Classification and Predication in Data Mining

Edit by **Dr. Khawla Hussein Ali**

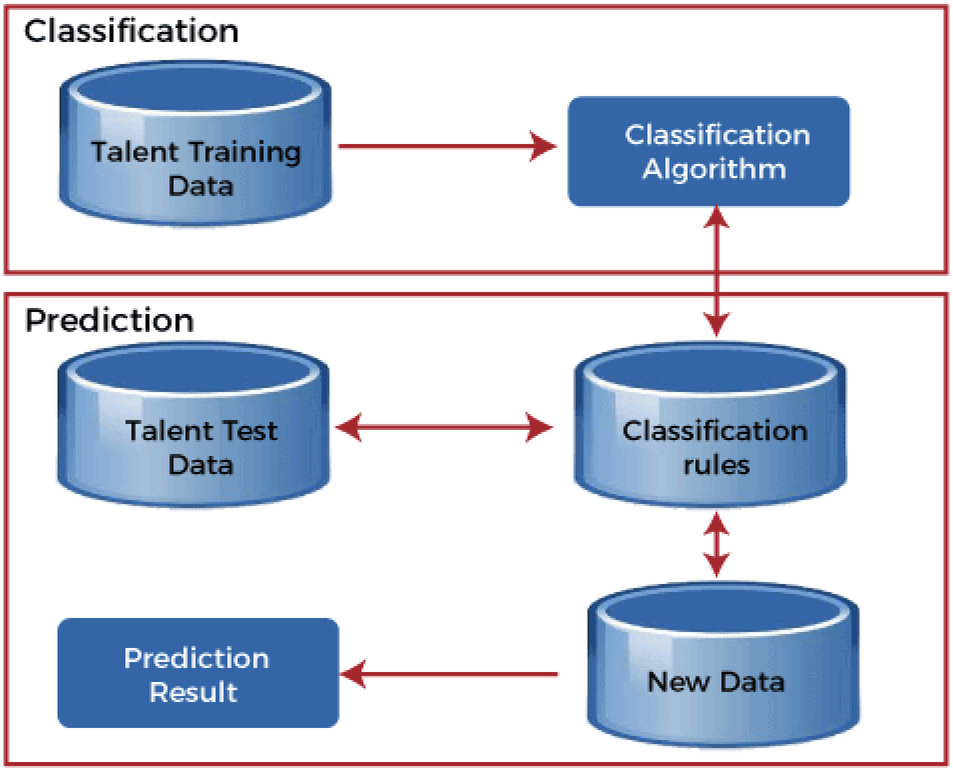
Dec 13, 2022

There are two forms of data analysis that can be used to extract models describing important classes or predict future data trends. These two forms are as follows:

1. Classification
2. Prediction

We use classification and prediction to extract a model, representing the data classes to predict future data trends. Classification predicts the categorical labels of data with the prediction models. This analysis provides us with the best understanding of the data at a large scale.

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.



Classification and Prediction Process

What is Classification?

Classification is to identify the category or the class label of a new observation. First, a set of data is used as training data. The set of input data and the corresponding outputs are given to the algorithm. So, the training data set includes the input data and their associated class labels. Using the training dataset, the algorithm derives a model or the classifier. The derived model can be a decision tree, mathematical formula, or a neural network. In classification, when unlabeled data is given to the model, it should find the class to which it belongs. The new data provided to the model is the test data set.

