# Unit No 5 Classification & Prediction

- **Classification:**
  1. **Defn:- Classification** is a classic **data mining** technique that assigns items in a collection to target categories or classes.
  2. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.
  3. Example:
     - A bank loan officer wants to analyze the data in order to know which customers (loan applicant) are risky or which are safe.
     - A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.
- **Prediction:**
  1. The goal of prediction is to forecast or deduce the value of an attribute based on values of other attributes.
  2. A model is first created based on the data distribution.
  3. The model is then used to predict future or unknown values.
  4. Example:
     - Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction.

- **Difference bet Classification & Prediction**

| | **Classification** | **Prediction** |
|---|---|---|
| Defn | Classification is the process of identifying the category or class label of the new observation which it belongs to. | Predication is the process of identifying the missing or unavailable numerical data for a new observation. |
| Goal | The goal of data classification is to organize and categorize data in distinct classes. | The goal of prediction is to forecast or predict the value of an attribute based on values of other attributes. |
| Accuracy | In classification, the accuracy depends on finding the class label correctly | the accuracy depends on how well a given predicator can guess the value of a predicated attribute for a new data. |
| Model | Classification predicts categorical (discrete, unordered) labels, | Prediction models continuous valued functions |
| Synonyms | Classification Model is classifier. | Prediction Model is Predictor |
| Use | Used for -forecasting **discrete** value | Used for -forecasting **Continuous** value |
| Example | A bank loan officer wants to analyze the data in order to know which customers (loan applicant) are risky or which are safe. | Marketing manager needs to predict how much a given customer will spend during a sale at his company. |

- **Classifiers Of Machine Learning**
  - Decision Trees
  - Bayesian Classifiers
  - Neural Networks
  - K-Nearest Neighbour
  - Support Vector Machines
  - Linear Regression
  - Logistic Regression

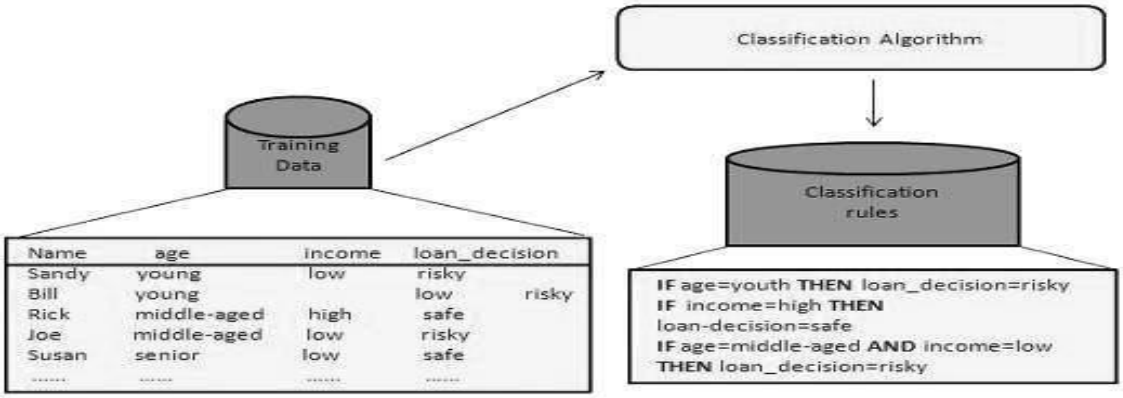
- **How Does Classification Works:**

The Data Classification process includes two steps −

1. Building the Classifier or Model
2. Using Classifier for Classification

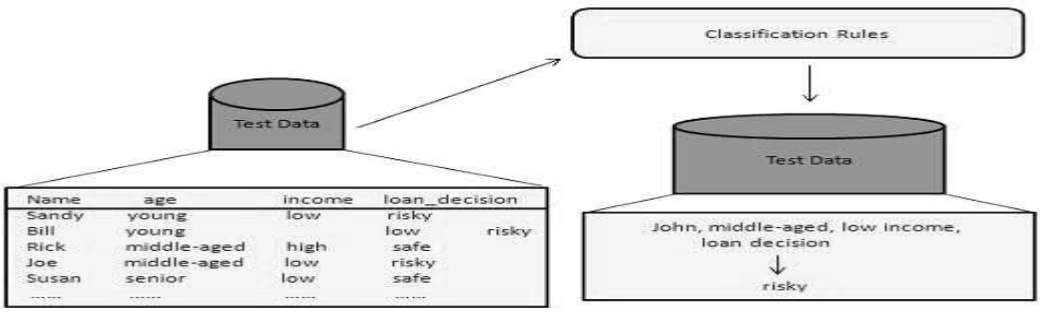
**1. Building the Classifier or Model**

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels..




**2. Using Classifier for Classification**

- In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules.
- The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



- **How Does Predication Works:**

  - Regression is generally used for predication.
  - Predicating the value of a house depending on the facts such as the number of rooms, the total area etc. is an example for predication.
  - A company might find the amount of money spent by the customer during a sale. That is also an example for prediction.

