Introduction to Data Mining

Prof. Dr. Necati Aras

Industry 4.0 – Digital Transformation

Digitization and digital transformation are the keys to producing high value-added services and products, as well as ensuring competitiveness at the international level by increasing efficiency in the industry.

In the past decade, great leaps have been made in information and communication technologies:

- Connecting digital devices, people, equipment, and materials to the Internet has enabled all resources to be constantly monitored and evaluated.
- The sensors that enable this and their communication network, also known as the "Internet of Things", created the "Big Data" paradigm by providing continuous data collection.

Industry 4.0 – Digital Transformation

In the past decade, great leaps have been made in information and communication technologies:

- Cloud computing offered a scalable and economical solution for storing and analyzing data. The analysis of this data with machine learning / Data Analytics methods has led to a transformation in all industries and “artificial intelligence” has become everyone's area of interest.

- Robots that live and work with humans are no longer the subject of science fiction; a near reality. On the other hand, the people of tomorrow will be "augmented people" with their smart phones, smart watches, augmented reality glasses.

- Edge-to-cloud computing capacity is increasing, offering scalable and unlimited computing power: The software-as-a-service, platform-as-a-service, and infrastructure-as-a-service paradigms have enabled the outsourcing of computing services, leading to the proliferation of digital services.

Digitization-Digitalization-Digital Transformation

- Digitization: It is the transfer of physically stored information to digital. It allows the information remaining in the archives to take place in the digital environment.
- Digitalization: Improving business processes by using the benefit of digitization
- Digital Transformation: Transforming business activities, processes, products and business models by using digital technologies in order to produce a more effective product/service.

Digitization-Digitalization-Digital Transformation

- In short, digitalization is about applying technology to the existing business.
- Digital transformation means doing things in a new, digital way. Digital transformation is a broader term than digitization. Digitization and digitization are parts of digital transformation.
- It includes all aspects of business such as digital transformation, customer insight and touchpoints, growth strategy, enterprise mobile apps, process digitization, employee enablement, performance, new business models and more.

Introduction to Data Analytics

Data Analytics

- Customer point of sale (POS) terminals (barcode readers, RFID systems, smart card readers)
- Web logs that keep the products analyzed and compared by customers during shopping from e-commerce sites on the Internet
- Hand terminals
- Records of conversations with customers in call centers

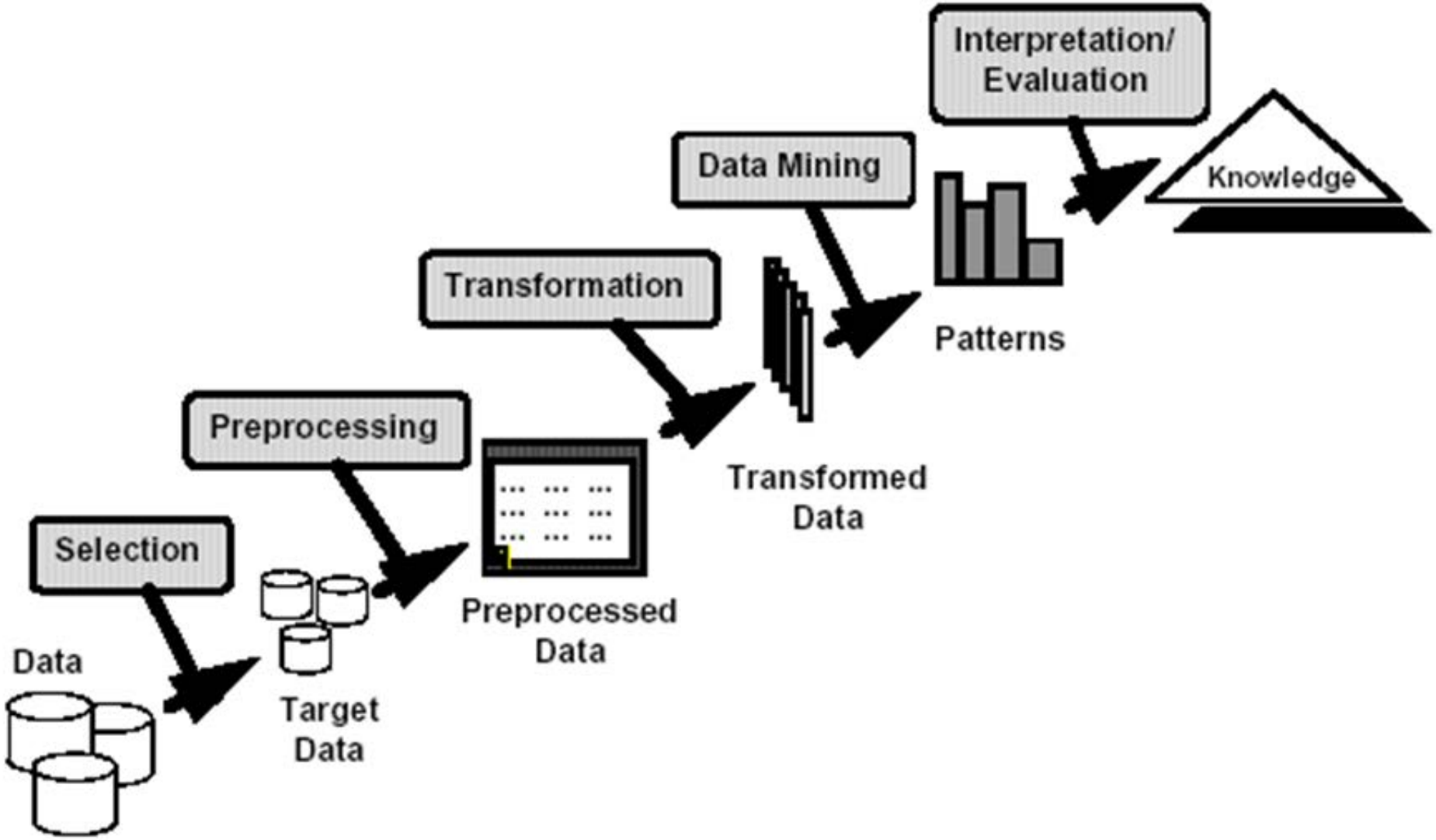
Data Analytics

- Data Analytics
- Data Science
- Data Mining
- Business Analytics
- Machine Learning
- Statistical Learning

Data Analytics

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- The process of employing one or more machine learning techniques to automatically analyze and extract knowledge from data.

Data Analytics



Data Analytics

What is not Data Analytics?

- Finding customers whose names start with the letter "A«
- Identification of credit card holders with a credit card expenditure of more than 15,000 TL in the last year
- Determining subscribers who talk on a mobile phone for less than 5 minutes on average in a week
- Identifying customers who order from the e-commerce site at least twice a week

Data Analytics

- (Cf): Classification Problem
- (R): Regression Problem
- (Cl): Clustering Problem
- (MBA): Market Basket Analysis
- (TS): Time Series Analysis
- (PCA): Principal Component Analysis

Data Analytics

- Finding out which of the customers who apply for a loan to a bank will have problems paying their debts regularly (Class.)
- Determining which of the current subscribers of a mobile phone operator or internet service provider company will cease to receive service and move to another company in the near future (Class.)
- Predicting the outcome of a football or basketball game (1st team winner, or 2nd team winner (Class.)
- Predicting which of the many potential customers will respond positively to a company's campaign offer (Class.)

Data Analytics

- Predicting the average monthly credit card expenditure of a customer (Regr.)
- Predicting the average monthly precepitation in Kuwait City (Regr.)
- Predicting the purchasing price of a apartment in Istanbul (Regr.)
- A mobile operator dividing existing subscribers into homogeneous groups in terms of various features and offering similar offers and new services to each group (Clus.)
- Grouping the branches of a supermarket chain or a bank in terms of profitability, number of customers, etc. (Clus.)

