

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 7



Predictive Analytics : Data Mining Process, Methods, and Algorithms

Lecture 7

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

1.1 Data Mining Process

1.2 Learn different methods and algorithms of data mining

1.3 Data Mining Software Tools

1.4 Python libraries

02

Data mining process and methods

Data mining tasks can be classified into three main categories: **prediction, association, and clustering**

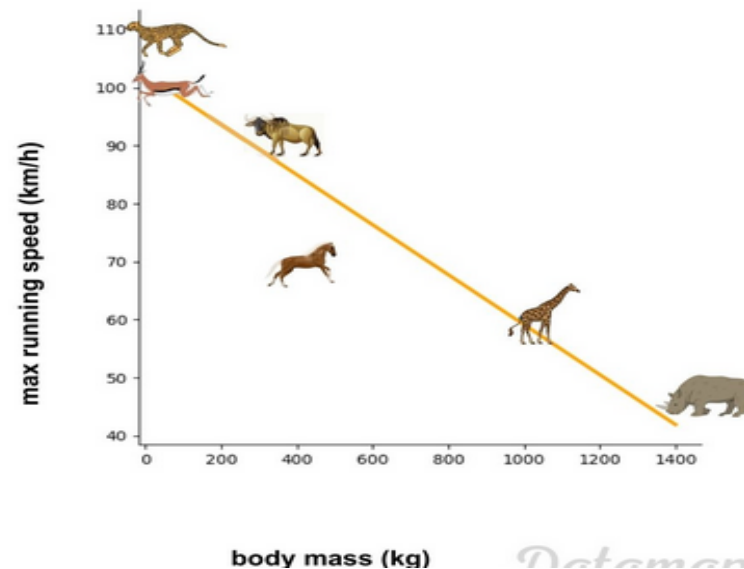
Prediction is commonly referred to as **the act of telling about the future**. It differs from simple guessing by taking into account the experiences, opinions, and other relevant information in conducting the task of foretelling. A term that is commonly associated with prediction is *forecasting*.

In data mining terminology, ***prediction* and *forecasting* are used synonymously**, and the term *prediction* is used as the common representation of the act. **Regression (where the predicted thing, such as tomorrow's temperature, is a real number, such as "65°F").**

03

Data mining process and methods

Animal	body mass (kg)	max running speed (km/h)
rhino	1400	45
horse	400	70
pronghorn	50	100
giraffe	1000	60
wildebeest	300	90
cheetah	60	110



linear regression example

Classification is perhaps the most common of all data mining tasks. The objective of classification is to analyze the historical data stored in a database and automatically generate a model that can predict future behavior.

04

Data mining process

Classification: This induced model consists of generalizations over the records of a training data set, which help distinguish predefined classes.

The hope is that the model **can then be used to predict the classes of other unclassified records and, more important, to accurately predict actual future events.**

Common classification tools include **neural networks and decision trees** (from machine learning), **logistic regression and discriminant analysis** (from traditional statistics), and emerging tools such as support vector machines (SVMs), and genetic algorithms.

Neural networks involve the development of mathematical structures (somewhat resembling the biological neural networks in the human brain) that have the capability to learn from past experiences presented in the form of well-structured data sets.

05

Data mining process

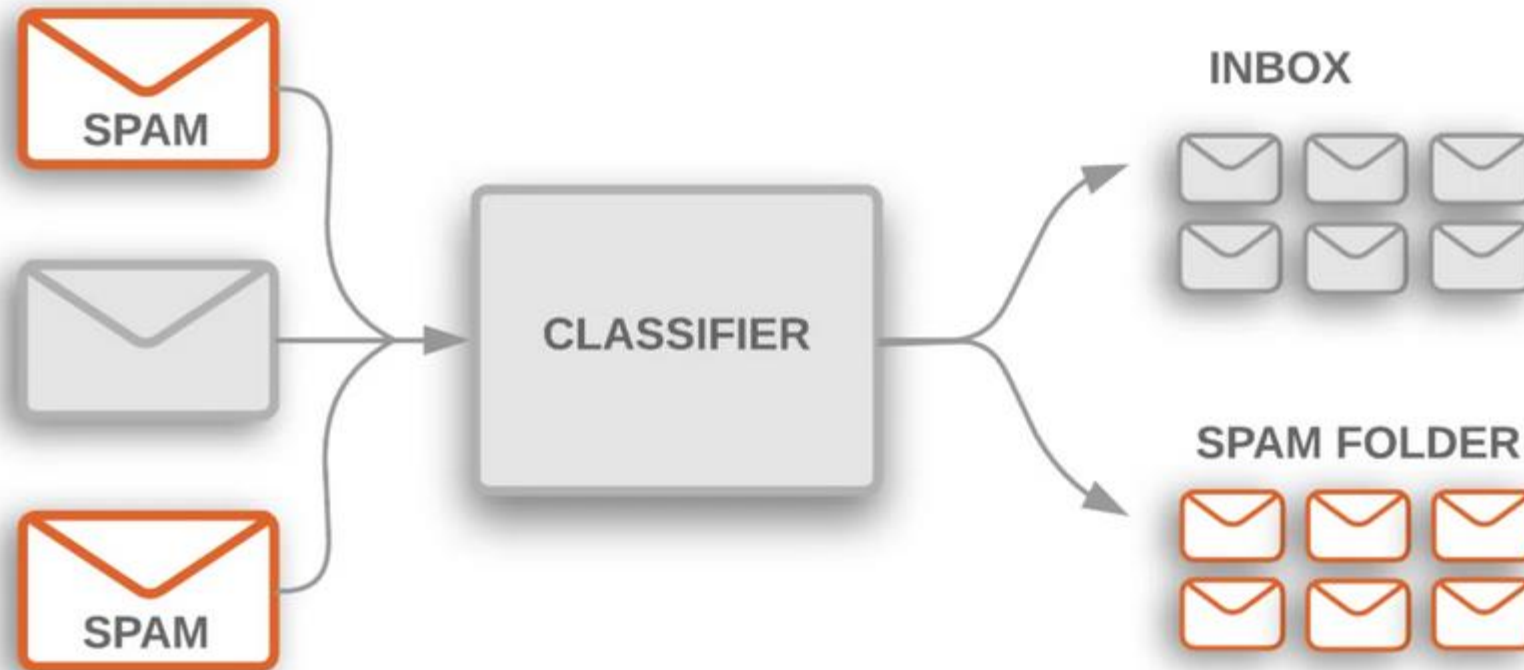
For example, one could use classification to predict whether the weather on a particular day will be “sunny,” “rainy,” or “cloudy.”

