- **Data** is collection of facts.
- **Dataset** is collection of data.
- **Mean deviation (μ)** is the mean of absolute differences between each value in a set of value and the mean of all the values of that set.
- **Standard deviation (σ)** measures the spread in the data values of a continuous variable.
- **Standard visualization techniques** are Bar Chart, Histogram, Line Graph, and Scatterplot.

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING I

In order to predict an unknown event, we require following two things:

1. Supervised Data: The data should clearly indicate two set of features namely, indicator variables and predictive variable (class variable). Indicator variables are the set of variables used to forecast the predictive variable.

2. Supervised Techniques: A suitable supervised technique that can be applied on supervised data to get predictions.

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING II

Supervised Data-Example

Consider a hypothetical data set in Figure 11 medical domain, Where, observation on five patients is presented with attributes Age, X-Ray and Cancer.

Age	X-ray	Cancer
67	Normal	No
62	Abnormal	Yes
58	Normal	No
51	Normal	No
59	Abnormal	Yes

Figure 11: Hypothetical data set on medical domain

Suppose a new patient comes with (age = 54, X-ray = Normal) and, domain expert wants to predict if person is suffering from Cancer.

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING III

Supervised Data-Example

This problem on last slide can be stated as a classification problem for the reason that objective is ""prediction". The data set given in Figure 12 is a supervised data because the objective variable (which is Cancer in the problem) is given in the data set.

Supervised techniques divides the feature space into two categories, i.e., indicator and predictor. In this problem, (Age, X-Ray) are considered as indicator variables whereas, Cancer is a predictor variable (also called as class or label). The indicator variables helps in predicting the "class" of new unknown event.

Indicator variables

Predictive/class variable

Age	X-ray	Cancer
67	Normal	No
62	Abnormal	Yes
58	Normal	No
51	Normal	No
59	Abnormal	Yes

Features/attributes/variables

Data points/Observations/ Events/Objects

Figure 12: Hypothetical data set on medical domain

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING IV

Supervised Data- Formal Definition

The basic intention of providing supervised data to the classification technique is for the reason that we want it to "learn the mapping between indicator variables to a class variable".

Consider a data set of size ($m \times n$) where, m represents data points and n is the number of dimensions.

The set of features $D = (d_1, d_2, \ldots, d_{(n-1)})$ represents indicator variables whereas, the n^{th} feature indicates the class variable C.

Given this supervised data, a supervised technique will find a mapping function as in Equation 4.

$$C = f(D) \qquad (4)$$

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING IV

Supervised Data- Formal Definition

$D = (d1, d2, d3 \ldots\ldots, d(n-1), C)$

$f(D) \Rightarrow C$

$f(d1, d2, d3 \ldots\ldots, d(n-1)) \Rightarrow C$

$C = f(D)$(mapping function)

Indicator variables | Predictive/class variable

Age	X-ray	Cancer
67	Normal	No
62	Abnormal	Yes
58	Normal	No
51	Normal	No
59	Abnormal	Yes

Features/attributes /variables

Data points/ Observations/ Events/Objects

Figure 12: Hypothetical data set on medical domain

$f(Age, X\text{-}ray) \Rightarrow Cancer$

$C = f(D)$ (4)