Data Analytics

- Displaying the type of news you want on news sites according to their content (economy, politics, travel, sports, magazine) (Clus.)
- Finding out which products are purchased together by customers in a store (Market Basket Analysis)

Classification

Input attributes

Output attribute

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Classification

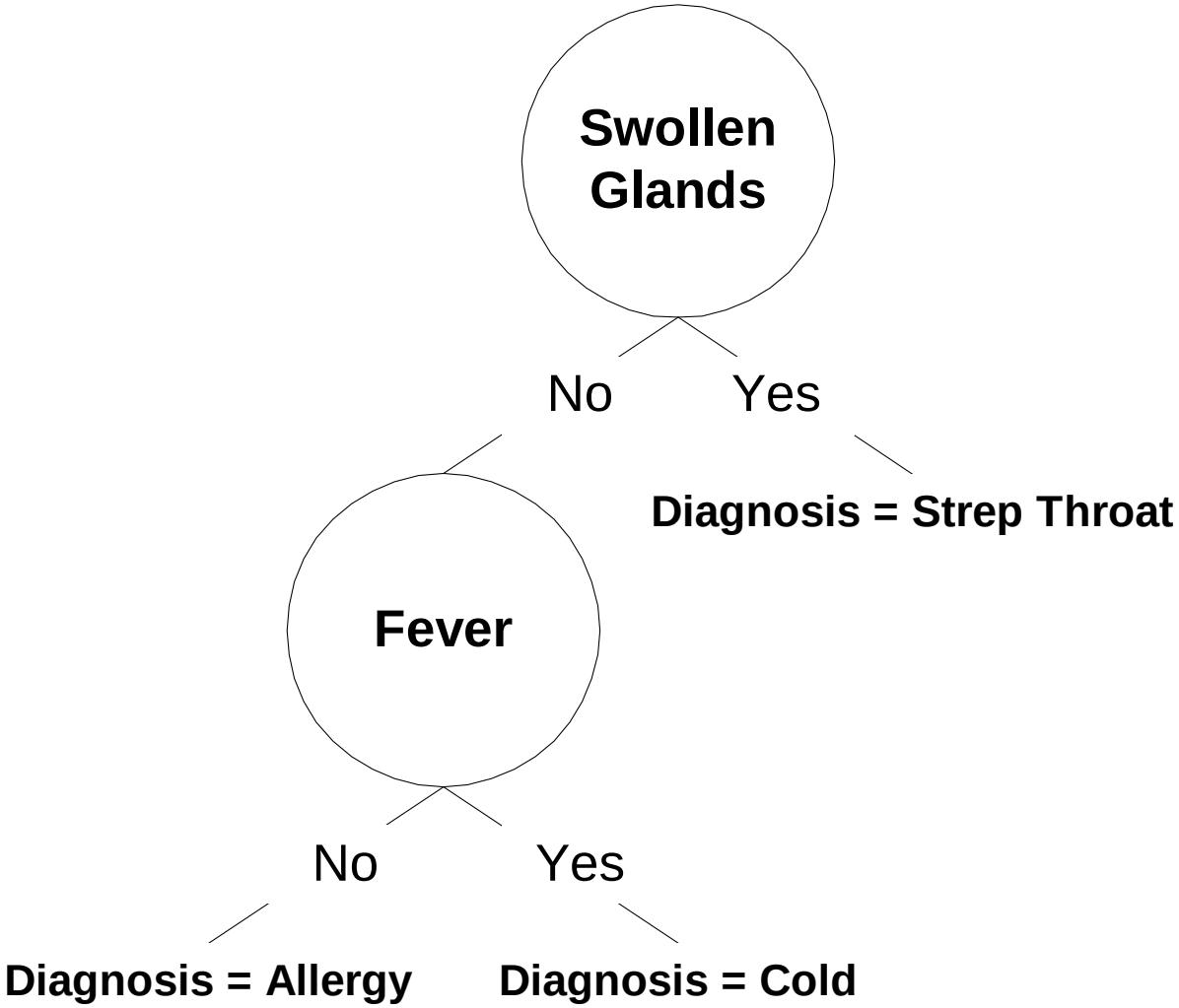
Input attributes

Output attribute

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?

Classification

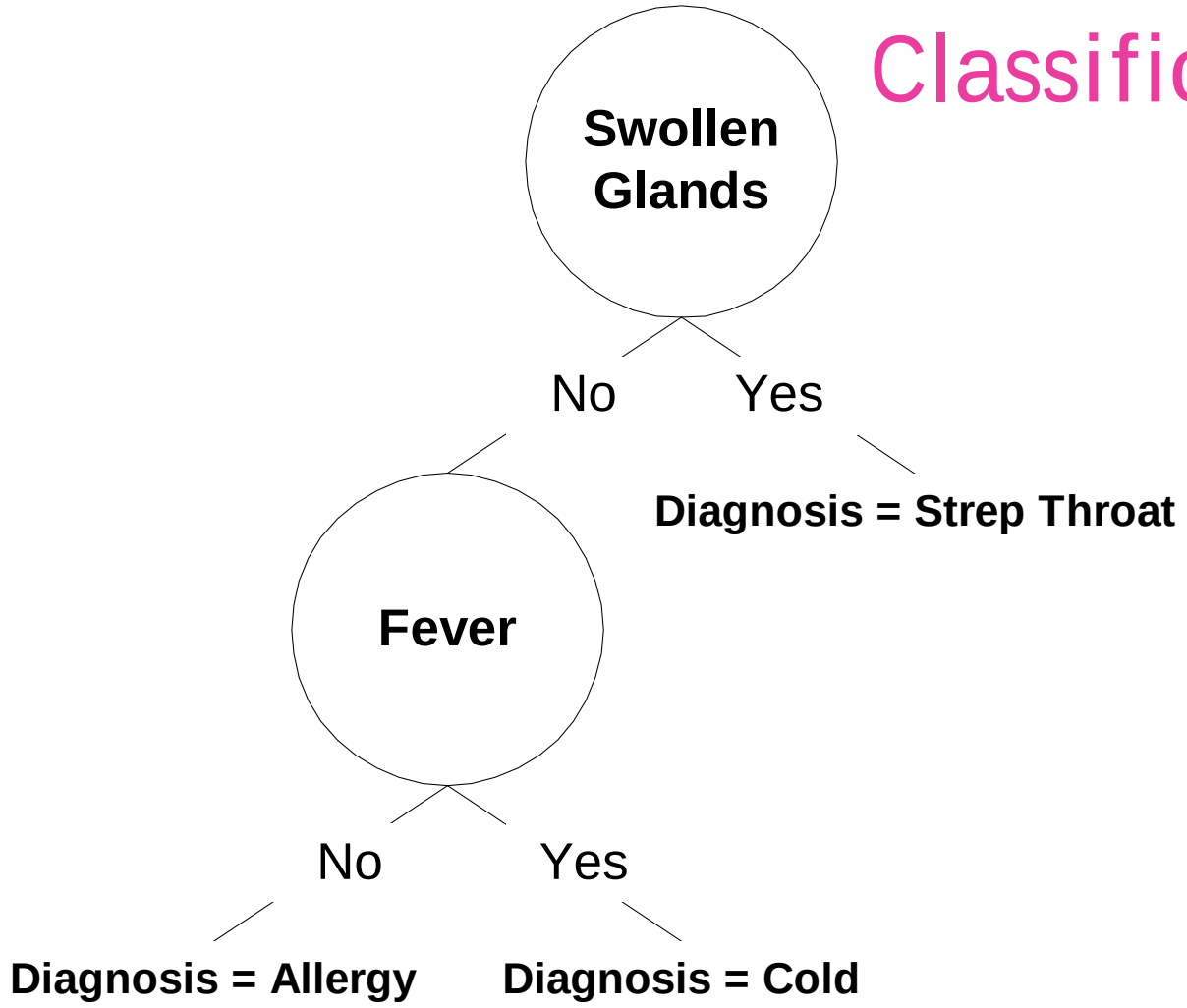


IF *Swollen Glands* = *Yes* THEN
Diagnosis = *Strep Throat*

IF *Swollen Glands* = *No* & *Fever* = *Yes* THEN
Diagnosis = *Cold*

IF *Swollen Glands* = *No* & *Fever* = *No* THEN
Diagnosis = *Allergy*

Classification



Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?

