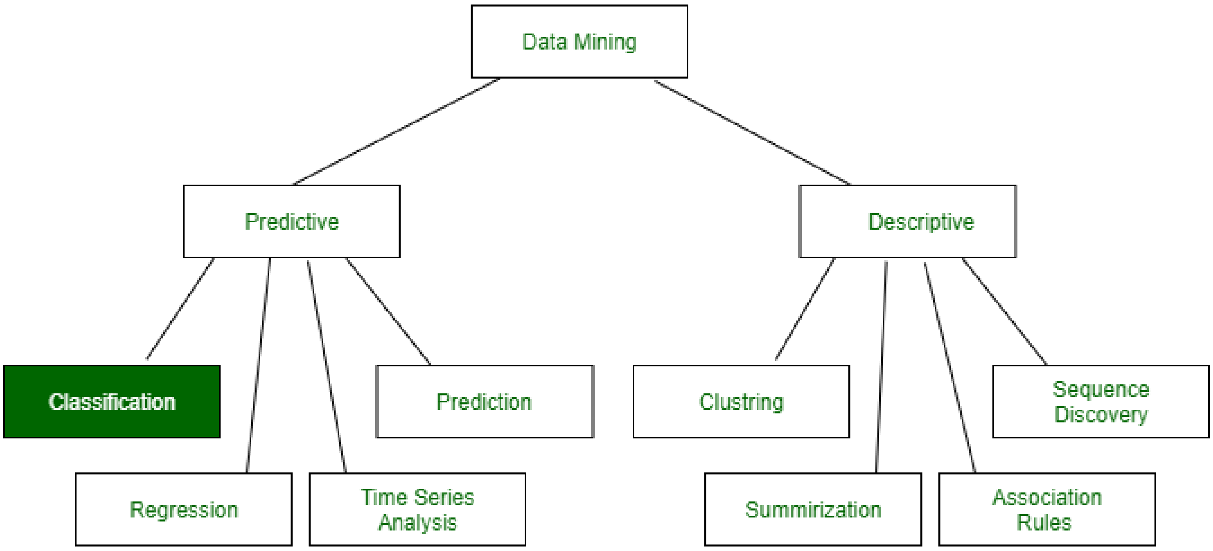
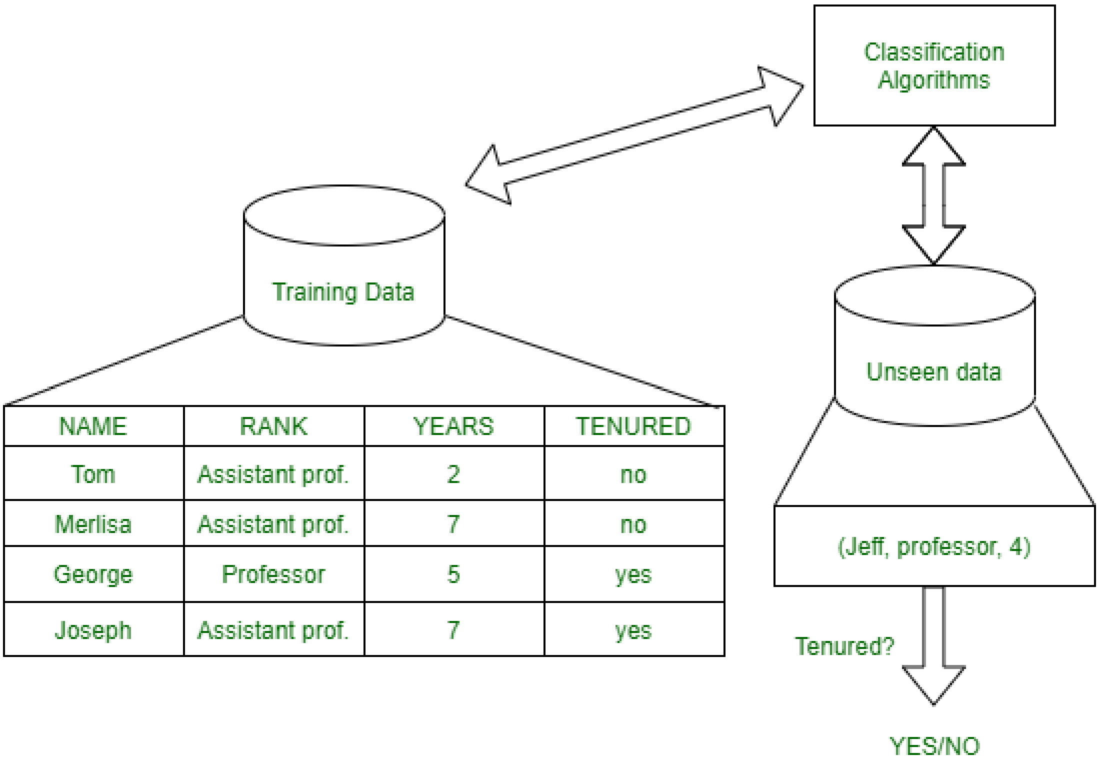
Basic Concept of Data Mining

Data Mining: Data mining in general terms means mining or digging deep into data that is in different forms to gain patterns, and to gain knowledge on that pattern. In the process of data mining, large data sets are first sorted, then patterns are identified and relationships are established to perform data analysis and solve problems.



