ARTIFICIAL INTELLIGENCE
&
MACHINE LEARNING

SESSION NO: 17

TOPIC: **LINEAR REGRESSION & LOGISTIC REGRESSION**

INDEX

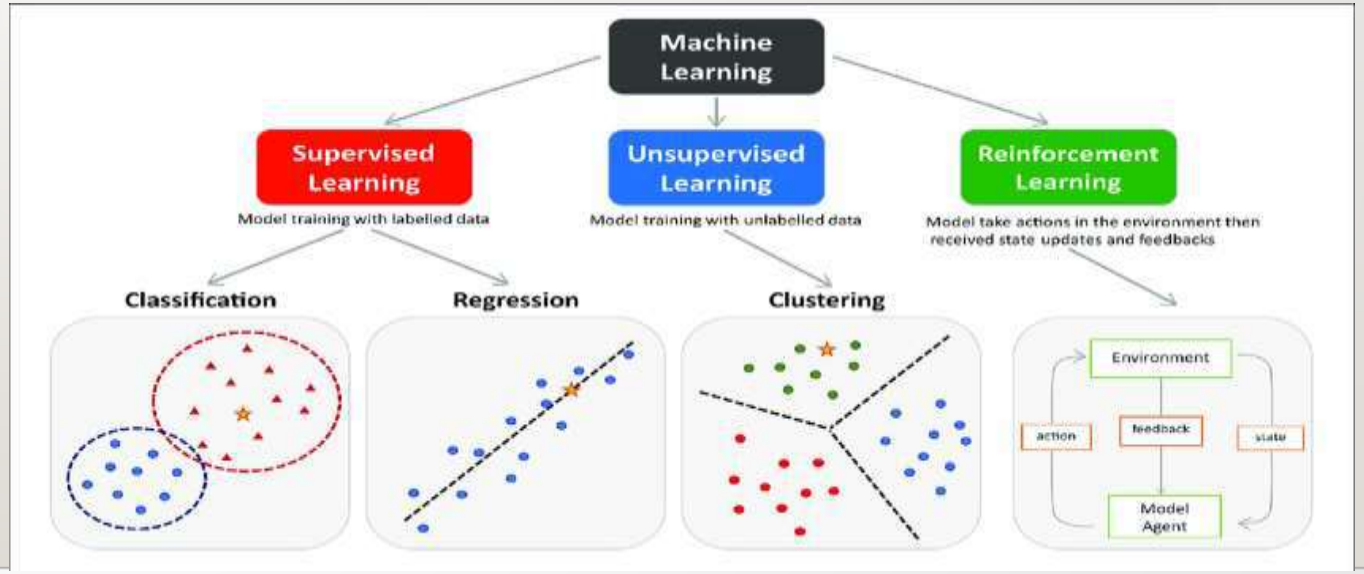
- Machine Learning and types – Introduction
- Classification VS Regression
- Types of Regression
- Linear Regression – Simple and Multiple
- Logistic Regression
- Linear VS Logistic
- Summary

MACHINE LEARNING - INTRODUCTION

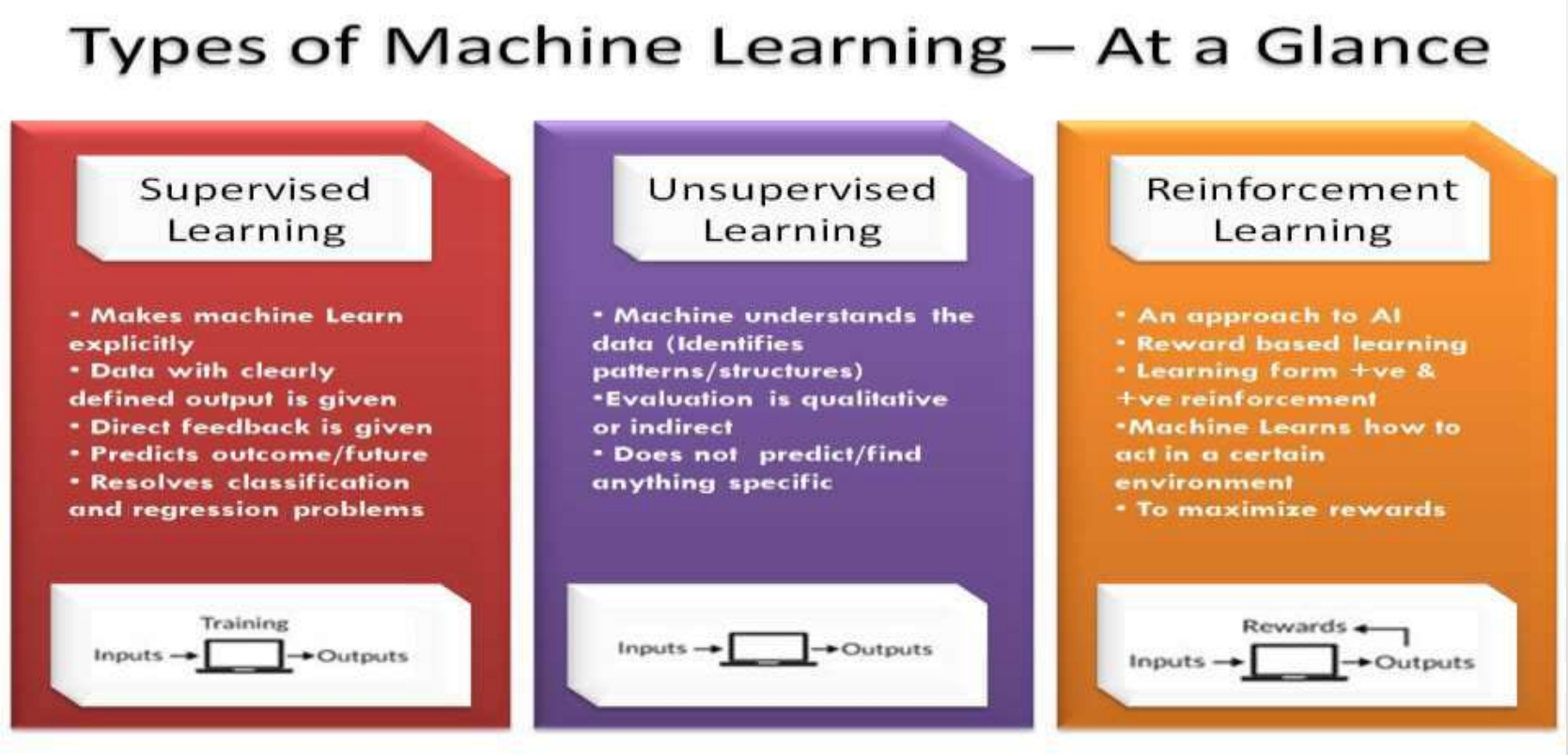
- **Machine learning** - branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy.

