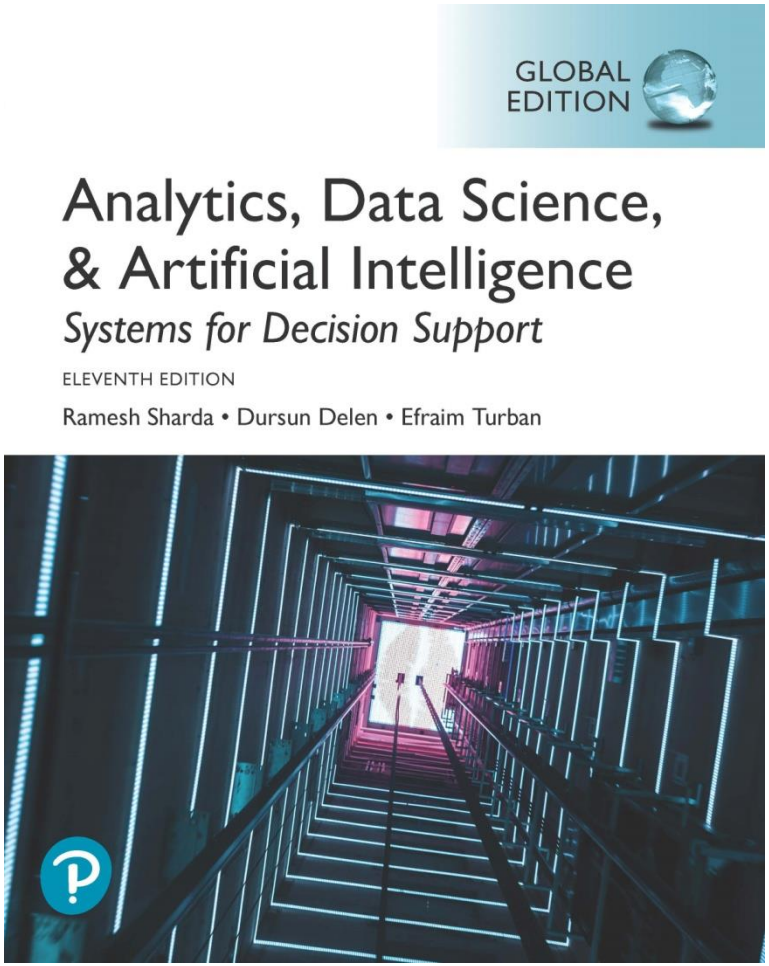
Analytics, Data Science and AI: Systems for Decision Support

Eleventh Edition, Global Edition



Chapter 4

Data Mining Process, Methods, and Algorithms

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Learning Objectives

4.1 Define data mining as an enabling technology for business analytics

4.2 Understand the objectives and benefits of data mining

4.3 Become familiar with the wide range of applications of data mining

4.4 Learn the standardized data mining processes

4.5 Learn different methods and algorithms of data mining

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Concepts and Definitions Why Data Mining?

- More intense competition at the global scale.
- Recognition of the value in data sources.
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses.
- The exponential increase in data processing and storage capabilities; and decrease in cost.

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Definition of Data Mining

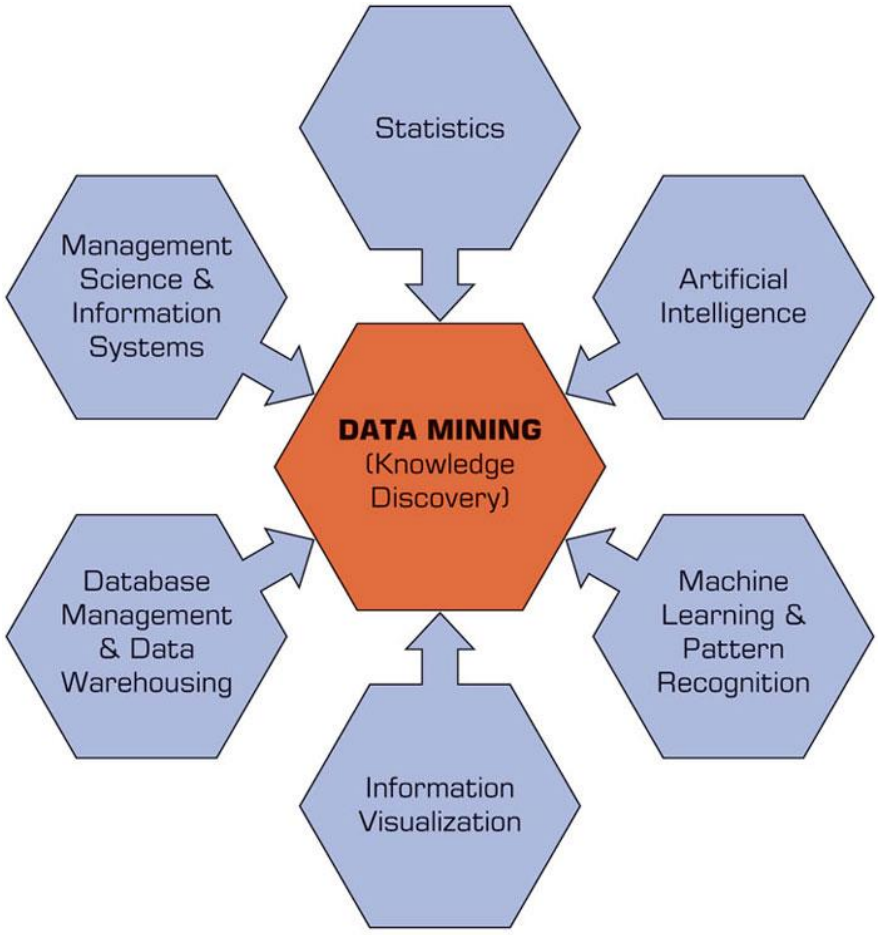
- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases. -- *Fayyad et al.,* (1996)
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable.
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,…

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Is a Blend of Multiple Disciplines

Figure 4.1 Data Mining Is a Blend of Multiple Disciplines.



Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Characteristics & Objectives

- Source of data for DM is often a consolidated data warehouse (not always!).
- Data is the most critical ingredient for DM which may include soft/unstructured data.
- The miner is often an end user
- Data mining tools' capabilities and ease of use are essential (Web, parallel processing, etc.)

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

How Data Mining Works

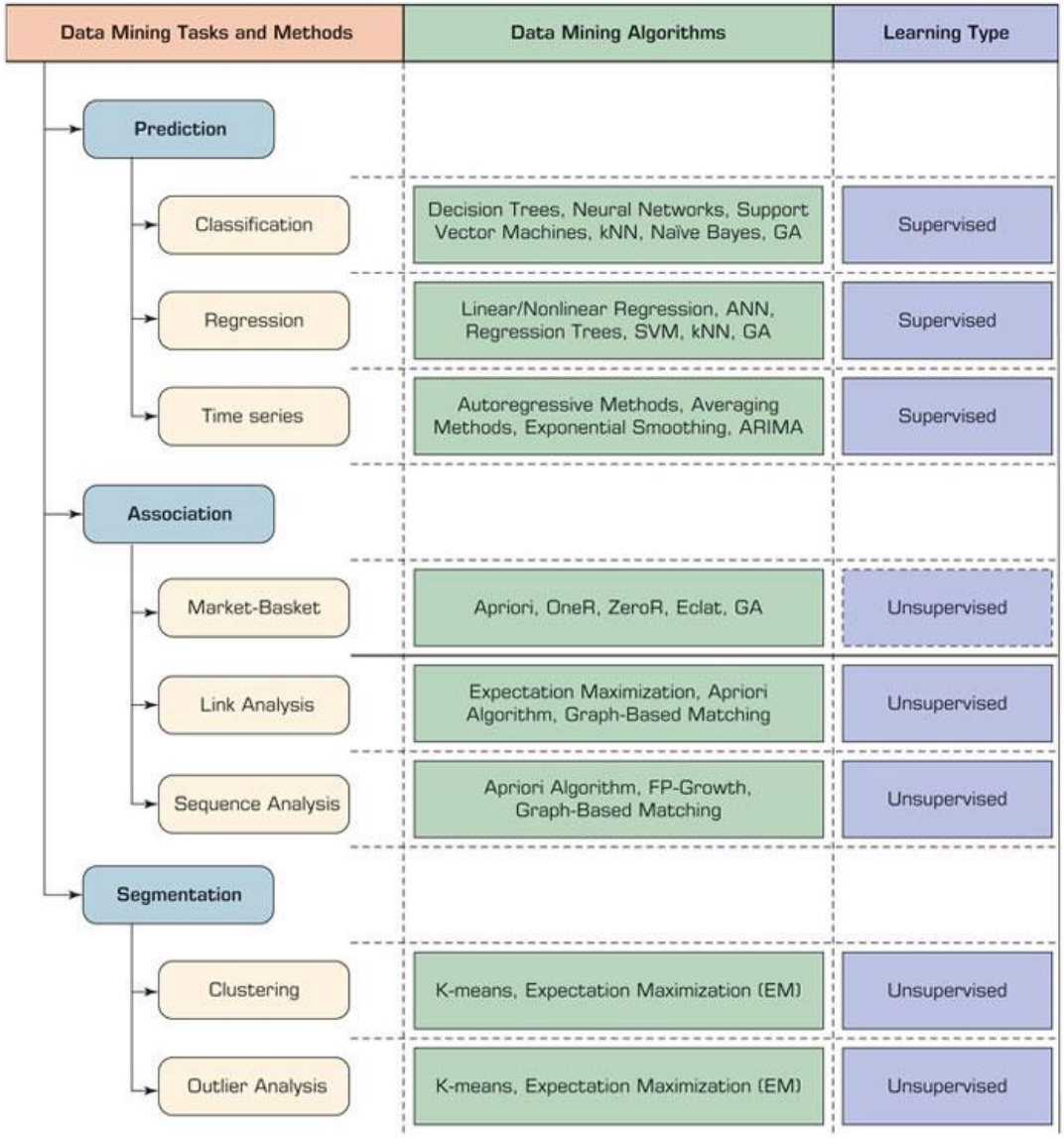
- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

A Taxonomy for Data Mining

Figure 4.2 A Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms.



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Other Data Mining Patterns/Tasks

- Time-series forecasting
  - Part of the sequence or link analysis?
- Visualization
  - Another data mining task?
  - Covered in Chapter 3
- Data Mining versus Statistics
  - Are they the same?
  - What is the relationship between the two?

