SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT II – REGRESSION AND REGRESSION LINE

REGRESSION

EMPATHY:

- In data science, regression is a supervised learning method used to model relationships between variables (independent/features) to predict a continuous dependent variable (target), helping find trends, forecast outcomes, and understand cause-and-effect, like predicting house prices based on size.

- Key types include Linear Regression (finding best-fit line) and advanced methods like Ridge Regression, with common goals being minimizing prediction errors (residuals) to build accurate predictive models for data-driven decisions.

DEFINE:

Goal:

- To find the best mathematical relationship (e.g., a line, curve) between input features (independent variables) and a continuous output (dependent variable).

How it Works:

- It fits a model to historical data, then uses that model to predict future values.

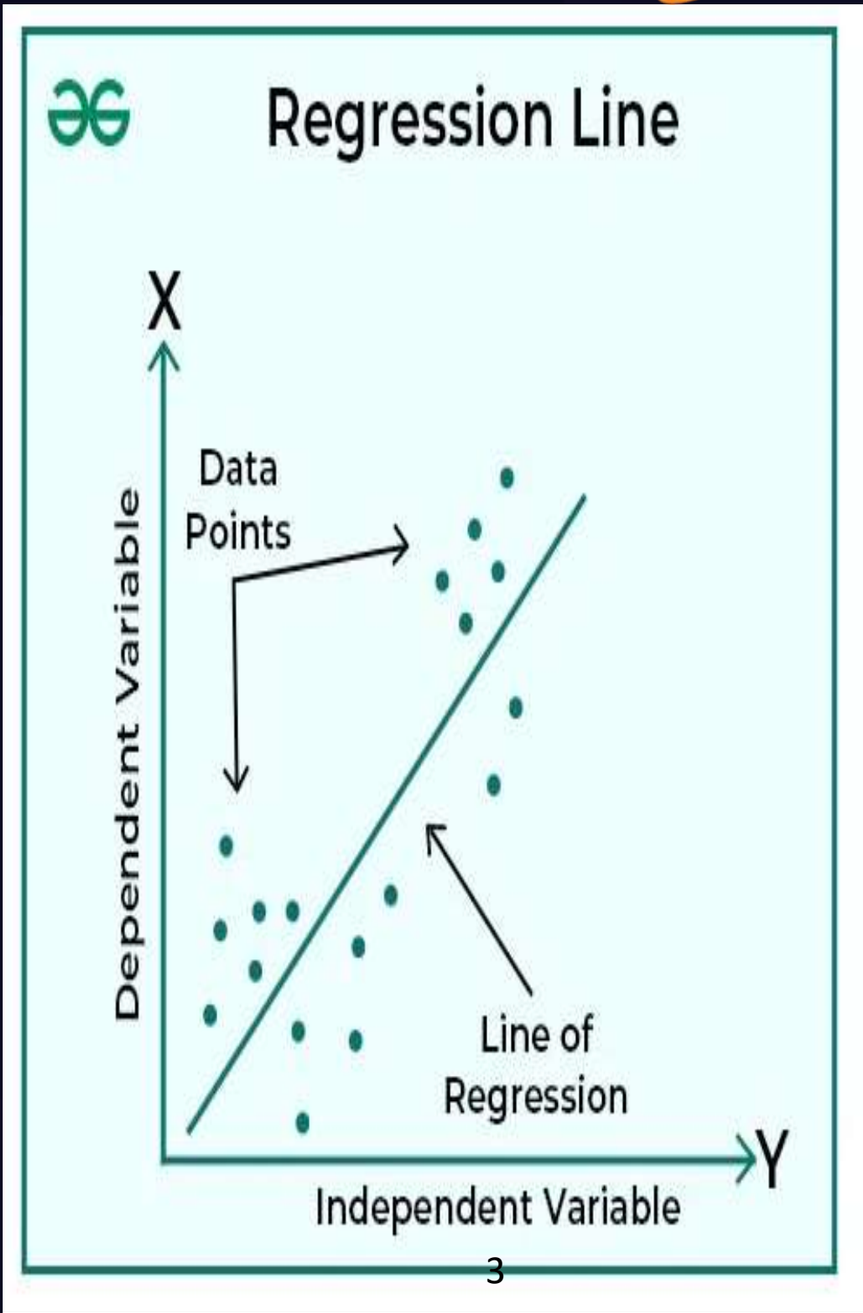
Example:

- Predicting sales (dependent) based on advertising spend (independent).

