## Features:

→A **feature** in a dataset is an individual measurable **property or characteristic of the data**. It represents a **variable that can be used for analysis or modeling**, typically appearing as a column in a dataset. Features provide the information needed to make predictions, classifications, or insights in machine learning and data analysis.
→A **feature** (also called an **attribute**, **variable**, or **column**) is an individual measurable property or characteristic of the data. Features are typically represented as columns in a table, and they describe different aspects of the data points (rows).

**Ex 1:** in a dataset of house prices, features might include **square footage, number of bedrooms, location**

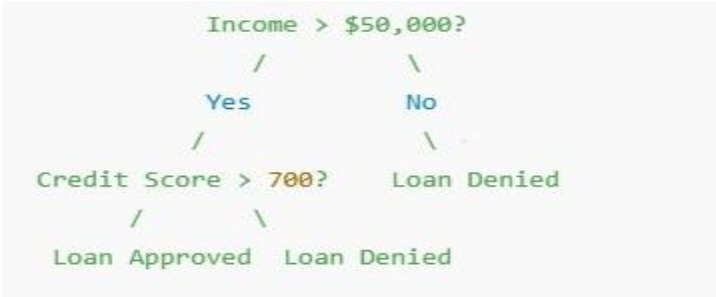
## Ex 2:

If a bank wants to approve loans based on customer profiles. A decision tree could use attributes such as **income, credit score, and debt-to-income ratio** to classify whether a loan should be **approved or denied**.

**Income > $50,000?**

Yes → **Credit Score > 700?**

- Yes → Loan Approved
- No → Loan Denied o No → Loan Denied




## Types of Features in a Dataset:

1. **Numerical Features (Continuous or Discrete)**
    - Ex: Age, Salary, Temperature

2. **Categorical Features**
    - Ex: Gender (Male/Female), Country (USA, India)
3. **Ordinal Features** (Categorical but with order)
    - Ex: Education Level (High School < Bachelor's < Master's < PhD)
4. **Binary Features**
    - Ex:: Loan Approval (Yes/No), Is Employee (0/1)
5. **Derived Features** (Engineered from existing ones)
    - Ex: BMI (from Weight & Height), Customer Score (from multiple behaviors)

## Confusion Matrix:

A Confusion Matrix is a table used to evaluate the performance of a classification model by comparing **Actual vs. Predicted values.**

## 2x2 Confusion Matrix:

| **Actual \ Predicted** | **Sentiment Analysis(SA) (Positive)** | **SA (Negative)** |
|---|---|---|
| **Sentiment Analysis(SA) (Positive)** | 40 (True Positive, TP) | 10 (False Negative, FN) |
| **(SA) (Negative)** | 5 (False Positive, FP) | 45 (True Negative, TN) |

## Definitions:

- **True Positive (TP) = 40**
- **False Negative (FN) = 10**

- **False Positive (FP) = 5**
- **True Negative (TN) = 45**

## Key Metrics:

- **Accuracy = (TP + TN) / (TP + TN + FP + FN) = (40 + 45) / 100 = 85%**
- **Precision = TP / (TP + FP) = 40 / (40 + 5) = 88.9%**
- **Recall = TP / (TP + FN) = 40 / (40 + 10) = 80%**
- **F1-score = 2 × (Precision × Recall) / (Precision + Recall) = 84.2%**

## 3x3 confusion Matrix:

### EX:

| Actual \ Predicted | Pred A | Pred B | Pred C |
|---|---|---|---|
| Actual A | TP_A | FP_BA | FP_CA |
| Actual B | FP_AB | TP_B | FP_CB |
| Actual C | FP_AC | FP_BC | TP_C |

## Business Problem :

A company wants to analyze customer feedback from online reviews to understand customer sentiment towards its products and services. Negative reviews can impact brand reputation and sales, while positive reviews can help identify strengths. However, manually analyzing thousands of reviews is time-consuming and inefficient.

## Business Objective :

Develop a sentiment analysis model that can automatically classify customer reviews as **positive, negative, or neutral**. The insights will help the company improve its products, address customer concerns, and enhance customer satisfaction.

## Business Constraints :

1. **Data Quality:**

Reviews may contain slang, misspellings, sarcasm, and mixed sentiments, making classification challenging.

2. **Class Imbalance:**

   The dataset may have more positive reviews than negative or neutral ones, leading to biased predictions.

3. **Computational Resources:**

   Training deep learning models for sentiment analysis requires significant computational power.