Popular classification tasks include credit approval (i.e., good or bad credit risk), store location (e.g., good, moderate, bad), target marketing (e.g., likely customer, no hope), fraud detection (i.e., yes/no), and telecommunication (e.g., likely to turn to another phone company, yes/no).

If what is being predicted is a class label (e.g., “sunny,” “rainy,” or “cloudy”), the prediction problem is called a classification, whereas if it is a numeric value (e.g., temperature, such as 68°F), the prediction problem is called a regression.

06

Data mining process



Example of
Classification

07

Data mining process

The most common two-step methodology of classification-type prediction involves **model development/training and model testing/deployment**. In the model **development** phase, **a collection of input data, including the actual class labels, is used**.

After a model has been trained, the model is tested **against the holdout sample for accuracy assessment** and eventually deployed for actual use where it is to predict classes of new data instances (where the class label is unknown).

Several factors are considered in assessing the model, including the following:

Predictive accuracy.: The model's ability to correctly predict the class label of new or previously unseen data. Prediction accuracy is the most commonly used assessment factor for classification models.

08

Data mining process

To compute this measure, **actual class labels of a test data** set are matched against the class labels predicted by the model. The accuracy can then be computed **as the *accuracy rate***, which is the **percentage of test data set samples correctly classified by the model**.

In classification problems, the primary source for accuracy estimation is the **confusion matrix** (also called **a classification matrix or a contingency table**).

Figure 7.1 shows a confusion matrix for a two-class classification problem. **The numbers along the diagonal from the upper left to the lower right represent correct decisions**, and **the numbers outside this diagonal represent the errors**.

09

Data mining process

		True/Observed Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Figure 7.1 A Simple Confusion Matrix

Table 7.1 provides equations for common accuracy metrics for classification models.

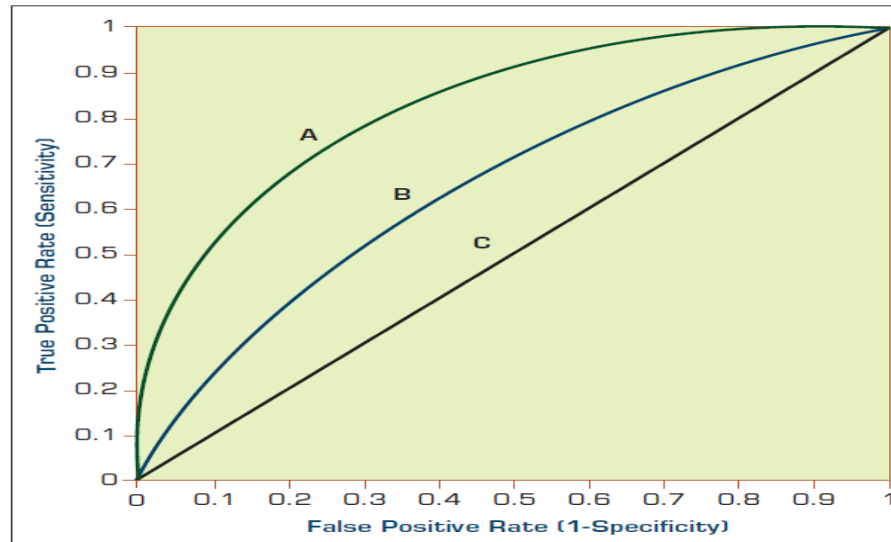
Metric	Description
$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	The ratio of correctly classified instances (positives and negative) divided by the total numbers of instances
$\text{True Positive Rate} = \frac{TP}{TP + FN}$	(a.k.a. Sensitivity) The ratio of correctly classified positives divided by the total positive count (i.e., hit rate or recall)
$\text{True Negative Rate} = \frac{TN}{TN + FP}$	(a.k.a. Specificity) The ratio of correctly classified negatives divided by the total negative count (i.e., false alarm rate)
$\text{Precision} = \frac{TP}{TP + FP}$	The ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified positives
$\text{Recall} = \frac{TP}{TP + FN}$	Ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified negatives

11

Data mining process

Area under the ROC curve: The area under the ROC curve is a graphical assessment technique where the true positive rate is plotted on the y-axis and the false positive rate is plotted on the x-axis.