- IBM

- **Categories of Machine Learning**
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

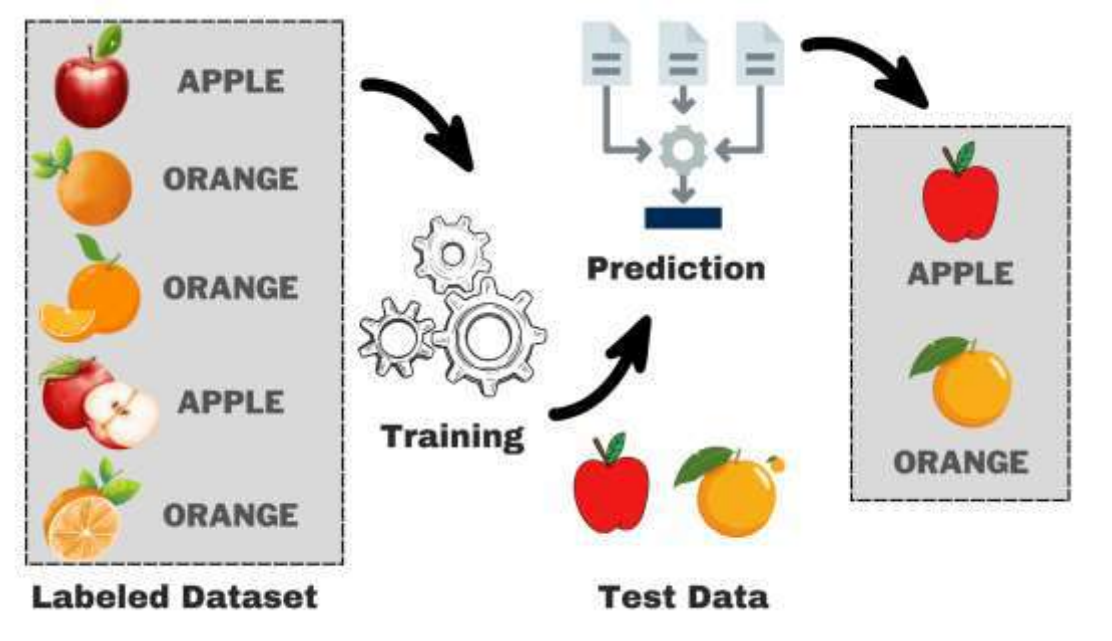


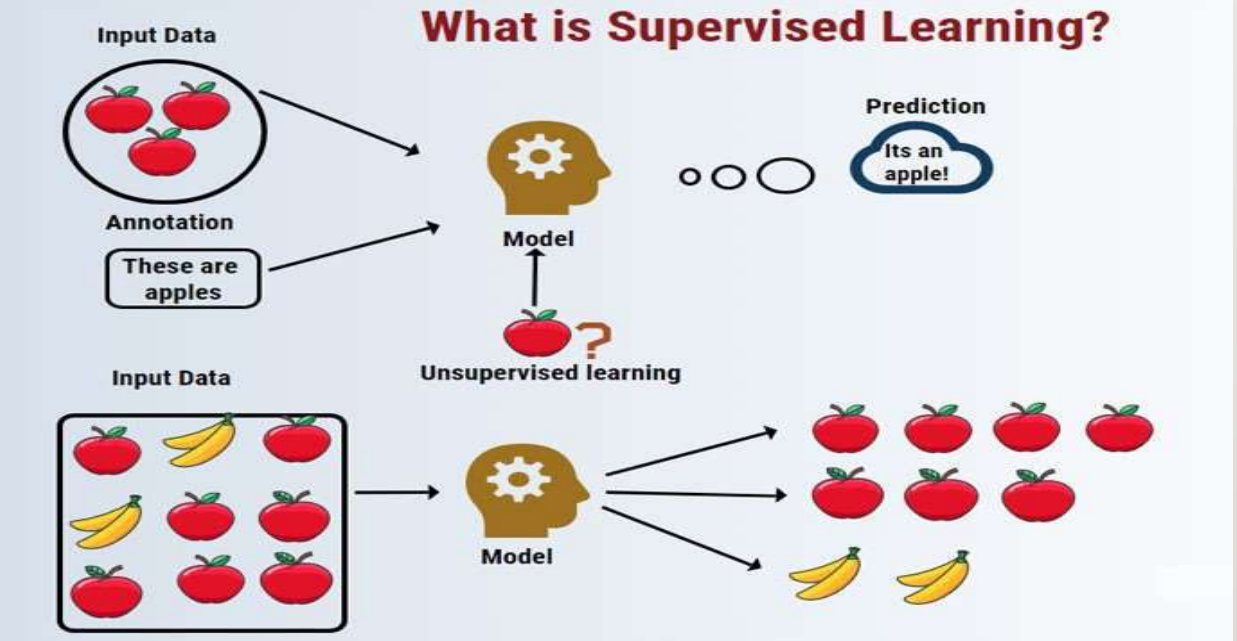
TYPES OF MACHINE LEARNING



SUPERVISED LEARNING

- **Supervised Learning** - category of machine learning that uses **labelled datasets** to train algorithms to predict outcomes and recognize patterns.