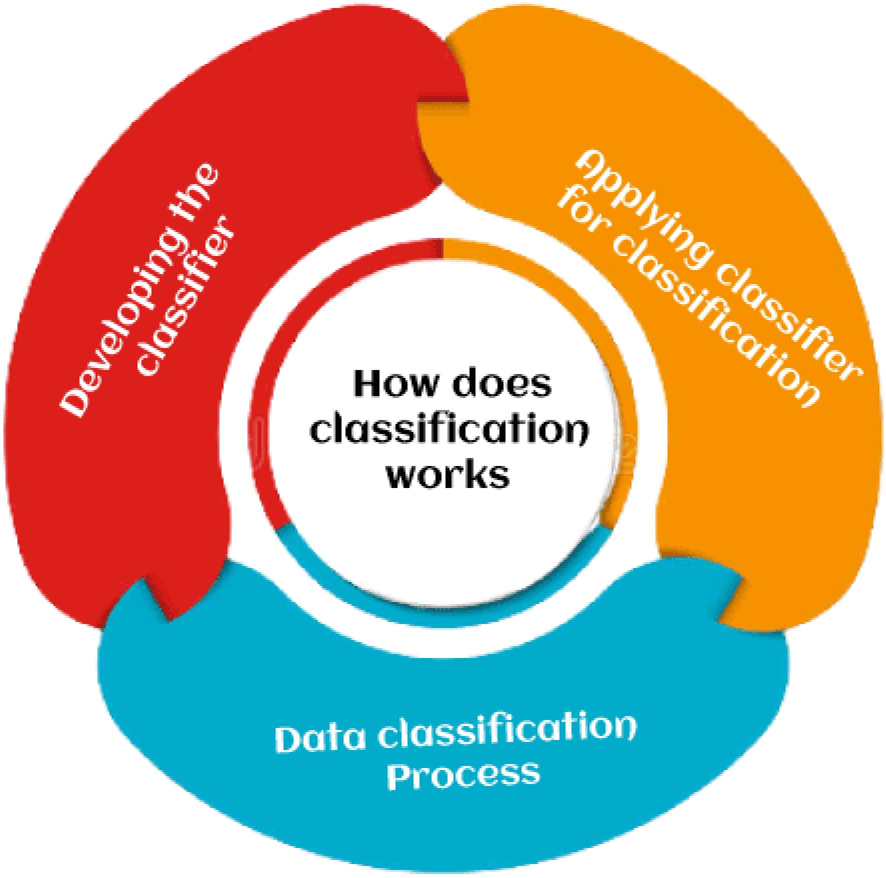
[redacted]

Classification is the process of classifying a record. One simple example of classification is to check whether it is raining or not. The answer can either be yes or no. So, there is a particular number of choices. Sometimes there can be more than two classes to classify. That is called ***multiclass classification***.

The bank needs to analyze whether giving a loan to a particular customer is risky or not. **For example**, based on observable data for multiple loan borrowers, a classification model may be established that forecasts credit risk. The data could track job records, homeownership or leasing, years of residency, number, type of deposits, historical credit ranking, etc. The goal would be credit ranking, the predictors would be the other characteristics, and the data would represent a case for each consumer. In this example, a model is constructed to find the categorical label. The labels are risky or safe.

How does Classification Works?

The functioning of classification with the assistance of the bank loan application has been mentioned above. There are two stages in the data classification system: classifier or model creation and classification classifier.