The area under the ROC curve determines the accuracy measure of a classifier: A value of 1 indicates a perfect classifier, whereas 0.5 indicates no better than random chance. Figure 7.2, A has a better classification performance than B, whereas C is not any better than the random chance of flipping a coin.



12

Data mining process

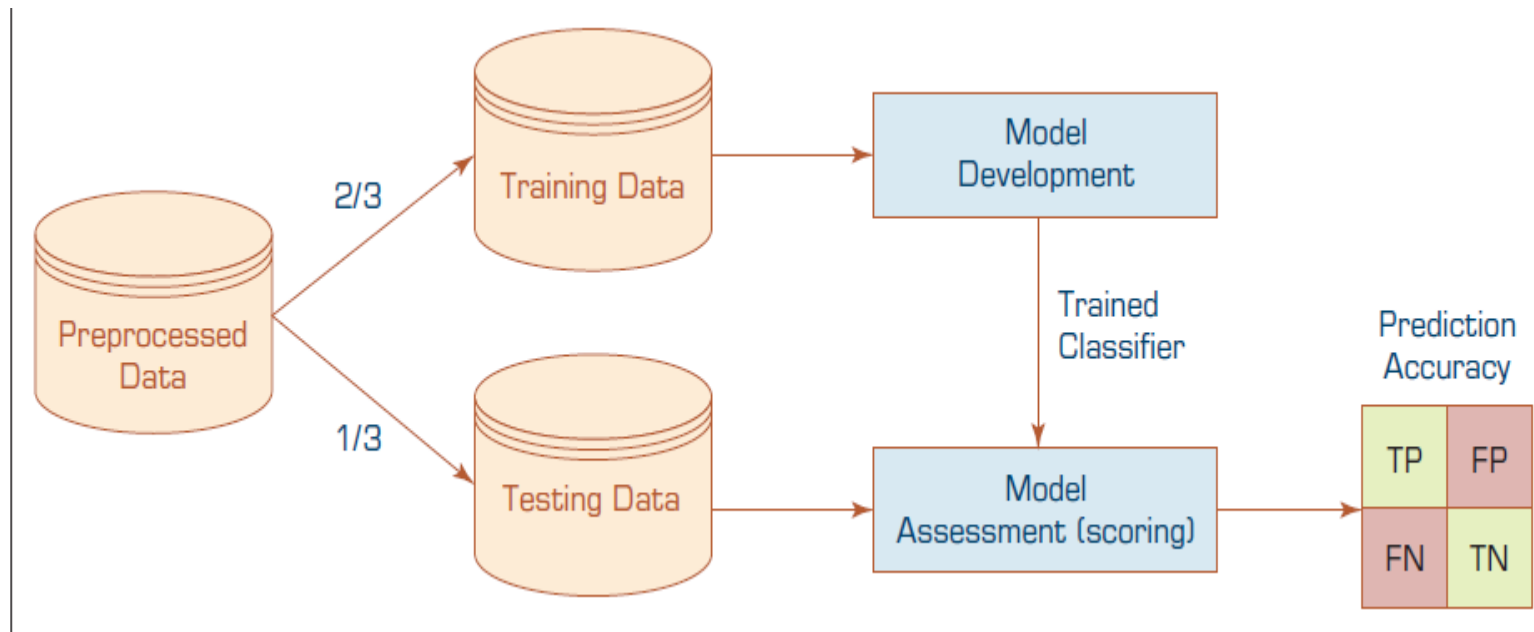
Estimating the accuracy of a classification model (or classifier) induced by a supervised learning algorithm is important for the following two reasons: First, it can be used to estimate its future prediction accuracy, which could imply the level of confidence one should have in the classifier's output in the prediction system. Second, it can be used for choosing a classifier from a given set (identifying the “best” classification model among the many trained). The following are among the most popular estimation methodologies used for classification-type data mining models.

The simple split (or holdout or test sample estimation) partitions the data into two mutually exclusive subsets called a *training set* and a *test set* (or *holdoutset*).

It is common to designate two-thirds of the data as the training set and the remaining one-third as the test set. The training set is used by the inducer (model builder), and the built classifier is then tested on the test set.