1. Basic **Predication** approaches:

   - Instance-based (nearest neighbor)
   - Statistical (naive bayes)
   - Bayesian networks
   - Regression (a kind of concept learning for continuous class)

- **Classification and Prediction Issues**

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities −

- **Data Cleaning** − Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

- **Relevance Analysis** − Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.

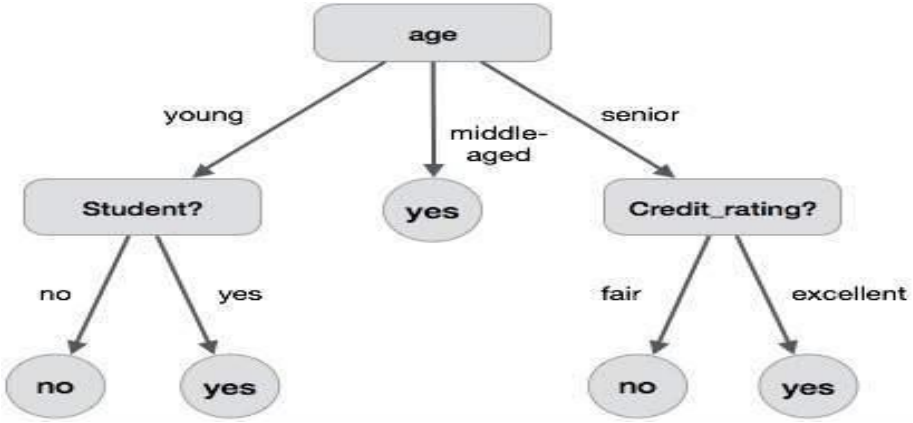
- **Data Transformation and reduction** − The data can be transformed by any of the following methods.

  - **Normalization** − The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.

  - **Generalization** − The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

- **Decision Tree Induction: one of Approach of classification**
  1. Decision tree is the most powerful and popular tool for classification and prediction.
  2. A Decision tree is a flowchart like tree structure, where **each internal node denotes a test on an attribute**, **each branch represents an outcome of the test,** and each **leaf node (terminal node) holds a class label.**

**Example:** The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.




- **Decision Tree Induction Algorithm**
  1. A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser).
  2. Pseudo Code:

```
ID3(Examples, Attributes, TargetAttribute)

Create RootNode for the tree
if all members of Examples are in the same class C
        then RootNode = single-node tree with label = C
else if Attributes is empty
        then RootNode = single-node tree with label = most common value of Target_attribute in
                                                      Examples;
else
        A = element in Attributes that maximizes InformationGain(Examples, A)
A is decision attribute for RootNode
for each possible value v of A
        add a Branch below RootNode, testing for A = v
        Examples_v = subset of Examples with A = v
        if Examples_v is empty
                then below Branch add Leaf with label = most common value
                                                        of Target_attribute in Examples;
        else
                below Branch add Subtree ID3(Examples_v, Attributes - {A}, TargetAttribute);

return RootNode;
```

- **Strengths and Weakness of Decision Tree approach**

  The strengths of decision tree methods are:

  1. Decision trees are able to generate understandable rules.
  2. Does not require domain Knowledge.
  3. Decision trees perform classification without requiring much computation.
  4. Decision trees are able to handle both continuous and categorical variables.
  5. Decision trees provide a clear indication of which fields are most important for prediction or classification.

  The weaknesses of decision tree methods:

  1. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
  2. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
  3. Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive.

- **Prediction:**
  1. **Predictive Data Mining is** also known as Regression.
  2. Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset.
  3. Predict real-valued output for given input – given a training set.
  4. Regression involves **predictor variable** (the values which are known) and **response variable** (values to be predicted).

  Examples:

  • Predict rainfall in cm for month

  • Predict stock prices in next day

  • Predict number of users who will click on an internet advertisement

- **The two basic types of regression are:**
  1. **Linear regression**
  - It is simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.
  - Linear regression attempts to find the mathematical relationship between variables.
  - If outcome is straight line then it is considered as linear model and if it is curved line, then it is a non linear model.

- The relationship between dependent variable is given by straight line and it has only one independent variable.

  $\mathbf{Y = \alpha + B\ X}$
- Model '**Y**', is a linear function of '**X**'.
- The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.

2. Multiple regression model
   - Multiple linear regression is an extension of linear regression analysis.
   - It uses two or more independent variables to predict an outcome and a single continuous dependent variable.

     $\mathbf{Y = a_0 + a_1\ X_1 + a_2\ X_2 + \ldots\ldots\ldots + a_k\ X_k + e}$

     **where,**

     '**Y**' is the response variable.

     $\mathbf{X_1 + X_2 + X_k}$ are the independent predictors.

     '**e**' is random error.

     $\mathbf{a_0, a_1, a_2, a_k}$ are the regression coefficients.