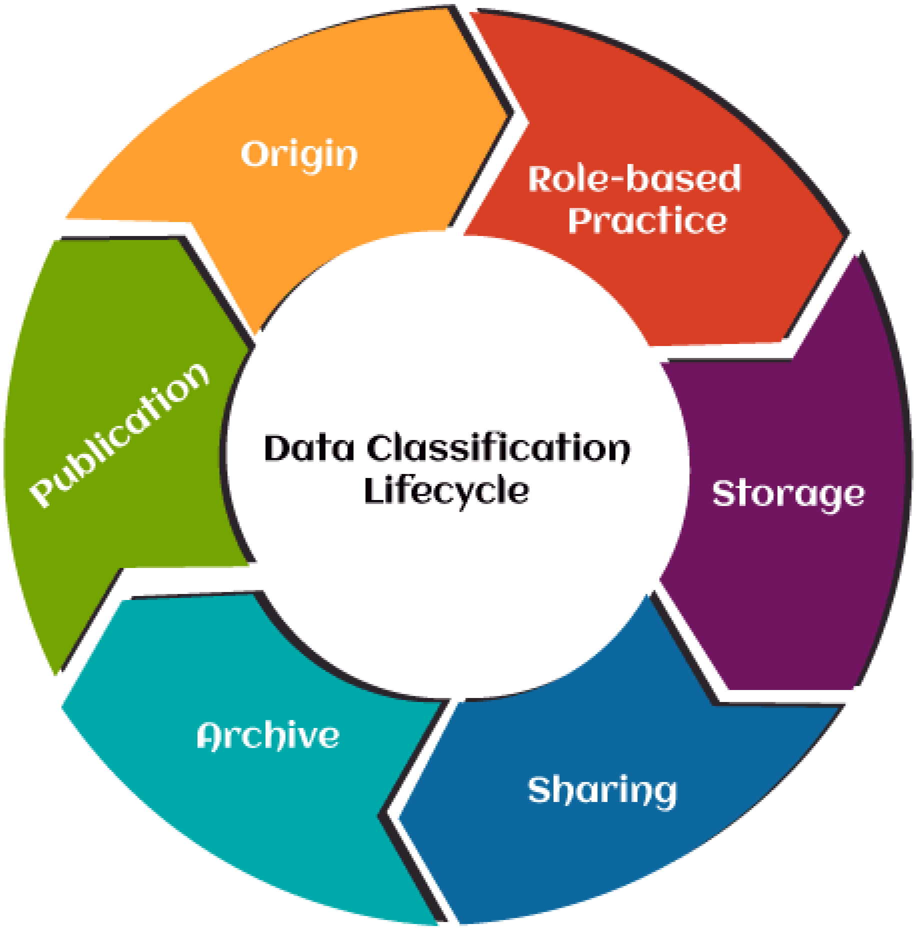

How does classification works
Developing the classifier
Applying classifier for classification
Data classification Process

1. **Developing the Classifier or model creation:** This level is the learning stage or the learning process. The classification algorithms construct the classifier in this stage. A classifier is constructed from a training set composed of the records of databases and their corresponding class names. Each category that makes up the training set is referred to as a category or class. We may also refer to these records as samples, objects, or data points.

2. **Applying classifier for classification:** The classifier is used for classification at this level. The test data are used here to estimate the accuracy of the classification algorithm. If the consistency is deemed sufficient, the classification rules can be expanded to cover new data records. It includes:
    - **Sentiment Analysis:** Sentiment analysis is highly helpful in social media monitoring. We can use it to extract social media insights. We can build sentiment analysis models to read and analyze misspelled words with advanced machine learning algorithms. The accurate trained models provide consistently accurate outcomes and result in a fraction of the time.
    - **Document Classification:** We can use document classification to organize the documents into sections according to the content. Document classification refers to text classification; we can classify the words in the entire document. And with the help of machine learning classification algorithms, we can execute it automatically.
    - **Image Classification:** Image classification is used for the trained categories of an image. These could be the caption of the image, a statistical value, a theme. You can tag images to train your model for relevant categories by applying supervised learning algorithms.
    - **Machine Learning Classification:** It uses the statistically demonstrable algorithm rules to execute analytical tasks that would take humans hundreds of more hours to perform.
3. **Data Classification Process:** The data classification process can be categorized into five steps:
    - Create the goals of data classification, strategy, workflows, and architecture of data classification.
    - Classify confidential details that we store.
    - Using marks by data labelling.
    - To improve protection and obedience, use effects.
    - Data is complex, and a continuous method is a classification.

What is Data Classification Lifecycle?

The data classification life cycle produces an excellent structure for controlling the flow of data to an enterprise. Businesses need to account for data security and compliance at each level. With the help of data classification, we can perform it at every stage, from origin to deletion. The data life-cycle has the following stages, such as:



1. **Origin:** It produces sensitive data in various formats, with emails, Excel, Word, Google documents, social media, and websites.
2. **Role-based practice:** Role-based security restrictions apply to all delicate data by tagging based on in-house protection policies and agreement rules.
3. **Storage:** Here, we have the obtained data, including access controls and encryption.
4. **Sharing:** Data is continually distributed among agents, consumers, and co-workers from various devices and platforms.
5. **Archive:** Here, data is eventually archived within an industry's storage systems.

6. **Publication:** Through the publication of data, it can reach customers. They can then view and download in the form of dashboards.

What is Prediction?

