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. **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. **Example:**

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

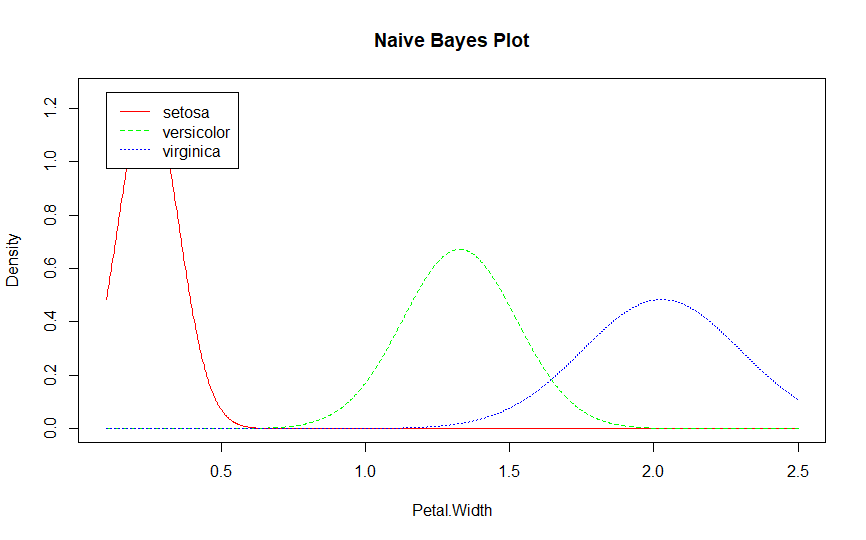
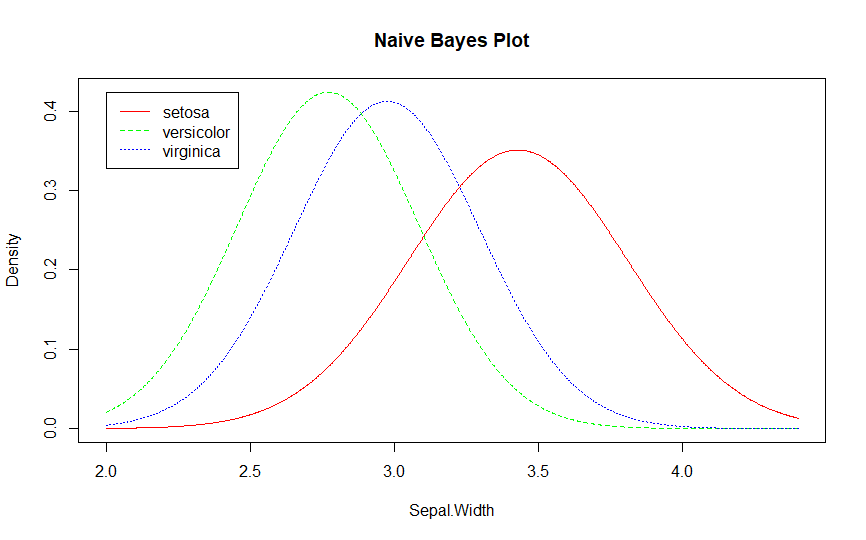
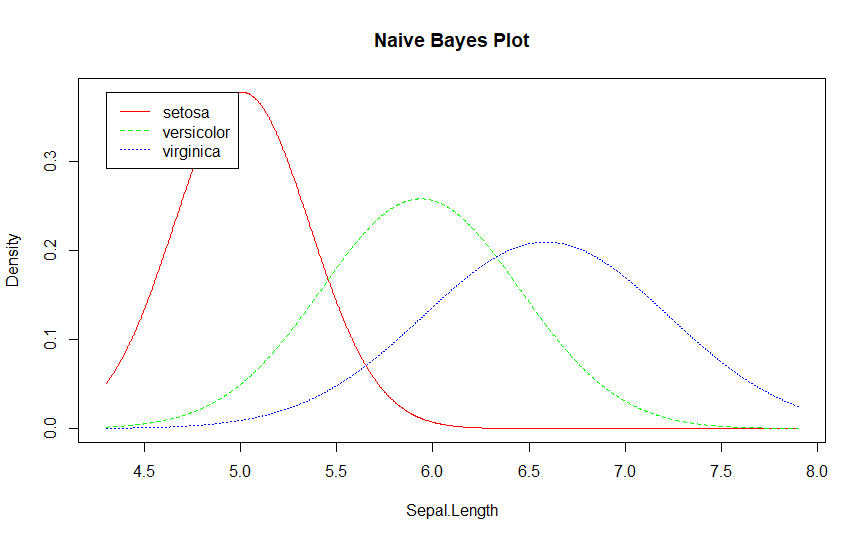


