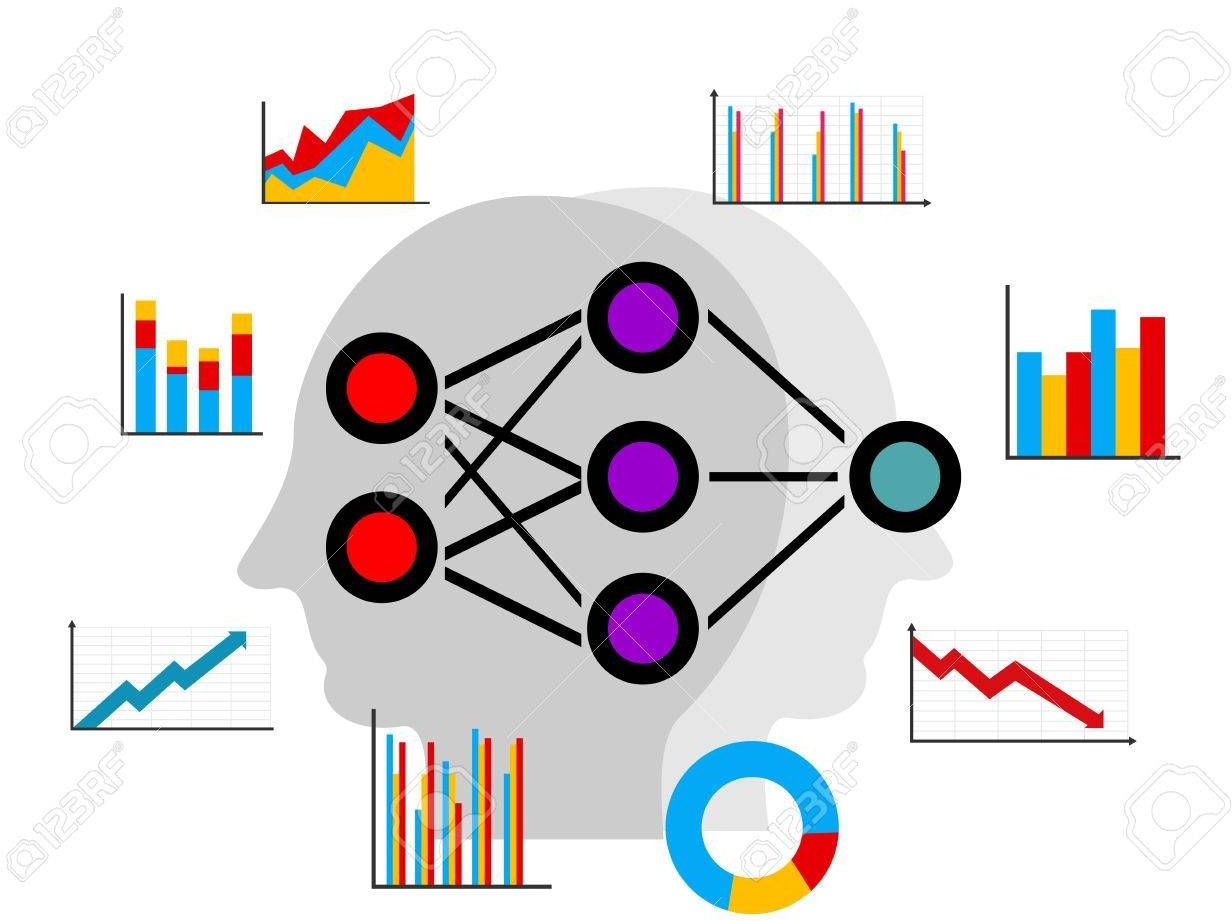
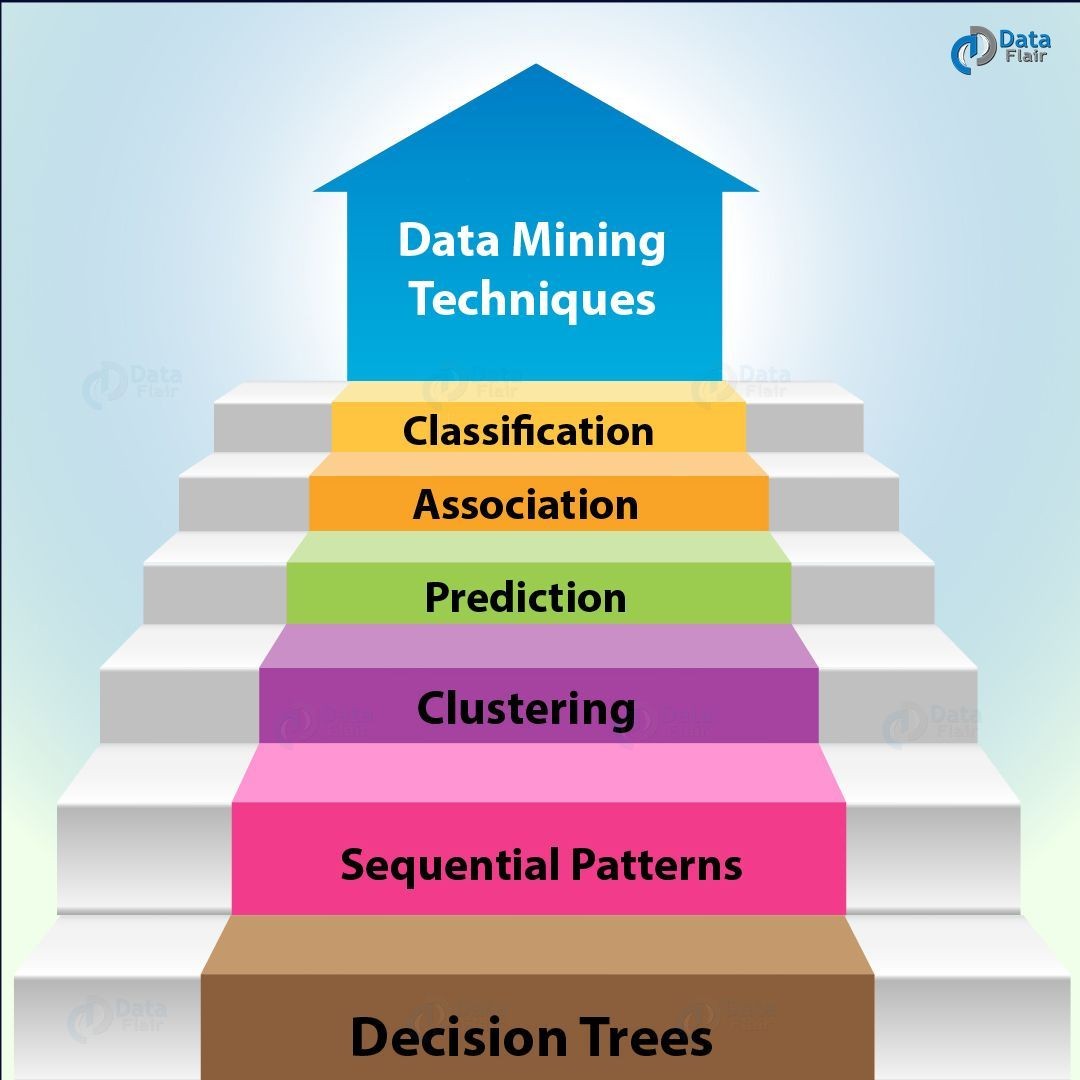
**CHAPTER 6 CLASSIFICATION AND PREDICTION**





