INTRODUCTION TO MACHINE LEARNING

CONTENTS

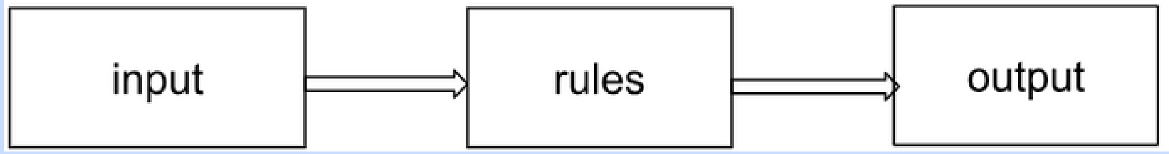
1. Introduction to Machine Learning (ML)
2. Understanding the basic concepts before diving into algorithms

TRADITIONAL PROGRAMMING VS. MACHINE LEARNING:

Traditional Programming:

- In Traditional Programming we define fix set of rules using a program and based on that program we will get output.
- For example we have given a number as input and we need to calculate whether it is even or odd.
- Then we will write some set of rules and find out information about even or add.

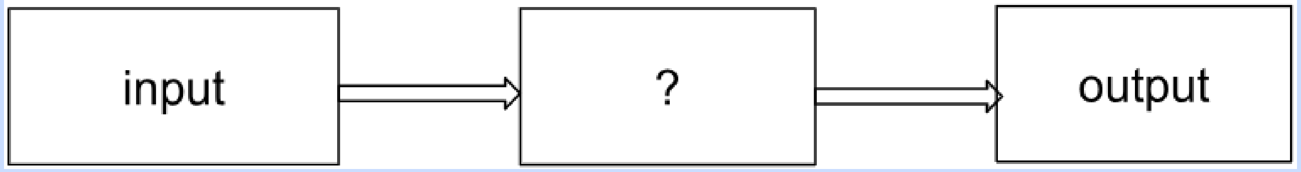
```
if(num%2 == 0){
cout<<“Even”<<endl;
}
else{
cout<<“Odd<<endl;
}
```



Machine Learning:

- Machine Learning (ML) is a branch of artificial intelligence (AI) that enables computers to learn and make decisions without explicit programming.
- In machine learning the machine learns patterns and relationships without explicit programming.

Input, output, and learning in between

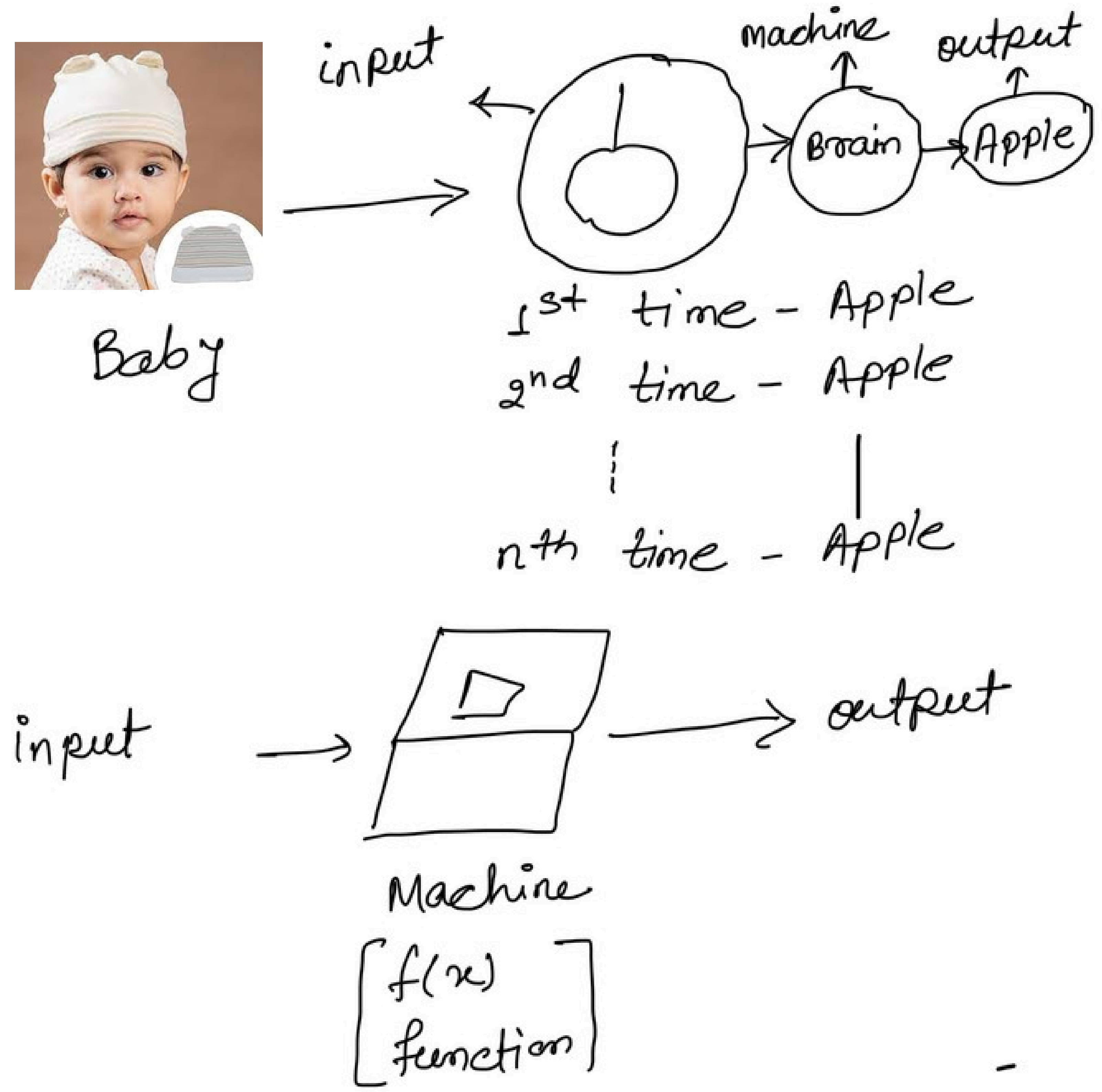


Model to learn
Or
Machine to learn
f(x)
Learning a function (e.g., f(x)) from data

HUMAN LEARNING ANALOGY:

- For example a baby learning to identify apples
- We need to tell Baby a lot of time (may be thousand of time) that it is an Apple, then he will understand about Apples.

- Similarly in case of machine we need to give a lot of data to machine than machine will understand how to proceed with further data.
- In the image you can see when we give input than machine will understand those input data and then give output for further data.



CLASSIFICATION OF MACHINE LEARNING

Feature	SUPERVISED LEARNING	SEMI-SUPERVISED LEARNING	UNSUPERVISED LEARNING
Definition	Learning with labeled data to predict output.	Combines a small amount of labeled data with a large amount of unlabeled data.	Learning with unlabeled data to find hidden patterns.
Input Data	Labeled data (features + labels).	Mostly unlabeled data with a few labeled samples.	Unlabeled data (features only).
Goal	Predict the outcome based on known labels.	Improve model performance by leveraging both labeled and unlabeled data.	Discover hidden patterns or structure.
Example Algorithms	Linear Regression, Decision Trees, SVM, KNN.	Semi-Supervised SVM, Label Propagation, Co-training.	K-Means, Hierarchical Clustering, PCA, t-SNE.
Accuracy	High if sufficient labeled data is available.	Typically more accurate than unsupervised learning but less than supervised.	Can vary depending on the complexity of the patterns.