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Applications (1 of 4)

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross-, up-selling)
  - Identify and treat most valued customers
- Banking & Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Maximize customer value (cross-, up-selling)
  - Optimizing cash reserves with forecasting

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Applications (2 of 4)

- Retailing and Logistics
  - Optimize inventory levels at different locations
  - Improve the store layout and sales promotions
  - Optimize logistics by predicting seasonal effects
  - Minimize losses due to limited shelf life
- Manufacturing and Maintenance
  - Predict/prevent machinery failures
  - Identify anomalies in production systems to optimize the use manufacturing capacity
  - Discover novel patterns to improve product quality

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Applications (3 of 4)

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading
- Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Applications (4 of 4)

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel, entertainment, sports
- Healthcare and medicine
- Sports,… virtually everywhere…

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process

- A manifestation of the best practices
- A systematic way to conduct DM projects
- Moving from Art to Science for DM project
- Everybody has a different version
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM (1 of 2)

- Cross Industry Standard Process for Data Mining
- Proposed in 1990s by a European consortium
- Composed of six consecutive steps
  - Step 1: Business Understanding
  - Step 2: Data Understanding
  - Step 3: Data Preparation
  - Step 4: Model Building
  - Step 5: Testing and Evaluation
  - Step 6: Deployment

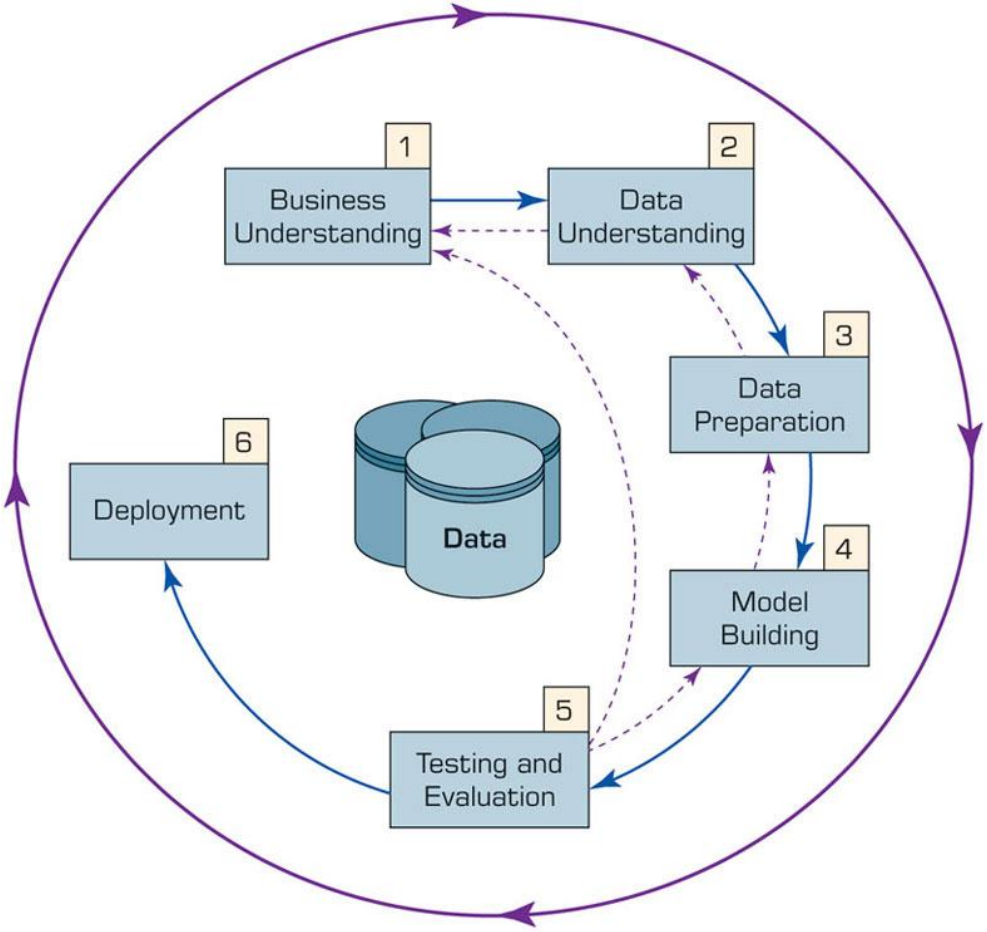
Accounts for ~85% of total project time

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM (2 of 2)

- **Figure 4.3** The Six-Step CRISP-DM Data Mining Process. →
- The process is highly repetitive and experimental (DM: art versus science?)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM

Step 1: Business Understanding

- Define the business problem.
  - Why sales fall, do we lost customers,….
- Who is going to collect and analyze the data for finding such knowledge

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM

Step 2: Data Understanding

- Identify the relevant data
- Identify the sources of such data
- Type of data: quantitative vs. qualitative

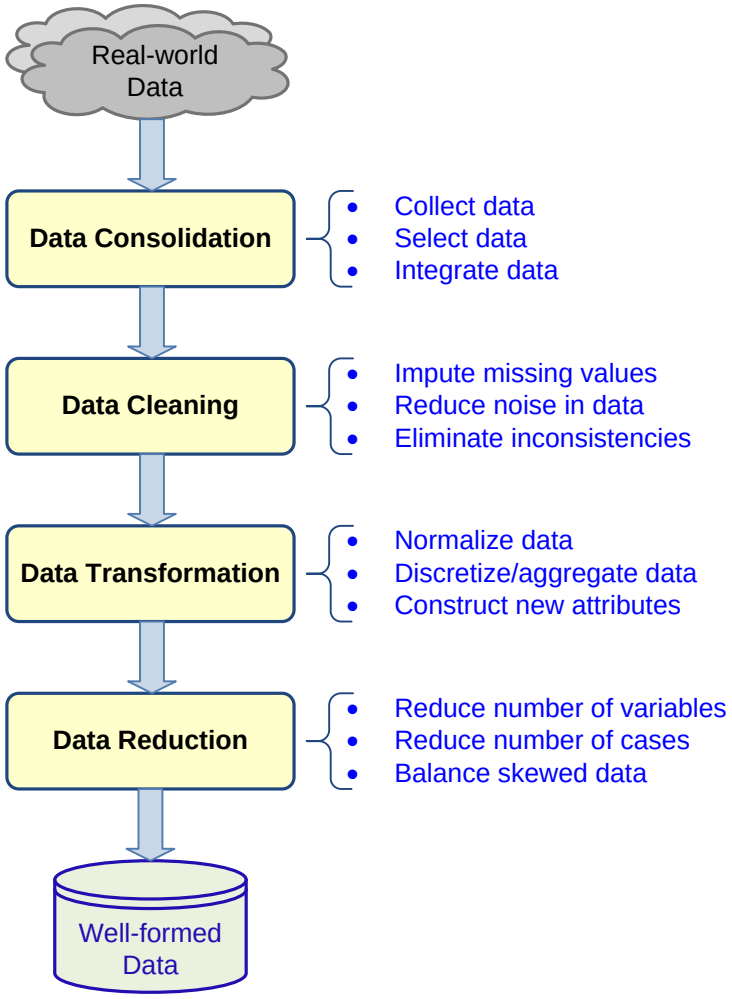
Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM

Step 3: Data Preparation

- Commonly called *preprocessing*



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM

Step 4: Model Building

- Depending on the business needs, build the right model.
  - Select the DM task: prediction, association or clustering
  - Use the right algorithm: decision tree, regression, ANN,….
  -

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: CRISP-DM

Step 5: Testing and Evaluation

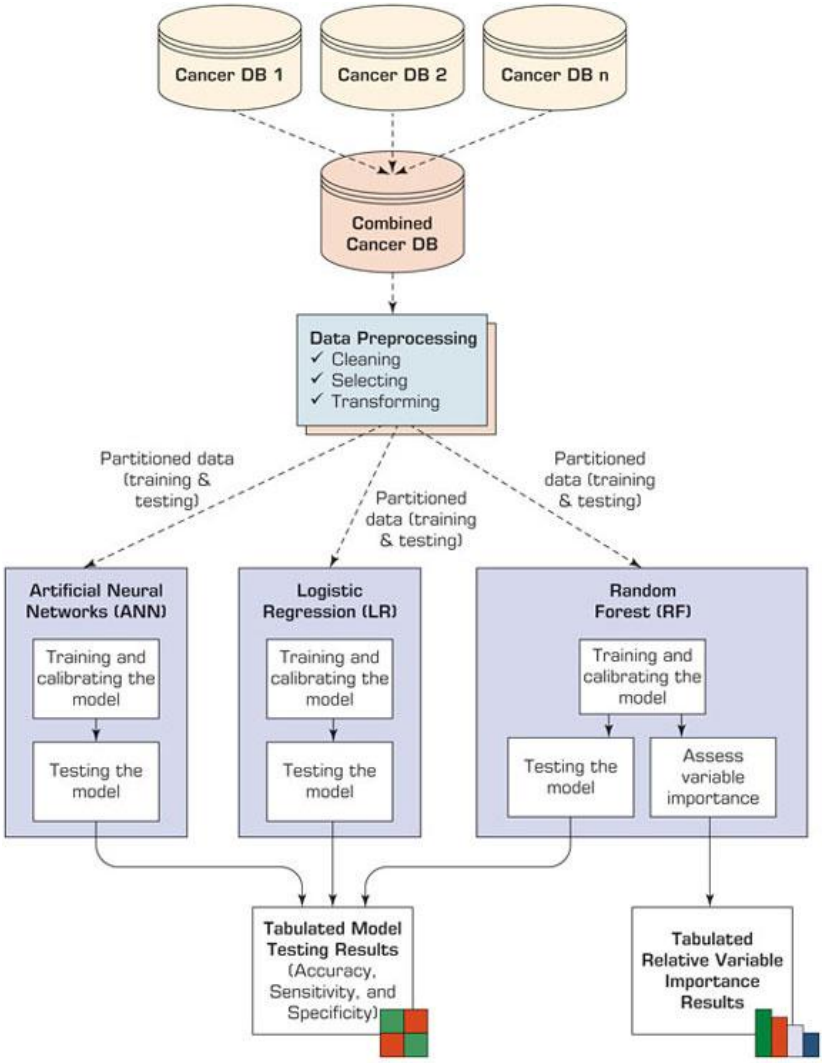
- The model is assessed and evaluated
- Data analyst and decision owner interact and validate the model

Step 6: Deployment

- Decision owner understands what actions to be taken in order to make use of the model outcomes

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.



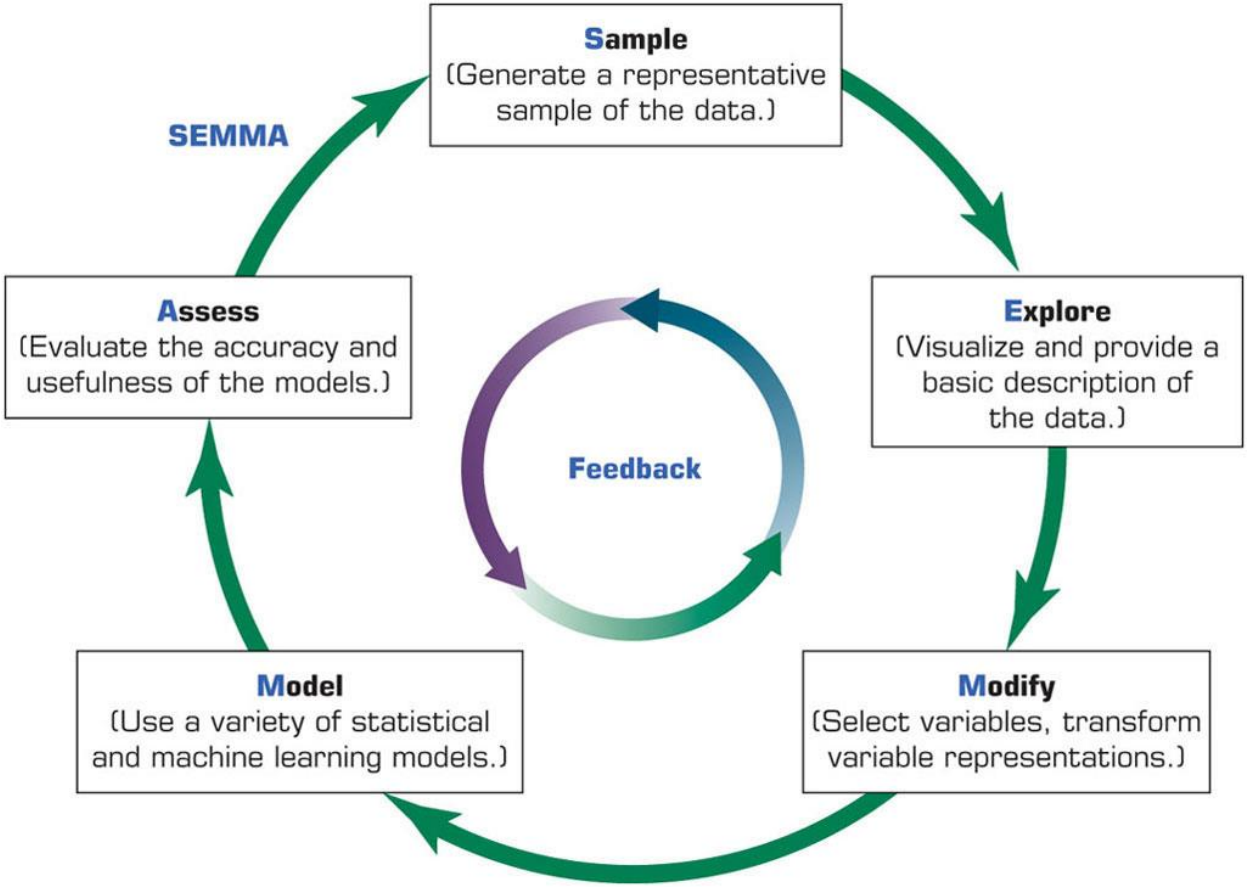
Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: SEMMA

Figure 4.5 SEMMA Data Mining Process.