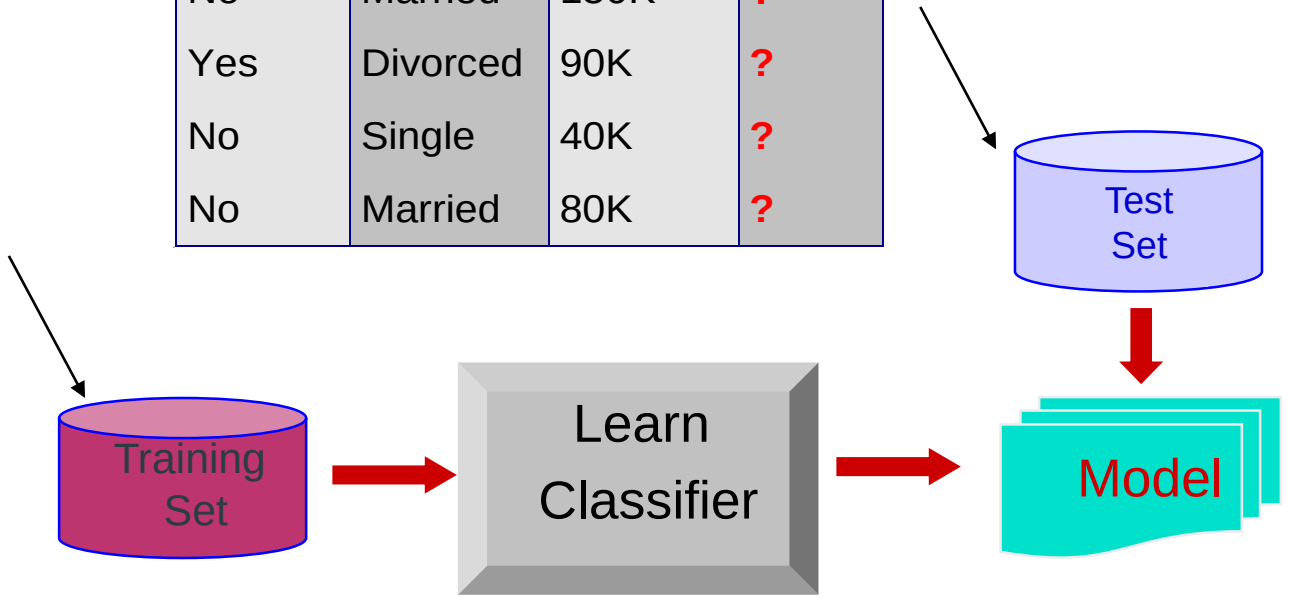
Input attributes, independent variables, predictors, features, covariates

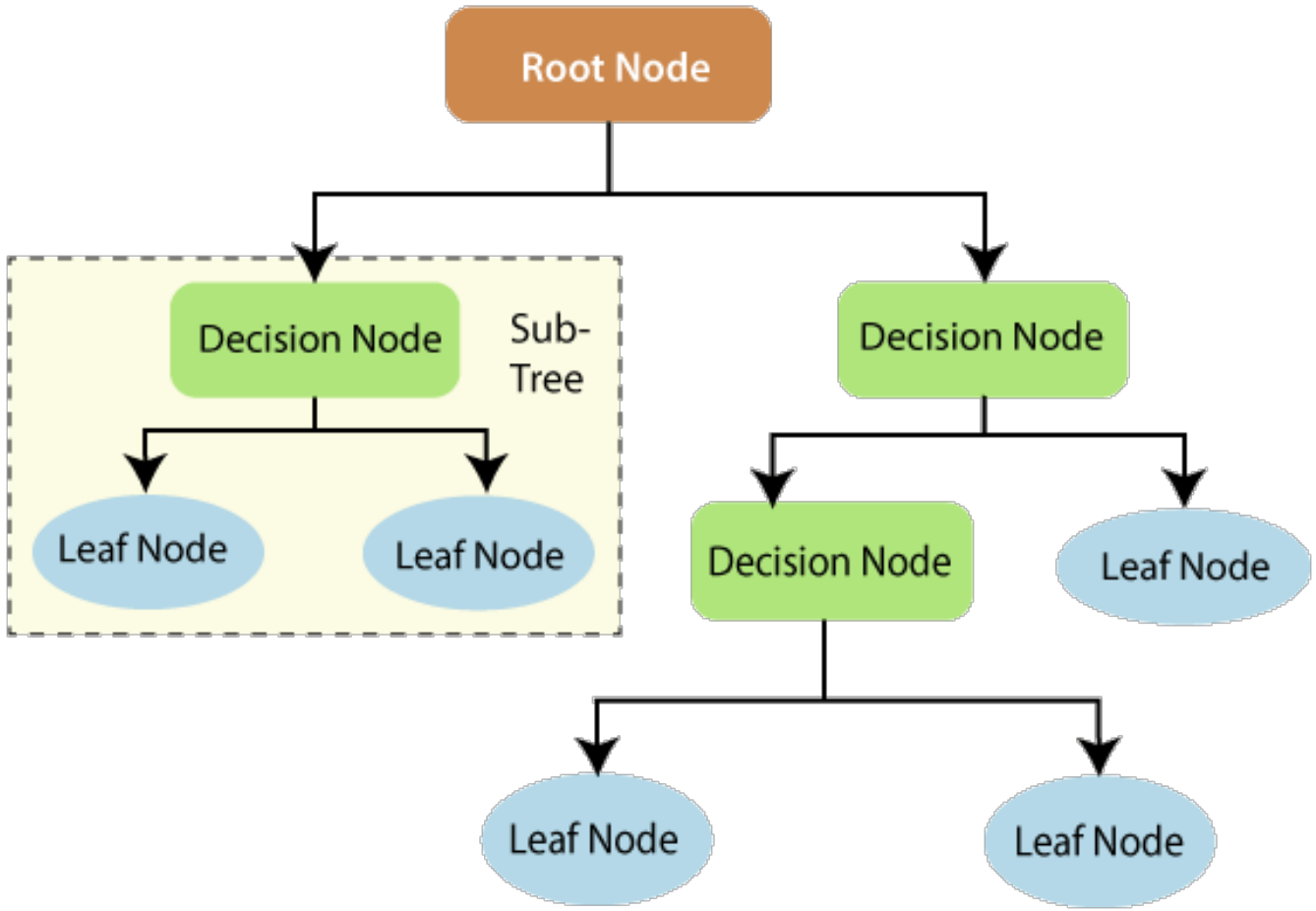
Output attribute, dependent variables, response, outcome