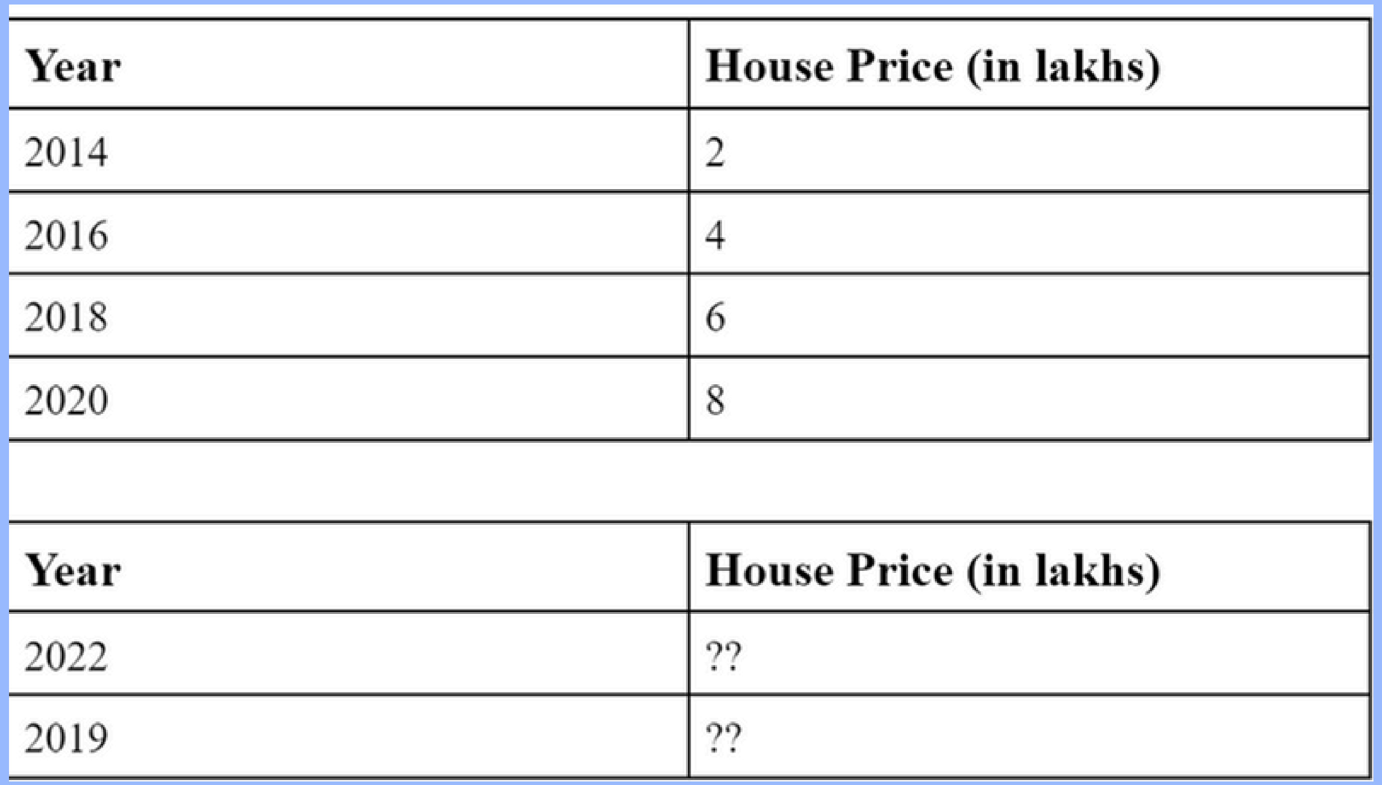
SUPERVISED LEARNING - REGRESSION :

PREDICTING HOUSE PRICES:

Scenario:

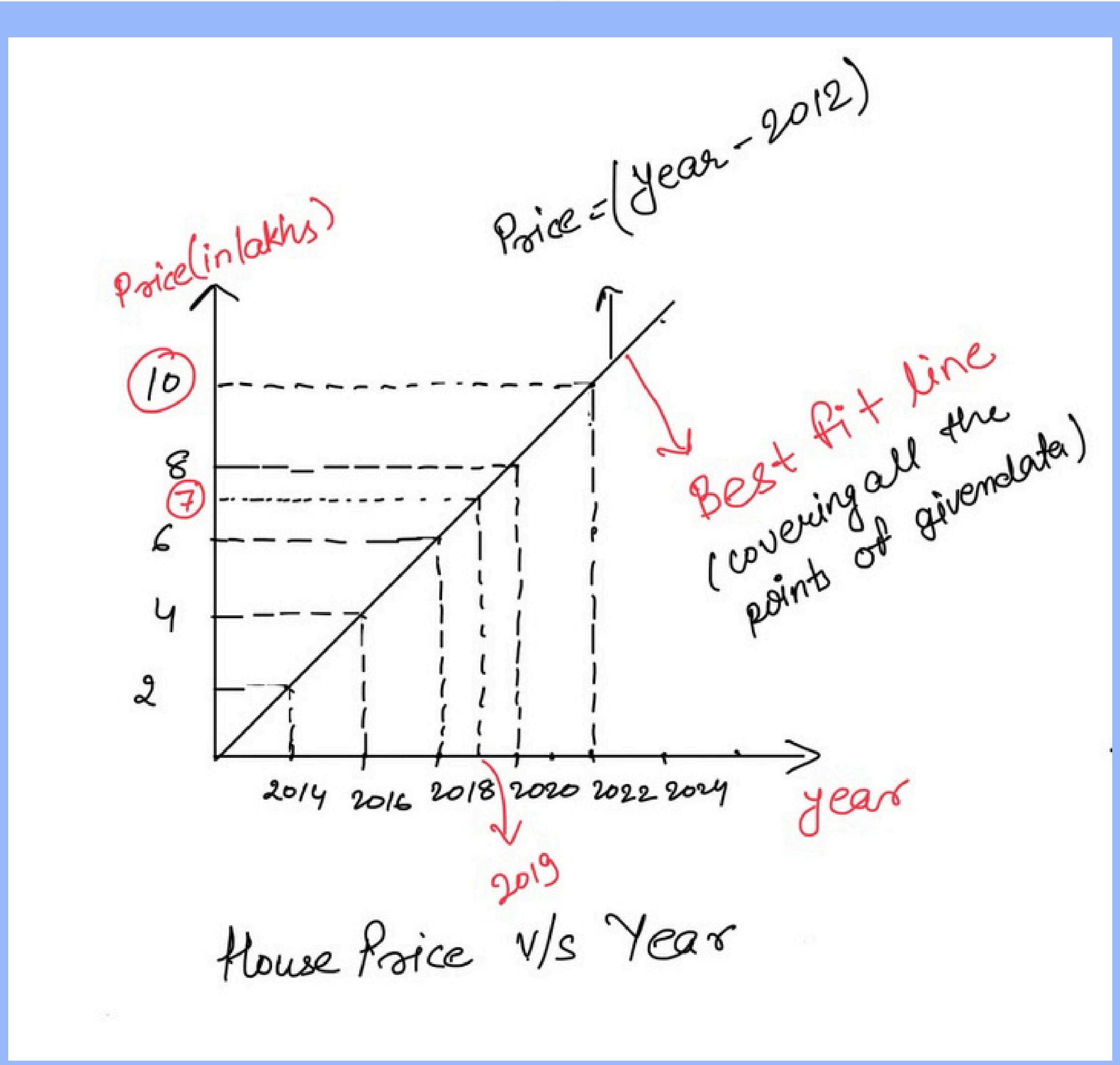
- Given historical data of house prices from 2014 to 2020
- Task is to predict the price for the year 2022 and 2019.

- Plotting the data on a graph
- Observing a linear relationship in the data



Year	House Price (in lakhs)
2014	2
2016	4
2018	6
2020	8

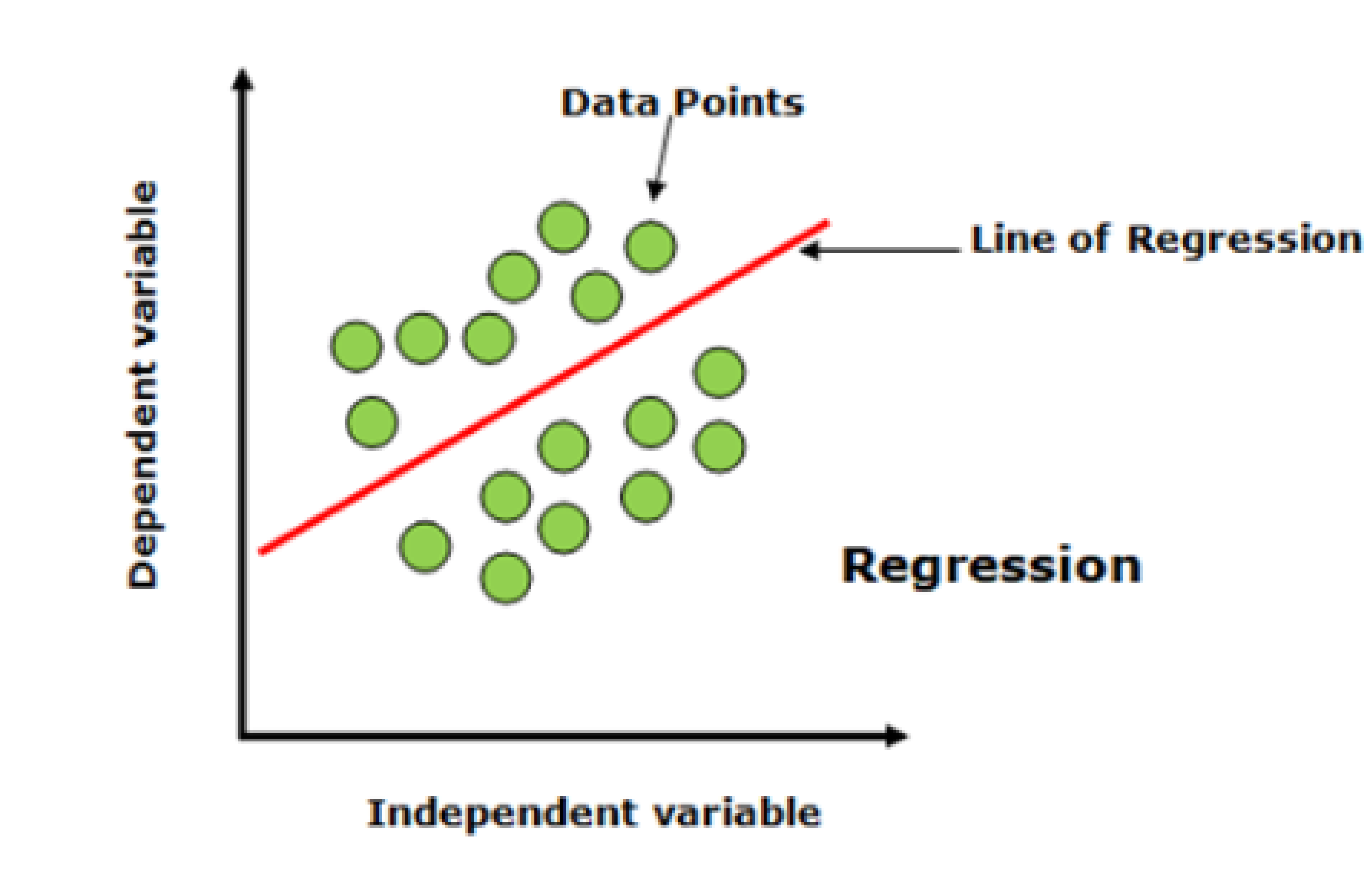
Year	House Price (in lakhs)
2022	??
2019	??



Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

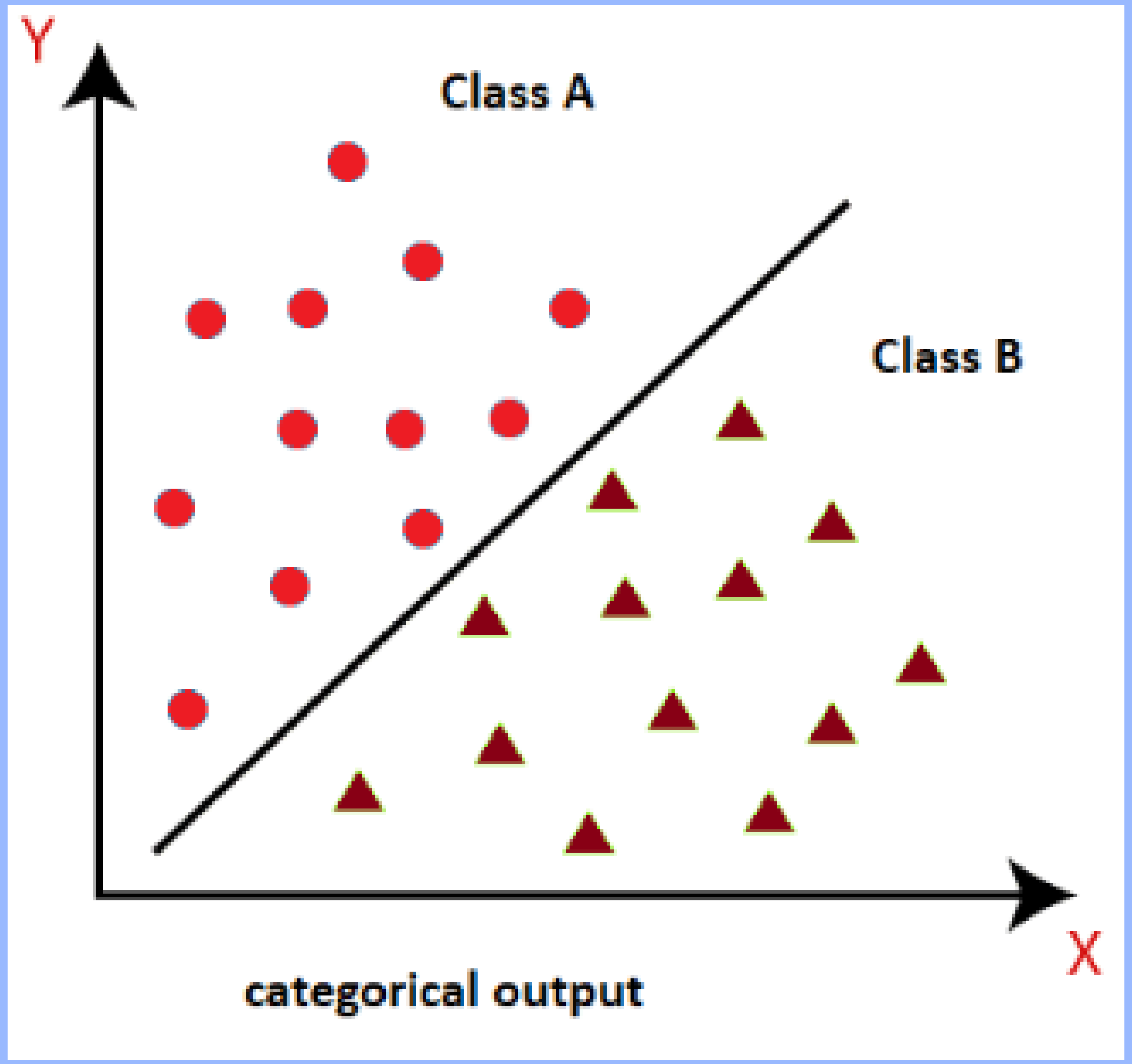


Classification

Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not. Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.

Some common classification algorithms include:

- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forests
- Naive Baye



Semi-Supervised Learning

- Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning.
- It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model.
- The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning.
- However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

Examples of Semi-Supervised Learning:

- Text classification
- Image classification
- Anomaly detection

Anomaly Detection

- Anomaly is the pattern in the data that does not confirm to the expected behavior
- Detecting outliers, exeptions or anomalies in a dataset.
- Identify instances that deviate significantly from the norm.

Example:

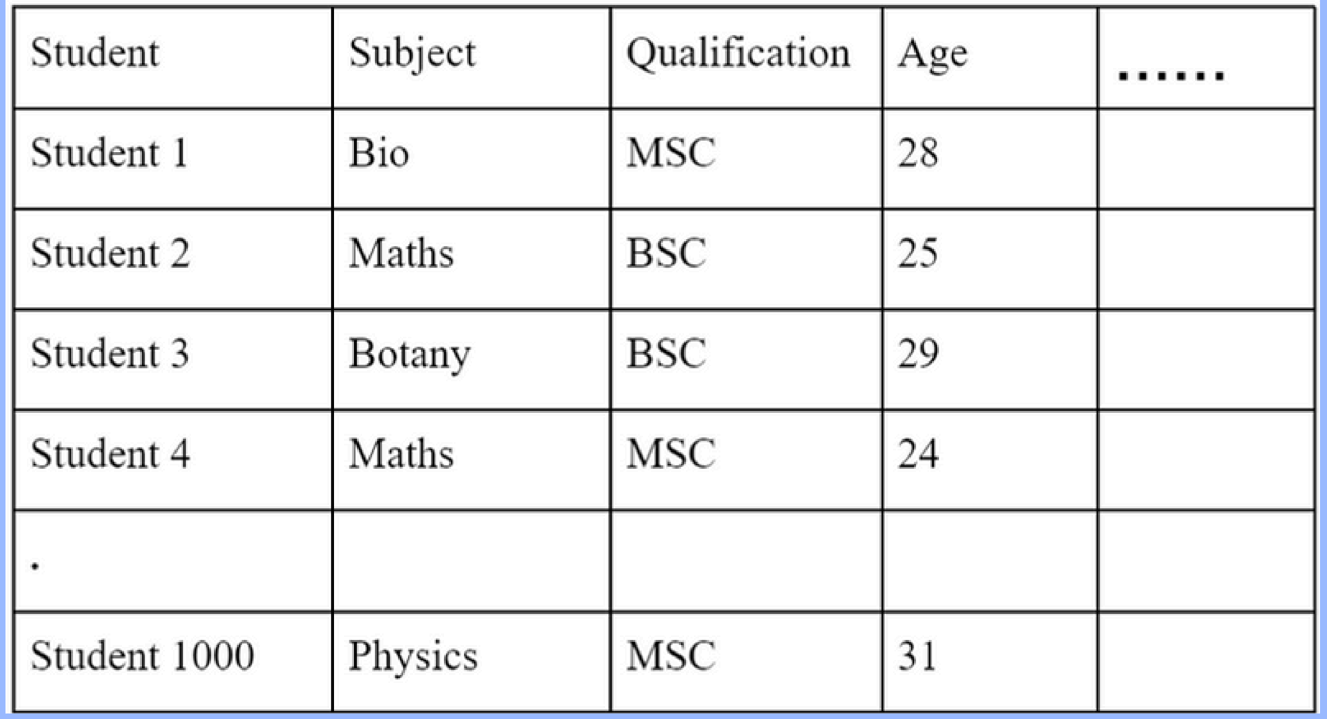
- Credit card transactions are monitored for anomalies to enhance security.
- User makes a credit card purchase in Mumbai.
- Shortly afterward, another transaction is attempted in China.
- Physically impossible for an individual to be in both Mumbai and China in such a short time frame.
- Significance of Anomaly: High likelihood of fraudulent activity due to improbable travel distance.