- Developed by SAS Institute



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: KDD

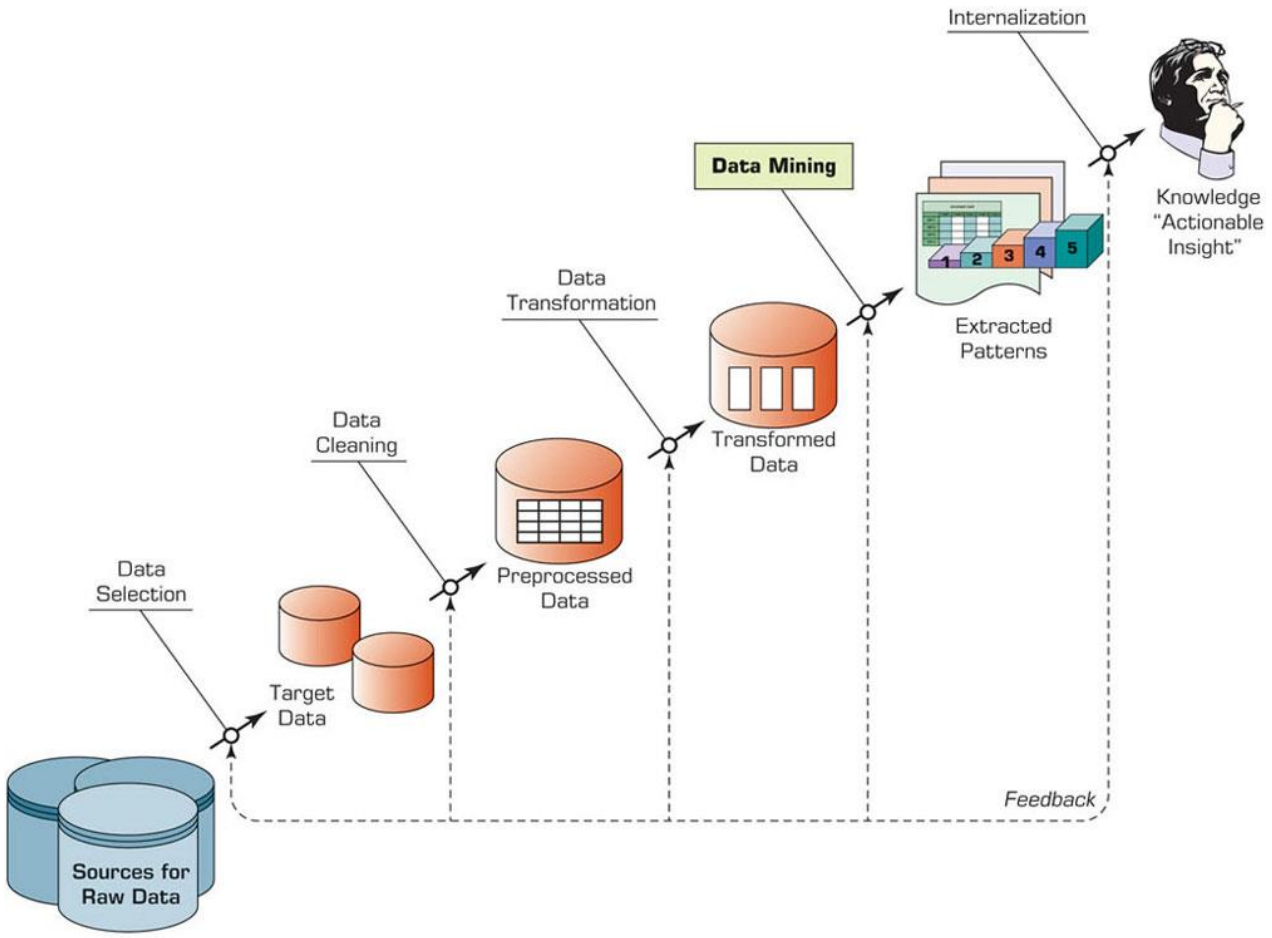
- Data selection
- Data transformation and cleaning
- Data mining to extract patterns or models
- Interpretation and evaluation

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Process: KDD

Figure 4.6 KDD (Knowledge Discovery in Databases) Process.

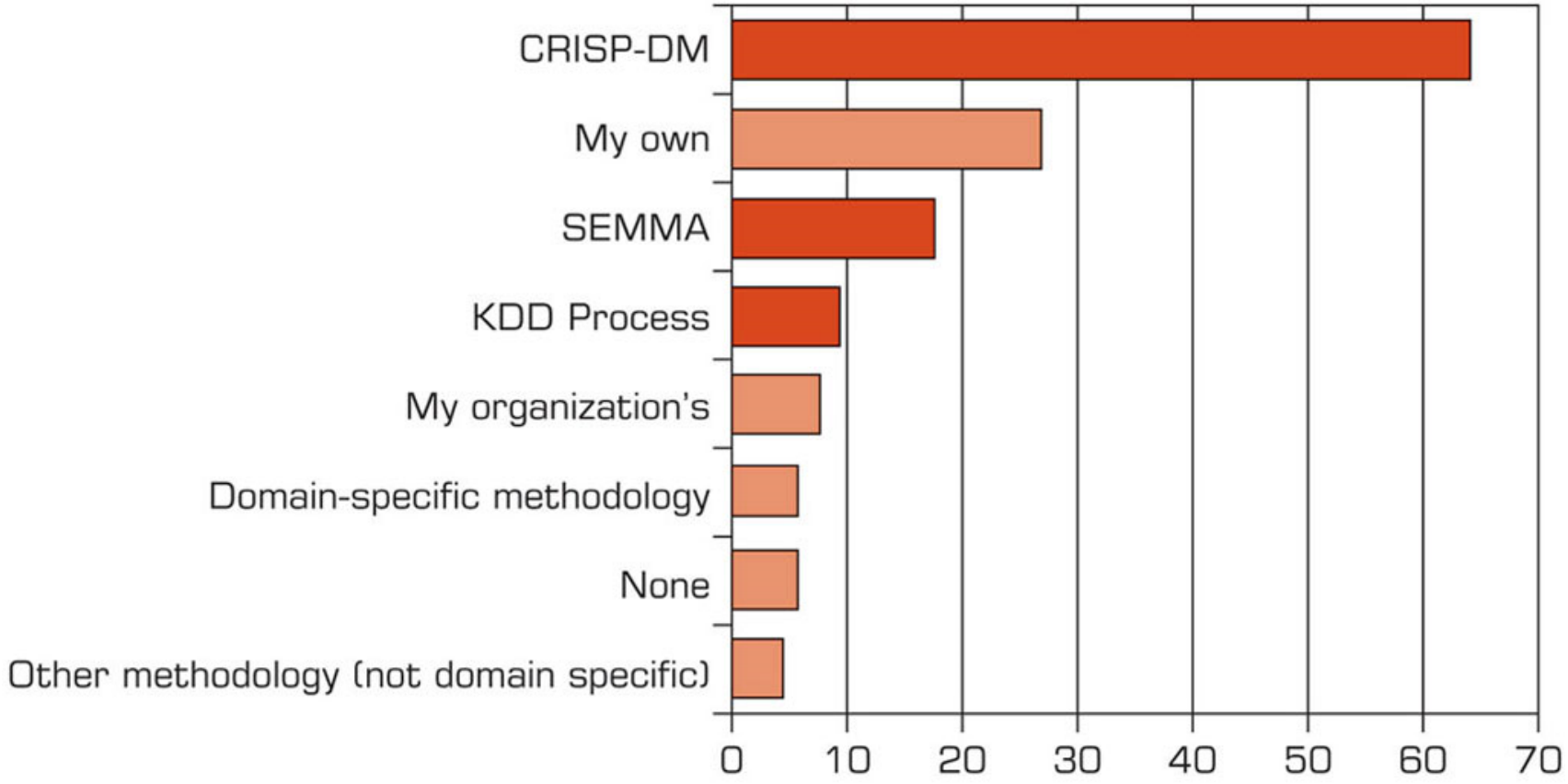


Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Which Data Mining Process is the Best?

Figure 4.7 Ranking of Data Mining Methodologies/Processes.



Source: Used with permission from **KDnuggets.com**.

Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Classification—A Two-Step Process

1) Model construction: describing a set of predetermined classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute (e.g. Spam or non-spam)
- The set of tuples used for model construction is training set
- The model is represented as classification rules, decision trees, or mathematical formulae

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Classification—A Two-Step Process

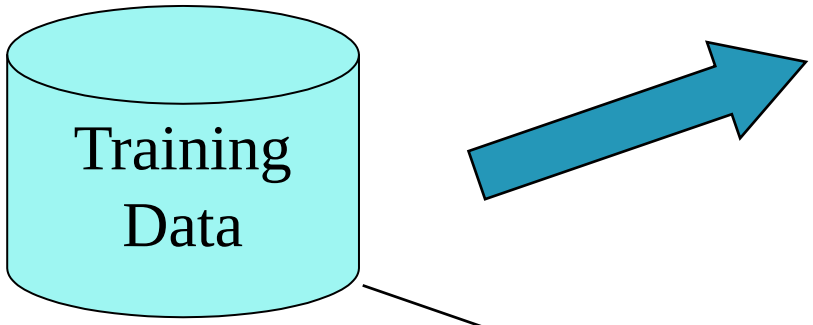
2) Model usage: for classifying future or unknown objects

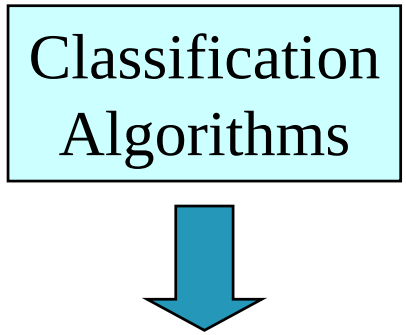
- Estimate accuracy of the model
  - The known label of test sample is compared with the classified result from the model
  - Accuracy rate is the percentage of test set samples that are correctly classified by the model
  - Test set is independent of training set
- If the accuracy is acceptable, use the model to classify new data

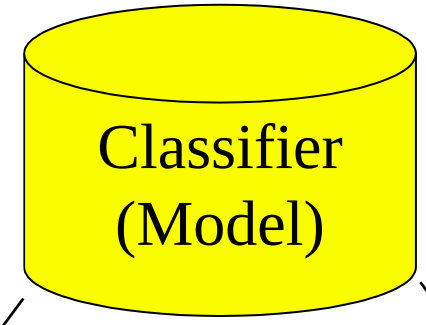
Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Process (1): Model Construction







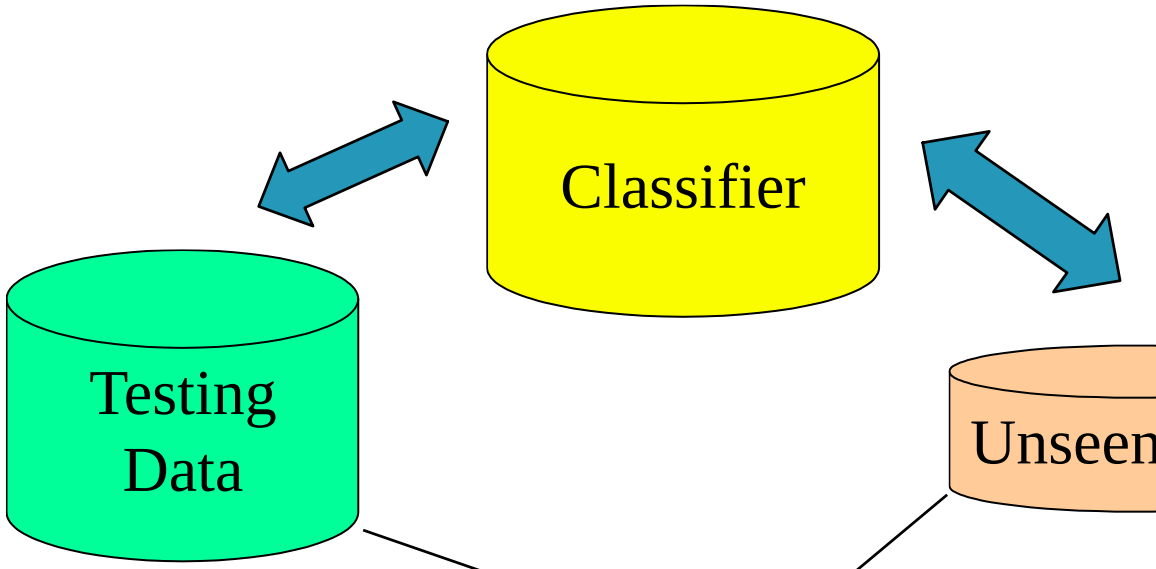
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > = 7
THEN tenured = ‘yes’

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(George, Professor, 5)

Tenured?