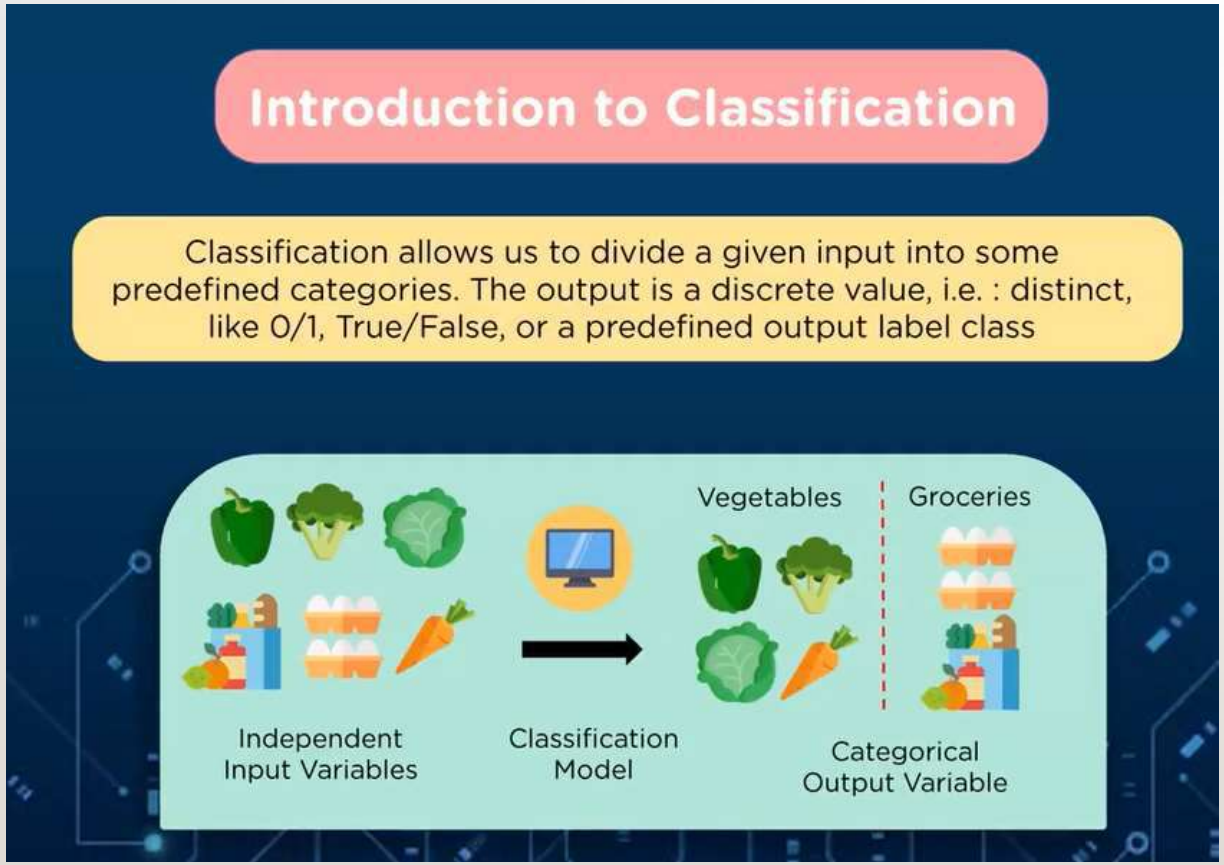


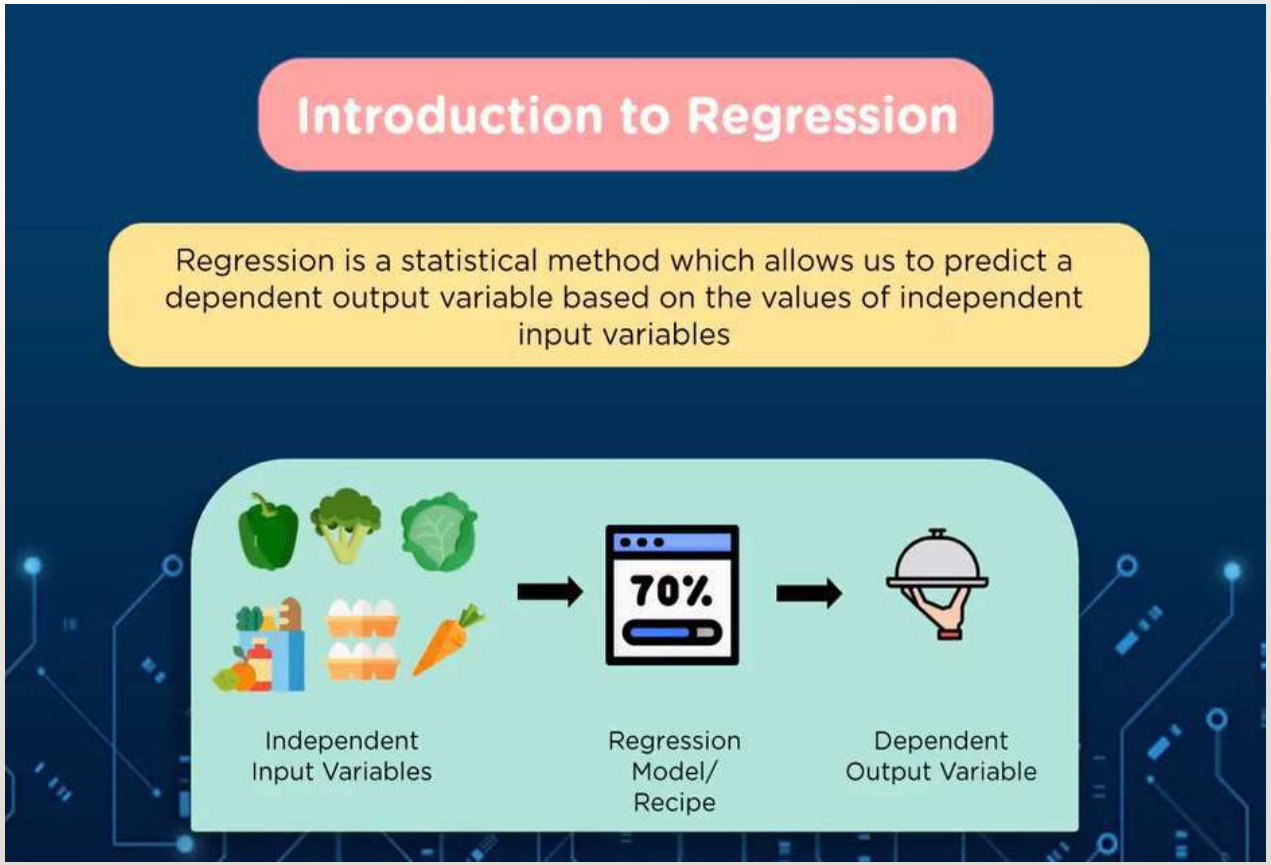


CLASSIFICATION VS REGRESSION (SUPERVISED LEARNING)

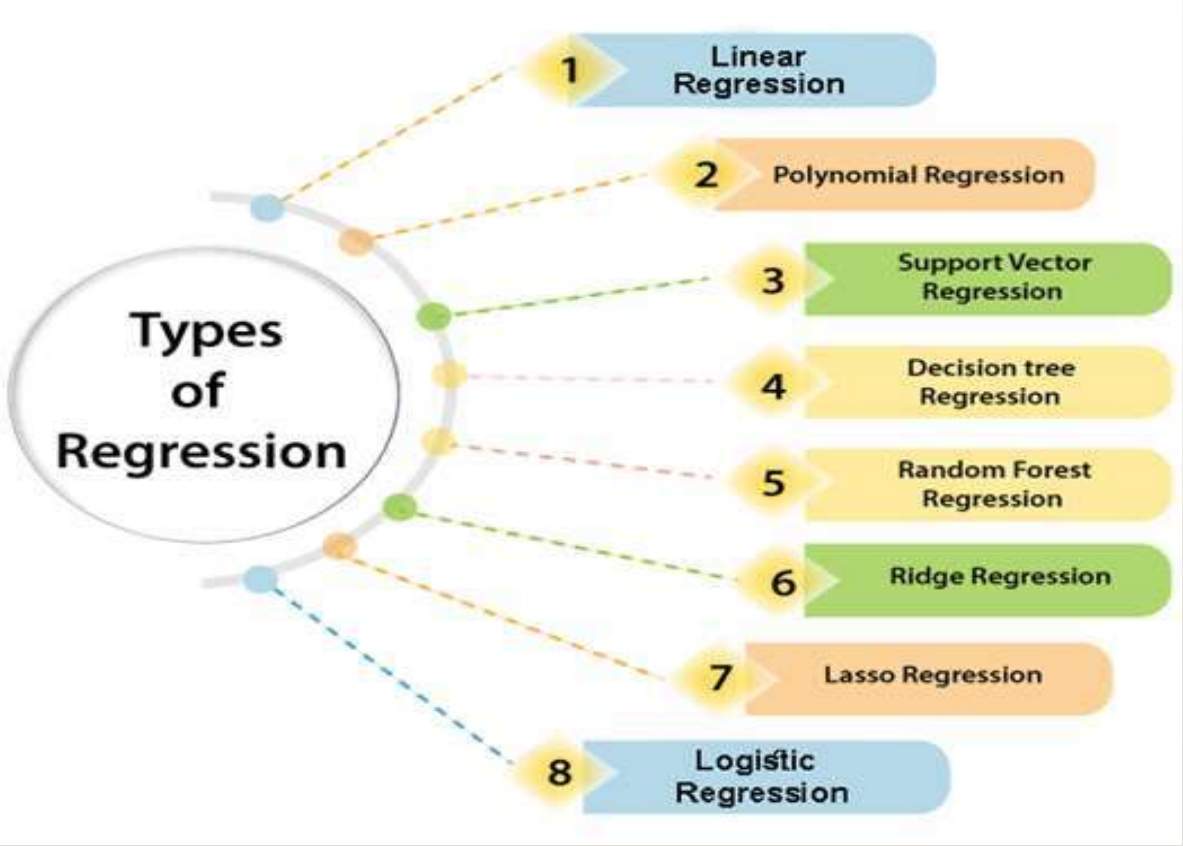
- **Classification - a supervised machine learning method where the model tries to predict the correct label of a given input data**.
- In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.
- **Regression - a supervised machine learning technique which is used to predict continuous values**.
- The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.
- The three main metrics that are used for evaluating the trained regression model are variance, bias and error.