UNSUPERVISED LEARNING - CLUSTERING:

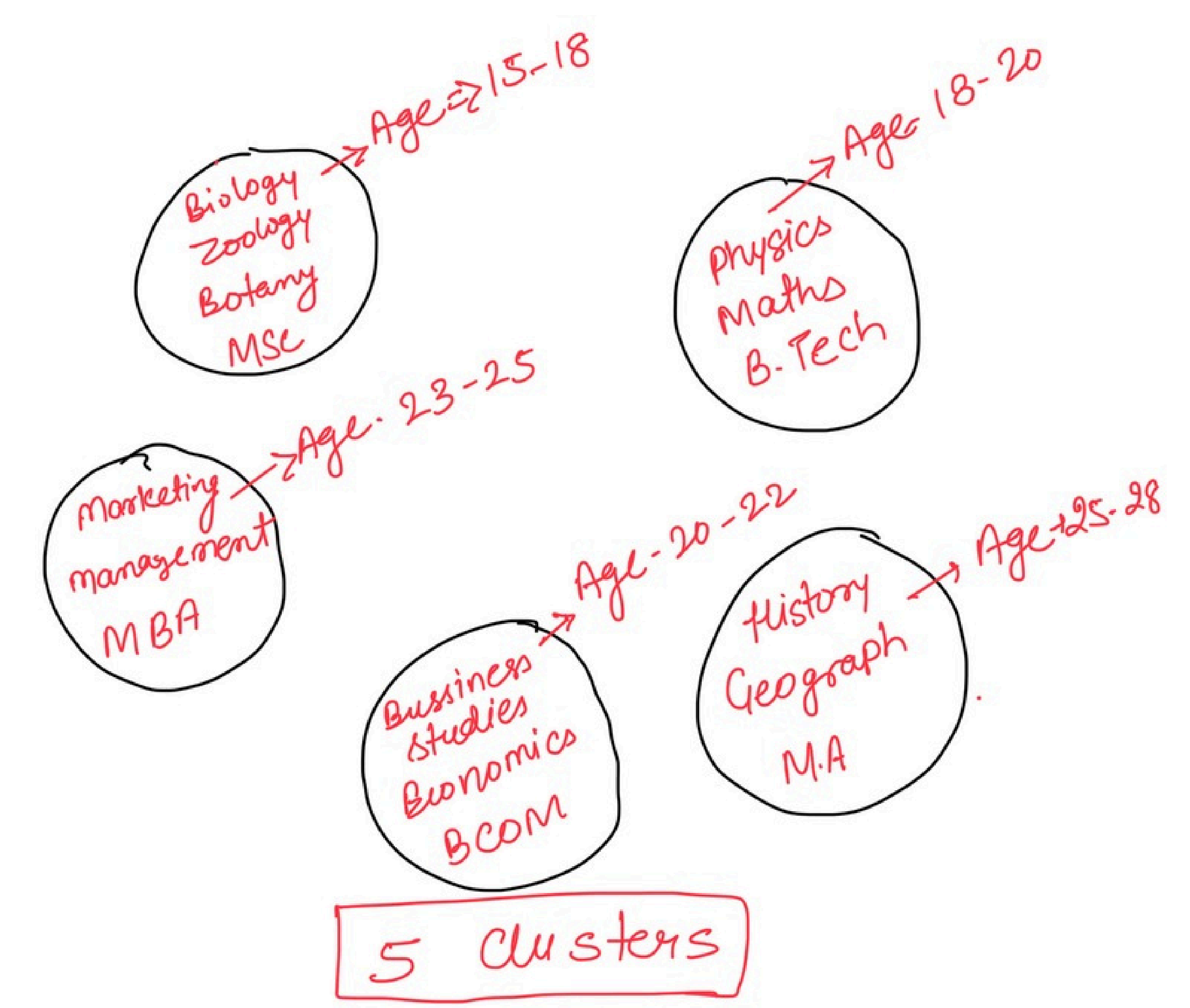
- Clustering involves grouping data points with similar characteristics.
- It is commonly used for tasks like customer segmentation or recommendation systems.
- K-means clustering is an example where data points are grouped into k clusters based on similarities.

Example - Recommendation System:

- A recommendation system suggests items based on what other users with similar preferences have chosen.
- This is an application of unsupervised learning and clustering.
- Users or items are grouped into clusters, and recommendations are made based on the preferences of users in the same cluster.

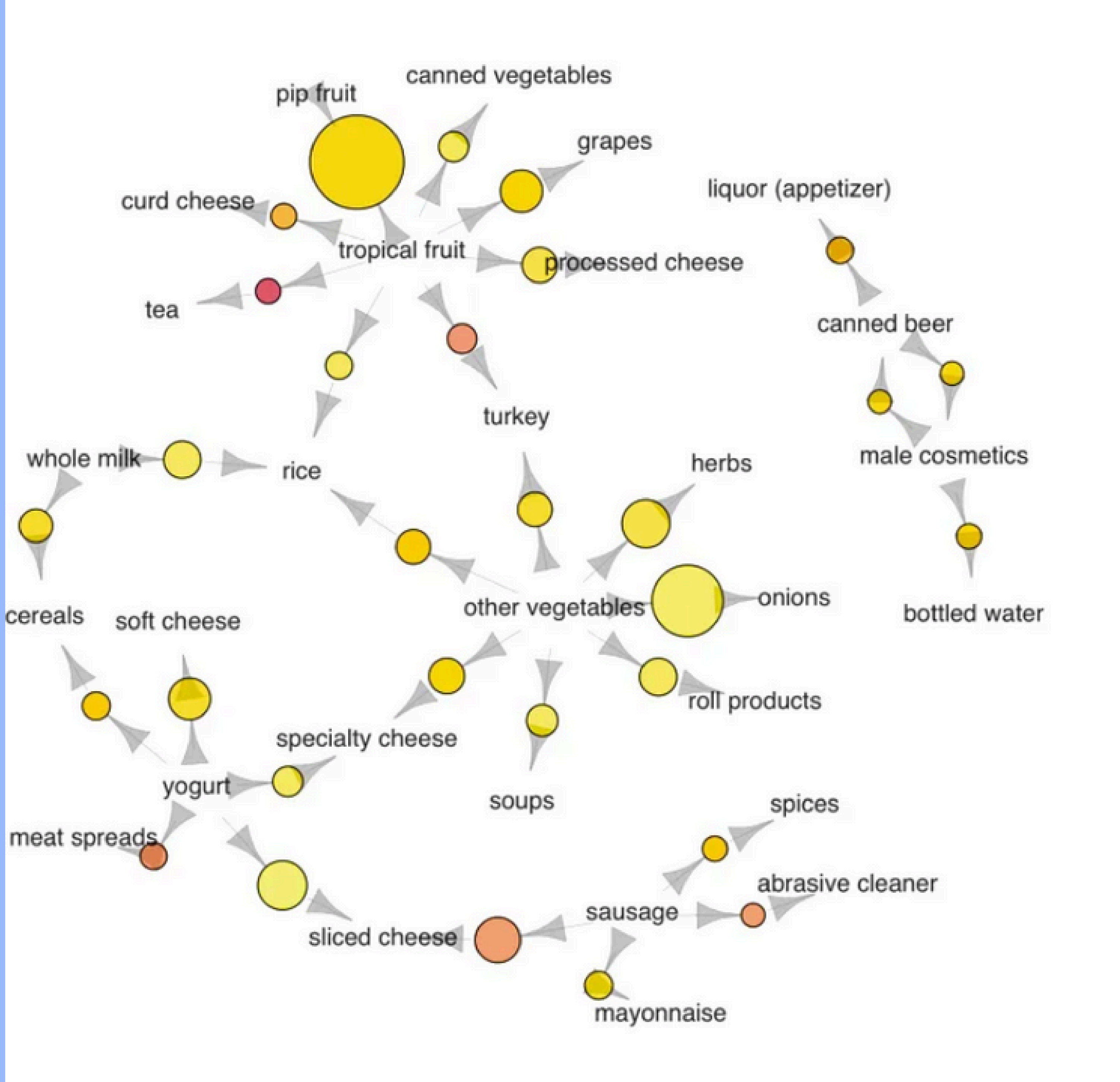


Student	Subject	Qualification	Age	......
Student 1	Bio	MSC	28	
Student 2	Maths	BSC	25	
Student 3	Botany	BSC	29	
Student 4	Maths	MSC	24	
.				
Student 1000	Physics	MSC	31	



