ASSIGNMENT-2

Task: Take any one domain and draw the graph (normal distribution using empirical formula).

1. Selected Domain: Employee Salary Data

Domain Description

Consider a real-time dataset of **monthly salaries of employees in a mid-sized IT company**.
Such salary data often follows a normal distribution because:

- Most employees earn around an average salary
- Very few employees earn extremely low or extremely high salaries

This type of data is commonly used in:

- Salary prediction models
- HR analytics
- Workforce planning systems

2. Assumptions for the Dataset

To apply the empirical rule, assume:

- Mean salary (μ) = **₹50,000**
- Standard deviation (σ) = **₹10,000**

These values help illustrate how employee salaries are distributed.

3. Normal Distribution in Machine Learning

A **normal distribution** is a symmetric, bell-shaped curve where:

- The centre represents the **mean (μ)**
- The spread is controlled by **standard deviation (σ)**

Importance in Machine Learning

- Used in **feature normalization**
- Helps in **outlier detection**
- Applied in **probabilistic models**
- Used for **confidence interval estimation**

4. Empirical Rule (68–95–97 Rule)

For a normally distributed dataset:

- **68%** of values lie within **μ ± 1σ**
- **95%** of values lie within **μ ± 2σ**
- **99.7%** of values lie within **μ ± 3σ**

5. Application of Empirical Rule to Salary Data

Calculations

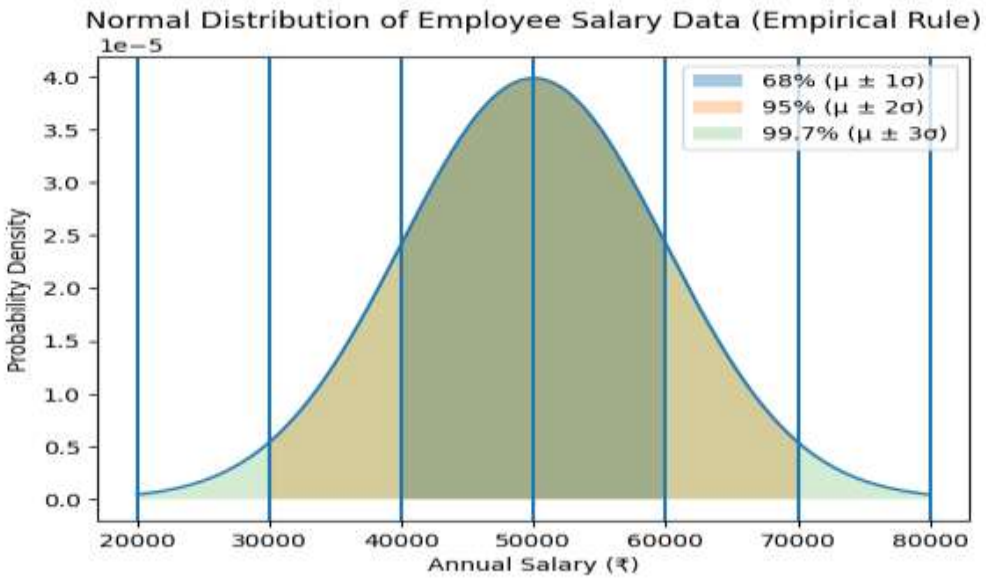
- **μ ± 1σ** → ₹50,000 ± ₹10,000 → **₹40,000 to ₹60,000**
- **μ ± 2σ** → ₹50,000 ± ₹20,000 → **₹30,000 to ₹70,000**
- **μ ± 3σ** → ₹50,000 ± ₹30,000 → **₹20,000 to ₹80,000**

Interpretation

- About **68%** of employees earn between **₹40,000 and ₹60,000**
- About **95%** of employees earn between **₹30,000 and ₹70,000**
- About **99.7%** of employees earn between **₹20,000 and ₹80,000**

Employees outside this range can be treated as **outliers** in machine learning models.

8. Normal Distribution Graph (Empirical Rule)



9. Significance in Machine Learning

Using the empirical rule in this domain helps to:

- Identify **salary outliers**
- Improve **data preprocessing**
- Design better **regression models**
- Avoid biased predictions due to extreme values