IDEATE:

What is Regression Line?

- Regression Line is defined as a statistical concept that facilitates and predicts the relationship between two or more variables.
- A regression line is a straight line that reflects the best-fit connection in a dataset between independent and dependent variables.
- The independent variable is generally shown on the X-axis and the dependent variable is shown on the Y-axis.
- The main purpose of developing a regression line is to predict or estimate the value of the dependent variable based on the values of one or more independent variables.



PROTOTYPING:

Equation of Regression Line

The equation of a simple linear regression line is given by:

$$Y = a + bX + \varepsilon$$

Here,

Y is the dependent variable
X is the independent variable
a is the y-intercept, which represents the value of Y when X is 0.
b is the slope, which represents the change in Y for a unit change in X
ε is residual error.

TESTING:

Example 2:

In continuation with the above example, the figures of three students are given as follows:

Student 1: Studied for 2 hours and scored 60 marks.

Student 2: Studied for 3 hours and scored 65 marks.

What will the marks scored by the 4th student in case he/she studies for 5 hours.
Solution:

The required equation of regression line as calculated in previous example is,

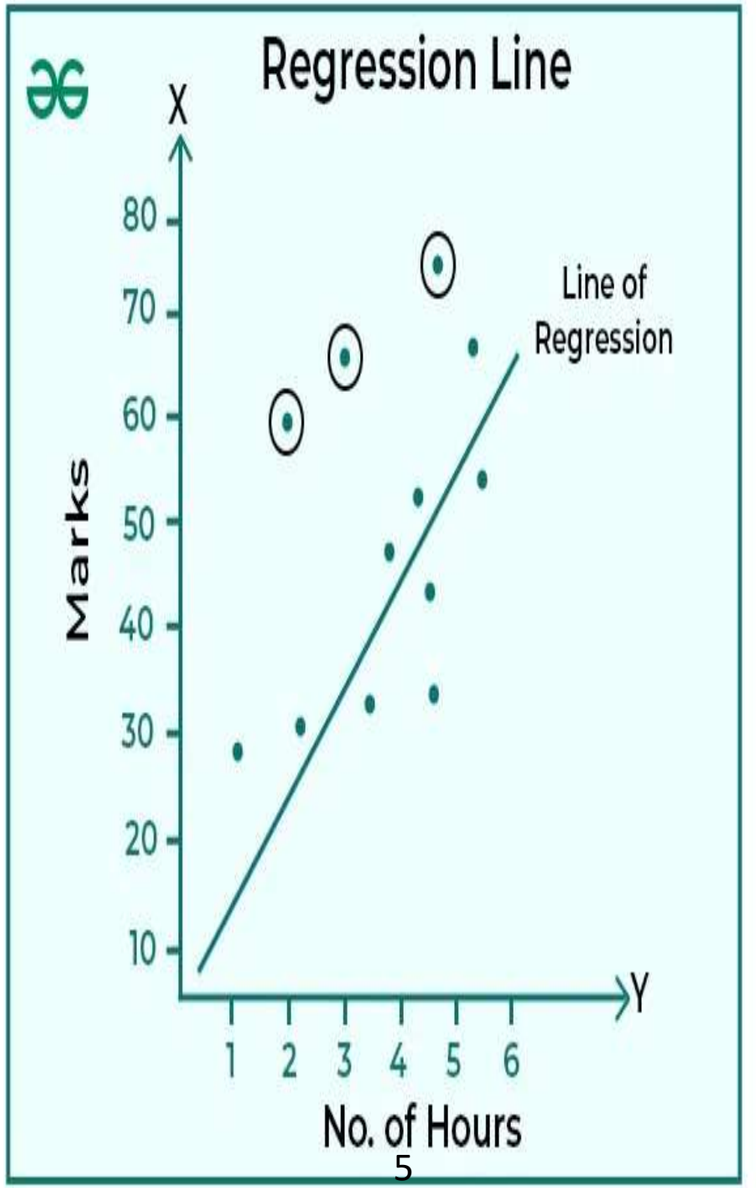
$Y = 50 + 5X$

In case of 4th student, who studies for 5 hours (X = 5), the marks scored by him will be calculated as,

$Y = 50 + 5X.$

$Y = 50 + 5(5)$

$Y = 75$ Marks



Regression-Line-2-copy