Yes

Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

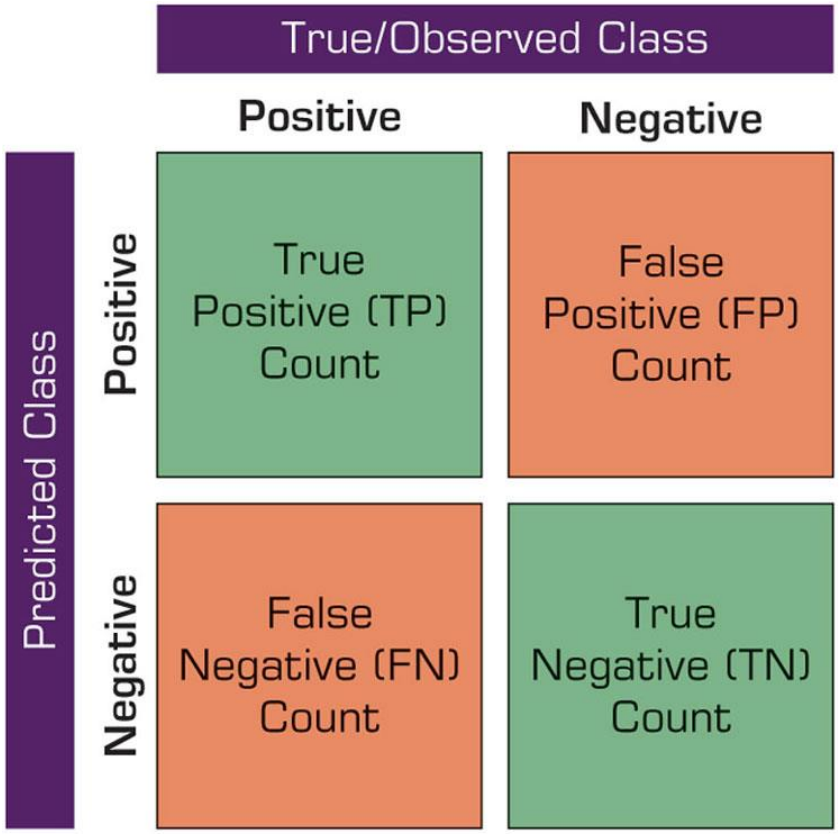
Assessment Methods for Classification

- Predictive accuracy
  - % of test data set samples correctly classified by the model
- Speed
  - Model building versus predicting/usage speed
- Robustness: model predictions with noisy data
- Scalability: model construction with large amount of data
- Interpretability
  - Level of understanding and insight provided by the model

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix



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

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Accuracy of Classification Models Example

Actual Class \ Predicted class	cancer = P	cancer = N	Total
cancer = P	**TP = 90**	**FN = 210**	300
cancer = N	**FP = 140**	**TN = 9560**	9700
Total	230	9770	10000

Accuracy (% of test set tuples that are correctly classified)= TP+TN/(TP+TN+FP+FN)=(90+6560)/10000 = 96.5%

Precision = (% of tuples that the classifier labeled as positive are actually positive) TP/(TP+FP) = 90/(90+140) = 39.13%

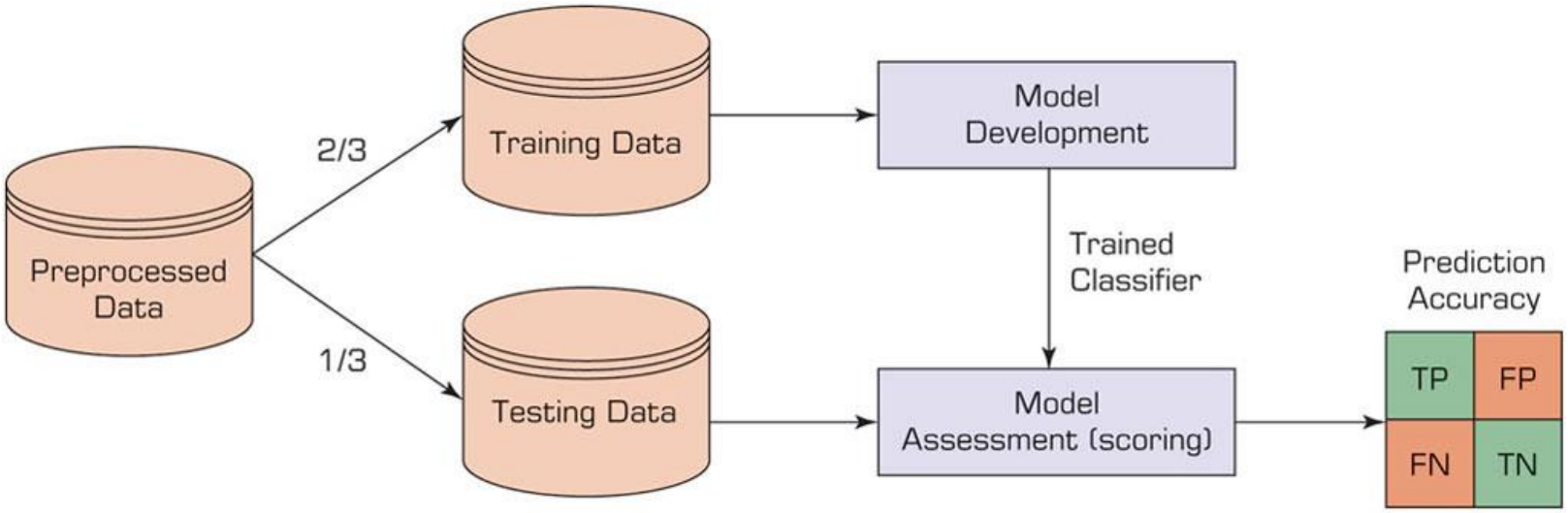
Recall (% of positive tuples that the classifier label as positive) TP/(TP+FN) = 90/300 = 30.00%

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Estimation Methodologies for Classification: Single/Simple Split

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets: training (~70%) and testing (30%)

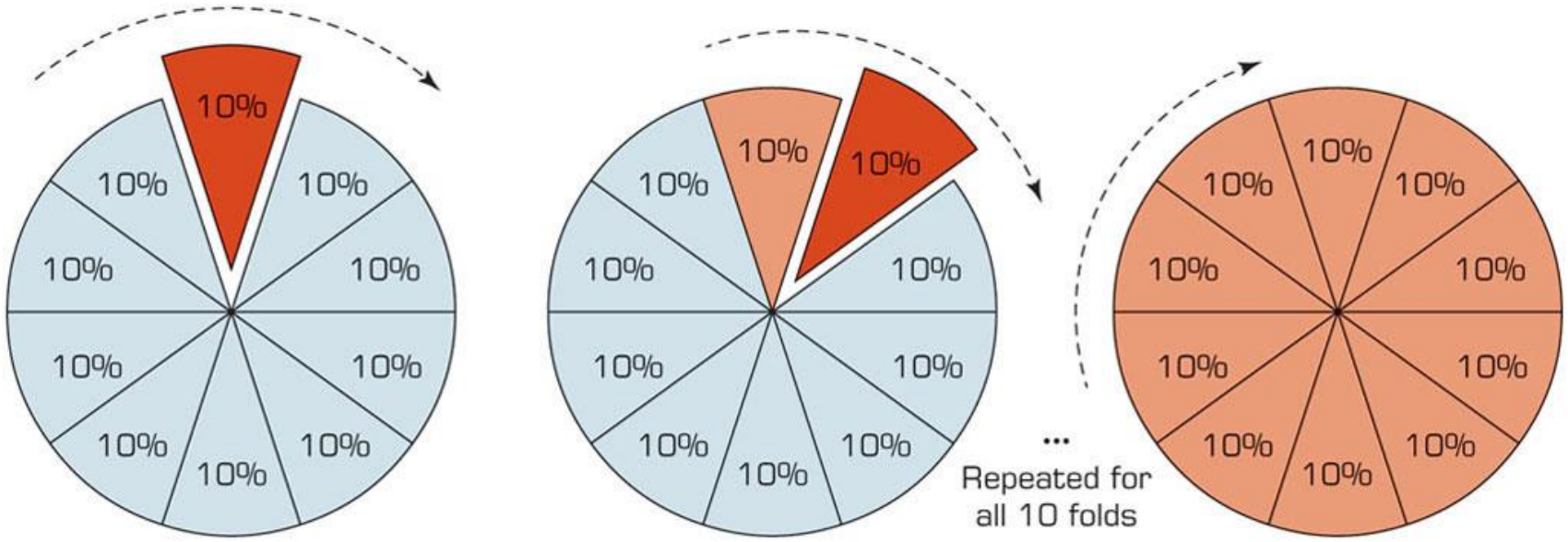


Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Estimation Methodologies for Classification: *k*-Fold Cross Validation

- Data is split into *k* mutual subsets and *k* number training/testing experiments are conducted

Figure 4.10 A Graphical Depiction of *k*-Fold Cross-Validation.



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Estimation Methodologies for Classification

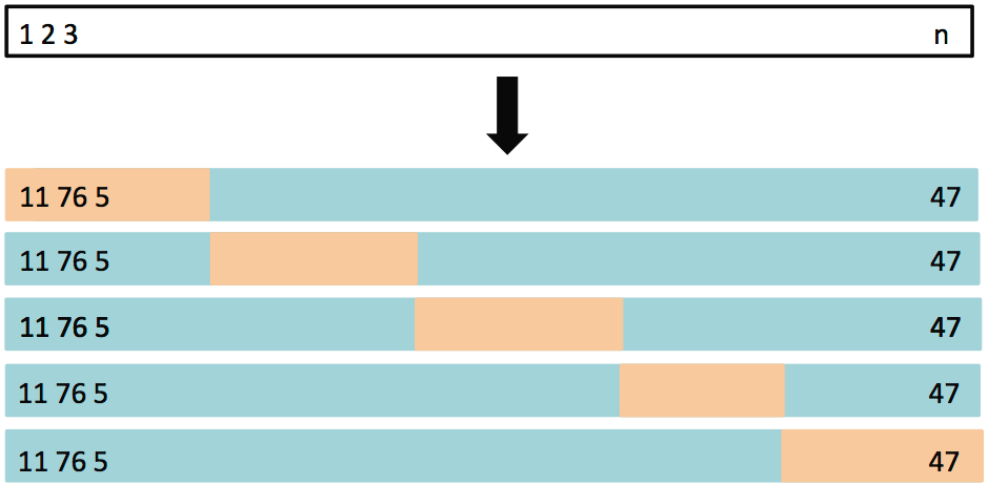
- *k*-Fold Cross Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets of approximate size
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Estimation Methodologies for Classification: *k*-Fold Cross Validation

We randomly divide the data set of into *K* folds (typically *K* = 5 or 10).