## WHAT IS CLASSIFICATION & PREDICTION?

* There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows −
  + Classification
  + Prediction
* Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

## WHAT IS CLASSIFICATION?

* Following are the examples of cases where the data analysis task is Classification

−

* + A bank loan officer wants to analyze the data in order to know which customer (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.
* In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

## WHAT IS PREDICTION?

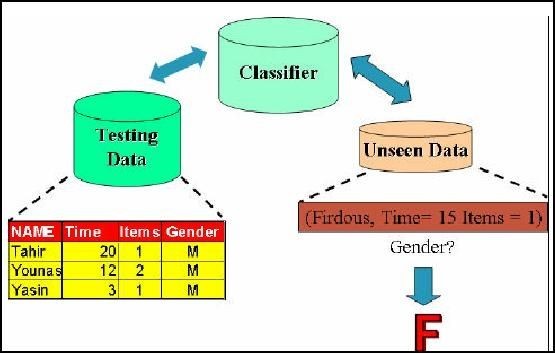
* Following are the examples of cases where the data analysis task is Prediction −
* Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.
* Note − Regression analysis is a statistical methodology that is most often used

for numeric prediction

## HOW DOES CLASSIFICATION WORKS?

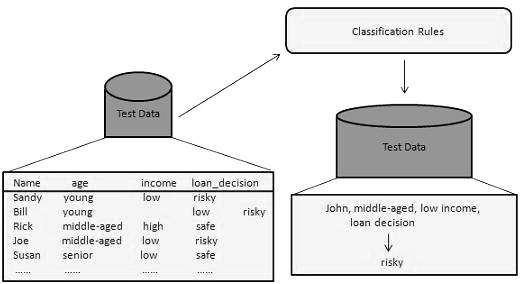
* With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps −
  + Building the Classifier or Model
  + Using Classifier for Classification
* Building the Classifier or Model
  + This step 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.
  + Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.





## 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.



## ISSUE REGARDING CLASSIFICATION & PREDICTION

* 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.

## ISSUE REGARDING CLASSIFICATION & PREDICTION

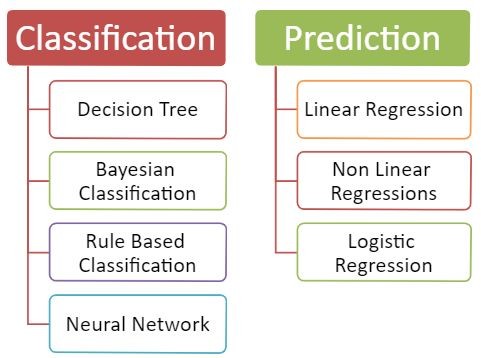
* + Data Transformation and reduction − The data can be transformed by any of the following methods.
    - Normalization − The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization 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.

COMPARISON OF CLASSIFICATION AND PREDICTION METHOD

### Here is the criteria for comparing the methods of Classification and Prediction −

* + Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
  + Speed − This refers to the computational cost in generating and using the classifier or predictor.
  + Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
  + Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.
  + Interpretability − It refers to what extent the classifier or predictor understands.

CLASSIFICATION & PREDICTION METHOD



# CLASSIFICATION METHODS

## DECISION TREE

* Decision Tree Mining is a type of data mining technique that is used to build Classification Models. It builds classification models in the form of a tree-like structure, just like its name. This type of mining belongs to supervised class learning.
* In supervised learning, the target result is

already known. Decision trees can be used for both categorical and numerical data. The categorical data represent gender, marital status, etc. while the numerical data represent age, temperature, etc.

## EXAMPLE OF DECISION TREE



WHAT IS THE USE OF DECISION TREE?

* Decision Tree is used to build classification and regression models. It is used to create data models that will predict class labels or values for the

decision-making process. The models are built from the training dataset fed to the system (supervised learning).

* Using a decision tree, we can visualize the decisions that make it easy to

understand and thus it is a popular data mining technique.

## CLASSIFICATION ANALYSIS

* Data Classification is a form of analysis which builds a model that describes important class variables. For example, a model built to categorize bank loan applications as safe or risky. Classification methods are used in machine learning, and pattern recognition.
* Application of classification includes fraud detection, medical diagnosis, target marketing, etc. The output of the classification problem is taken as “Mode” of all observed values of the terminal node.
* **A two-step process is followed, to build a classification model.**
  + In the first step i.e. learning: A classification model based on training data is built.
  + In the second step i.e. Classification, the accuracy of the model is checked and then the model is used to classify new data. The class labels presented here are in the form of discrete values such as “yes” or “no”, “safe” or “risky”.

## GENERAL APPROACH TO BUILD CLASSIFICATION TREE

REGRESSION ANALYSIS

* Regression analysis is used for the prediction of numeric attributes.
* Numeric attributes are also called continuous values. A model built to predict the continuous values instead of class labels is called the regression model. The output of regression analysis is the “Mean” of all observed values of the node.

## HOW DOES A DECISION TREE WORK?

* A decision tree is a supervised learning algorithm that works for both discrete and continuous variables. It splits the dataset into subsets on the basis of the most significant attribute in the dataset. How the decision tree identifies this attribute and how this splitting is done is decided by the algorithms.
* The most significant predictor is designated as the root node, splitting is done to form sub-nodes

called decision nodes, and the nodes which do not split further are terminal or leaf nodes.

* In the decision tree, the dataset is divided into homogeneous and non-overlapping regions. It follows a top-down approach as the top region presents all the observations at a single place which splits into two or more branches that further split. This approach is also called a *greedy approach* as it only considers the current node between the worked on without focusing on the future nodes.
* The decision tree algorithms will continue running until a stop criteria such as the minimum

number of observations etc. is reached.