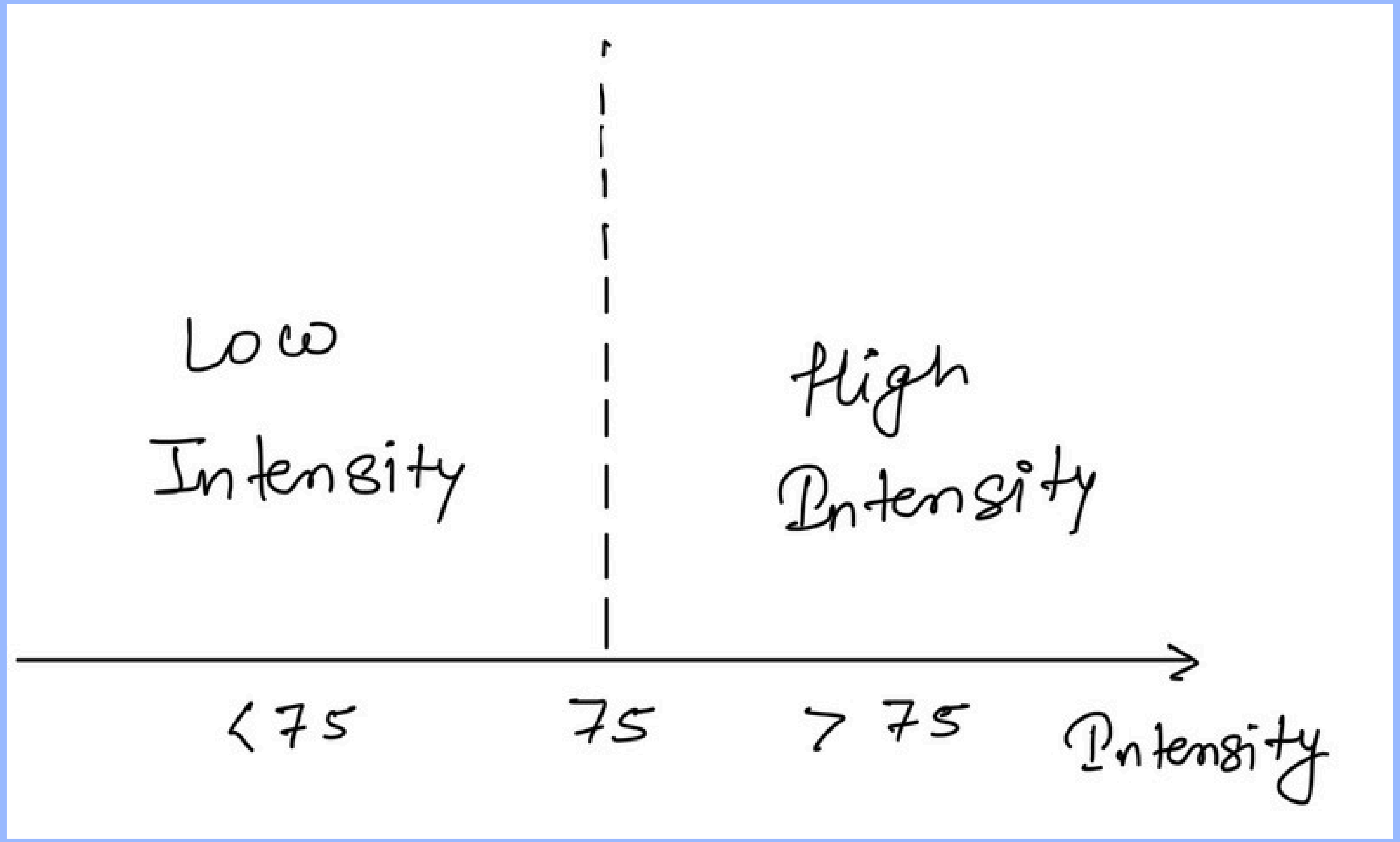
Association

- An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database.
- It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item.
- A typical example of Association rule is Market Basket Analysis.

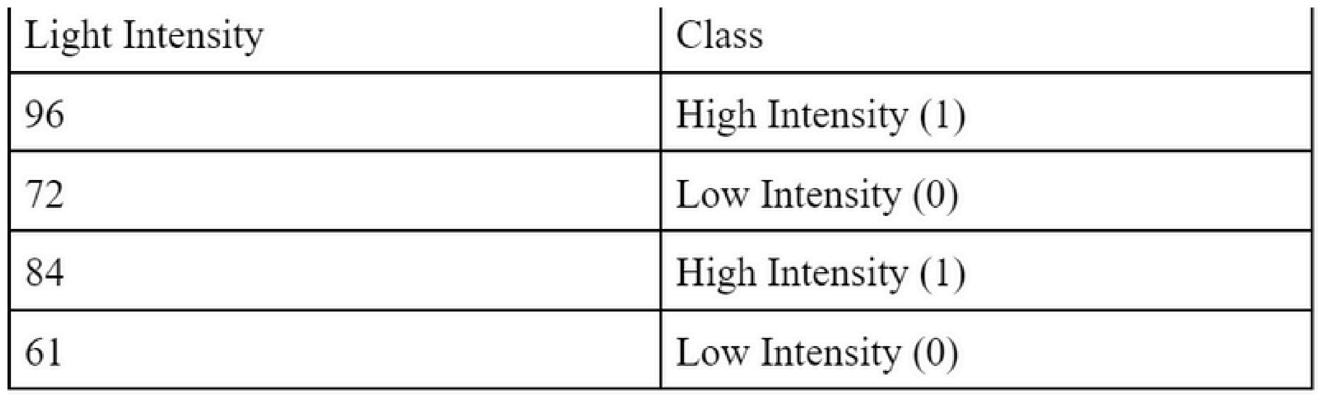


CLASSIFICATION BASED ON CLASS

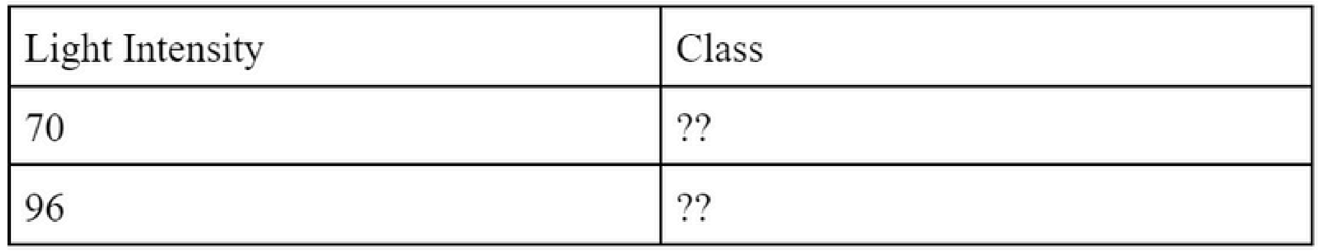
Classification involves predicting the category or class an instance belongs to.
Binary classification has two classes (0 or 1), while multi-class classification involves more than two classes (e.g., 0, 1, 2).
Eg., Light Intensity Binary Classification