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

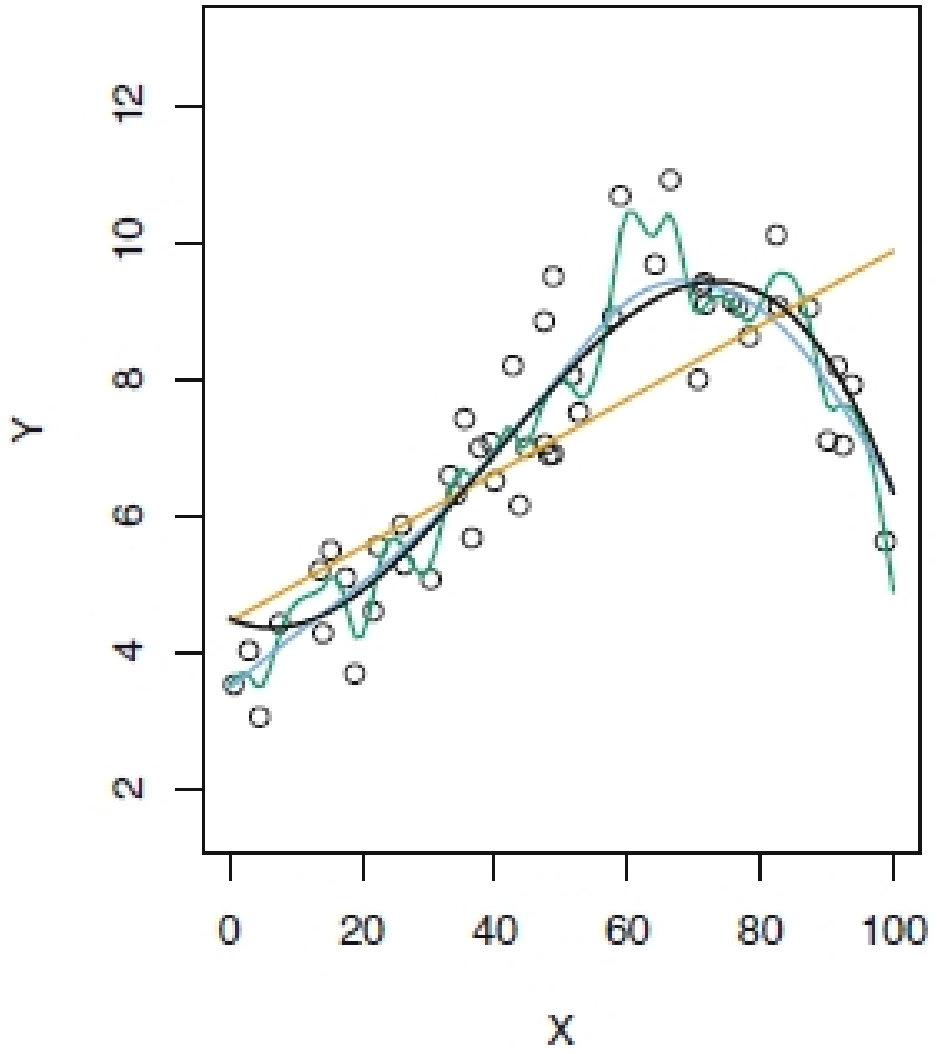


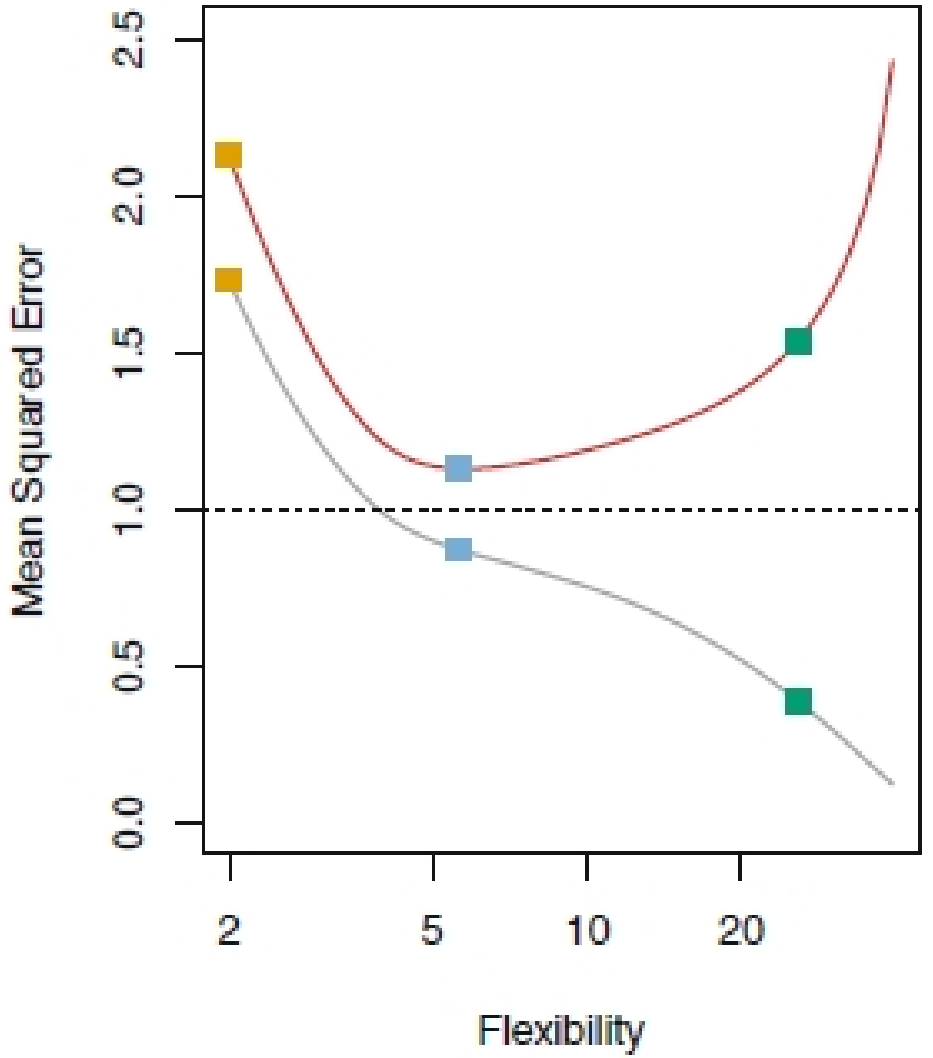

Root Node
Decision Node
Sub-
Tree
Leaf Node
Leaf Node
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node

Classification

- The method generally followed in determining the best classification tree for a dataset is to use some of the dataset as the "training dataset" and the rest as the "test dataset".

- The first set is used to build the tree, and the second set is used to measure success. In some algorithms, the dataset also has a "validation dataset" section, where the data in this section is used to prune the tree after it is first created and to find the smallest possible tree without sacrificing success.





Source: An Introduction to Statistical Learning with Applications in R
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 7. edition

Classification Methods

Classification Tree

Random Forest

Boosting Trees

Logistic Regression

K-Nearest Neighbors

Support Vector Machines

Artificial Neural Networks

Naive Bayes Classification

Classification

	Predicted Class		
Actual Class		Negative Class	Positive Class
	Neg. Class	TN	FP
	Pos. Class	FN	TP

TP (true positive)

FN (false negative)

FP (false positive)

TN (true negative)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Classification

	Predicted Class		
Actual Class		Negative Class	Positive Class
	Neg. Class	90	10
	Pos. Class	20	80

$$\text{Accuracy} = ?$$

$$\text{Sensitivity} = ?$$

$$\text{Specificity} = ?$$

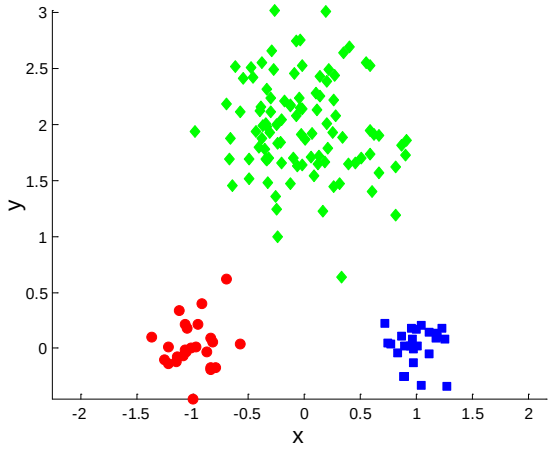
$$\text{Precision} = ?$$

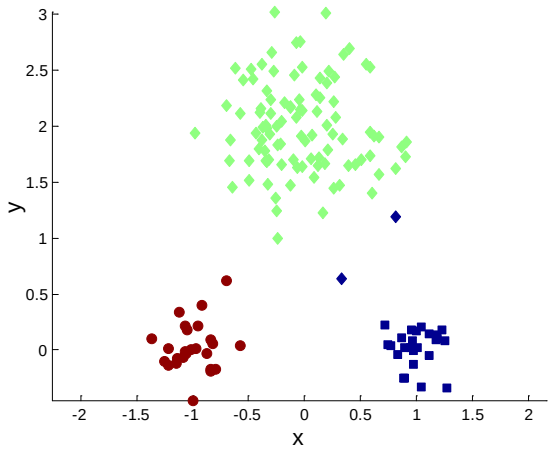
Clustering

Cluster analysis is a technique whose main purpose is to group observations (products, customers) based on their characteristics.