CLASSIFICATION VS REGRESSION



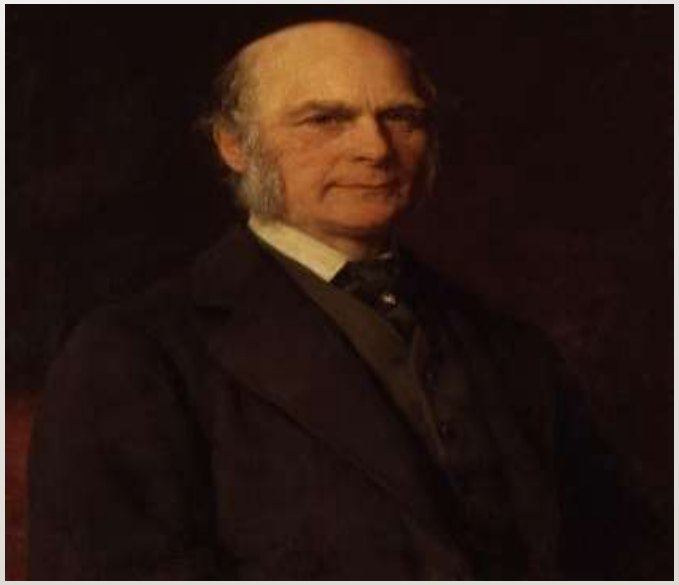


TYPES OF REGRESSION



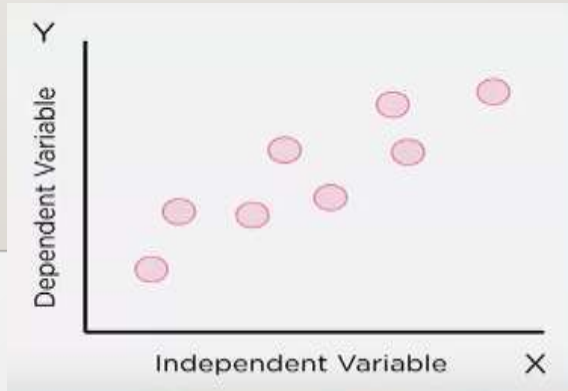
LINEAR REGRESSION - HISTORY

- Sir Francis Galton (1885) introduced the idea of "regression" to the research community in a study examining the relationship of fathers' and sons' heights.
- In his study he observed that **sons do not tend toward their fathers' heights but instead "regress to" the mean of the population**.



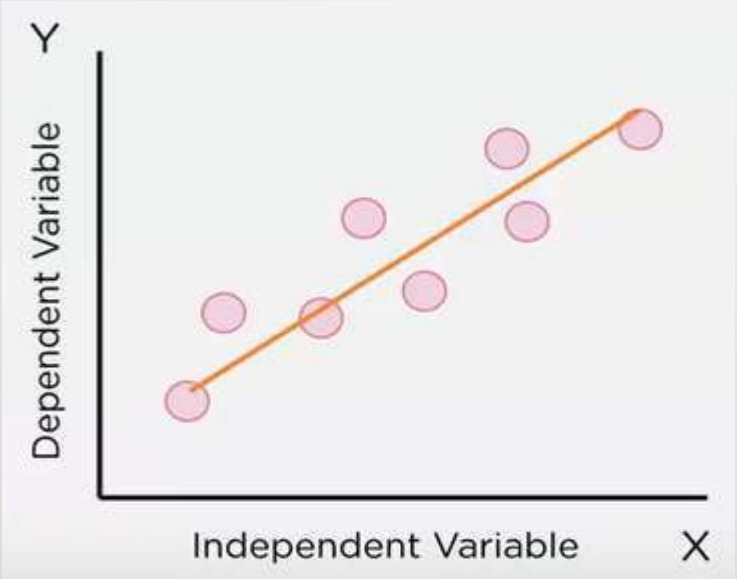
LINEAR REGRESSION - DISCUSSION

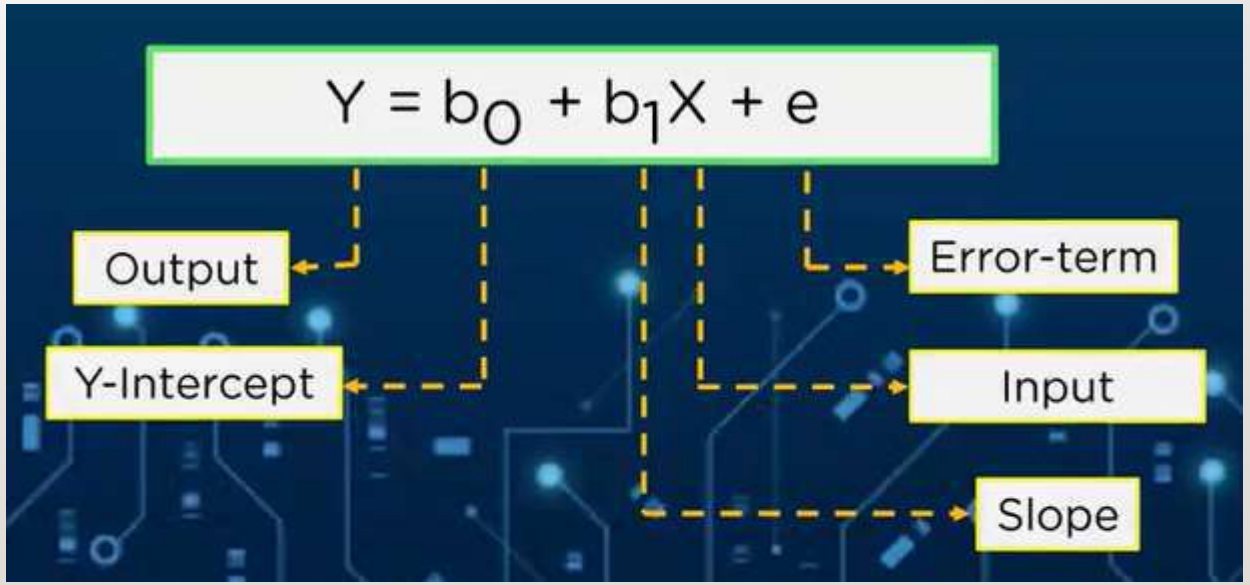
- **Linear regression** – machine learning model that is used to predict the values of output variables based on the value of input variables.
- It predicts the relationship between two variables by assuming they have a straight-line connection.
- It finds the best line that minimizes the differences between predicted and actual values.
- Linear regression is a quiet and the simplest statistical regression technique used for predictive analysis in machine learning.
- Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent (output) variable i.e. Y-axis.