Binary classification

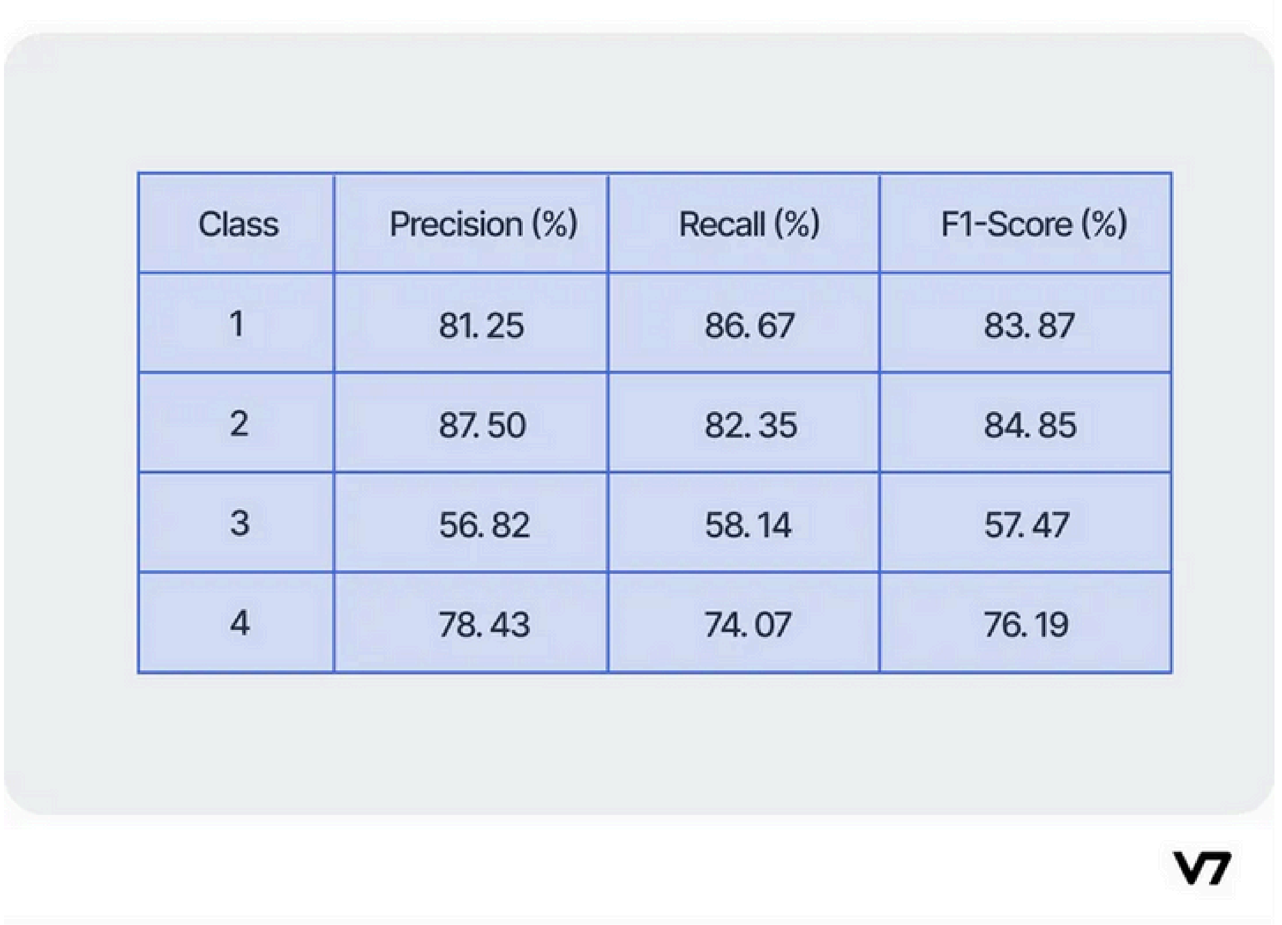


Light Intensity	Class
96	High Intensity (1)
72	Low Intensity (0)
84	High Intensity (1)
61	Low Intensity (0)



Light Intensity	Class
70	??
96	??

Multiclass Classification



Class	Precision (%)	Recall (%)	F1-Score (%)
1	81. 25	86. 67	83. 87
2	87. 50	82. 35	84. 85
3	56. 82	58. 14	57. 47
4	78. 43	74. 07	76. 19

V7

TYPES OF DATA:

Structured Data:

- Rows and columns, e.g., tabular data.

Unstructured Data:

- No specific structure, e.g., images, text, videos.

Natural Language Processing (NLP) Examples:

Google Lens:

- Combines image and text for information extraction.

Tesla Autopilot:

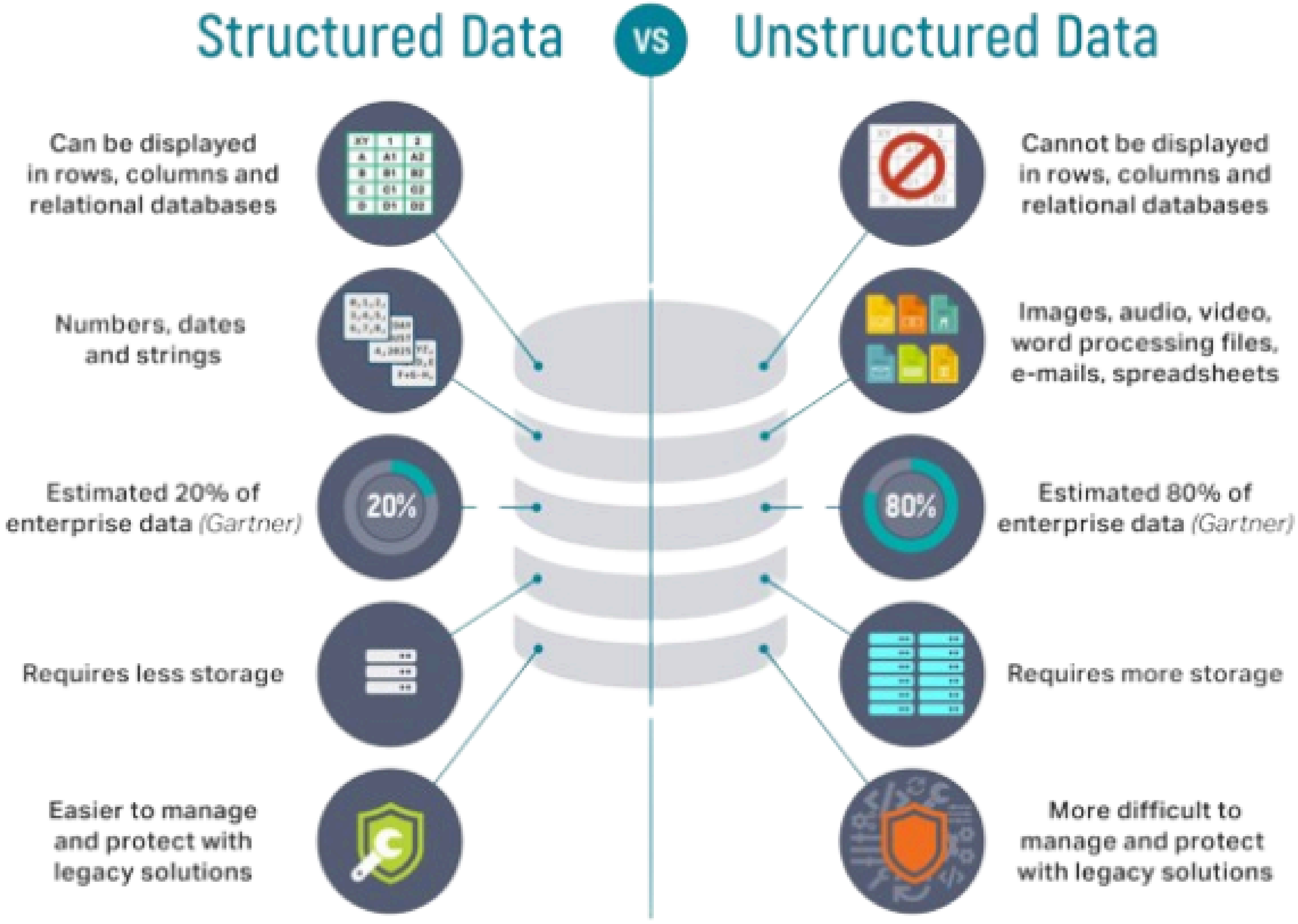
- Navigates based on image (video) data.

Gmail Autocomplete:

- Uses NLP for predicting next words in emails.

Google Search Engine:

Utilizes NLP for autocomplete in search queries.



Structured Data VS Unstructured Data

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data *(Gartner)*

Requires more storage

More difficult to manage and protect with legacy solutions

DEPENDENT AND INDEPENDENT:

- Dependent Variable (Y): What we want to predict (e.g., house price).
- Independent Variables (X): Features influencing the prediction independently.

Record/Sample/Data Point:

- Rows in the dataset, each representing a record, sample, or data point.