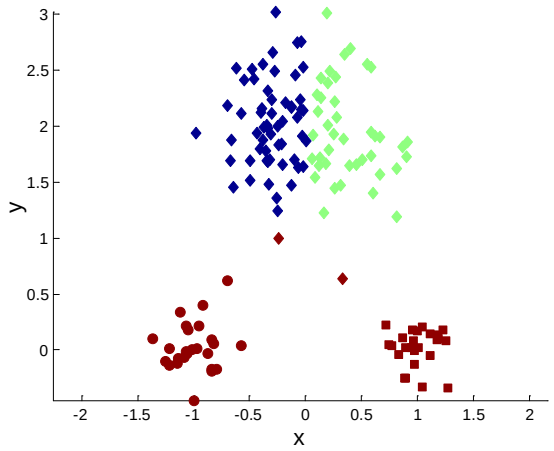
It divides *n* objects into clusters that are as homogeneous as possible within the cluster and as different as possible among clusters.

Clustering



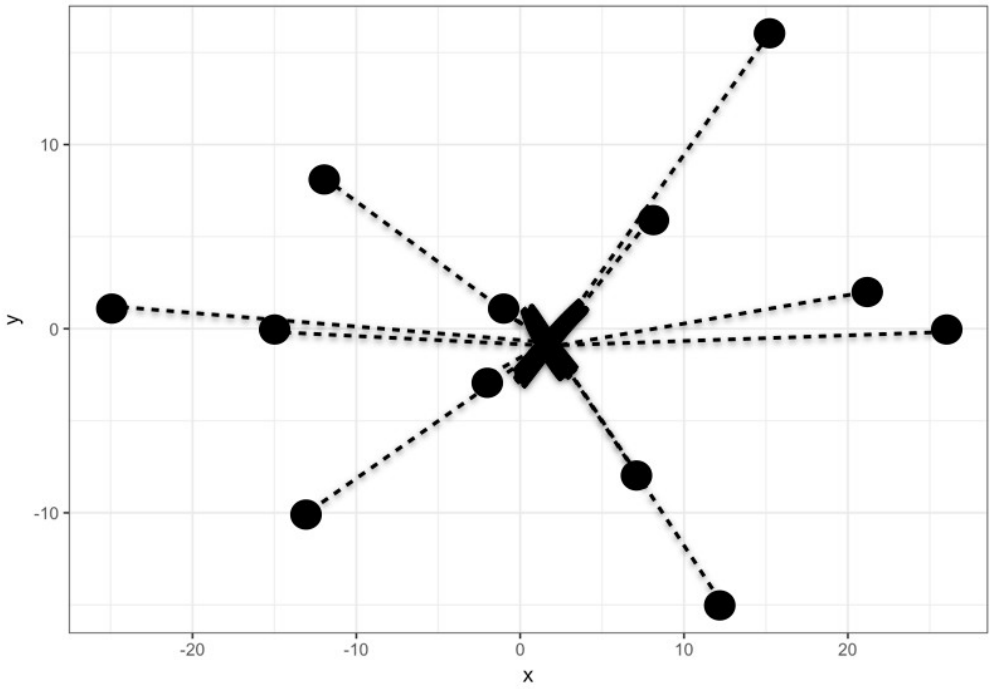


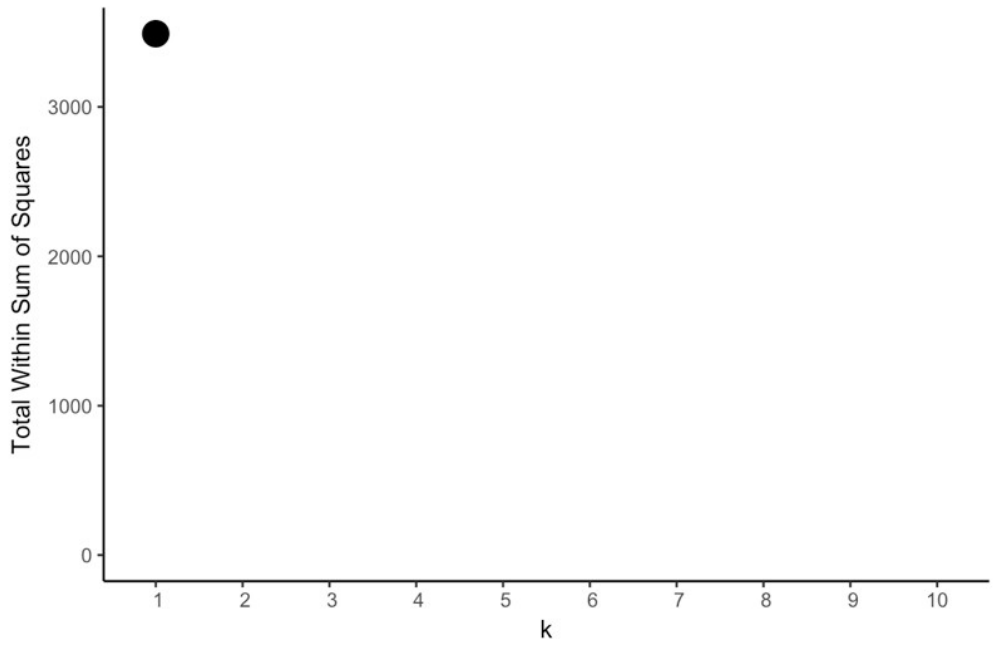
Optimal Clustering



Nonoptimal clustering

Clustering Performance

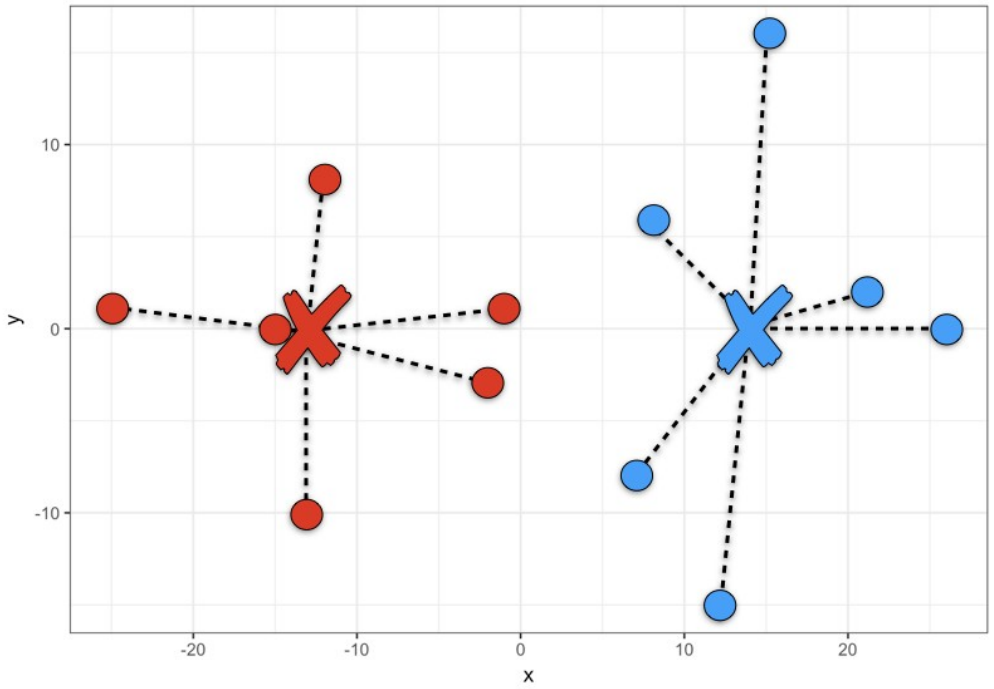


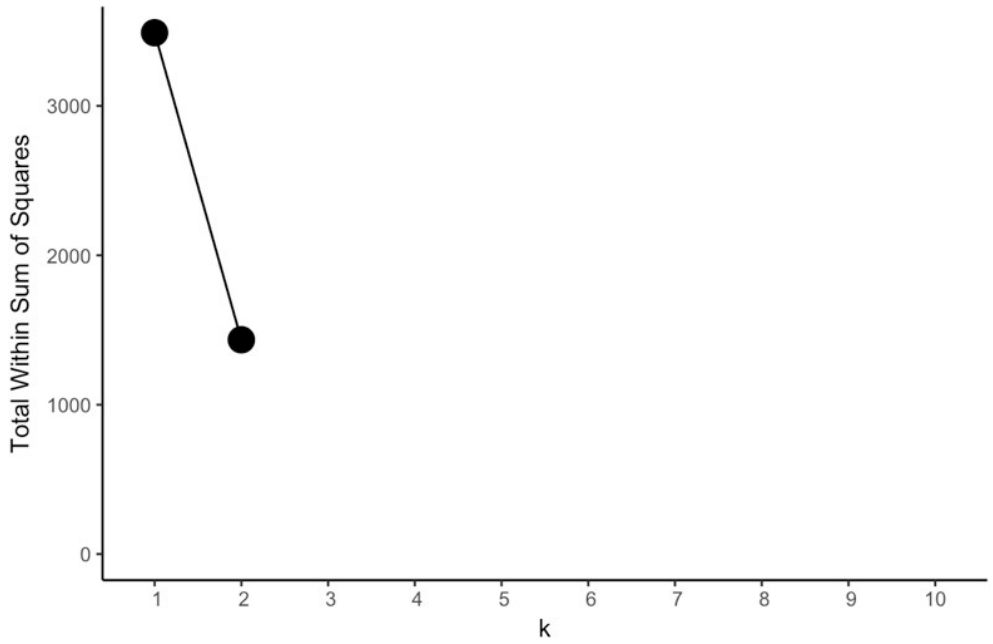


$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

k=1

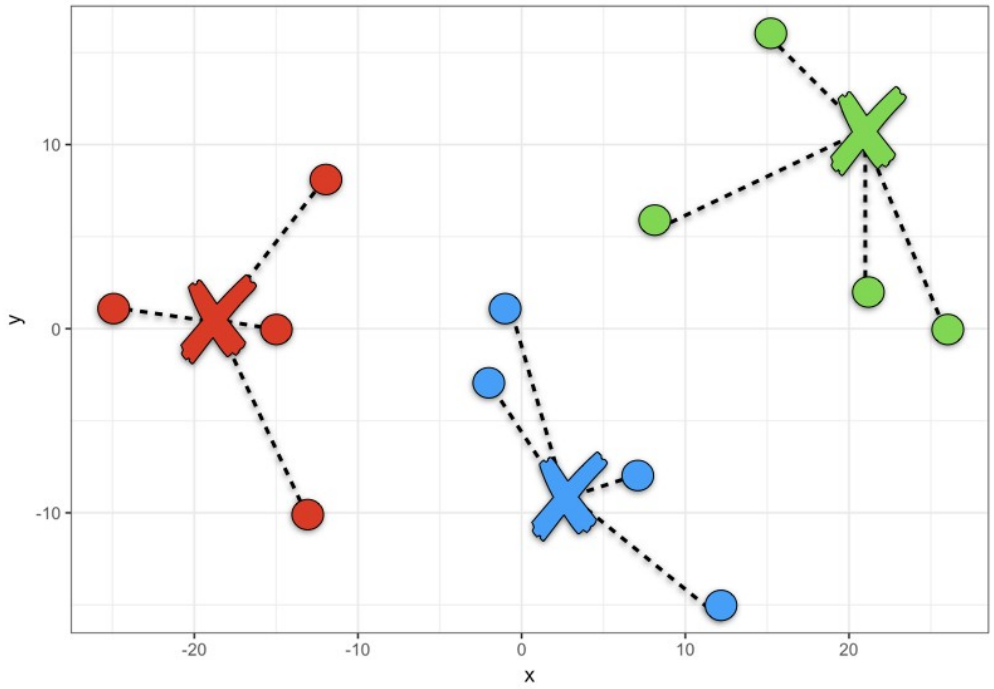
Clustering Performance

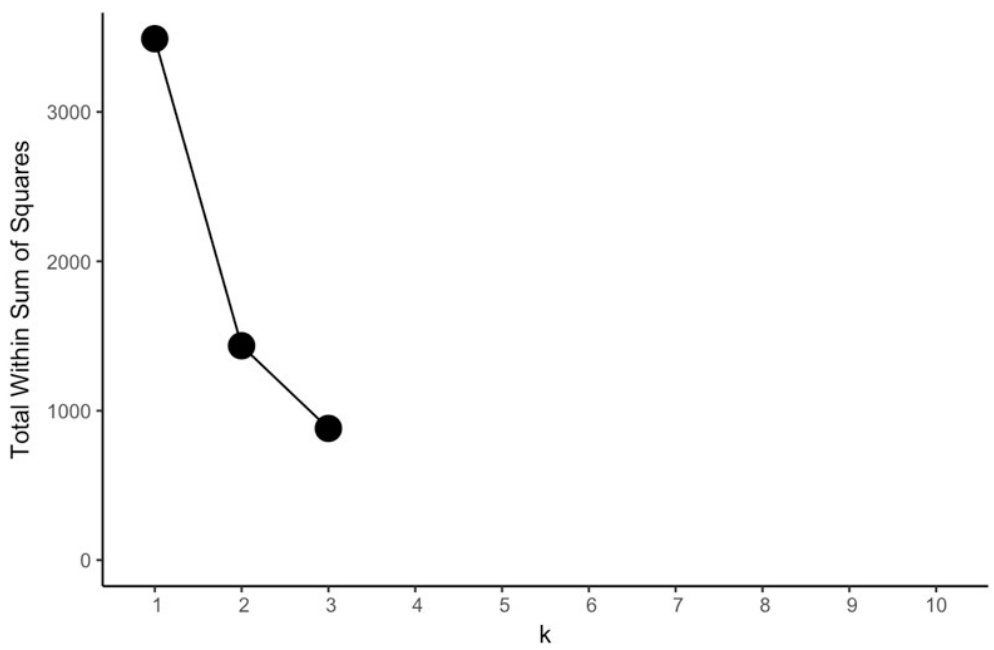




k=2

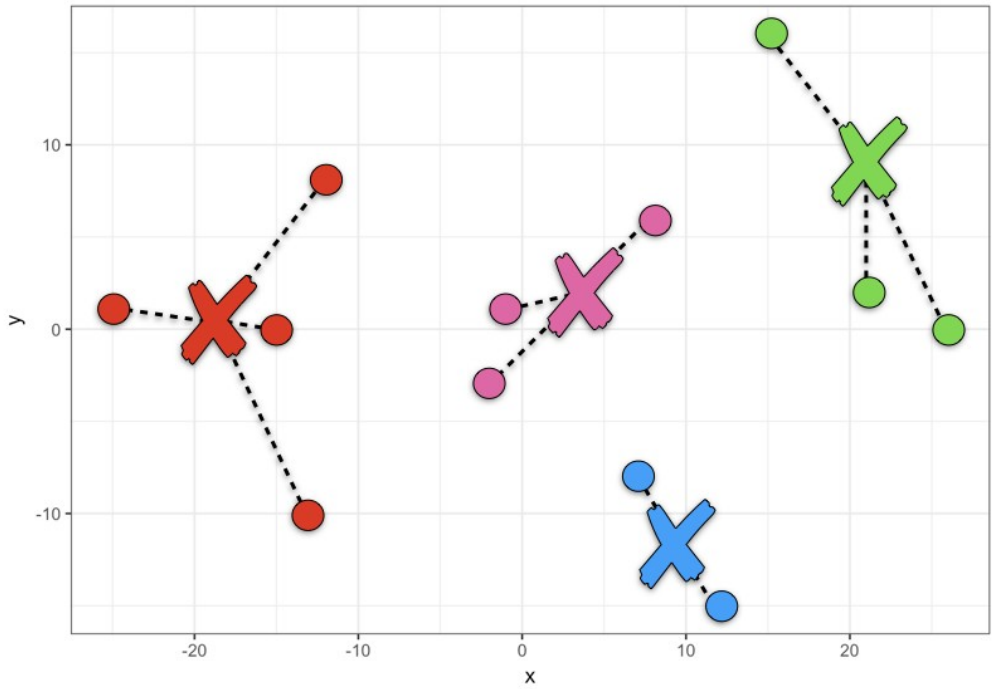