Fig: 3

Fig: 2

Fig: 1

* 1. **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. **Example**

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

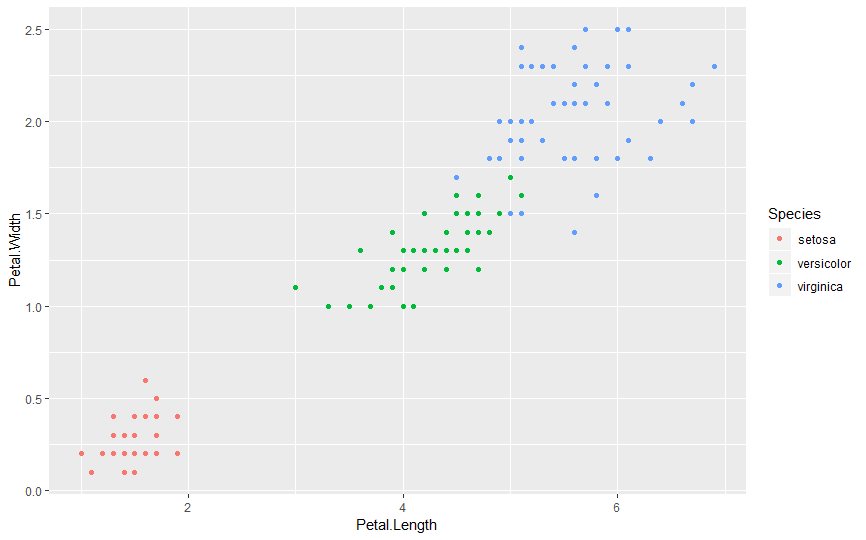


Fig: 4

1. **Data mining using the Census dataset**
   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.

1. fnlwgt: continuous.
2. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-
3. acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th,
4. Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated,
7. Widowed, Married-spouse-absent, Married-AF-spouse.
8. 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

1. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative,

Unmarried.

1. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
2. sex: Female, Male.
3. capital-gain: continuous.
4. capital-loss: continuous.
5. hours-per-week: continuous.
6. 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.

* 1. **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

Naïve Bayes is usually used when the dataset is large with several independent features. Since the census dataset satisfies all these conditions, we have chosen to implement the Naïve Bayes classifier.

* 1. **Naïve Bayes classifier**

The Naïve Bayes classifier assumes that there is a naïve independence between the features. The model assigns class labels to problem instances and uses the Bayes’ theorem in order to classify data.

**2.3.1 Bayes’ theorem**

The Bayes’ theorem finds the probability of an event based on prior knowledge of conditions that affect the outcome of an event.

Bayes’ theorem is given by the following equation:

where:

* **P(A | B)** = likelihood of A occurring given that B is TRUE
* **P(B | A)** = likelihood of B occurring given that A is TRUE
* **P(A)** = probability of the occurrence of A
* **P(B)** = probability of the occurrence of B

**2.3.2 Bayes classifier using Bayes’ theorem**

The Bayes classifier uses the Bayes’ theorem to obtain the conditional probability. The following equation represents this probability.

where:

* **p(Ck  | x)** = conditional probability
* **p(Ck)** = probability of prior event
* **p(x|Ck)** = likelihood of outcome
* **p(x)** = evidence supporting the outcome

The Bayes classifier combines this conditional probability along with a decision rule in order to classify data. The most common decision rule is the *MAP* (maximum a posteriori) decision rule.