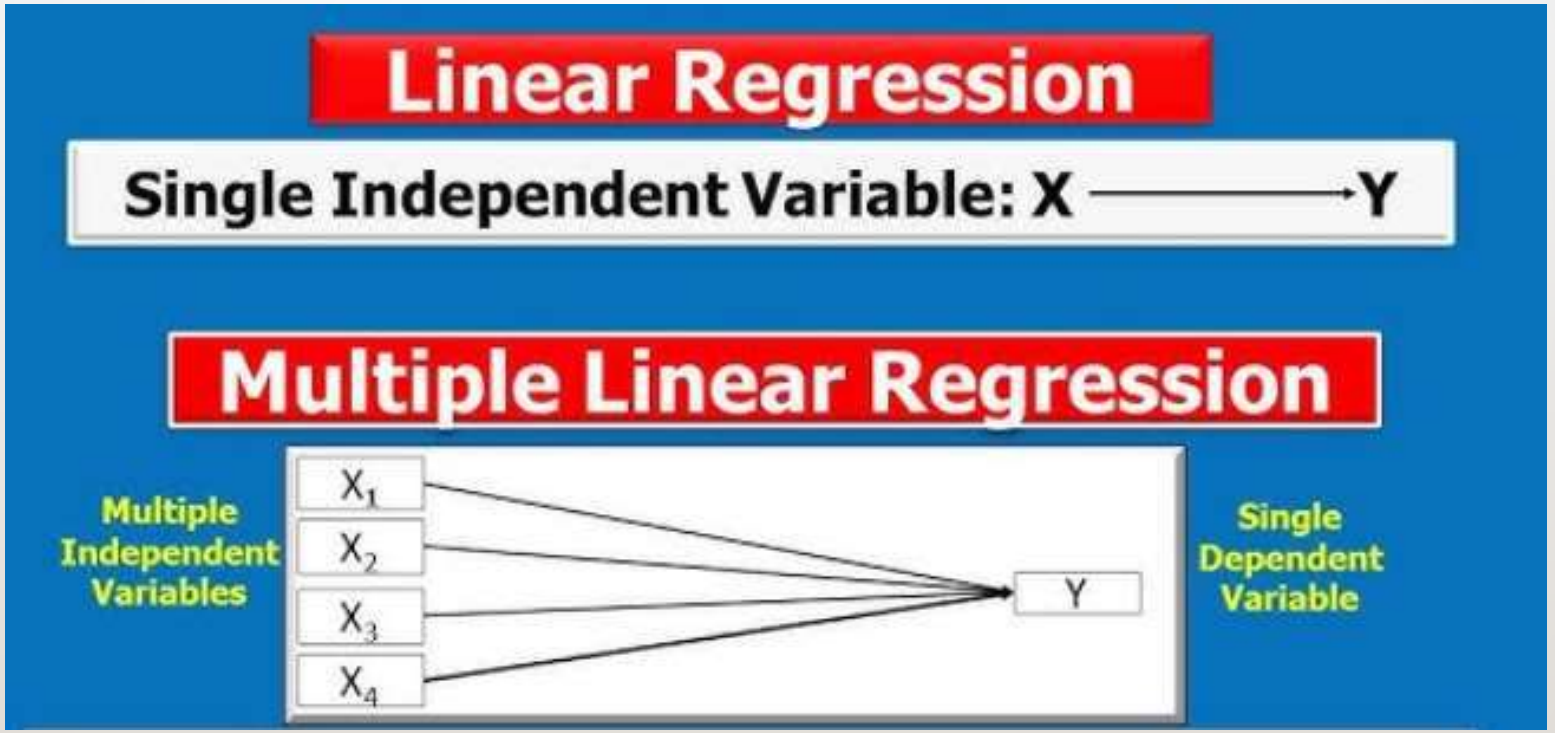
LINEAR REGRESSION - DISCUSSION

- The equation which can be used to fit a line is the equation of straight line.
- This equation gives the output variable based on the input variable and inclination of the line.





SIMPLE LINEAR REGRESSION
VS
MULTIPLE LINEAR REGRESSION



MULTIPLE LINEAR REGRESSION

- Unlike simple linear regression, multiple linear regression can include two or more independent variables.
- The goal is to estimate one variable based on several other variables.
- The variable to be estimated is called the **dependent variable (criterion).** The variables that are used for prediction are called **independent variables (predictors)**.
- Multiple linear regression is often used in empirical social research as well as in market research. In both areas it is of interest to find out what influence different factors have on a variable.

MULTIPLE LINEAR REGRESSION

Simple linear Regression

$$\hat{y} = b \cdot x + a$$

Multiple linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

- The coefficients are interpreted similarly to the linear regression equation.
- If all independent variables are 0, the value a is obtained.
- If an independent variable changes by one unit, the associated coefficient b indicates by how much the dependent variable changes.

LINEAR REGRESSION – EXAMPLE PROBLEM

- The following table shows the Mid Term and Final Exam grades obtained by the students in database course.
- Use the method of Least Squares to find the equation for the prediction of student's final exam grade based on the student's midterm grade in the course.
- Predict the final exam grade of a student who received an 86 in the midterm exam.

