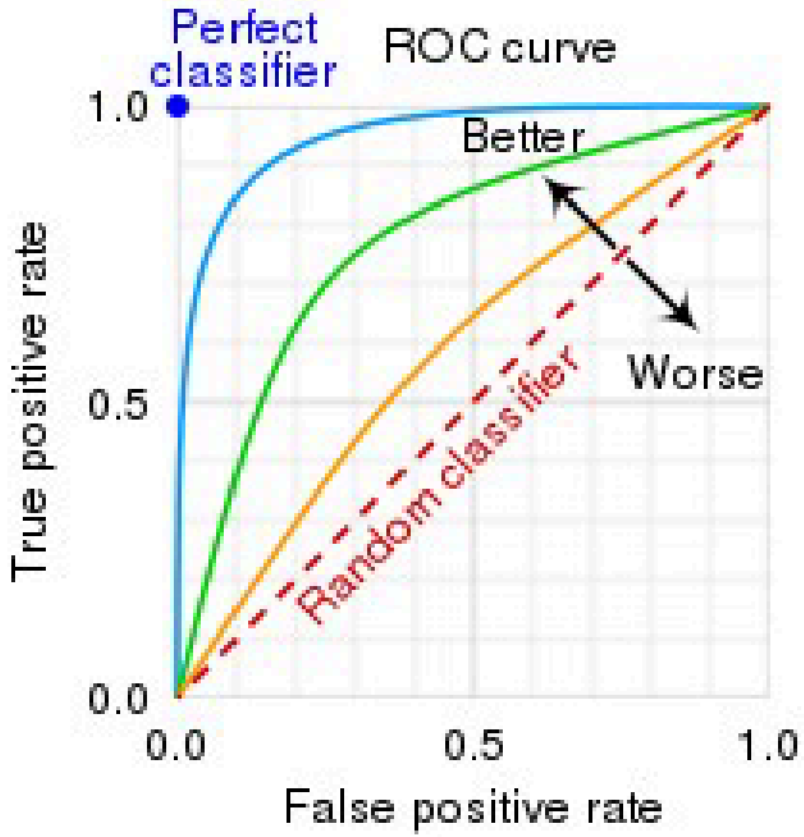
ROC curve

A **receiver operating characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$



AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

Problem 1:: confusion matrix

- TP = 30, TN = 930, FP = 30, FN = 10

What is the accuracy, precision, recall?

https://kavita-ganesan.com/precision-and-recall-machine-learning/