* Once a decision tree is built, many nodes may represent outliers or noisy data. Tree pruning method is applied to remove unwanted data. This, in turn, improves the accuracy of the classification model.
* To find the accuracy of the model, a test set consisting of test tuples and

class labels is used. The percentages of the test set tuples are correctly classified by the model to identify the accuracy of the model. If the model is found to be accurate then it is used to classify the data tuples for which the class labels are not known.

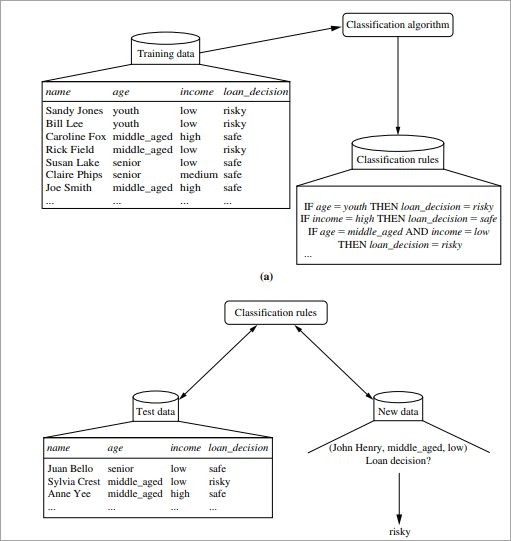
* Some of the decision tree algorithms include Hunt’s Algorithm, ID3,

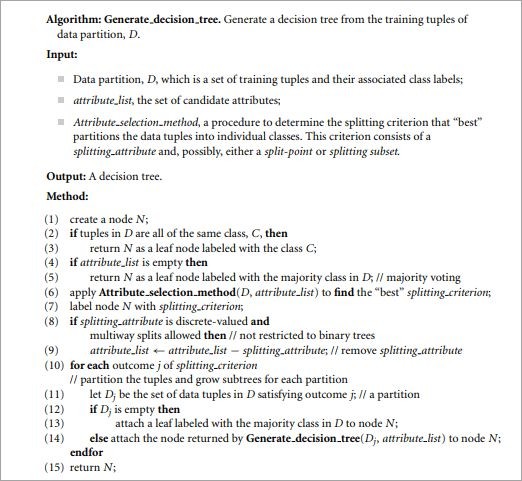
CD4.5, and CART.

## EXAMPLE OF CREATING A DECISION TREE

**#1) Learning Step:** The training data is fed into the system to be analyzed by a classification algorithm. In this example, the class label is the attribute i.e. “loan decision”. The model built from this training data is represented in the form of decision rules.

**#2) Classification:** Test dataset are fed to the model to check the accuracy of the classification rule. If the model gives acceptable results then it is applied to a new dataset with unknown class variables.

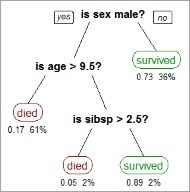




DECISION TREE INDUCTION ALGORITHM

## DECISION TREE INDUCTION

* Decision tree induction is the method of learning the decision trees from the training set. The training set consists of attributes and class labels. Applications of decision tree induction include astronomy, financial analysis, medical diagnosis, manufacturing, and production.
* A decision tree is a flowchart tree-like structure that is made from training set tuples. The dataset is broken down into smaller subsets and is present in the form of nodes of a tree. The tree structure has a root node, internal nodes or decision nodes, leaf node, and branches.
* The root node is the topmost node. It represents the best attribute selected for classification. Internal nodes of the decision nodes represent a test of an attribute of the dataset leaf node or terminal node which represents the classification or decision label. The branches show the outcome of the test performed.
* Some decision trees only have *binary nodes*, that means exactly two branches of a node, while some decision trees are non-binary.



**The image above shows the decision tree for the Titanic dataset to predict whether the passenger will survive or not.**

## CART

* CART model i.e. Classification and Regression Models is a decision tree algorithm for building models. Decision Tree model where the target values have a discrete nature is called classification models.
* A discrete value is a finite or countably infinite set of values, **For Example,** age, size, etc. The models where the target values are represented by continuous values are usually numbers that are called Regression Models. Continuous variables are floating-point variables. These two models together are called CART.
* CART uses Gini Index as Classification matrix.

## DECISION TREE INDUCTION FOR ID3

* In the late 1970s and early 1980s, J.Ross Quinlan was a researcher who built a decision tree algorithm for machine learning. This algorithm is known as **ID3, Iterative Dichotomiser**. This algorithm was an extension of the concept learning systems described by E.B Hunt, J, and Marin.
* ID3 later came to be known as C4.5. ID3 and C4.5 follow a greedy top-down

approach for constructing decision trees. The algorithm starts with a training dataset with class labels that are portioned into smaller subsets as the tree is being constructed.

## DECISION TREE INDUCTION FOR ID3

**#1)** Initially, there are three parameters i.e. **attribute list, attribute selection method and data partition**. The attribute list describes the attributes of the training set tuples.

**#2)** The attribute selection method describes the method for selecting the best attribute for discrimination among tuples. The methods used for attribute selection can either be Information Gain or Gini Index.

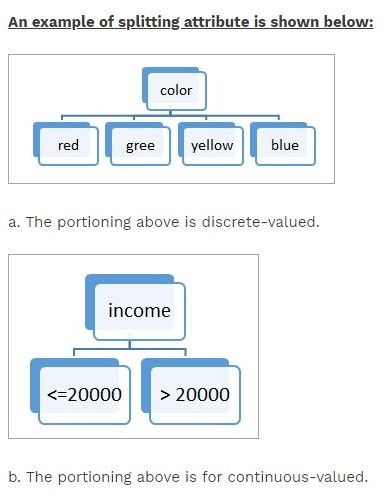
**#3)** The structure of the tree (binary or non-binary) is decided by the attribute selection method.

**#4)** When constructing a decision tree, it starts as a single node representing the tuples.

## DECISION TREE INDUCTION FOR ID3

**#5)** If the root node tuples represent different class labels, then it calls an attribute selection method to split or partition the tuples. The step will lead to the formation of branches and decision nodes.

**#6)** The splitting method will determine which attribute should be selected to partition the data tuples. It also determines the branches to be grown from the node according to the test outcome. The main motive of the splitting criteria is that the partition at each branch of the decision tree should represent the same class label.



## DECISION TREE INDUCTION FOR ID3

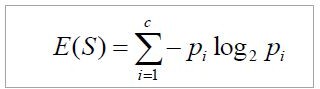
**#7)** The above partitioning steps are followed recursively to form a decision tree for the training dataset tuples.

**#8)** The portioning stops only when either all the partitions are made or when the remaining tuples cannot be partitioned further.

**#9)** The complexity of the algorithm is described by **n \* |D| \* log |D|** where n is the number of attributes in training dataset D and |D| is the number of tuples.