X Mid Term	Y Final Exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

LINEAR REGRESSION – EXAMPLE PROBLEM

- Here, X is independent variable and y is dependent variable.
- Equation of linear regression looks like:

$$y = a + bx$$

- a is the intercept and b is the slope or coefficient
- $b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$
- $\bar{y} = a + b\bar{x}$

X Mid Term	Y Final Exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

LINEAR REGRESSION – EXAMPLE PROBLEM

X Mid Term	Y Final Exam	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$	$(x - \bar{x})^2$
72	84	-0.17	10	-1.7	0.03
50	63	-22.17	-11	243.87	491.51
81	77	8.83	3	26.49	77.97
74	78	1.83	4	7.32	3.35
94	90	21.83	16	349.28	476.55
86	75	13.83	1	13.83	191.27
59	49	-13.17	-25	329.25	173.45
83	79	10.83	5	54.15	117.29
65	77	-7.17	3	-21.51	51.41
33	52	-39.17	-22	861.74	1534.29
88	74	15.83	0	0	250.59
81	90	8.83	16	141.28	77.97
$\bar{x} = 72.17$	$\bar{y} = 74$			2004	3445.67

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

$$b = \frac{2004}{3445.67}$$

$$b = 0.5816$$

$$\bar{y} = a + b\bar{x}$$

$$a = \bar{y} - b\bar{x}$$

$$a = 74 - 0.5816 * 72.17$$

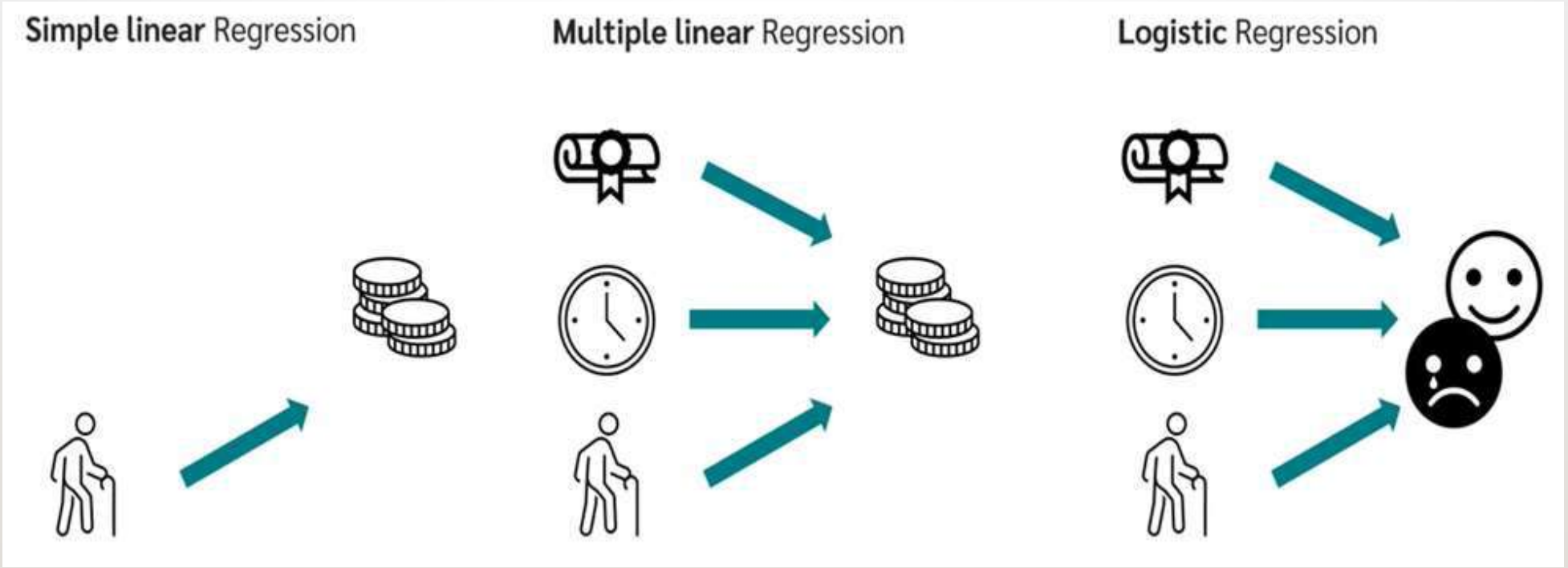
$$a = 32.02$$

$$y = a + bx$$

$$y = 32.02 + 0.5816x$$

$$y = 82.04$$

FORMS OF REGRESSION ANALYSIS



LOGISTIC REGRESSION

Logistic regression is a special case of regression analysis and is calculated when the dependent variable is nominally or ordinally scaled.

Business example:

For an online retailer, you need to predict which product a particular customer is most likely to buy. For this, you receive a data set with past visitors and their purchases from the online retailer.

Medical example:

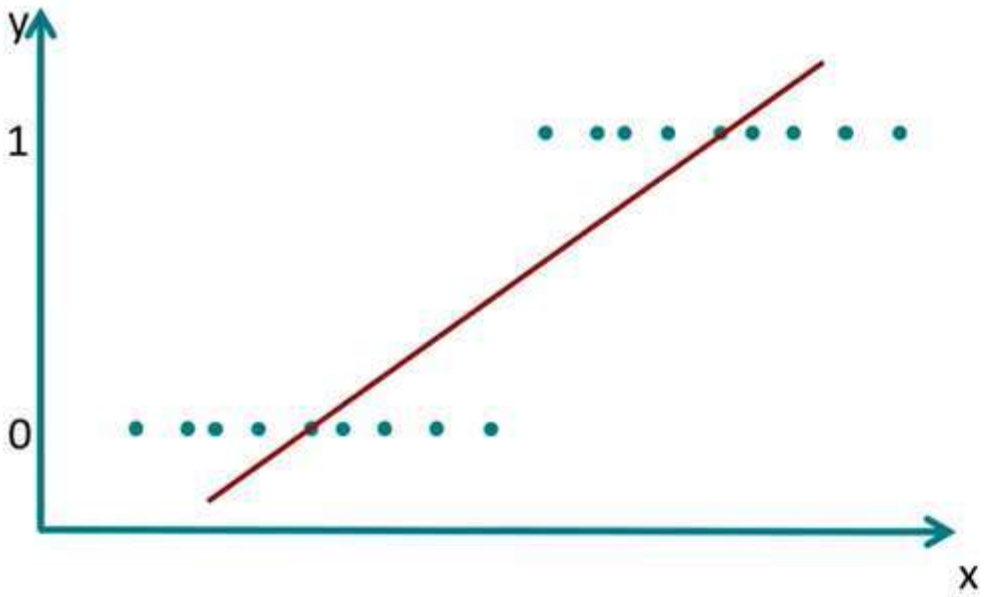
You want to investigate whether a person is susceptible to a certain disease or not. For this purpose, you receive a data set with diseased and non-diseased persons as well as other medical parameters.