Clustering Performance

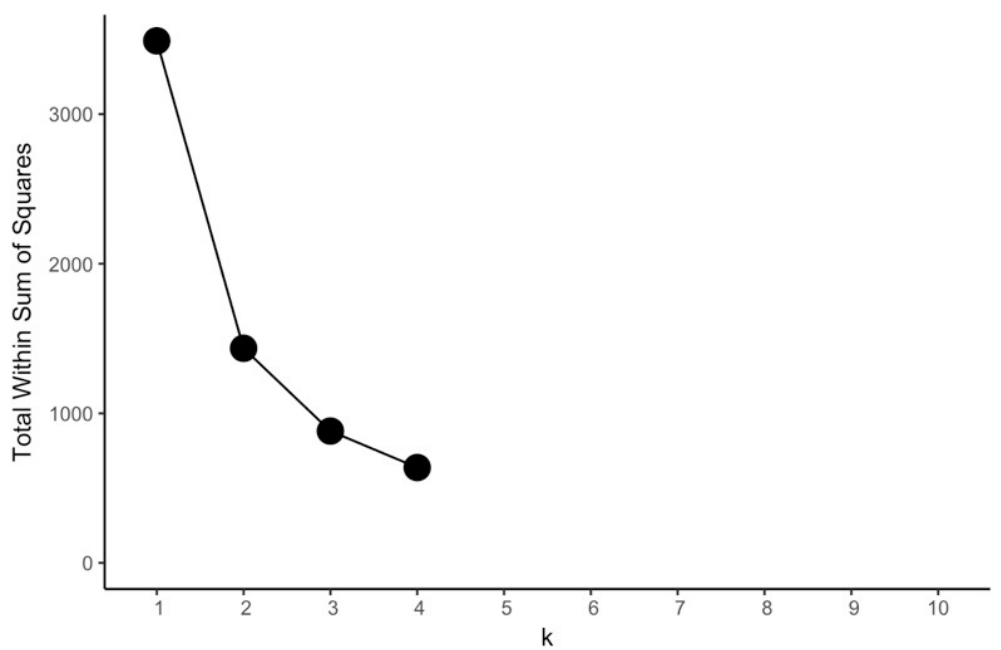




k=3

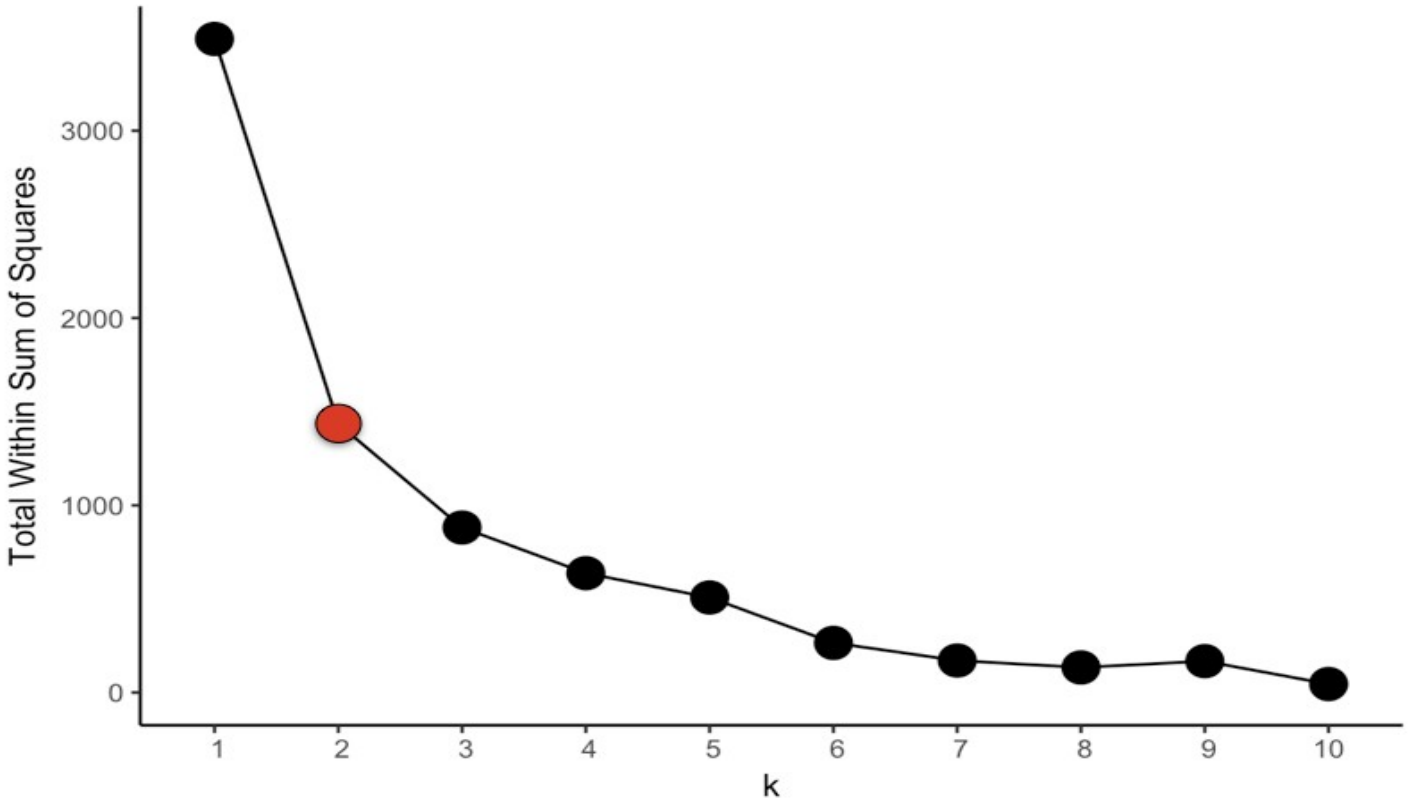
Clustering Performance





k=4

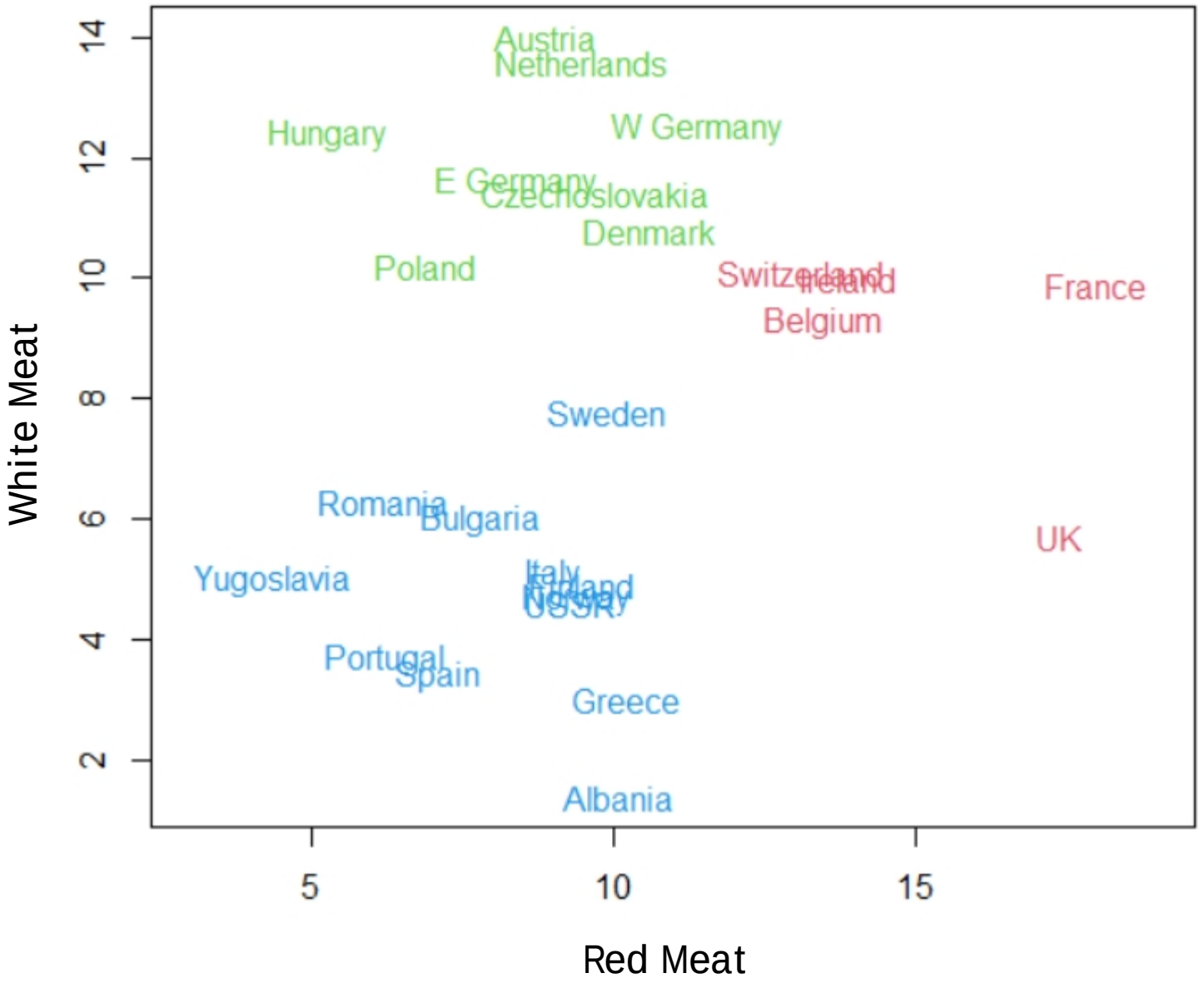
Clustering Performance



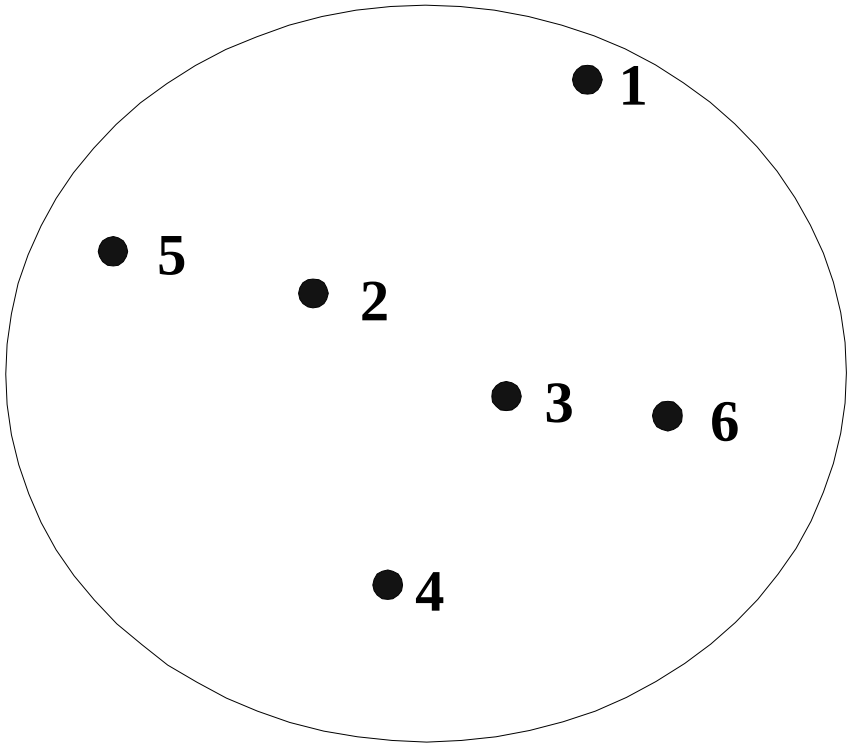
Clustering-Example

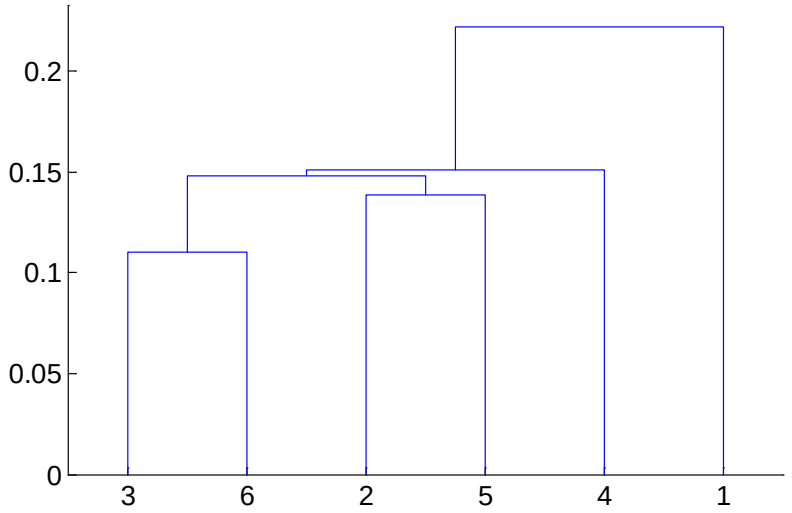
Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Clustering-Example



Hierarchical Clustering





Dendrogram

Market Basket Analysis

	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Sample Association Rules

{Diaper} → {Beer}
{Milk, Bread} → {Eggs,Coke}
{Beer, Bread} → {Milk}

Minimum support = 3