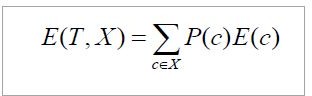
## INFORMATION GAIN

* This method is the main method that is used to build decision trees. It reduces the information that is required to classify the tuples. It reduces the number of tests that are needed to classify the given tuple. The attribute with the highest information gain is selected.
* The original information needed for classification of a tuple in dataset D is

given by:

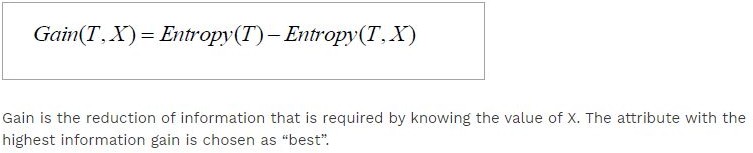
Where p is the probability that the tuple belongs to class C. The information is encoded in bits, therefore, log to the base 2 is used. E(s) represents the average amount of information required to find out the class label of dataset D. This information gain is also called **Entropy**.

The information required for exact classification after portioning is given by the formula:



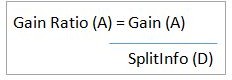
Where P (c) is the weight of partition. This information represents the information needed to classify the dataset D on portioning by X.

Information gain is the difference between the original and expected information that is required to classify the tuples of dataset D.

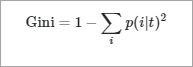


## GAIN RATIO

Information gain might sometimes result in portioning useless for classification. However, the Gain ratio splits the training data set into partitions and considers the number of tuples of the outcome with respect to the total tuples. The attribute with the max gain ratio is used as a splitting attribute.



## GINI INDEX

Gini Index is calculated for binary variables only. It measures the impurity in training tuples of dataset D, as

P is the probability that tuple belongs to class C. The Gini index that is calculated for binary split dataset D by attribute A is given by:

## WHAT IS TREE PRUNING?

* Pruning is the method of removing the unused branches from the decision tree. Some branches of the decision tree might represent outliers or noisy data.
* Tree pruning is the method to reduce the unwanted branches of the tree. This will reduce the complexity of the tree and help in effective predictive analysis. It reduces the overfitting as it removes the unimportant branches from the trees.

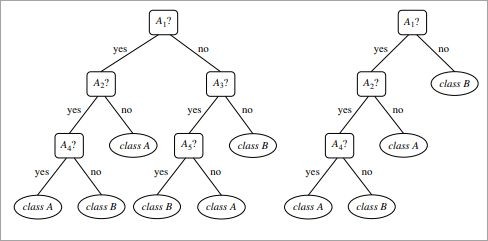
**#1) Pre Pruning**: In this approach, the construction of the decision tree is stopped early. It means it is decided not to further partition the branches. The last node constructed becomes the leaf node and this leaf node may hold the most frequent class among the tuples.

The attribute selection measures are used to find out the weightage of the split. Threshold values are prescribed to decide which splits are regarded as useful. If the portioning of the node results in splitting by falling below threshold then the process is halted.

**#2) Post Pruning**: This method removes the outlier branches from a fully grown tree. The unwanted branches are removed and replaced by a leaf node denoting the most frequent class label. This technique requires more computation than prepruning, however, it is more reliable.

The pruned trees are more precise and compact when compared to unpruned trees but they carry a disadvantage of replication and repetition.

Repetition occurs when the same attribute is tested again and again along a branch of a tree. *Replication* occurs when the duplicate subtrees are present within the tree. These issues can be solved by multivariate splits.

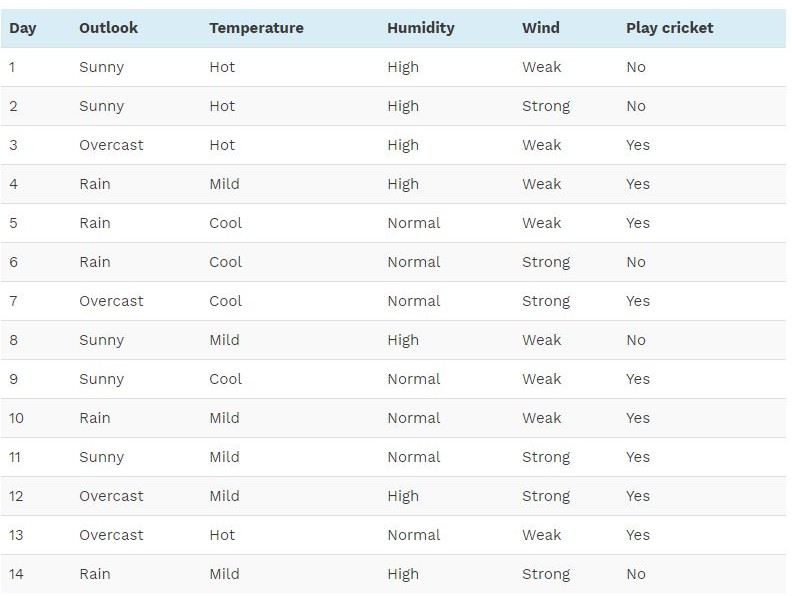


**The above image shows an unpruned and pruned tree.**

## EXAMPLE OF DECISION TREE

**Constructing a Decision Tree**

Let us take an example of the last 10 days weather dataset with attributes outlook, temperature, wind, and humidity. The outcome variable will be playing cricket or not. We will use the ID3 algorithm to build the decision tree.



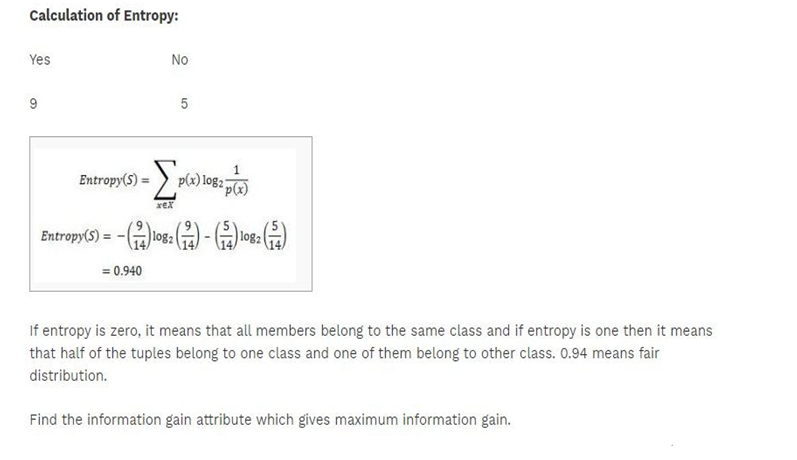
**Step1:** The first step will be to create a root node.