The first fold is treated as a test set, and the method is fit on the remaining *K* – 1 folds (training sets)

Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Estimation Methodologies for Classification: *k*-Fold Cross Validation

The cross-validation of the overall accuracy of a model is calculated by averaging the k individual accuracy measures:

$$\mathrm{CVA} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{A}_i$$

CVA stands for cross-validation accuracy, K is the number of folds and A is the accuracy measure

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

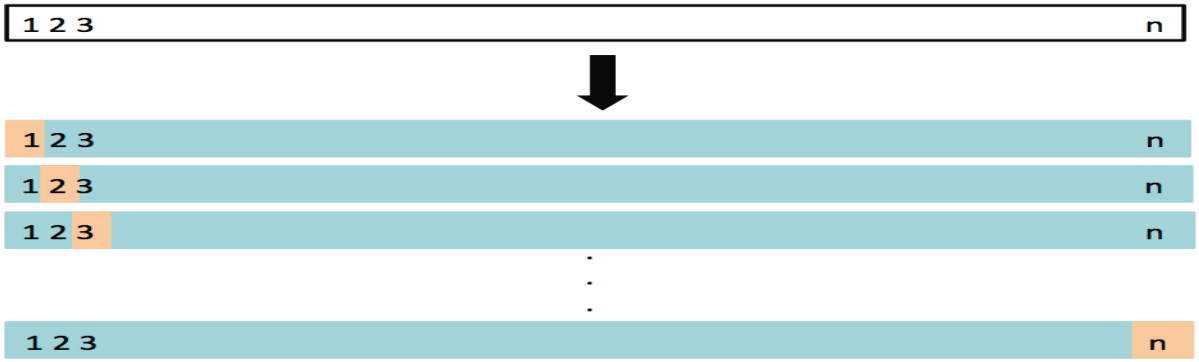
Additional Estimation Methodologies for Classification

- Leave-one-out
  - Similar to *k*-fold where *k* = number of samples
- Bootstrapping
  - Random sampling with replacement
- Jackknifing
  - Similar to leave-one-out
- Area Under the ROC Curve (AUC)
  - ROC: receiver operating characteristics (a term borrowed from radar image processing)

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Leave-One-Out

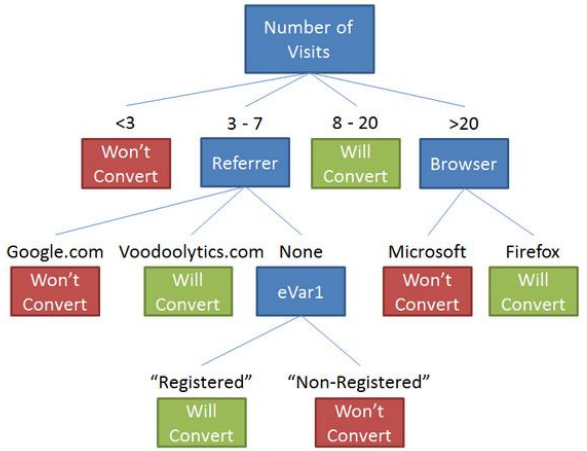
- Split the entire data set of size *n* into:
  - Blue = training data set
  - Beige = test data set
- Fit the model using the training data set
- Evaluate the model using test set and compute the corresponding CVA.
- Repeat this process *n* times, producing *n* CVAs. The average of these *n CVs e*stimates the overall CVA.

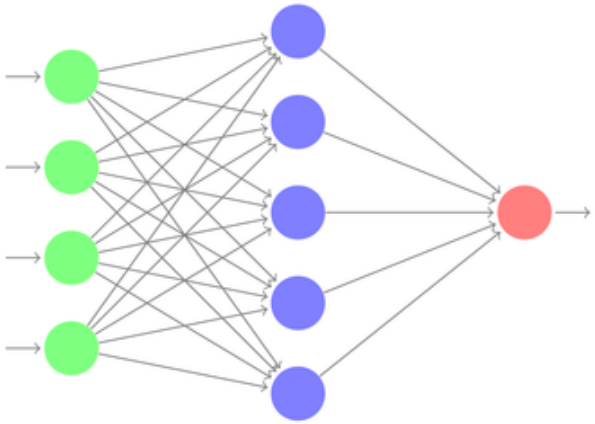


Pearson
Rights Reserved.

Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets





Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Cluster Analysis for Data Mining (1 of 4)

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- There is not an output/target variable
- In marketing, it is also known as segmentation

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Cluster Analysis for Data Mining (2 of 4)

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Cluster Analysis for Data Mining (3 of 4)

- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on.
  - Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms
- How many clusters?

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Cluster Analysis for Data Mining (4 of 4)

- *k*-Means Clustering Algorithm
  - *k*: pre-determined number of clusters
  - Algorithm (Step 0: determine value of *k*)

  Step 1: Randomly generate *k* random points as initial cluster centers (centroids).

  Step 2: Assign each point to the nearest cluster center.

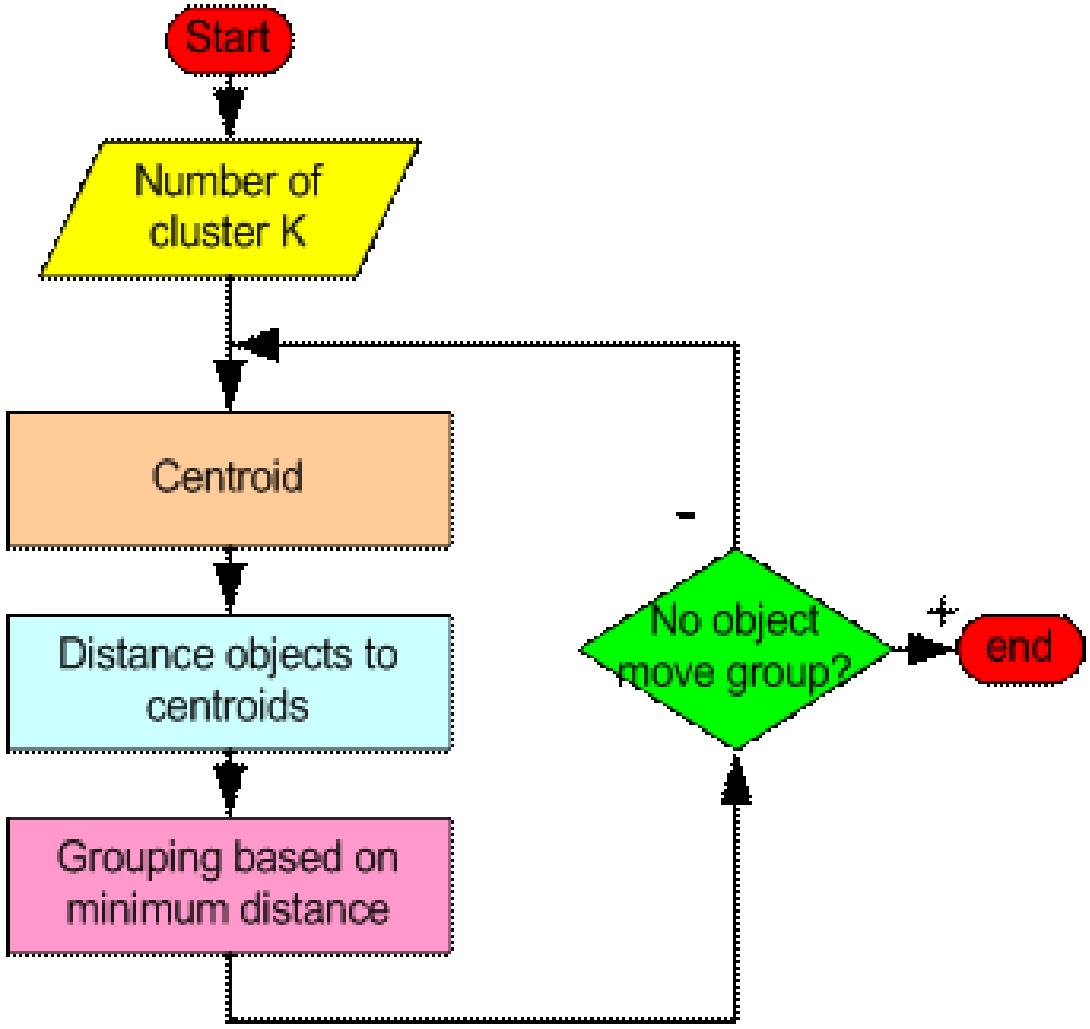
  Step 3: Re-compute the new cluster centers.

  Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Cluster Analysis for Data Mining



Pearson

Copyright © 2021 by Pearson Education, Inc. All Rights Reserved

An Example of K-Means Clustering

We have 4 medicines as our training data points object and each medicine has 2 attributes. Each attribute represents coordinate of the object. We have to determine which medicines belong to cluster 1 and which medicines belong to the other cluster.*

Object/medicine	Attribute 1 (X): weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

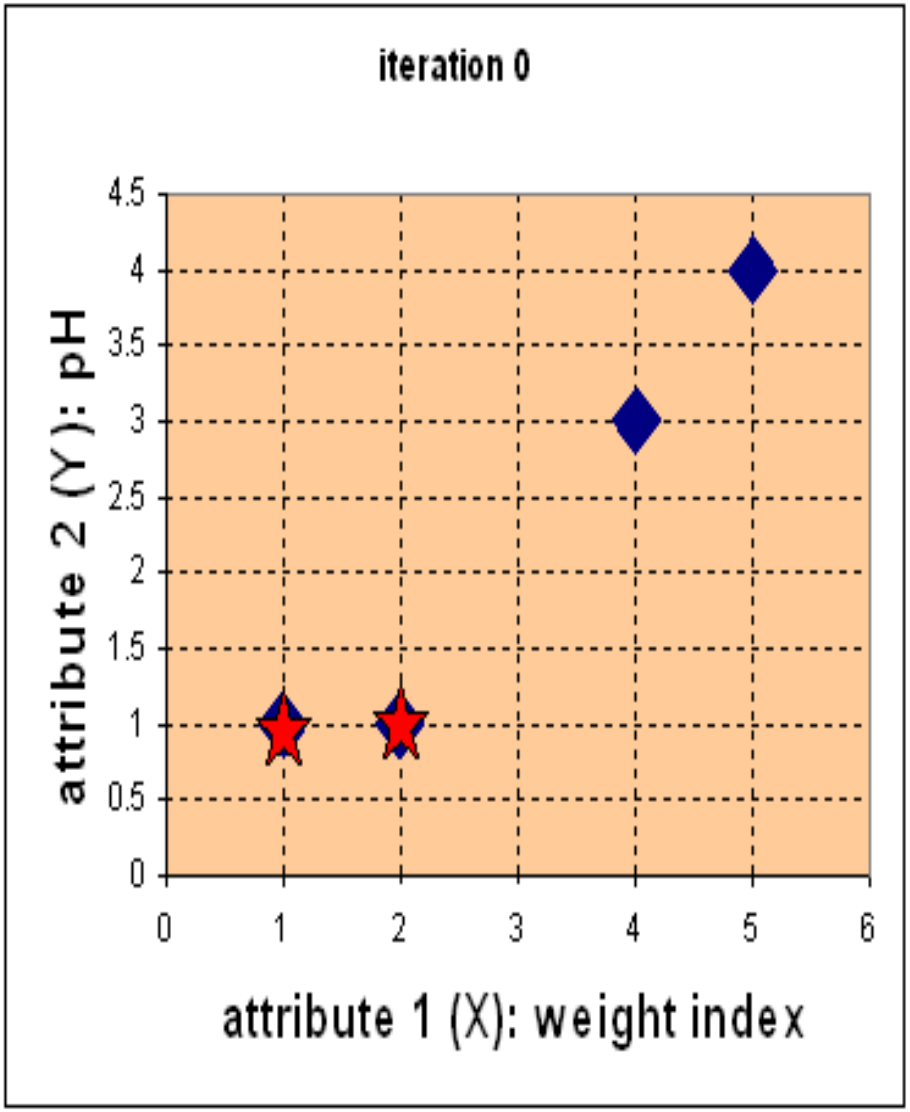
***https://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm**

Pearson

Copyright © 2021 by Pearson Education, Inc. All Rights Reserved

Step 1:

- **Initial value of centroids**: Suppose we use medicine A and medicine B as the first centroids.
- Let c_1 and c_2 denote the coordinate of the centroids, then c_1=(1,1) and c_2=(2,1)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

- **Objects-Centroids distance** : Calculate the distance between cluster centroid to each object, using Euclidean distance.
- Distance D from the centroid

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

- Each column in the distance matrix symbolizes the object.
- The first row of the distance matrix corresponds to the distance of each object to the first centroid and the second row is the distance of each object to the second centroid.
- For example, distance from medicine C = (4, 3) to the first centroid C_1=(1,1) is $\sqrt{(4-1)^2+(3-1)^2} = 3.61$ and its distance to the second centroid C_2=(2,1) is $\sqrt{(4-2)^2+(3-1)^2} = 2.83$ etc.