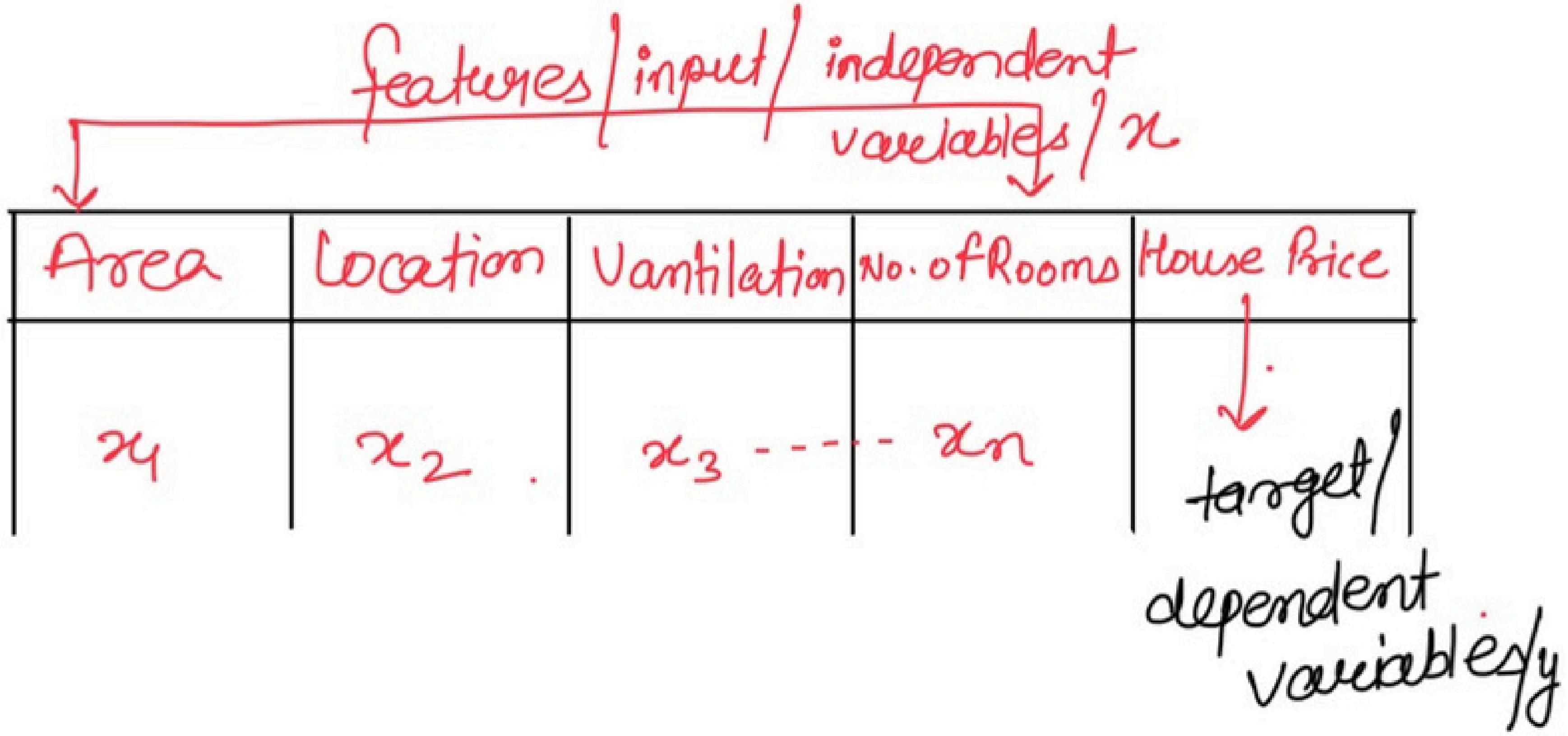
Notations:

- X1, X2, X3, ..., Xn: Features (independent variables).
- Y: Dependent variable.

Common Usage:

- X and Y are used interchangeably with features and target variables.

Eg., House Price Prediction


features/input/ independent varelables/x
target/ dependent variables/y

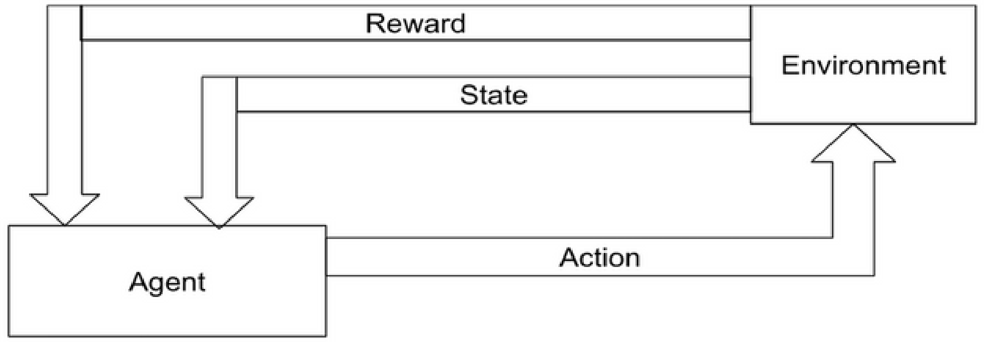
Area	Location	Vantilation	No. of Rooms	House Price
x_1	x_2 .	x_3 - - -	- x_n	

REINFORCEMENT LEARNING

Involves an agent learning from actions, states, and rewards in an environment.

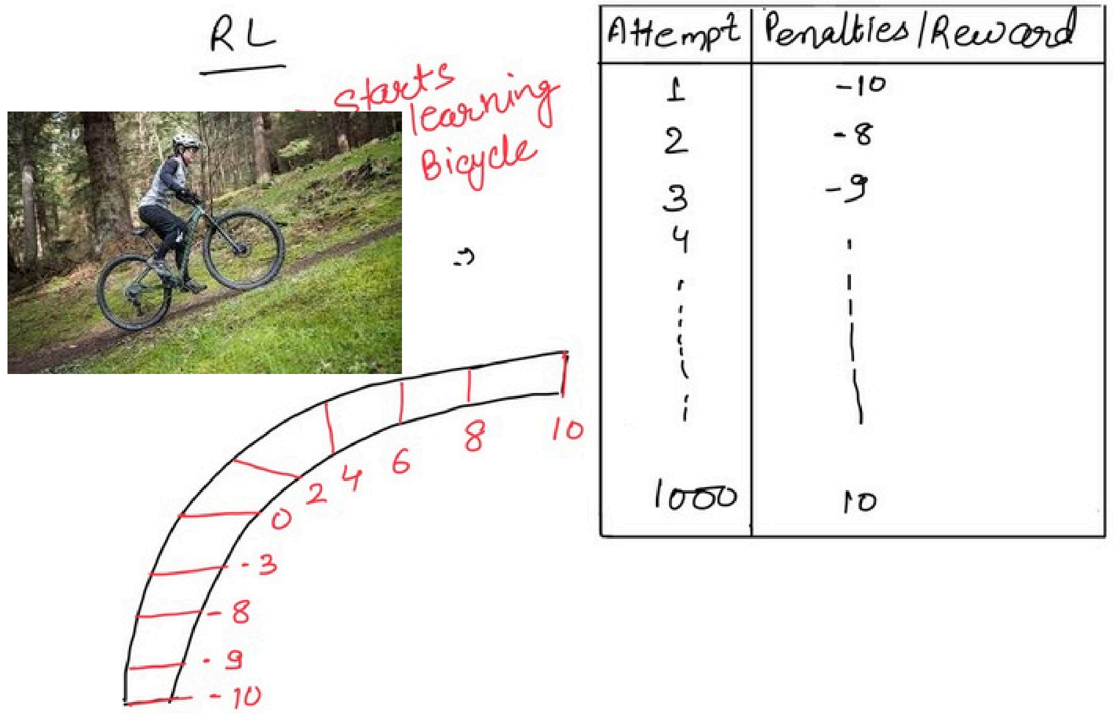
RL Components:

- **Agent**: Entity performing actions in an environment.
- **Environment**: Surroundings where actions occur.
- **State**: Resulting condition or situation from an action.
- **Reward**/Penalty: Consequences assigned to actions; positive for desired outcomes, negative for undesired ones.



RL Example:

- A person wants to learn how to ride a bicycle. Initially, he doesn't know how to ride a bicycle.
- He starts attempting rides again and again to learn how to ride a bicycle.
- He has been assigned a task to turn his bicycle, and based on his falling point in any attempt, he will receive rewards and penalties.
- Illustrated in image below:



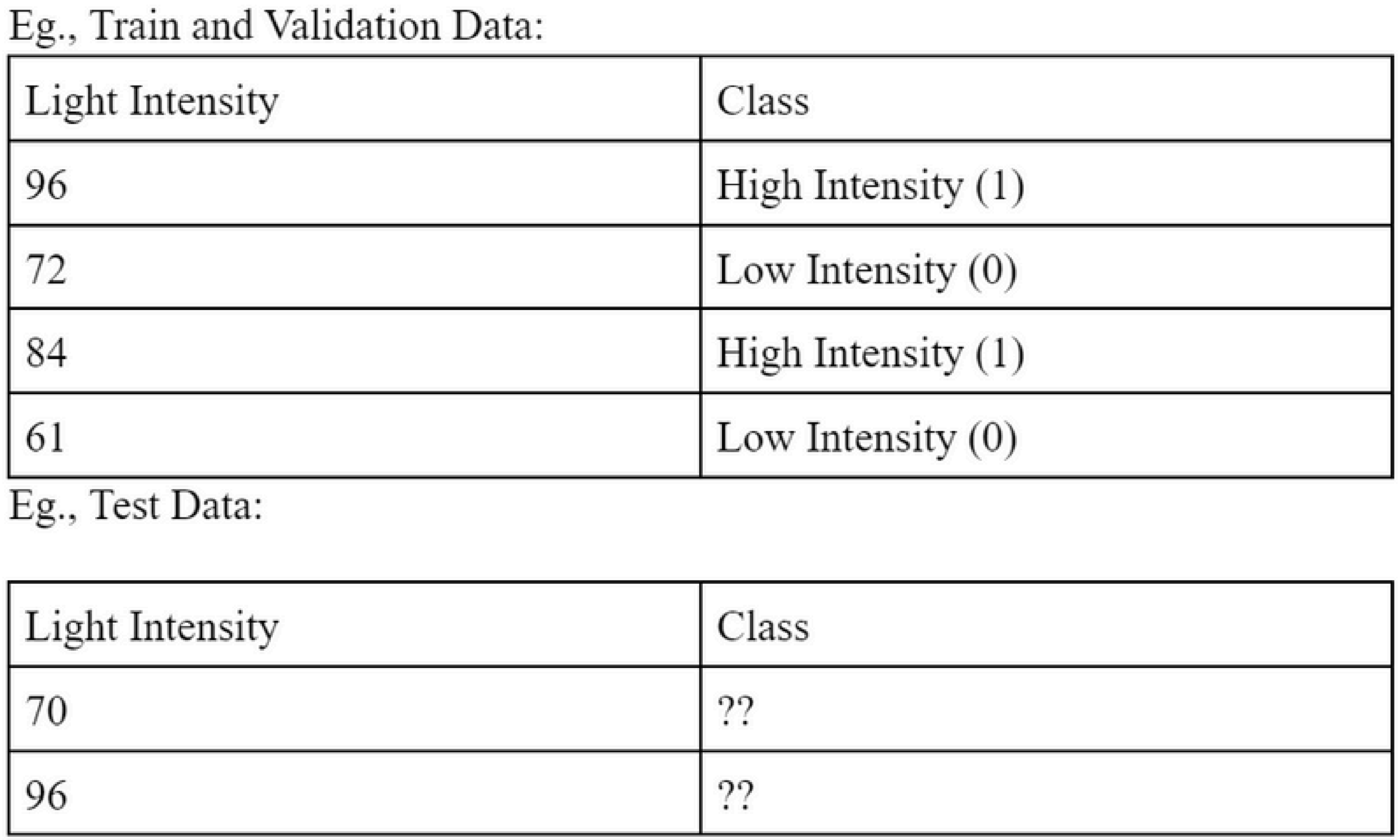
Attempt	Penalties/Reward
1	-10
2	-8
3	-9
4	⋮
⋮	⋮
1000	10

IMPORTANT TERMINOLOGIES IN ML:

- Train Data: The data used to train a machine learning model.
- Test Data: The data used to evaluate the model's performance.
- Validation Data: Optionally used to fine-tune the model during training.

Features:

- Features are the independent variables affecting the output.
- Examples of features for house price prediction: area, location, square feet, ventilation, number of rooms.



Eg., Train and Validation Data:

Light Intensity	Class
96	High Intensity (1)
72	Low Intensity (0)
84	High Intensity (1)
61	Low Intensity (0)

Eg., Test Data:

Light Intensity	Class
70	??
96	??

Machine Learning Model Flow

- Structured data with rows and columns.

Pre-processing:

- Data cleaning and transformation.

Train-Validation Split:

- Splits data into training and validation sets.

Model Training:

- Algorithms learn from training data.

Validation and Error Evaluation:

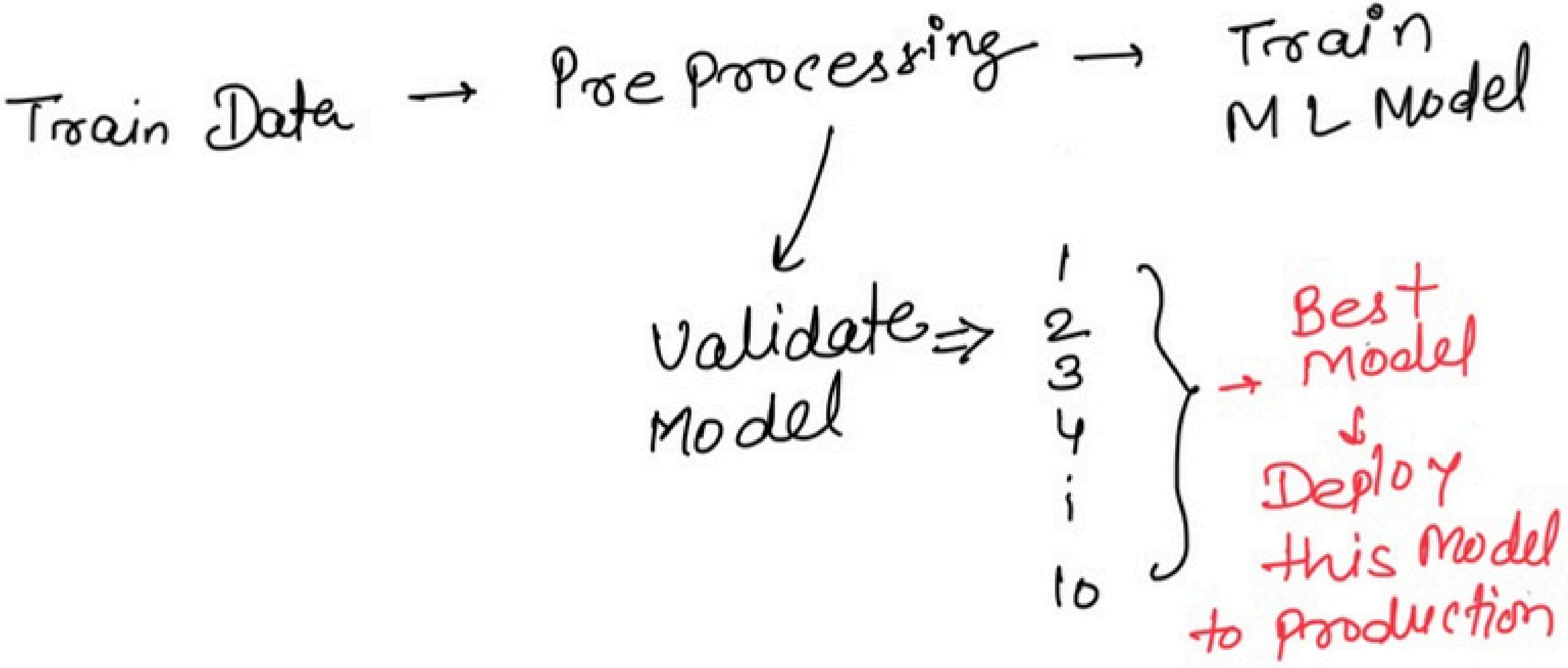
- Model validated on a separate dataset.
- Error metrics used for evaluation.

Model Comparison:

- Multiple models compared based on error metrics.

Deployment to Production:

- Best-performing model deployed for real-world use.



Model Pipeline for House Price Prediction:

- Input Data: Features representing different aspects of a house, such as square feet, area, and room numbers.
- Target Variable (Y): Predicting the price of the house based on the input features.
- Train Data Set: Initial dataset used for model training.

Data Pre-processing:

- Handling Irregular Values: Identifying and handling irregularities, e.g., filtering out houses with zero square feet.
- Feature Correlation: Assessing correlation between features; dropping redundant features for better model performance

Data Splitting:

- Train-Validation Split: Splitting the dataset into a training set (70%) and a validation set (30%) for model training and evaluation.

.

Model Training:

- Training Process: Utilizing the training set to train the machine learning model (e.g., linear regression).
- Best Fit Line: Finding the best fit line that represents the relationship between features and house prices.

Model Evaluation:

- Predictions: Applying the trained model to the validation set, generating predictions for house prices.
- Error Calculation: Calculating errors by comparing actual prices with predicted prices for each house.
- Error Metrics: Understanding different evaluation metrics like Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE).
- Model Comparison: Comparing models based on error metrics; lower error indicating better model performance.
- Loss Functions: Interpreting error metrics as loss functions, where a lower loss signifies a better-performing model.

Feature Correlation:

- Examining whether certain features are highly correlated is essential.
- Dropping one of two highly correlated features may improve model efficiency.

Standard Deviation:

- A measure of the amount of variation or dispersion in a set of values.

Error Calculation:

- In this context, simple error calculation is done by subtracting predicted values from actual values.

Loss Function:

- Error metrics, such as RMSE, are sometimes referred to as loss functions.
- A lower loss indicates better model performance.