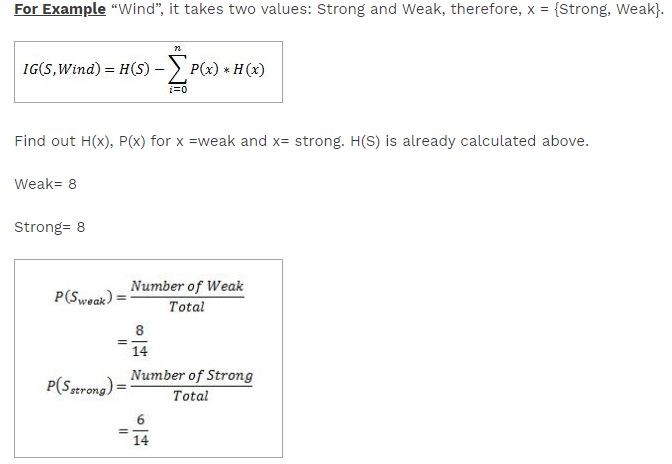
**Step2:** If all results are yes, then the leaf node “yes” will be returned else the leaf node “no” will be returned.

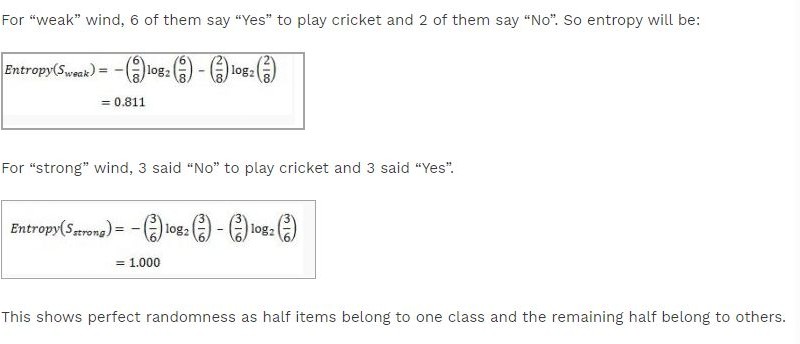
**Step3:** Find out the Entropy of all observations and entropy with attribute “x” that is E(S) and E(S, x).

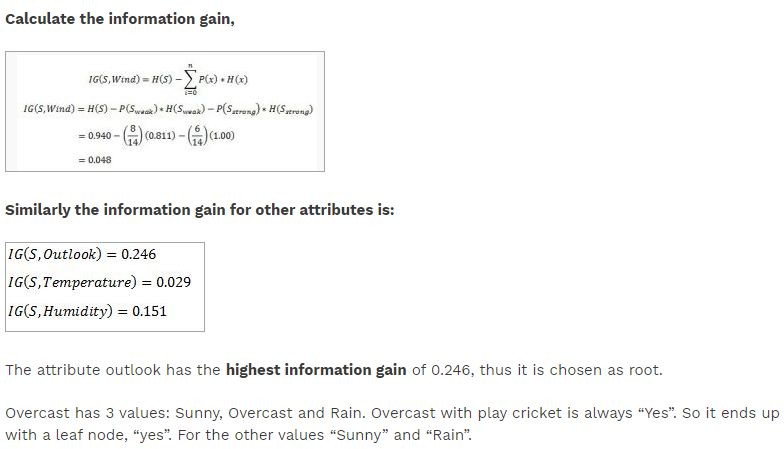
**Step4:** Find out the information gain and select the attribute with high information gain.

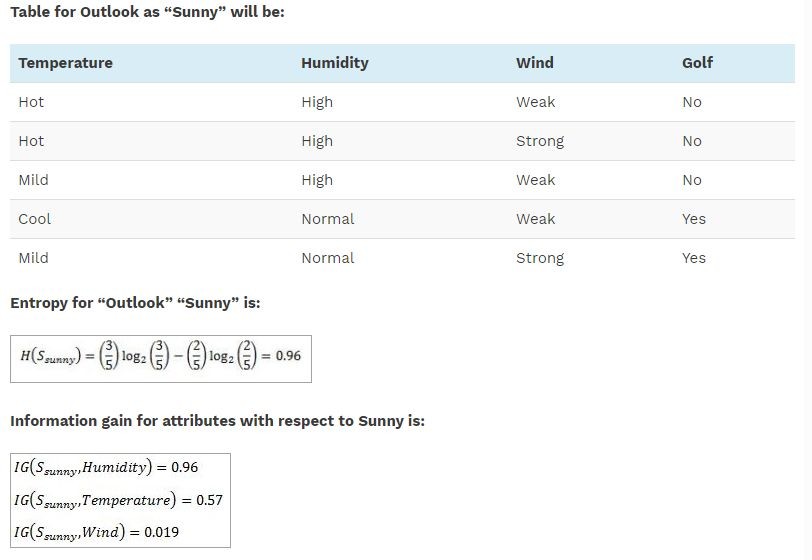
**Step5:** Repeat the above steps until all attributes are covered.

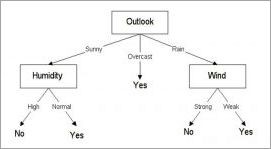












Advantages Of Decision Tree Classification

Enlisted below are the various merits of Decision Tree Classification:

1. Decision tree classification does not require any domain knowledge, hence, it is appropriate for the knowledge discovery process.
2. The representation of data in the form of the tree is easily understood by humans and it is intuitive.
3. It can handle multidimensional data.
4. It is a quick process with great accuracy.

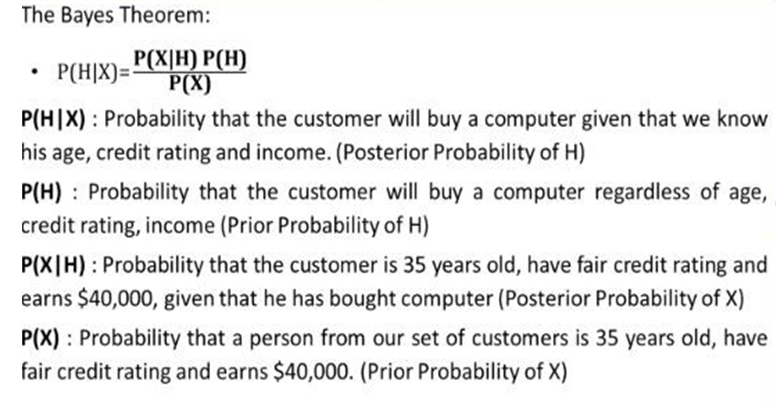
Disadvantages Of Decision Tree Classification

Given below are the various demerits of Decision Tree Classification:

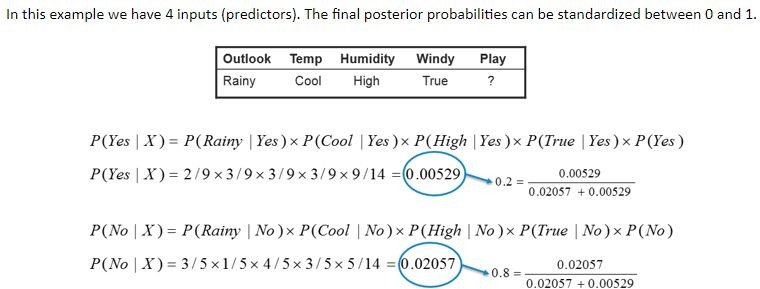
1. Sometimes decision trees become very complex and these are called overfitted trees.
2. The decision tree algorithm may not be an optimal solution.
3. The decision trees may return a biased solution if some class label dominates it.