13

Data mining process



Data Splitting

14

Data mining process

Clustering partitions a collection of things (e.g., objects, events, presented in a structured data set) into segments (or natural groupings) whose members share similar characteristics. **Unlike in classification, in clustering the class labels are unknown.**

The objective is to sort cases (e.g., people, things, events) into groups, or clusters, so that the degree of association is strong among members of the same cluster and weak among members of different clusters.

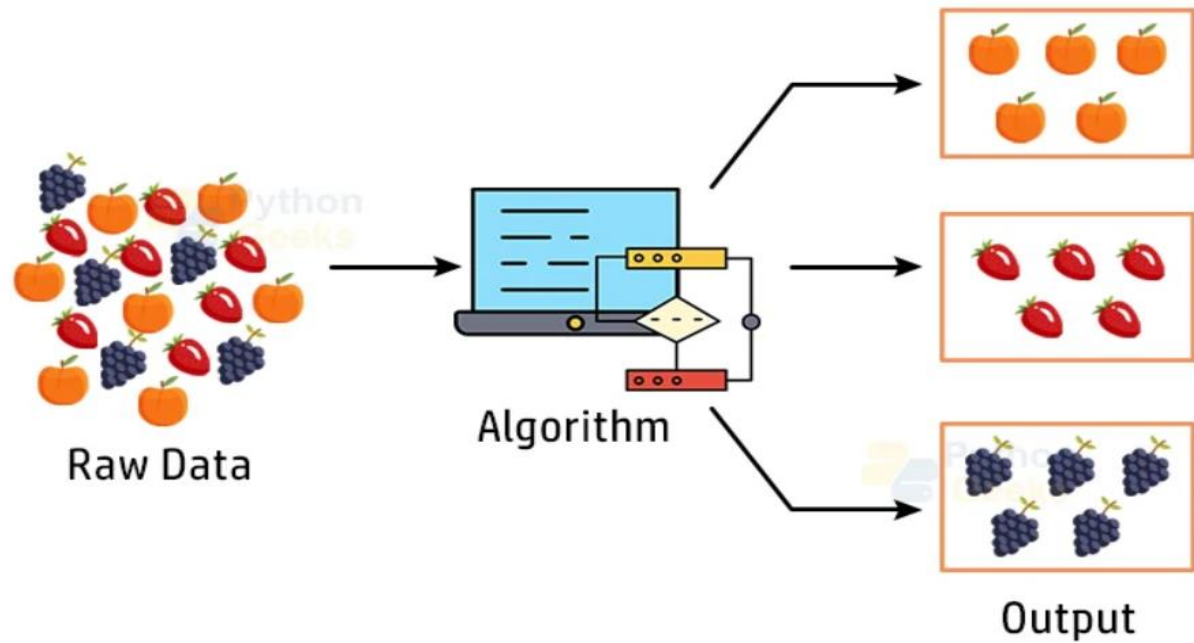
Each cluster describes the class to which its members belong.

Analysis Methods Cluster analysis may be based on one or more of the following general methods:

- Statistical methods such as k-means or k-modes
- Neural networks (with the architecture called self-organizing map)
- Fuzzy logic (e.g., fuzzy c-means algorithm)
- Genetic algorithms

15

Data mining process



Clustering Example

16

Data mining process and methods

Data mining researchers and practitioners have proposed several processes (workflows or simple step-by-step approaches) to maximize the chances of success in conducting data mining projects.

One such standardized process, arguably the most popular one, Cross-Industry Standard Process for Data Mining—CRISP-DM.

Figure 7.3 illustrates this proposed process, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfies the specific business need.