	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Item	Number
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

(1 itemsets)

Item Sets	Number
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

(2 itemsets)

Item Sets	Number
{Bread,Milk,Diaper}	3

(3 itemsets)

(No need for generating the itemsets including Coke and Eggs

Market Basket Analysis

Item Sets	Number
{Bread,Milk}	**3**
{Bread,Beer}	**2**
{Bread,Diaper}	**3**
{Milk,Beer}	**2**
{Milk,Diaper}	**3**
{Beer,Diaper}	**3**

Item Sets	Number
{Bread,Milk,Diaper}	**3**

{Bread} → {Milk}, {Milk} → {Bread}

{Bread} → {Diaper}, {Diaper} → {Bread}

{Milk} → {Diaper} , {Diaper}→ {Milk}
{Beer} → {Diaper} , {Diaper} → {Beer}

{Bread} → {Milk, Diaper} {Bread, Milk} → {Diaper}

{Milk} → {Bread, Diaper} {Bread, Diaper} → {Milk}

{Diaper} → {Milk, Bread} {Milk, Diaper} → {Bread}

Market Basket Analysis

Performance of the rules

- Support (s): The percentage of the transactions that include the items in the rule (an indication of how frequently the itemset appears in the dataset)
- Confidence (c): The percentage of all transactions satisfying the antecendent that also satisfy the consequent of the rules

{Milk} → {Bread}: s=0.6, c=0.75

{Bira} → {Diaper}: s=0.6, c=1

{Milk,Diaper} → {Beer}: s=0.4, c=0.67