## BAYESIAN CLASSIFICATION

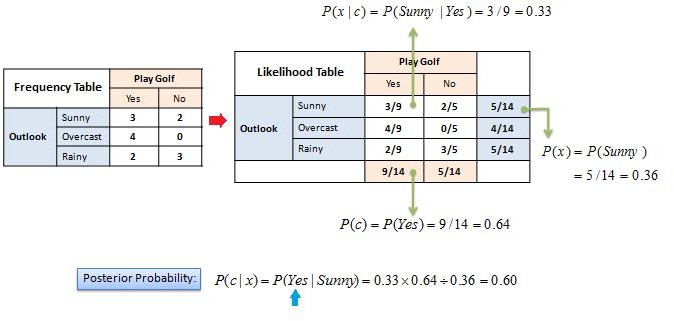
* Thomas Bayes, who proposed the Bayes Theorem so, it named Bayesian theorem.
* It is statistical method & supervised learning method for classification.
* **It can solve problems involving both categorical and continuous valued attributes**.
* Bayesian classification is used to find conditional probabilities.

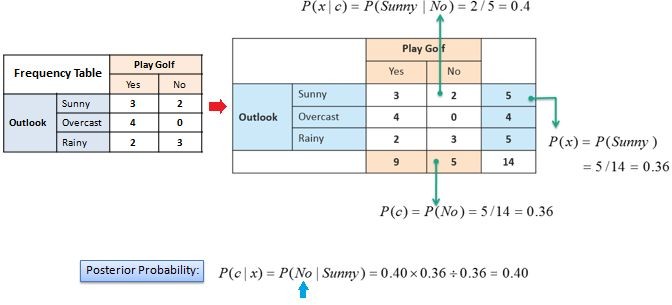


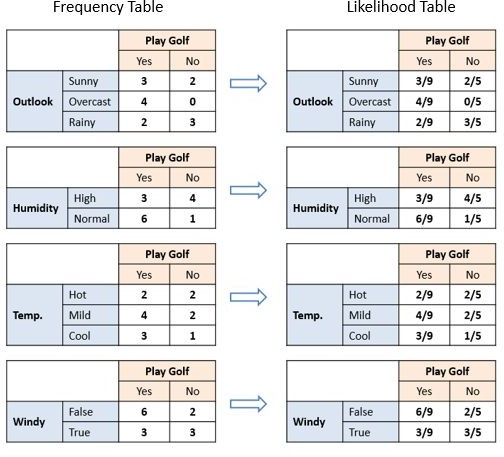
## NAIVE BAYES CLASSIFIER EXAMPLE











## RULE BASED

It is featured by building rules based on an object attributes.

Rule-based classifier makes use of a set of **IF-THEN rules** for classification. We can express a rule in the following from

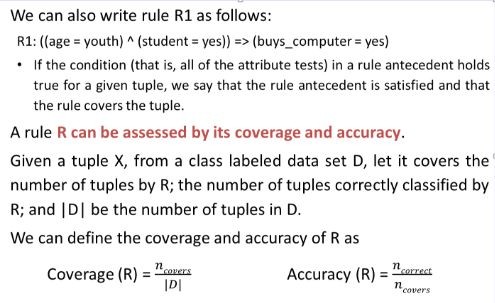
IF condition THEN conclusion Let us consider a rule R1,

R1: **IF age=youth AND student=yes THEN buy\_computer=yes**

The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent (conclusion).

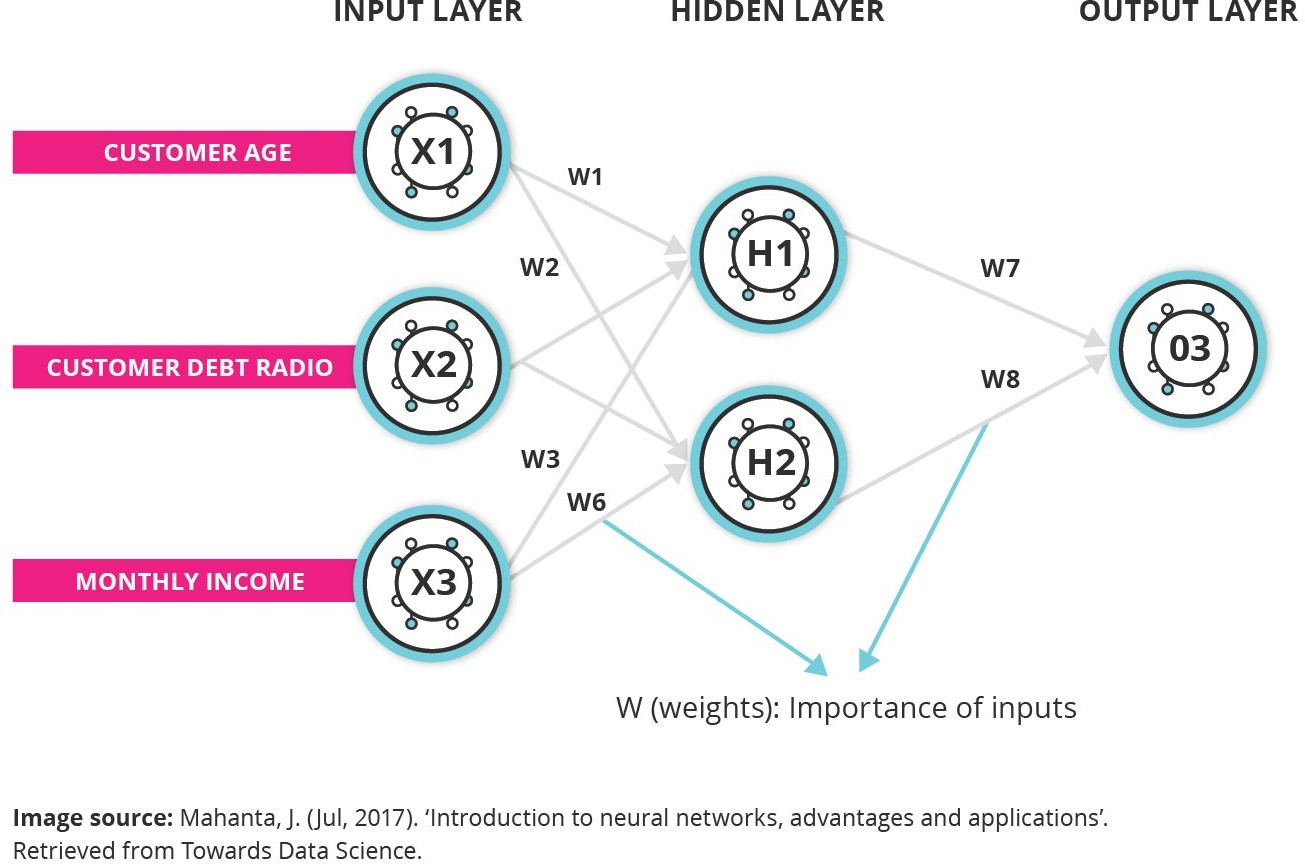
The antecedent (IF) part the condition consist of one or more attribute tests and these tests are logically ANDed.