17

Data mining process and methods

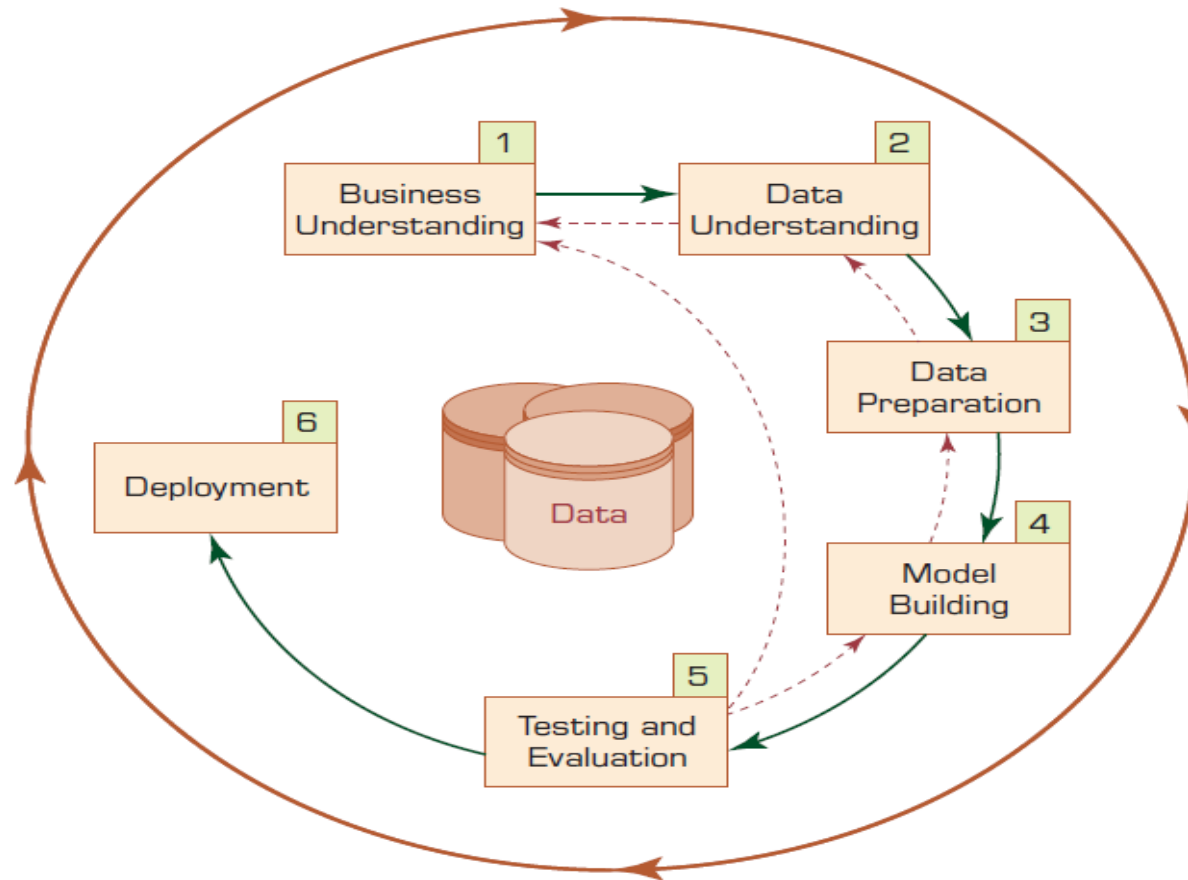


Figure 7.3 The Six-Step CRISP-DM Data Mining Process

18

Data mining process and methods

Even though these steps are sequential in nature, there is usually a great deal of backtracking.

Because the data mining is driven by experience and experimentation, depending on the problem situation and the knowledge/ experience of the analyst, the whole process can be very iterative (i.e., one should expect to go back and forth through the steps quite a few times) and time-consuming. Because later steps are built on the outcomes of the former ones, one should pay extra attention to the earlier steps in order not to put the whole study on an incorrect path from the onset.

Step 1: Business Understanding :

The key element of any data mining study is to know what the study is for. Answering such a question begins with a thorough understanding of the managerial need for new knowledge and an explicit specification of the business objective regarding the study to be conducted.

19

Data mining process and methods

Specific goals such as “What are the common characteristics of the customers we have lost to our competitors recently?” or “What are typical profiles of our customers, and how much value does each of them provide to us?” are needed. Then a project plan for finding such knowledge is developed that specifies the people responsible for collecting the data, analyzing the data, and reporting the findings.

Step 2: Data Understanding:

A data mining study is specific to addressing a well-defined business task, and different business tasks require different sets of data. Following the business understanding, the main activity of the data mining process is to identify the relevant data from many available databases. Some key points must be considered in the data identification and selection phase. First and foremost, the analyst should be clear and concise about the description of the data mining task so that the most relevant data can be identified.

20

Data mining process and methods

For example, a retail data mining project may seek to **identify spending behaviors of female shoppers who purchase seasonal clothes based on their demographics, credit card transactions attributes**. Furthermore, the analyst should build an intimate understanding of the data sources (e.g., where the relevant data are stored and in what form; what the process of collecting the data is—automated versus manual; who the collectors of the data are and how often the data are updated) and the variables (e.g., What are the most relevant variables?)

To better understand the data, **the analyst often uses a variety of statistical and graphical techniques**.

A careful **identification and selection of data sources and the most relevant variables can make it easier** for data mining algorithms to quickly discover useful knowledge patterns. **Data sources for data selection can vary**.

21

Data mining process and methods

Traditionally, data sources for business applications include **demographic data** (such as income, education, number of households, and age), **sociographic data** (such as hobby, club membership, and entertainment), **transactional data** (sales record, credit card spending, issued checks), and so on. Nowadays, data sources also use **external (open or commercial) data repositories**, social media, and machine-generated data.

Step 3: Data Preparation:

The purpose of data preparation (more commonly called data preprocessing) is to take the data identified in the previous step and prepare it for analysis by data mining methods.

Compared to the other steps in CRISP-DM, **data preprocessing consumes the most time and effort**; most believe that this step **accounts for roughly 80% of the total time spent on a data mining project**.

22

Data mining process and methods

Step 4: Model Building:

In this step, **various modeling techniques** are selected and applied to an already prepared data set to address the specific business need.

The model-building step also encompasses the assessment and comparative analysis of the various models built. Because there is not a **universally known *best* method or algorithm for a data mining task**, one should use a **variety of viable model types along with a well-defined experimentation and assessment strategy** to identify the “best” method for a given purpose.

Even for **a single method or algorithm**, a number of parameters need to be calibrated to obtain optimal results.