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING V

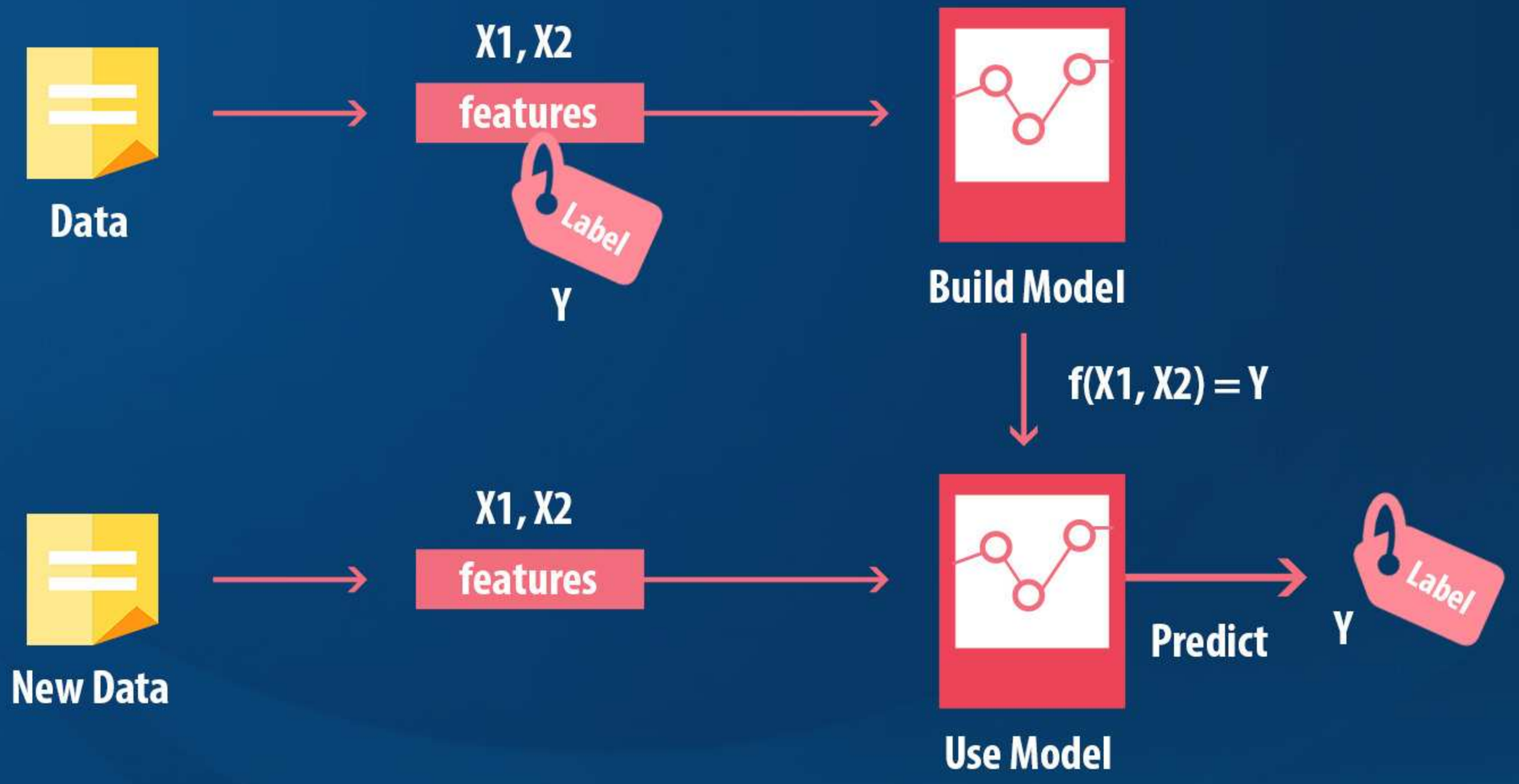

Figure 13: Learning mapping function f from the data set to predict an outcome Y

- **Steps involved in Supervised Techniques:**
  - Input data
  - Applying test data
  - Classification of data
  - Prediction
  - Learning the Mapping of data

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING VI

Supervised Techniques

Machine learning offers several supervised techniques in order to predict an unknown event. Supervised techniques are further divided into two categories namely, Classification and Regression. The objective of Classification and Regression is same, i.e., Prediction. However, they differ in their approaches of discovering mapping function and on the nature of class variable.

Table 6: Categories of Supervised Techniques

Classification	Regression
Objective: Prediction	
Differ	
1. Mapping function: $C = f(D)$ 2. Class variable is categorical/discrete in nature	1. Mapping function : $C = f(D)$ 2. Class variable is continuous/numerical in nature
Popular Techniques	
1. Naive Bayes 2. Support Vector Machines (SVM) 3. Decision Tree (DT) 4. K- Nearest Neighbour (KNN) 5. Random Forest 5. Logistic Regression	1. Linear Regression 2. Multiple Regression

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING VII

Nature of Indicator and Predictive variables

The variables or features in the data set can take up different type of values such as, integer, real, text, binary and more. The most common type of data values encountered in daily life are: continuous and categorical.

The continuous variable is always numerical in nature for eg. income, height. It can take any value in range of [$-\infty$, $+\infty$].

However, a categorical variable contains values of fixed category or group.

1. Categorical

- Gender : it takes value from a group of [male, female]
- Marital status : it values from a fixed pool of categories [married, unmarried, divorced]
- Ball Color : It may takes values in group of [red, blue, green,..,]
- Dog breed: The values of this variable may range in [collie, shepherd, terrier]

2. Continuous

- Age : A numerical integer value
- Salary: A numerical real value
- Height: A numerical integer/real value
- Population: A numerical real value

INTRODUCTORY CONCEPTS OF SUPERVISED LEARNING VIII

Nature of Indicator and Predictive variables

Consider Figure 14. In Data set 1, class variable is categorical whereas, in second Data set 2 it is continuous.

Categorical	Continuous	Categorical
Position	Salary (in K)	Gender
Manager	123.89	Male
Client	134.23	Female
Senior Executive	137.10	Male
Branch Head	165.67	Female
Assistant Manager	166.78	Female

Data Set 1 — Class

Categorical	Continuous	Continuous
Country	Production (in K)	Sales (in K)
India	1123.89	67.78
Australia	1334.23	78.90
Japan	1897.10	67.67
China	2695.67	34.12
Europe	2666.78	69.45

Data Set 2 — Class

Figure 14: Hypothetical data sets showcasing categorical and continuous features

Classification techniques are applicable on data set of first kind where class label is "categorical". While Regression is suitable on second data set for the reason that class is "continuous" in nature.