Pearson

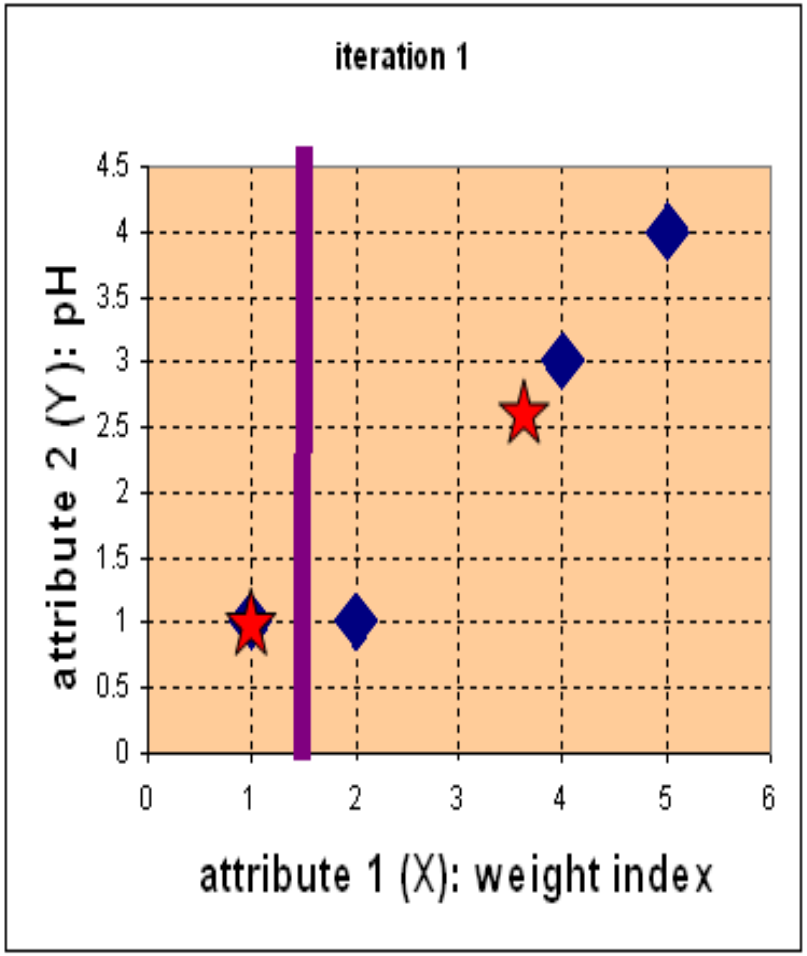
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Step 2

- **Objects clustering** : We assign each object based on the minimum distance.
- Medicine A is assigned to group 1, medicine B to group 2, medicine C to group 2 and medicine D to group 2.
- The elements of Group matrix below is 1 if and only if the object is assigned to that group.

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$



Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Step 3

Iteration-1, Objects-Centroids distances:

- The next step is to compute the distance of all objects to the new centroids.
- Similar to step 2, we have distance matrix at iteration 1 is

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

- **Iteration-1, Objects clustering**: Based on the new distance matrix, we move the medicine B to Group 1 while all the other objects remain. The Group matrix is shown below

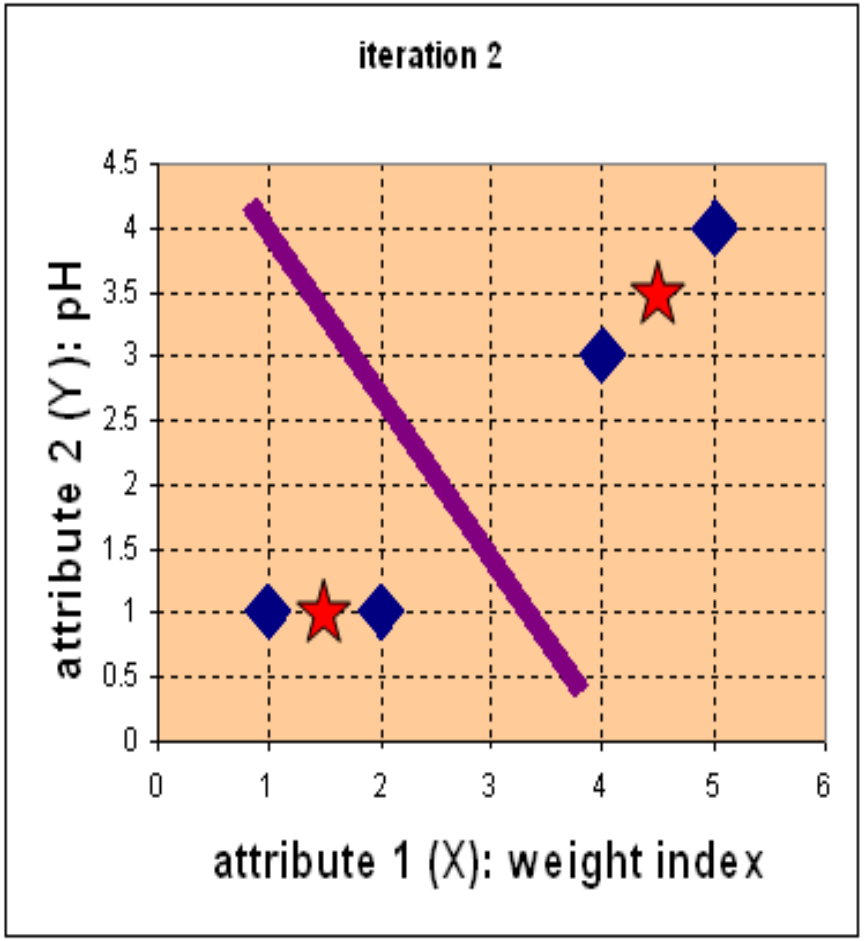
$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

- **Iteration 2, determine centroids:** Now calculate the new centroids coordinate based on the clustering of previous iteration. Group1 and group 2 both has two members, thus the new centroids are

$$c_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\tfrac{1}{2}, 1) \quad \text{and}$$

$$c_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$



Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

- **<u>Iteration-2, Objects-Centroids distances</u>** :
- Repeat step 3 again, we have new distance matrix at iteration 2 as:

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) & group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

- **Iteration-2, Objects clustering:**
- Assign each object based on the minimum distance.

$$\mathbf{G}^2 = \begin{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{matrix} group-1 \\ group-2 \end{matrix} \\ A \quad B \quad C \quad D & \end{matrix}$$

- We obtain result that $\mathbf{G}^2 = \mathbf{G}^1$. Comparing the grouping of last iteration and this iteration reveals that the objects does not move group anymore.
- Thus, the computation of the k-mean clustering has reached its stability and no more iteration is needed..

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

We get the final grouping as the results as:

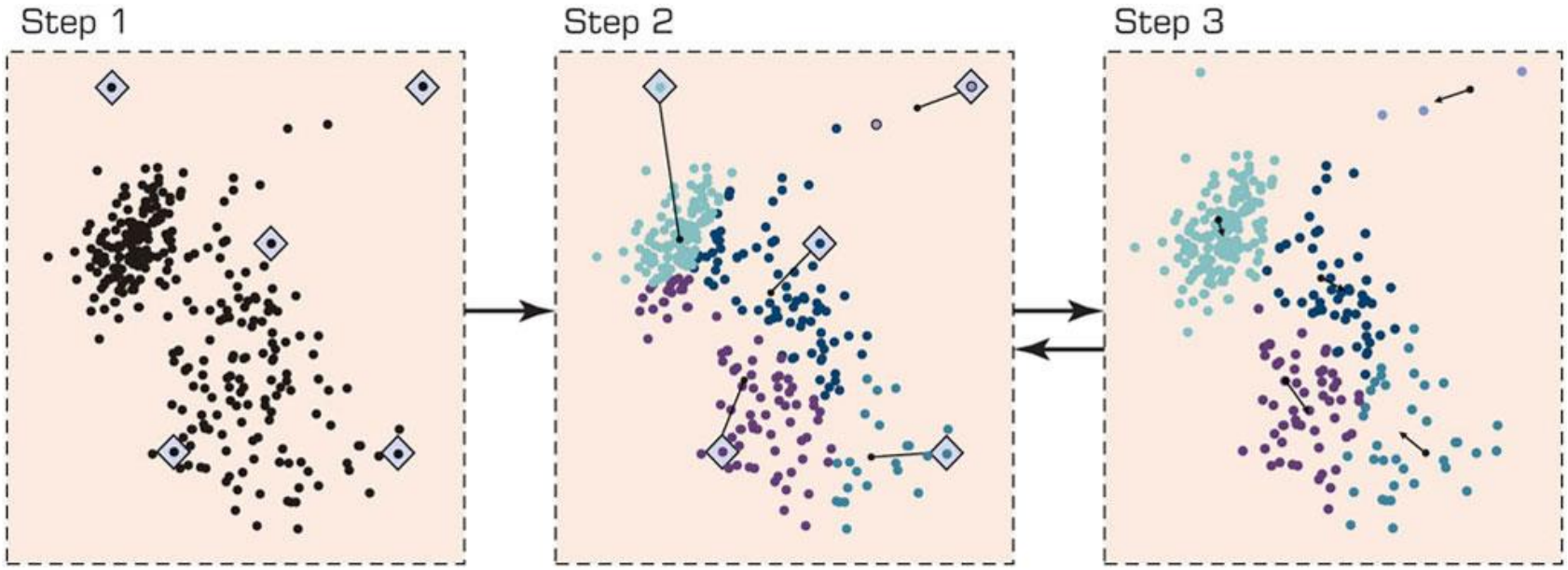
Object	Attribute 1(X): weight index	Attribute 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

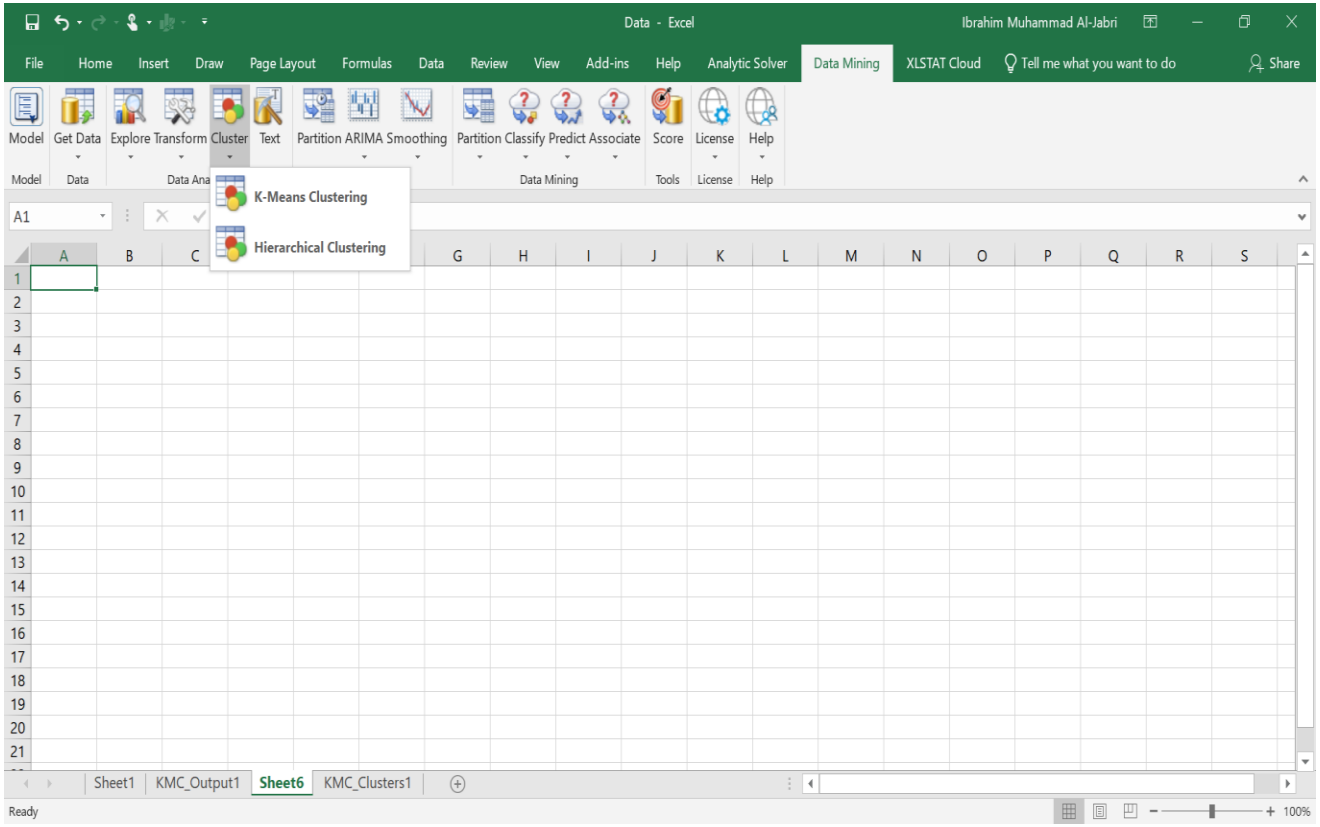
Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm

Figure 4.13 A Graphical Illustration of the Steps in the *k*-Means Algorithm.



Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

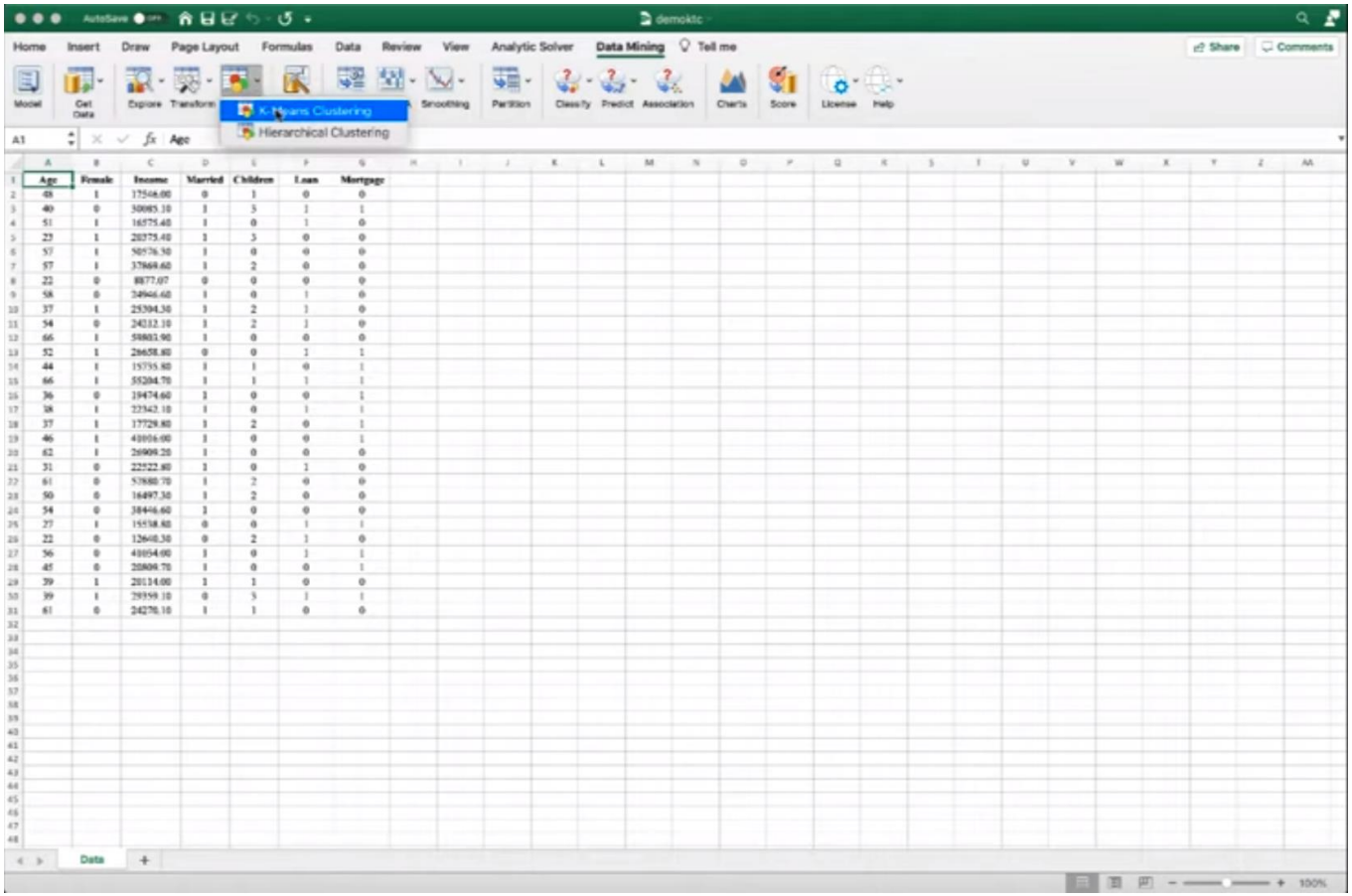
Data Mining in Excel (1 of 5)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

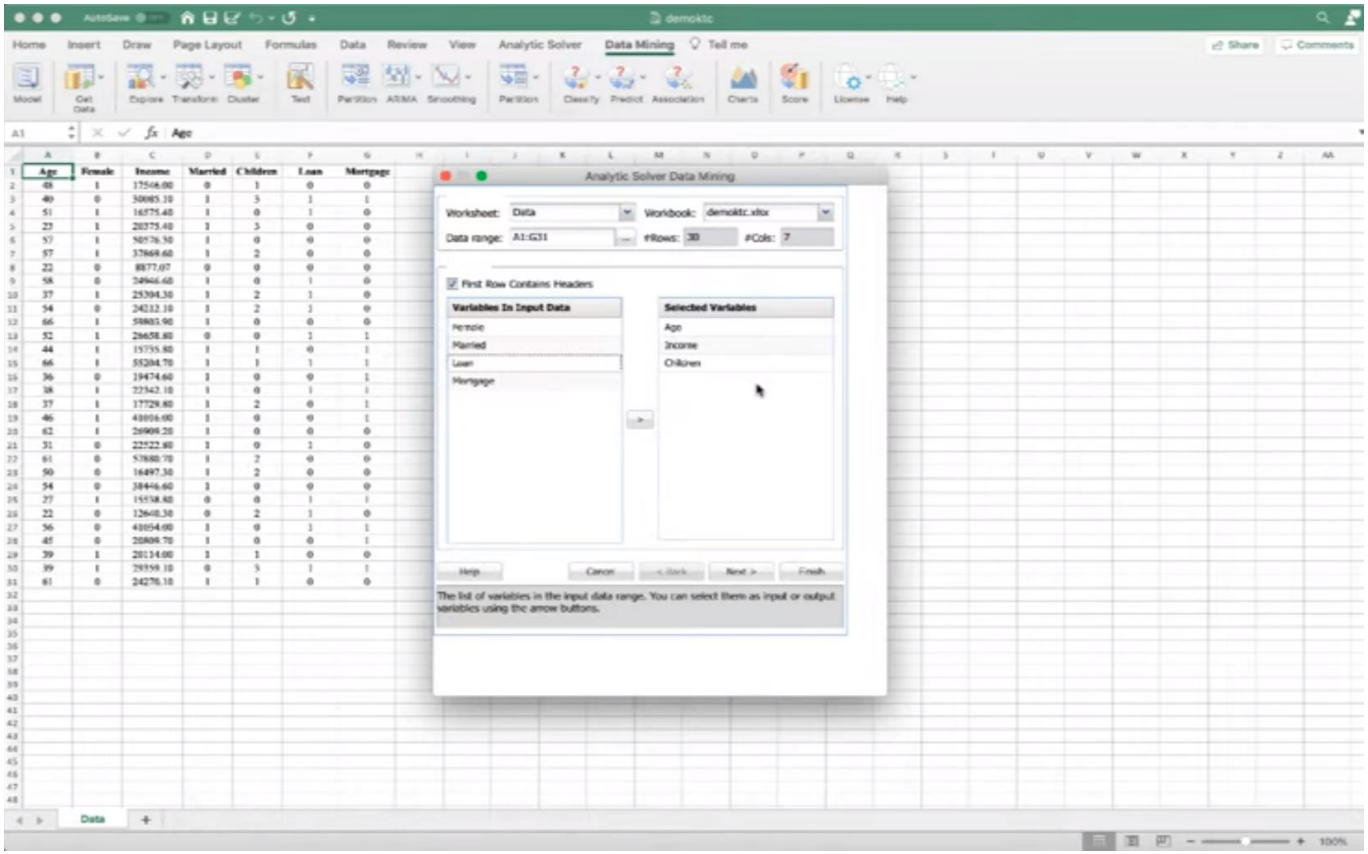
Data Mining in Excel (2 of 5)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

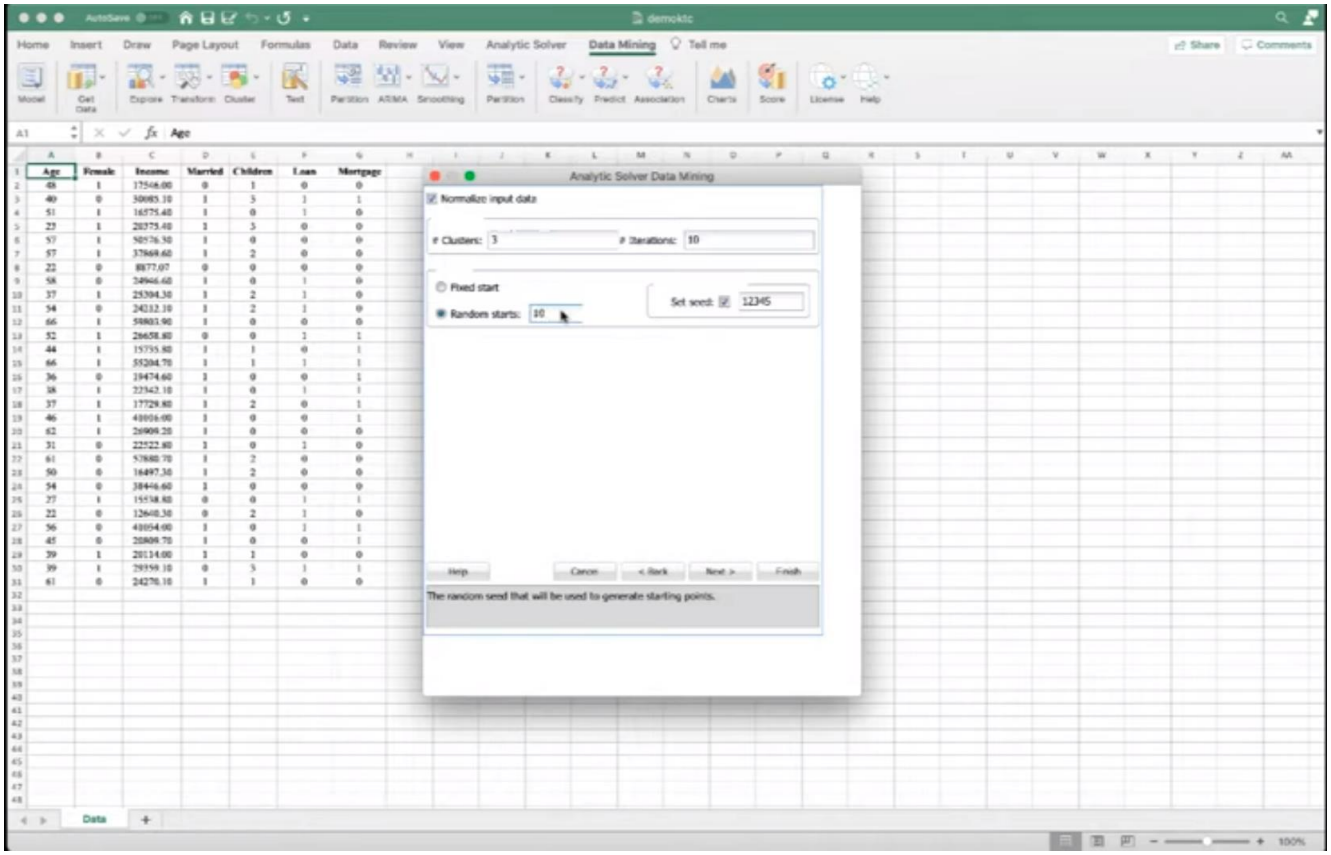
Data Mining in Excel (3 of 5)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

