1. Data mining:

Data mining is the process of learning existing data, scanning it for patterns and predicting outcomes for a given set of inputs.

The various techniques used to perform data mining are:

- Classification
- Clustering
- Regression
- Association rules
- Outer detection
- Sequential patterns
- Prediction

Out of these, Classification and Clustering are the most popular techniques.

1.1 Classification

Classification retrieves important information about the given data and classifies it into different classes. This data mining technique is usually used when various class labels are present.

1.1.1 Example:

The following images show the plot of an Iris dataset using a Naïve Bayes classifier with a cross validation of 10.

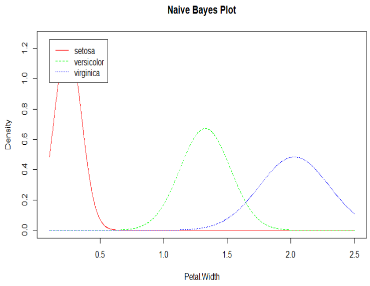

Fig: 1

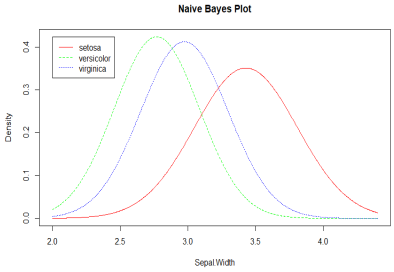

Fig: 2

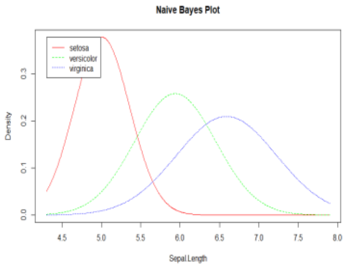

Fig: 3

1.2 Clustering

Clustering analysis data, identifies similar data and groups them together. This data mining technique is usually used when there are no specific class labels in the given data.

1.2.1 Example

The following image shows the plot of an Iris dataset using a k-means clustering method.

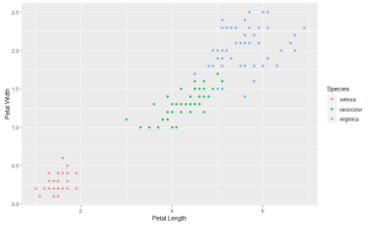

Fig: 4

2. Data mining using the Census dataset

2.1 Why data mining?

The Census data set consists of was extracted from the census bureau database, 1994. The data set consists of the following features:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-
5. acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th,
6. Doctorate, 5th-6th, Preschool.
7. education-num: continuous.
8. marital-status: Married-civ-spouse, Divorced, Never-married, Separated,
9. Widowed, Married-spouse-absent, Married-AF-spouse.
10. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm- clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
11. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
12. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
13. sex: Female, Male.
14. capital-gain: continuous.
15. capital-loss: continuous.
16. hours-per-week: continuous.
17. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

We perform data mining on this set so that when given the data about a random person, we could predict whether or not that person makes over 50K a year.

2.2 Data mining technique

Since the given data set contains binary class labels (i.e; ‘<=50K’ and ‘> 50K’) we perform the classification technique in order to predict the outcome.

There are various classification techniques. Some of them are:

- Perceptron
- Naïve Bayes
- Decision Tree

However the data set does not contain independent features. Hence to get the best accuracy we use an ensemble technique called Random Forest.

2.3 Random Forest

Random forest is an ensemble learning technique that constructs a multitude of decision trees while training and outputs the class that is the mode of classes.

2.3.1 Ensemble learning technique

Ensemble learning is a technique that uses multiple learning algorithms in order to obtain a better prediction than any of the constituent learning algorithms alone. An ensemble can be categorized as a supervised learning algorithm with a flexibility that enables them to over-fit training data more than a single model. To combat this we use a method called bagging.

2.3.1 Bagging

Bagging repeatedly selects a random sample with replacement from the training set and fits trees to these samples. After training, predictions for unseen samples can be made by averaging the predictions from all the individual trees:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

An estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}$$

2.3.2 Random forest using bagging

Random forest uses a modified tree learning algorithm that selects a random subset of features. This process is called feature bagging. For a classification problem with p features, $\sqrt{p}$ features are used in each split.