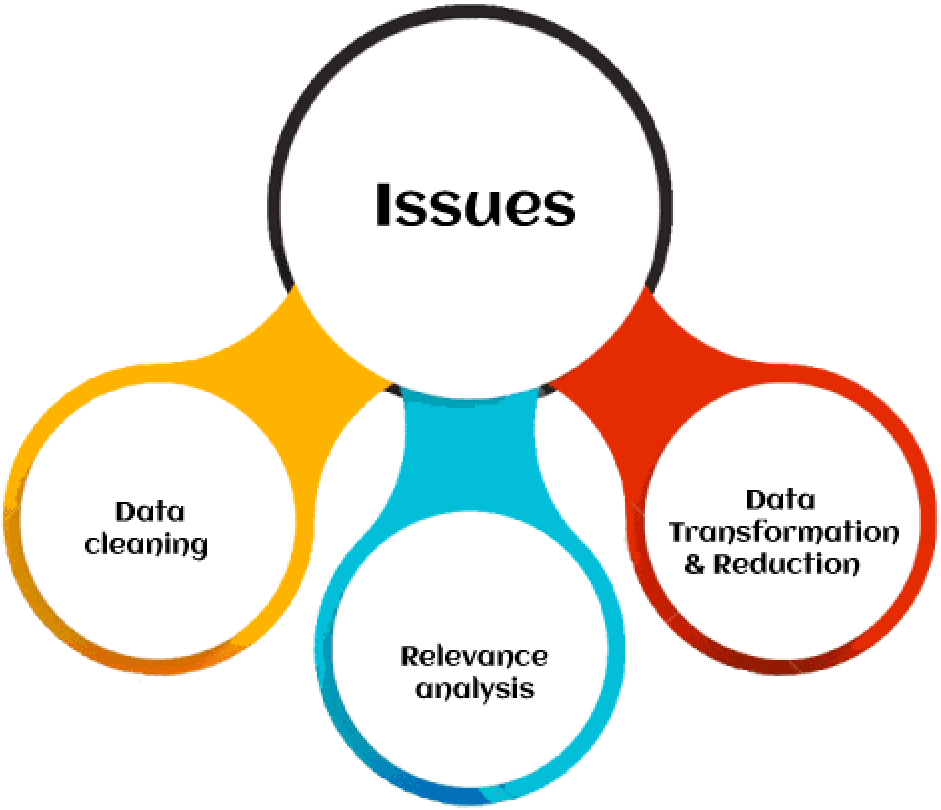
Another process of data analysis is prediction. It is used to find a numerical output. Same as in classification, the training dataset contains the inputs and corresponding numerical output values. The algorithm derives the model or a predictor according to the training dataset. The model should find a numerical output when the new data is given. Unlike in classification, this method does not have a class label. The model predicts a continuous-valued function or ordered value.

Regression is generally used for prediction. Predicting the value of a house depending on the facts such as the number of rooms, the total area, etc., is an example for prediction.

For example, suppose the marketing manager needs to predict how much a particular customer will spend at his company during a sale. We are bothered to forecast a numerical value in this case. Therefore, an example of numeric prediction is the data processing activity. In this case, a model or a predictor will be developed that forecasts a continuous or ordered value function.

Classification and Prediction Issues

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities, such as:

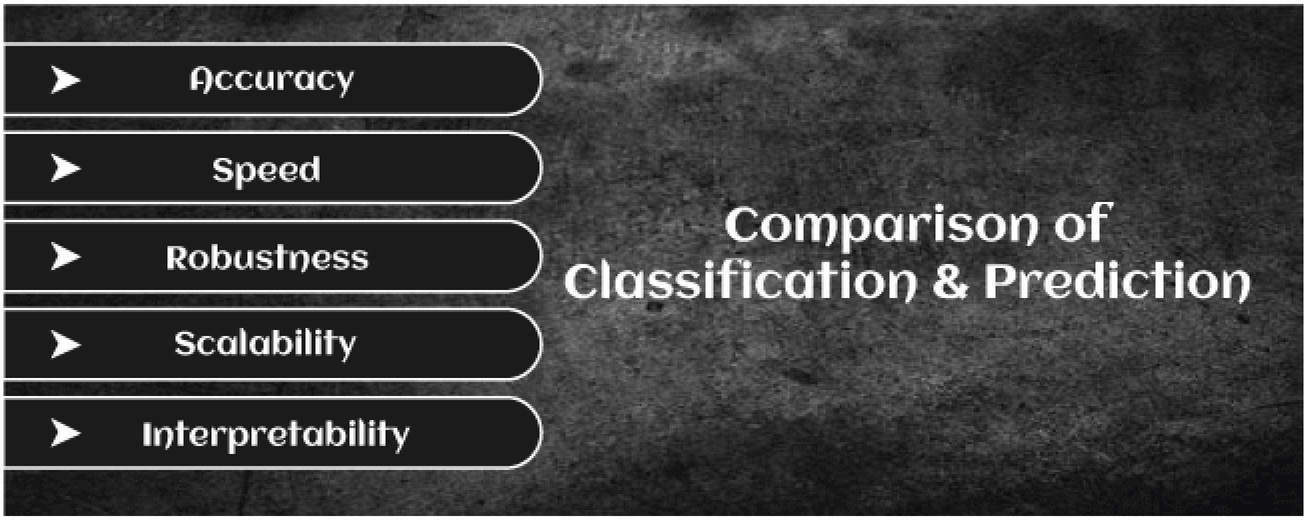


1. **Data Cleaning:** Data cleaning involves removing the noise and treatment of missing values. T noise is removed by applying smoothing techniques, and the problem of missing values is solv by replacing a missing value with the most commonly occurring value for that attribute.
2. **Relevance Analysis:** The database may also have irrelevant attributes. Correlation analysis is us to know whether any two given attributes are related.
3. **Data Transformation and reduction:** The data can be transformed by any of the followi methods.
   - **Normalization:** The data is transformed using normalization. Normalization invol scaling all values for a given attribute to make them fall within a small specified ran Normalization is used when the neural networks or the methods involving measureme are used in the learning step.
   - **Generalization:** The data can also be transformed by generalizing it to the higher conce For this purpose, we can use the concept hierarchies.