Classification: It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.
Example: Before starting any project, we need to check its feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two-step process such as:

1. **Learning Step (Training Phase)**: Construction of Classification Model Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.
2. **Classification Step**: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.



NAME	RANK	YEARS	TENURED
Tom	Assistant prof.	2	no
Merlisa	Assistant prof.	7	no
George	Professor	5	yes
Joseph	Assistant prof.	7	yes

Test data are used to estimate the accuracy of the classification rule

Training and Testing:
Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should get aside in order not to get hurt. So, this is his training part to move away. While Testing if the person sees any heavy object coming towards him or falling on him and moves aside then the system is tested positively and if the person does not move aside then the system is negatively tested.
The same is the case with the data, it should be trained in order to get the accurate and best results.

There are certain data types associated with data mining that actually tells us the format of the file (whether it is in text format or in numerical format).

Attributes – Represents different features of an object. Different types of attributes are:

1. **Binary**: Possesses only two values i.e. True or False
   Example: Suppose there is a survey evaluating some products. We need to check whether it’s useful or not. So, the Customer has to answer it in Yes or No.
   Product usefulness: Yes / No
   - **Symmetric**: Both values are equally important in all aspects
   - **Asymmetric**: When both the values may not be important.
2. **Nominal**: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.
   Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.
   Different Colors: Red, Green, Black, Yellow

- **Ordinal**: Values that must have some meaningful order.
  Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
  Grades: A, B, C, D
- **Continuous**: May have an infinite number of values, it is in float type
  Example: Measuring the weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
  Weight: 50, 51, 52, 53
- **Discrete**: Finite number of values.
  Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90
  Marks: 65, 70, 75, 80, 90

Syntax:

- Mathematical Notation: Classification is based on building a function taking input feature vector "X" and predicting its outcome "Y" (Qualitative response taking values in set C)
- Here Classifier (or model) is used which is a Supervised function, can be designed manually based on the expert's knowledge. It has been constructed to predict class labels (Example: Label – "Yes" or "No" for the approval of some event).
-

Classifiers can be categorized into two major types:
Discriminative: It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.
Example: Logistic Regression

1. **Generative**: It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.
   Example: Naive Bayes Classifier
   Detecting Spam emails by looking at the previous data. Suppose 100 emails and that too divided in 1:4 i.e. Class A: 25%(Spam emails) and Class B: 75%(Non-Spam emails). Now if a user wants to check that if an email contains the word cheap, then that may be termed as Spam.
   It seems to be that in Class A(i.e. in 25% of data), 20 out of 25 emails are spam and rest not.
   And in Class B(i.e. in 75% of data), 70 out of 75 emails are not spam and rest are spam.
   So, if the email contains the word cheap, what is the probability of it being spam ?? (= 80%)

Classifiers Of Machine Learning:

1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour
5. Support Vector Machines
6. Linear Regression

7. Logistic Regression

Associated Tools and Languages: Used to mine/ extract useful information from raw data.

- **Main Languages used**: R, SAS, Python, SQL
- **Major Tools used**: RapidMiner, Orange, KNIME, Spark, Weka
- **Libraries used**: Jupyter, NumPy, Matplotlib, Pandas, ScikitLearn, NLTK, TensorFlow, Seaborn, Basemap, etc.

Real–Life Examples :

- **Market Basket Analysis:**
  It is a modeling technique that has been associated with frequent transactions of buying some combination of items.
  Example: Amazon and many other Retailers use this technique. While viewing some products, certain suggestions for the commodities are shown that some people have bought in the past.
- **Weather Forecasting:**
  Changing Patterns in weather conditions needs to be observed based on parameters such as temperature, humidity, wind direction. This keen observation also requires the use of previous records in order to predict it accurately.

Advantages:

- Mining Based Methods are cost-effective and efficient
- Helps in identifying criminal suspects
- Helps in predicting the risk of diseases
- Helps Banks and Financial Institutions to identify defaulters so that they may approve Cards, Loan, etc.

Disadvantages:

Privacy: When the data is either are chances that a company may give some information about their customers to other vendors or use this information for their profit.

Accuracy Problem: Selection of Accurate model must be there in order to get the best accuracy and result.

APPLICATIONS:

- Marketing and Retailing
- Manufacturing
- Telecommunication Industry
- Intrusion Detection
- Education System
- Fraud Detection

GIST OF DATA MINING :

1. Choosing the correct classification method, like decision trees, Bayesian networks, or neural networks.

2. Need a sample of data, where all class values are known. Then the data will be divided into two parts, a training set, and a test set.

Now, the training set is given to a learning algorithm, which derives a classifier. Then the classifier is tested with the test set, where all class values are hidden. If the classifier classifies most cases in the test set correctly, it can be assumed that it works accurately also on the future data else it may be the wrong model chosen.