The consequent (THEN) part consists of class prediction.



## NEURAL NETWORK

*The Artificial Neural Network (ANN) bases its assimilation of data on the way that the human brain processes information. The brain has billions of cells called neurons that process information in the form of electric signals. External information, or stimuli, is received, after which the brain processes it, and then produces an output.*



* In Professor Arackioraj’s paper, “Applications of neural networks in data mining”, he notes that finding information that is hidden in data is as difficult as it is important.22 Neural networks mine data in areas such as bioinformatics, banking, and retail. Using neural networks, data warehousing organisations can harvest information from datasets to help users make more informed decisions through neural network’s ability to handle complex relationships, cross-pollination of data, and machine learning. Neural networks and [AI technologies](https://www.getsmarter.com/blog/market-trends/the-future-of-ai-4-industries-that-will-be-most-affected/) can carry out many business purposes with unstructured data, from tracking and documenting real-time communications, to finding new customers that automate follow-ups and flag warm leads.23
* Until recently, decision-makers had to rely primarily on extracted data from structured, highly

organised data sets, as these are easier to analyse. Unstructured data like emails and copy, are more difficult to analyse, and so have gone unutilised or simply ignored. Neural networks can now provide decision-makers with much deeper insight into the ‘why’ of a customer’s behaviour, which goes beyond what is provided in more structured data.24

* In healthcare, an example of how neural networks are successfully mining data is shown by Imperial

College London, where ANNs are used to produce optimal patient care recommendations for patients with sepsis.

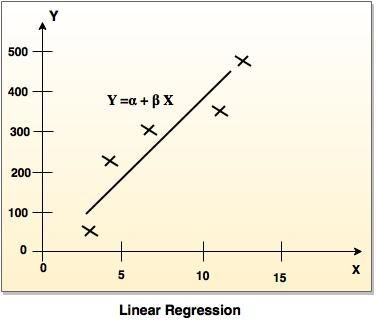
# PREDICTION METHODS

## LINEAR AND NONLINEAR 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.

**Y = α + Β 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.



## MULTIPLE LINEAR REGRESSION

* 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.

**Y = a0 + a1 X1 + a2 X2 +. +ak Xk +e**

**where,**

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

**X1 + X2 + Xk** are the independent predictors.

**'e'** is random error.

**a0, a1, a2, ak** are the regression coefficients.

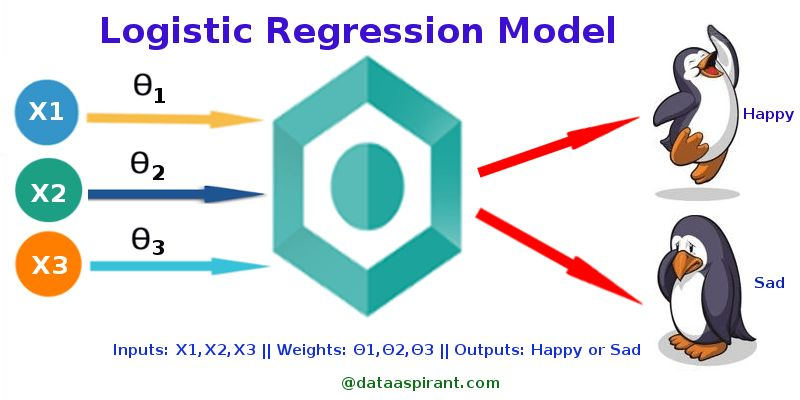
## LOGISTIC REGRESSION

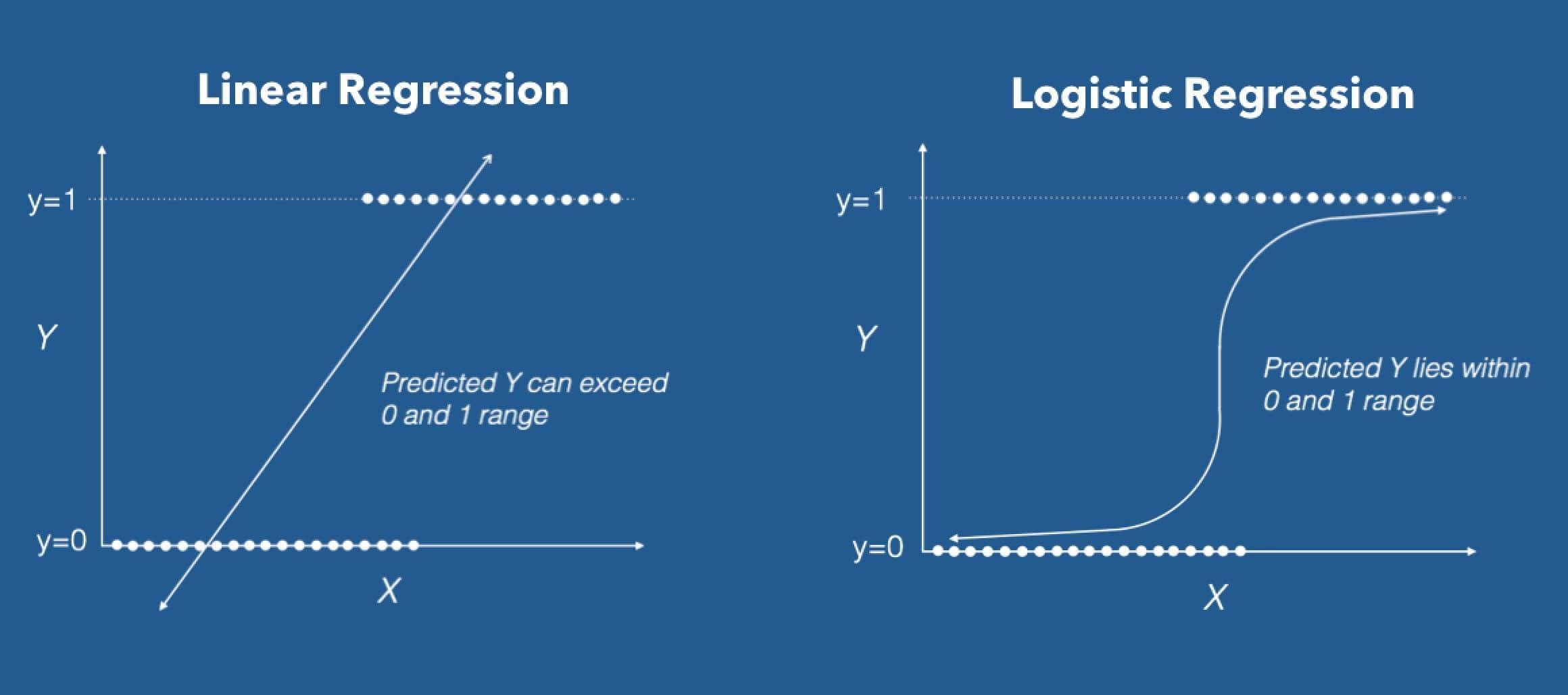
* Logistic Regression was used in the biological sciences in early twentieth century. It was then