23

Data mining process and methods

Step 5: Testing and Evaluation

In step 5, the developed models are assessed and evaluated for their accuracy and generality.

This step assesses the degree to which the selected model (or models) meets the business objectives and, if so, to what extent (i.e., Do more models need to be developed and assessed?).

Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit. The testing and evaluation step is a critical and challenging task. No value is added by the data mining task until the business value obtained from discovered knowledge patterns is identified and recognized.

Step 6: Deployment

Development and assessment of the models is not the end of the data mining project. Even if the purpose of the model is to have a simple exploration of the data, the knowledge gained from such exploration will need to be organized and presented in a way that the end user can understand and benefit from.

24

Data mining process and methods

This step involves deploying the model into the production environment. This includes integrating the model into existing systems and processes to make predictions in real-time. This step is important because it allows the model to be used in a practical setting and to generate value for the organization.

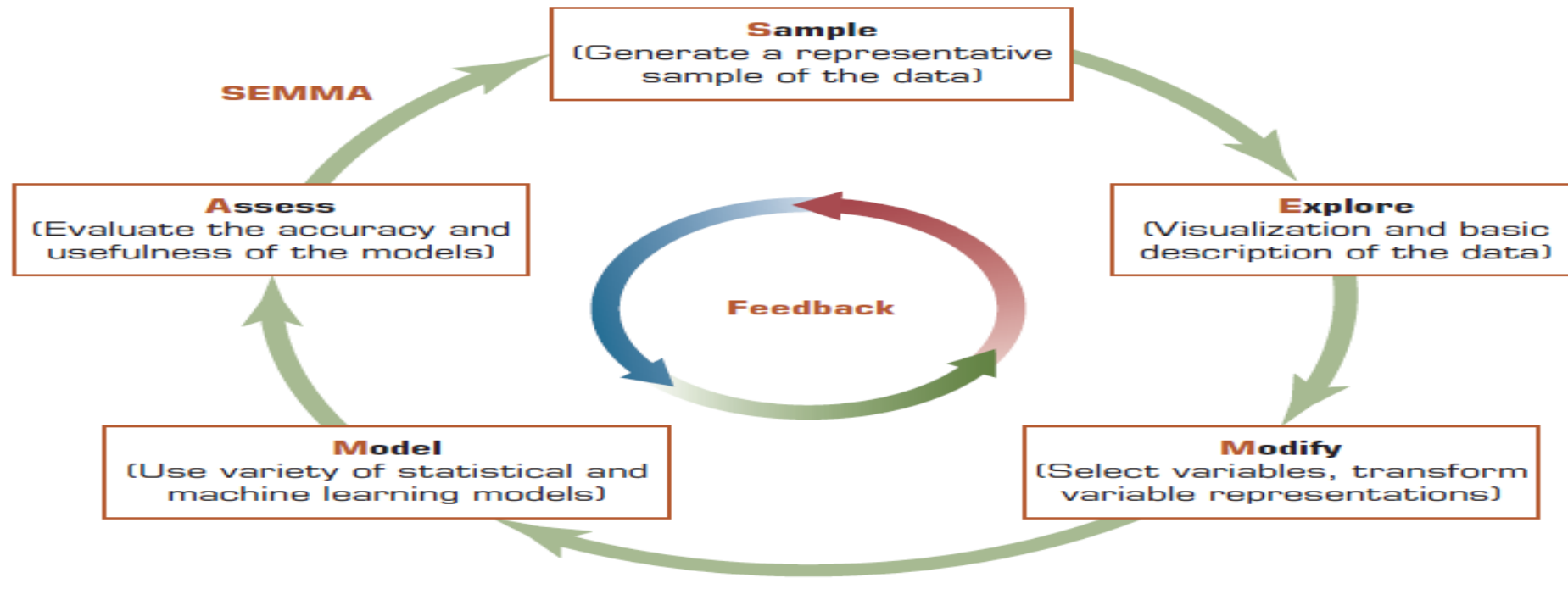
Other Data Mining Standardized Processes and Methodologies:

There is another well-known methodology developed by the SAS Institute, called SEMMA (2009). The acronym SEMMA stands for “sample, explore, modify, model, and assess.”