NOTE: Data can also be reduced by some other methods such as wavelet transformation, binning, histogra analysis, and clustering.

Comparison of Classification and Prediction Methods

Here are the criteria for comparing the methods of Classification and Prediction, such as:



Accuracy: The accuracy of the classifier can be referred to as the ability of the classifier to predict t class label correctly, and the accuracy of the predictor can be referred to as how well a given predict can estimate the unknown value.

Speed: The speed of the method depends on the computational cost of generating and using t classifier or predictor.

Robustness: Robustness is the ability to make correct predictions or classifications. In the context of da mining, robustness is the ability of the classifier or predictor to make correct predictions from incomi unknown data.

Scalability: Scalability refers to an increase or decrease in the performance of the classifier or predict based on the given data.

Interpretability: Interpretability is how readily we can understand the reasoning behind predictions classification made by the predictor or classifier.

Difference between Classification and Prediction

The decision tree, applied to existing data, is a classification model. We can get a class prediction applying it to new data for which the class is unknown. The assumption is that the new data comes fro a distribution similar to the data we used to construct our decision tree. In many instances, this is a corr assumption, so we can use the decision tree to build a predictive model. Classification of prediction is t

Classification	Prediction
Classification is the process of identifying which category a new observation belongs to based on a training data set containing observations whose category membership is known.	Predication is the process of identifying the missing or unavailable numerical data for a new observation.
In classification, the accuracy depends on finding the class label correctly.	In prediction, the accuracy depends on how well a given predictor can guess the value of a predicated attribute for new data.
In classification, the model can be known as the classifier.	In prediction, the model can be known as the predictor.
A model or the classifier is constructed to find the categorical labels.	A model or a predictor will be constructed that predicts a continuous-valued function or ordered value.
For example, the grouping of patients based on their medical records can be considered a classification.	**For example**, We can think of prediction as predicting the correct treatment for a particular disease for a person.

process of finding a model that describes the classes or concepts of information. The purpose is to pred the class of objects whose class label is unknown using this model. Below are some major differenc between classification and prediction.

Classification of data mining

These are given some of the important data mining classification methods:

Logistic Regression Method

The logistic Regression Method is used to predict the response variable.

K-Nearest Neighbors Method

Classification	Clustering
Classification is a supervised learning approach where a specific label is provided to the machine to classify new observations. Here the machine needs proper testing and training for the label verification.	Clustering is an unsupervised learning approach where grouping is done on similarities basis.
Supervised learning approach.	Unsupervised learning approach.
It uses a training dataset.	It does not use a training dataset.
It uses algorithms to categorize the new data as per the observations of the training set.	It uses statistical concepts in which the data set is divided into subsets with the same features.
In classification, there are labels for training data.	In clustering, there are no labels for training data.
Its objective is to find which class a new object belongs to form the set of predefined classes.	Its objective is to group a set of objects to find whether there is any relationship between them.
It is more complex as compared to clustering.	It is less complex as compared to clustering.

B

Data Mining Bayesian Classifiers

In numerous applications, the connection between the attribute set and the class variable is non- deterministic. In other words, we can say the class label of a test record can't be assumed with certainty even though its attribute set is the same as some of the training examples. These circumstances may emerge due to the noisy data or the presence of certain confusing factors that influence classification, but it is not included in the analysis. For example, consider the task of predicting the occurrence of whether an individual is at risk for liver illness based on individuals eating habits and working efficiency. Although most people who eat healthy and exercise consistently having less probability of occurrence of liver disease, they may still do so due to other factors. For example, due to consumption of the high-calorie street foods and alcohol abuse. Determining whether an individual's eating routine is healthy or the workout efficiency is sufficient is also subject to analysis, which in turn may introduce vulnerabilities into the leaning issue.

Bayesian classification uses Bayes theorem to predict the occurrence of any event. Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings. The theory expresses how a level of belief, expressed as a probability.