Political example:

Would a person vote for party A if there were elections next weekend?

Why not just use linear regression?

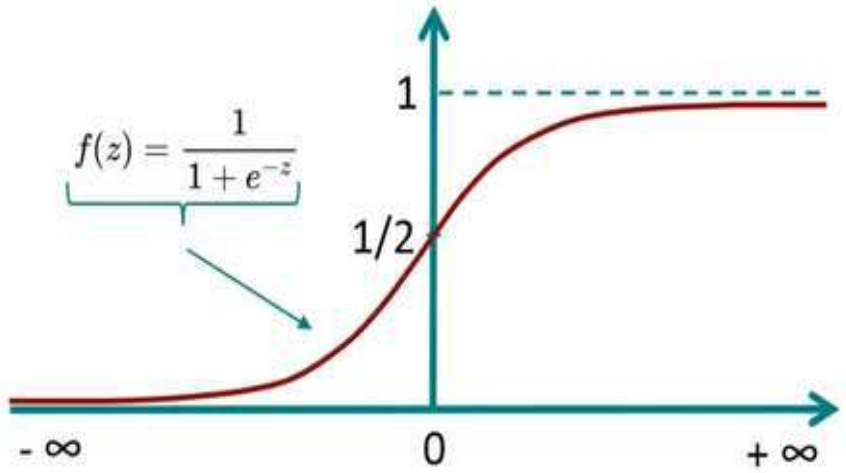
$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$



- The graph shows that values between **plus and minus infinity** can now occur.
- The goal of logistic regression is to estimate the **probability of occurrence**, not the value of the variable itself.
- The range of values for the prediction is restricted to the range between 0 and 1.
- To ensure that only values between 0 and 1 are possible, the logistic function f is used.

Logistic function

The logistic model is based on the logistic function.
The important thing about the logistic function is, that only values **between 0 and 1** are possible.



$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

LINEAR REGRESSION VS LOGISTIC REGRESSION

Linear Regression	Logistic Regression
Used to predict a dependent output variable based on independent input variable	Used to classify a dependent output variable based on independent input variable
Accuracy is measured using Least squares estimation	Accuracy is measured using Maximum Likelihood estimation
The best fit line is a straight line	The best fit is given by a curve
The output is a predicted integer value	The output is a binary value between 0 and 1 value
Used in business domain, forecasting stocks	Used for classification, image processing

SUMMARY

- Introduction to Supervised Learning
- Introduction to Regression
- Discussion about Linear and Logistic Regression

SELF ASSESSMENT QUESTIONS

1. Which of the following statements about linear regression is true?

(a) Linear regression can only model linear relationships.
(b) Linear regression is always the best model to use.
(c) Linear regression assumes that the relationship between the independent and dependent variables is linear.
(d) Linear regression does not require any assumptions about the distribution of errors.

2. In the context of linear regression, what does the term "residual" refer to?

(a) The difference between the predicted and actual values.
(b) The slope of the regression line.
(c) The intercept of the regression line.
(d) The coefficient of determination.

SELF ASSESSMENT QUESTIONS

3. Which of the following metrics is commonly used to evaluate the goodness-of-fit for a linear regression model?

(a) Confusion matrix
(b) R-squared (R^2)
(c) ROC curve
(d) Mean Absolute Percentage Error (MAPE)

4. In a simple linear regression model, the equation is given by $y=\beta0+\beta1x+\epsilon$. What does $\beta1$ represent?

(a) The intercept of the regression line.
(b) The slope of the regression line.
(c) The error term.
(d) The predicted value.

SELF ASSESSMENT QUESTIONS

5. Which of the following statements about logistic regression is true?

(a) Logistic regression is used for predicting continuous outcomes.
(b) Logistic regression requires the dependent variable to be binary.
(c) Logistic regression assumes a linear relationship between the independent and dependent variables.
(d) Logistic regression is a type of unsupervised learning.

6. Which of the following methods is commonly used to estimate the parameters in logistic regression?

(a) Ordinary least squares
(b) Maximum likelihood estimation
(c) Gradient descent
(d) Principal component analysis

TERMINAL QUESTIONS

- Explain the logistic regression model and how it differs from linear regression.
- Given a dataset with multiple predictors, describe the process of building and validating a linear regression model.
- Describe the steps involved in evaluating the performance of a logistic regression model.

REFERENCE BOOKS AND WEB LINKS

Text Books:

1) Russel and Norvig, 'Artificial Intelligence', third edition, Pearson Education, PHI, (2015)
2) Elaine Rich & Kevin Knight, 'Artificial Intelligence', 3rd Edition, Tata McGraw Hill Edition, Reprint (2008)

Web links:

1. https://www.youtube.com/watch?v=nifpRZhfHA0
2. https://www.youtube.com/watch?v=QWYkQDvCo4Y
3. https://www.youtube.com/watch?v=C5268D9t9Ak
4. https://www.youtube.com/watch?v=nwD5U2WxTdk&list=PLuhqtP7jdD8AFocJuxC6_Zz0HepAWL9cF
5. https://www.youtube.com/watch?v=0m-rs2M7K-Y