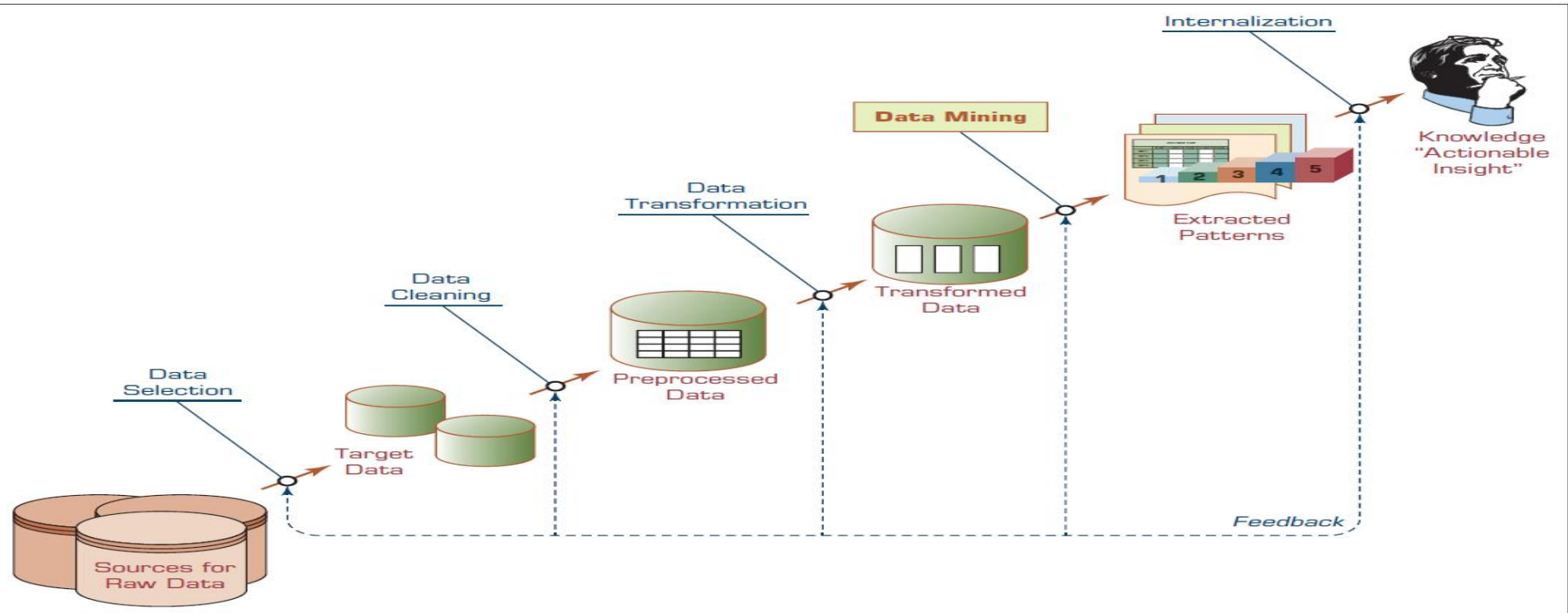
Some practitioners commonly use the term knowledge discovery in databases (KDD) as a synonym for data mining.

25

Data mining process and methods



SEMMA Data Mining Process



**Knowledge Discovery in Databases
Process**

Data Mining Software Tools

Product Name	Web Site (URL)
IBM SPSS Modeler	www-01.ibm.com/software/analytics/spss/products/modeler/
IBM Watson Analytics	ibm.com/analytics/watson-analytics/
SAS Enterprise Miner	sas.com/en_id/software/analytics/enterprise-miner.html
Dell Statistica	statsoft.com/products/statistica/product-index
PolyAnalyst	megaputer.com/site/polyanalyst.php
CART, RandomForest	salford-systems.com
Insightful Miner	solutionmetrics.com.au/products/iminer/default.html
XLMiner	solver.com/xlminer-data-mining
SAP InfiniteInsight (KXEN)	help.sap.com/ii
GhostMiner	fqs.pl/ghostminer
SQL Server Data Mining	msdn.microsoft.com/en-us/library/bb510516.aspx
Knowledge Miner	knowledgeminer.com
Teradata Warehouse Miner	teradata.com/products-and-services/teradata-warehouse-miner/
Oracle Data Mining (ODM)	oracle.com/technetwork/database/options/odm/
FICO Decision Management	fico.com/en/analytics/decision-management-suite/
Orange Data Mining Tool	orange.biolab.si/
Zementis Predictive Analytics	zementis.com

28

Python for data mining

Kaggle is an online platform for data science , data mining and machine learning.

Here's what it offers:

- Datasets – Thousands of public datasets you can download and analyze.
- Competitions – Companies and researchers host challenges where participants build the best predictive models to win prizes.
- Notebooks – Free cloud-based Jupyter notebooks for running Python or R code without setup.
- Courses – Short, beginner-friendly lessons on data science, machine learning, and AI.
- Community – A large forum where data scientists share code, solutions, and ideas.

Kaggle: <https://www.kaggle.com/>

[Competitions](#)[Datasets](#)[Models](#)[Code](#)[Discussions](#)[Blog](#)[Courses](#)[...](#)[Sign In](#)[Register](#)

Level up with the largest AI & ML community

Join over 25M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

 [Register with Google](#)[Register with Email](#)

30

Python for data mining

Python libraries commonly used in data mining:**❑ Pandas**

Purpose: Data manipulation and analysis.

Key Features:

- Provides DataFrame and Series objects for handling tabular data.
- Easy data cleaning, filtering, grouping, merging, and reshaping.
- Handles missing data gracefully.

Why it's important for data mining: Most raw datasets need heavy cleaning and transformation before modeling.

31

Python for data mining

Python libraries commonly used in data mining:

❑ **NumPy**

Purpose: Numerical computations and array processing.