used in many social science applications. Logistic Regression is used when the dependent

variable(target) is categorical.

* For example,
  + To predict whether an email is spam (1) or (0)
  + Whether the tumor is malignant (1) or not (0)



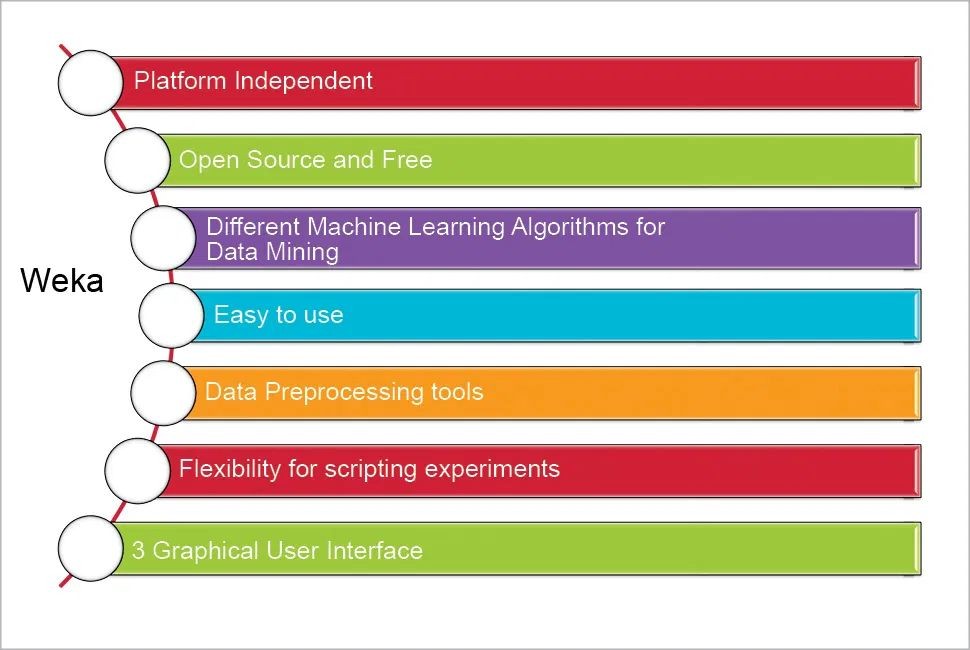


# INTRODUCTION OF TOOLS





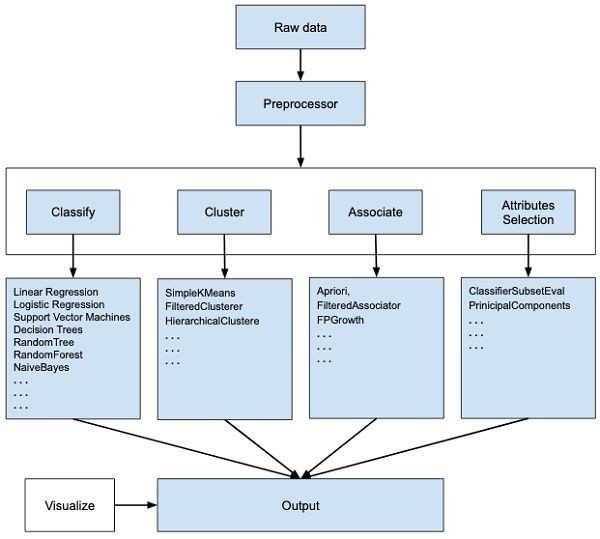
## WEKA



WEKA

* WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to

real-world data mining problems. What WEKA offers is summarized in the following diagram −



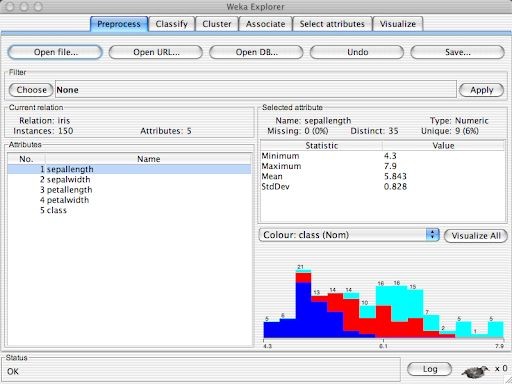
## WEKA

### If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning −

* First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

### Then, you would save the preprocessed data in your local storage for applying ML algorithms.

* Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as Classify, Cluster, or Associate. The Attributes Selection allows the automatic selection of features to create a reduced dataset.



## DBMINER

* A data mining system, DBMiner, has been developed for interactive mining of multiple-level knowledge in large relational databases. The system implements a wide spectrum of data mining functions, including generalization, characterization, association, classification, and prediction.
* A data mining system, DBMiner, has been developed for interactive mining of multiple-level

knowledge in large relational databases and data warehouses. The system implements a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering. By incorporating several interesting data mining techniques, including OLAP and attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta-rule guided mining, the system provides a user-friendly, interactive data mining environment with good performance.