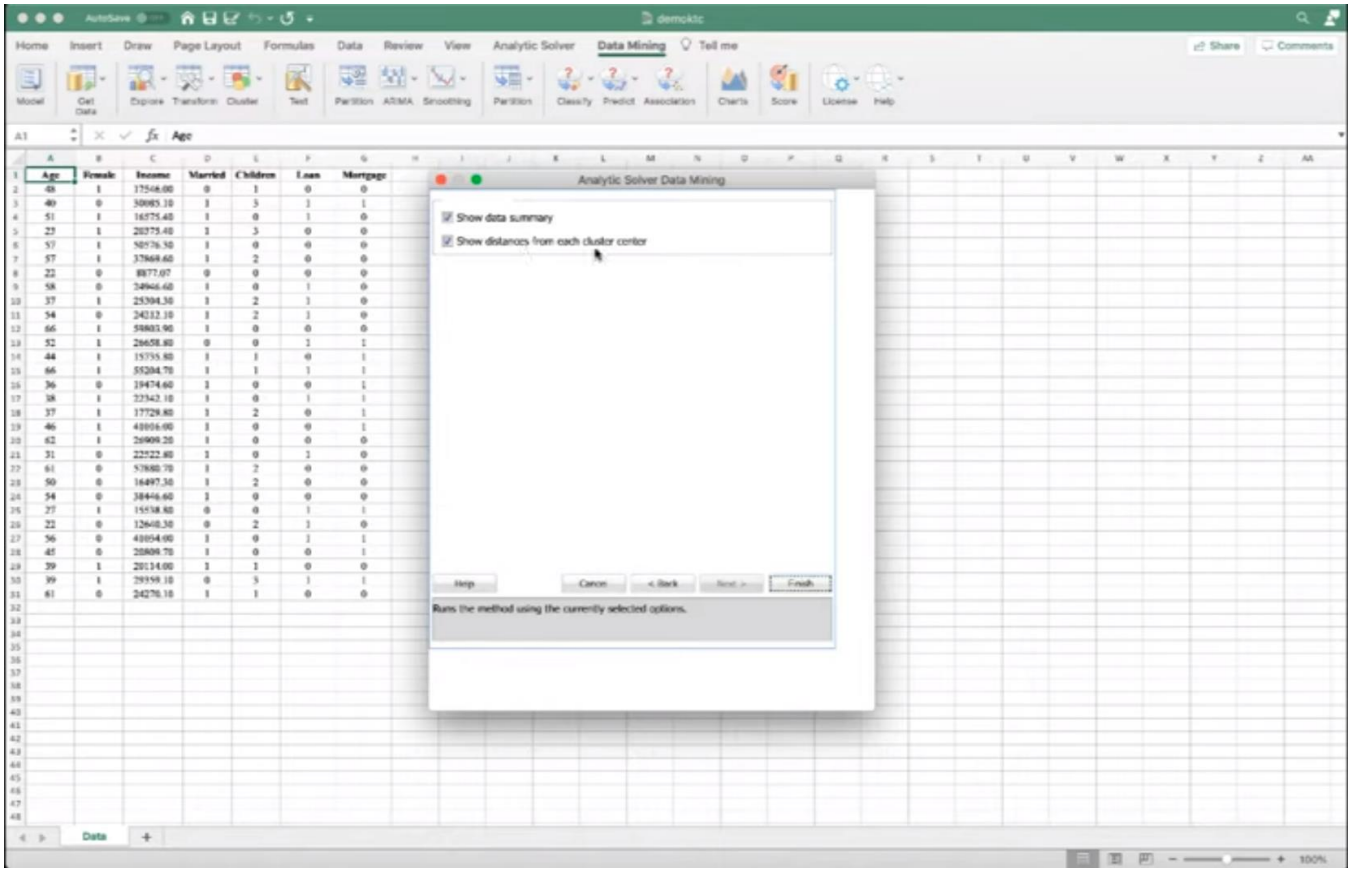
Data Mining in Excel (4 of 5)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining in Excel (5 of 5)



Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining (1 of 6)

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Supermarkets place items next to each other depending on the customers buying habits (purchased together)

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining (2 of 6)

- Input: the simple point-of-sale transaction data
- Output: Most frequent affinities among items
- Example: according to the transaction data…

  "Customer who bought a lap-top computer and a virus protection software, also bought extended service plan 70 percent of the time."

- How do you use such a pattern/knowledge?
  - Put the items next to each other
  - Promote the items as a package
  - Place items far apart from each other!

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining (3 of 6)

- A representative applications of association rule mining include
  - In business: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - In medicine: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)
  - …

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining (4 of 6)

- Several algorithms are developed for discovering (identifying) association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining (5 of 6)

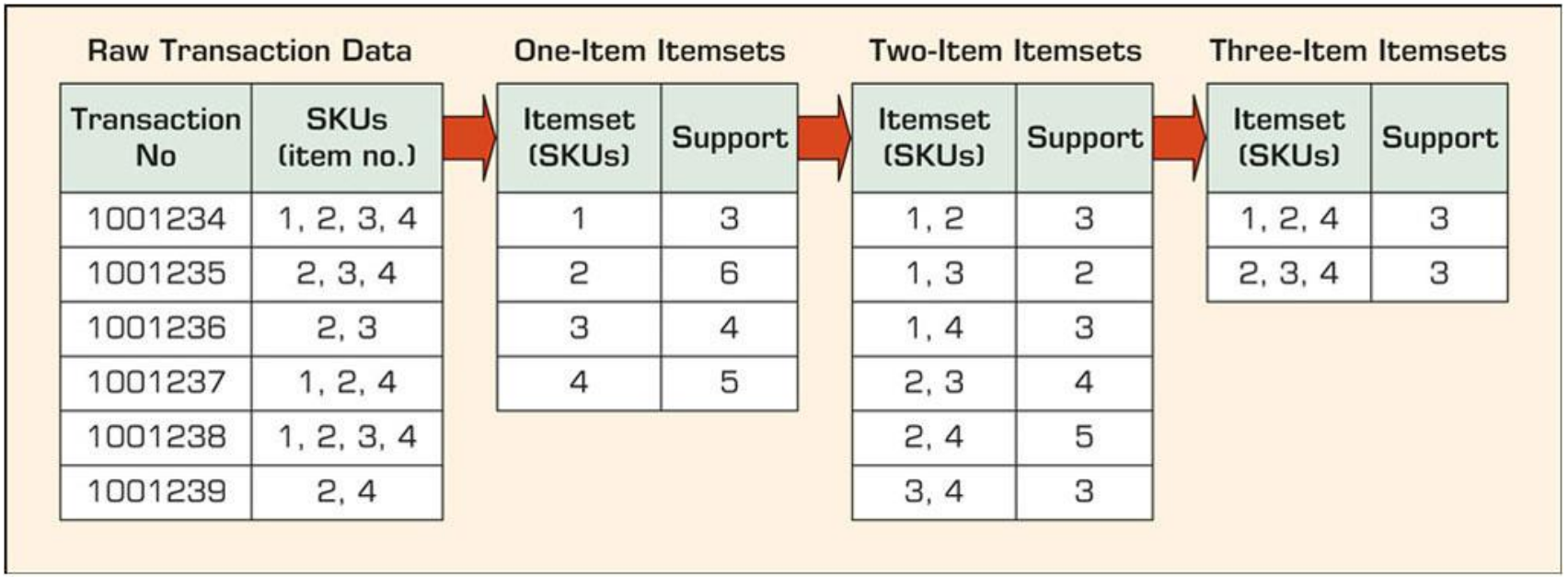
- Apriori Algorithm
  - Finds subsets that are common to at least a minimum number of the itemsets
  - Uses a bottom-up approach
    - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    - groups of candidates at each level are tested against the data for minimum support. (*see the figure*) → --

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Association Rule Mining Apriori Algorithm (6 of 6)

Figure 4.14 Identification of Frequent Itemsets in the Apriori Algorithm



Raw Transaction Data

Transaction No	SKUs (item no.)
1001234	1, 2, 3, 4
1001235	2, 3, 4
1001236	2, 3
1001237	1, 2, 4
1001238	1, 2, 3, 4
1001239	2, 4

One-Item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-Item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-Item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

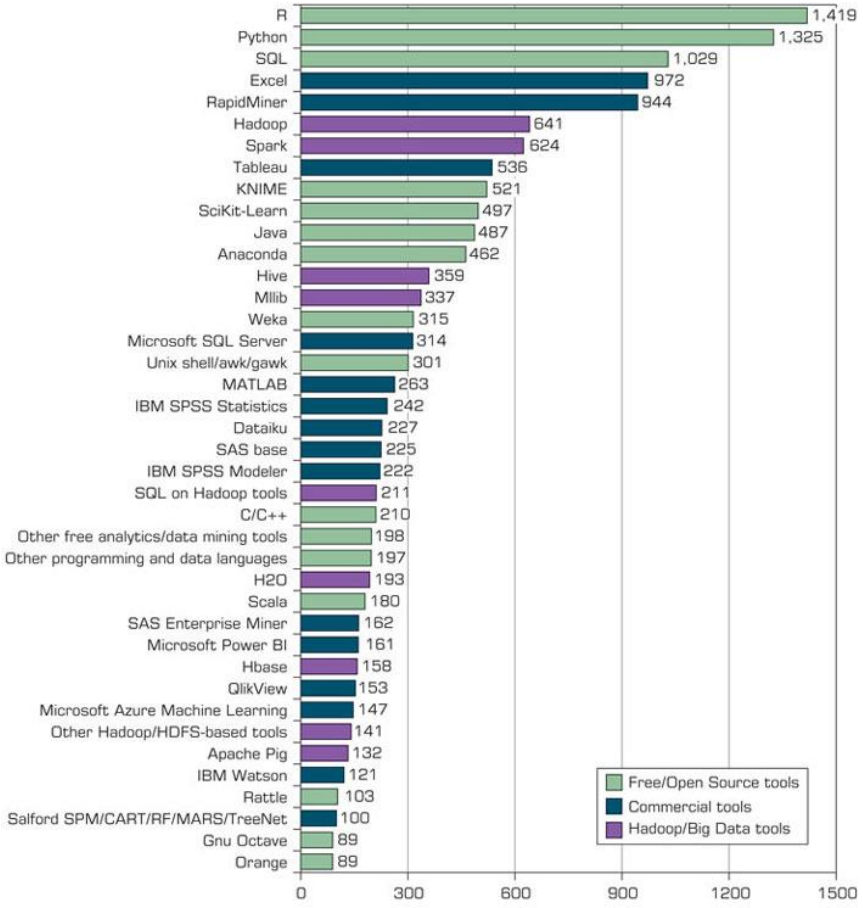
Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Software Tools

Figure 4.15 Popular Data Mining Software Tools (Poll Results).

- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS Enterprise Miner
  - Statistica - Dell/Statsoft
  - … many more
- Free and/or Open Source
  - KNIME
  - RapidMiner
  - Weka
  - R, …



Source: Used with permission from **KDnuggets.com**.

Pearson
Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Data Mining Mistakes

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Beginning without the end in mind.
4. Not leaving insufficient time for data acquisition, selection and preparation
5. Looking only at aggregated results and not at individual records/predictions
6. … 10 more mistakes… in your book

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.

Q & A

Pearson

Copyright © 2021 Pearson Education Ltd. All Rights Reserved.