Key Features:

- Powerful n-dimensional arrays.
- Fast mathematical operations (linear algebra, Fourier transforms, random number generation).
- Basis for many other libraries like Pandas and Scikit-learn.

Why it's important: Efficient data storage and manipulation for large datasets.

32

Python for data mining

Python libraries commonly used in data mining:

❑ Matplotlib & Seaborn

Purpose: Data visualization.

Key Features:

- **Matplotlib:** Highly customizable plots (line, bar, scatter, histogram, etc.).
- **Seaborn:** Simplified statistical graphics built on Matplotlib, with beautiful default styles.

Why it's important: Helps in **exploratory data analysis (EDA)** by visualizing patterns, correlations, and distributions.

33

Python for data mining

Python libraries commonly used in data mining:**❑ Scikit-learn**

Purpose: Machine learning for data mining.

Key Features:

- Implements most classic algorithms: clustering (K-means, DBSCAN), classification (SVM, Random Forest), regression
- Tools for model evaluation, cross-validation, and feature selection.

Why it's important: Provides a standardized framework for building and testing predictive models.

34

Python for data mining

notebook60b6ac80e5 Draft saved

File Edit View Run Settings Add-ons Help

+ - ✂ 📄 📁 ▶ ▶▶ Run All Code ▾ ● Draft Session off (run a cell to start) ⏻ ↺ ⋮

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a new version of your notebook.
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

+ Code + Markdown

35

Python for data mining

Notebook

Input

+ Add Input

⬆ Upload



No input attached

Attach a Kaggle dataset, model, or competition

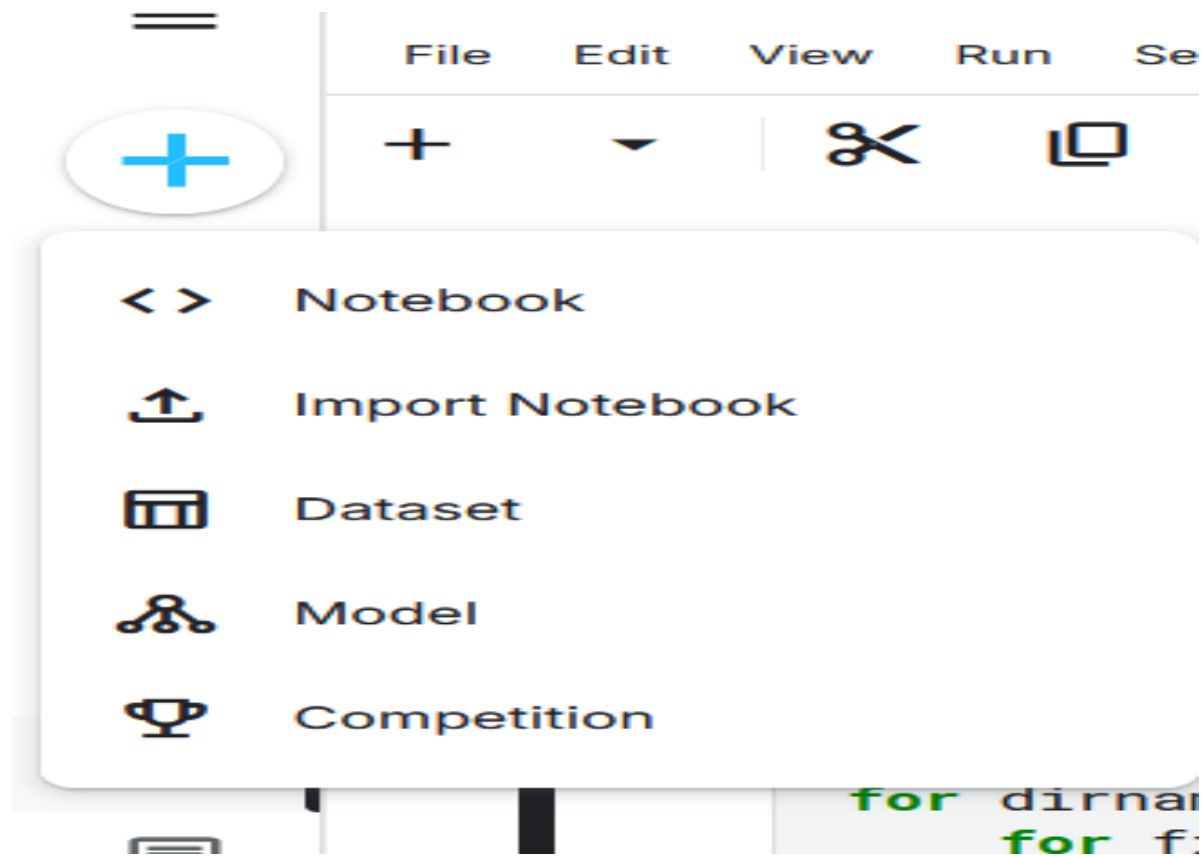
Output

▸ /kaggle/working

Table of contents

36

Python for data mining



37

Summary

1. Data mining techniques.
2. Data mining process.
3. Python libraries.



Thank You