Bayes theorem came into existence after Thomas Bayes, who first utilized conditional probability to provide an algorithm that uses evidence to calculate limits on an unknown parameter.

Bayes's theorem is expressed mathematically by the following equation that is given below.

[redacted]

$$P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}$$

Where X and Y are the events and P (Y) ≠ 0

P(X/Y) is a **conditional probability** that describes the occurrence of event **X** is given that **Y** is true.

P(Y/X) is a **conditional probability** that describes the occurrence of event **Y** is given that **X** is true.

P(X) and P(Y) are the probabilities of observing X and Y independently of each other. This is known as the **marginal probability**.

Bayesian interpretation:

In the Bayesian interpretation, probability determines a "**degree of belief**." Bayes theorem connects the degree of belief in a hypothesis before and after accounting for evidence. For example, lets us consider an example of the coin. If we toss a coin, then we get either heads or tails, and the percent of occurrence of either heads and tails is 50%. If the coin is flipped numbers of times, and the outcomes are observed, the degree of belief may rise, fall, or remain the same depending on the outcomes.

For proposition X and evidence Y,

- P(X), the prior, is the primary degree of belief in X
- P(X/Y), the posterior is the degree of belief having accounted for Y.
- The quotient $\frac{P(Y/X)}{P(Y)}$ represents the supports Y provides for X.

Bayes theorem can be derived from the conditional probability:

Bayes theorem can be derived from the conditional probability:

$$P(X/Y) \;=\; \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) \;=\; \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

Where P (X∩Y) is the **joint probability** of both X and Y being true, because

$$P\ (Y \cap X) = P\ (X \cap Y)$$

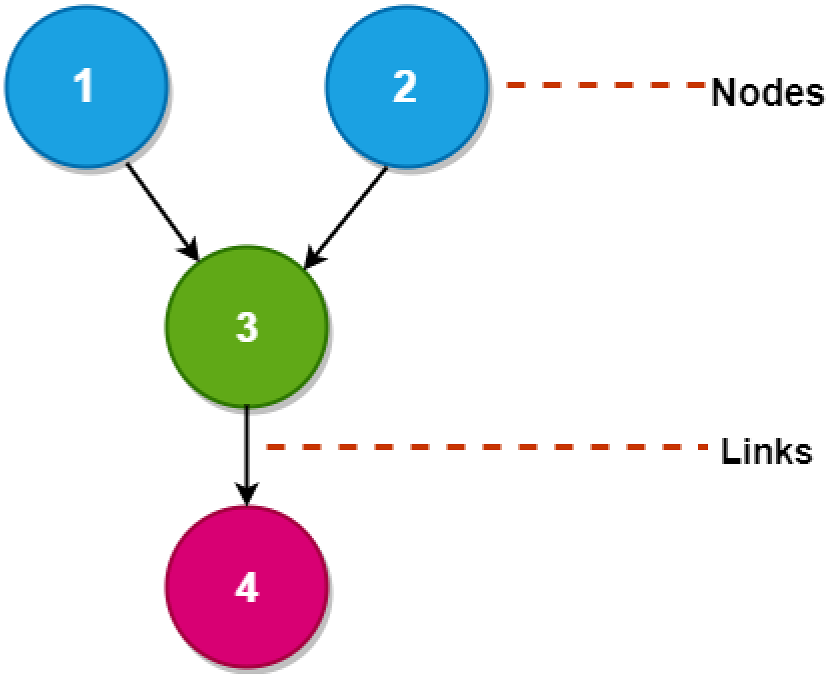
$$\text{or, } P\ (X \cap Y) = P(X/Y)P(Y) = P(Y/X)P(X)$$

$$\text{or, } P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}, \text{ if } P(Y) \neq 0$$

Bayesian network:

A Bayesian Network falls under the classification of Probabilistic Graphical Modelling (PGM) procedure that is utilized to compute uncertainties by utilizing the probability concept. Generally known as **Belief Networks, Bayesian Networks** are used to show uncertainties using **Directed Acyclic Graphs** (DAG)

A **Directed Acyclic Graph** is used to show a Bayesian Network, and like some other statistical graph, a DAG consists of a set of nodes and links, where the links signify the connection between the nodes.



The nodes here represent random variables, and the edges define the relationship between these variables.

A DAG models the uncertainty of an event taking place based on the Conditional Probability Distribution (CDP) of each random variable. A **Conditional Probability Table** (CPT) is used to represent the CPD of each variable in a network.