cct College Dublin
Computing • IT • Business

Machine Learning for Data Analysis
MSc in Data Analytics
CCT College Dublin

Introduction to Machine Learning
Week 1

Lecturer: Dr. Muhammad Iqbal*

Email: miqbal@cct.ie

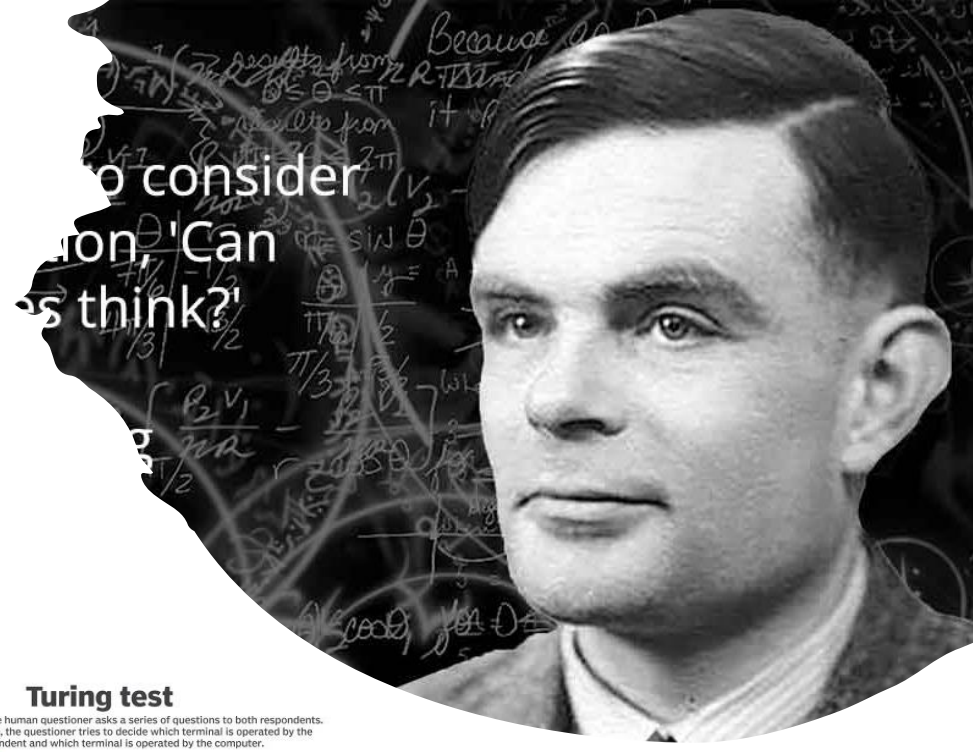
Agenda

cct College Dublin
Computing • IT • Business

- Introduction to Machine Learning (ML)
- Traditional Programming and Machine Learning
- What is Machine Learning?
- Application Areas of ML
- Branches of Machine Learning
- Supervised, Unsupervised, Semi-supervised and Reinforcement Learning
- ML Algorithm Selection
- Cross Industry Standard Process (CRISP-DM)

Mr Hendricks. Are you sure
you are in the right room?
This is MachineLearning 101.

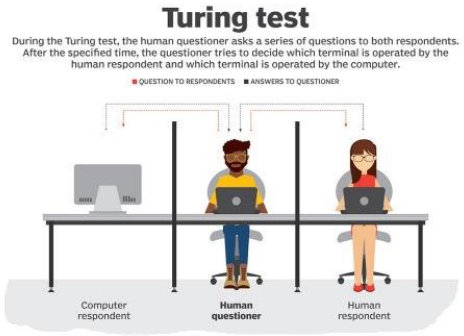
Introduction to ML



- In 1950, Alan Turing asked a question in his paper, “Computing Machinery and Intelligence”,

“Can machines think?” or “Can machines behave intelligently?”

- The paper describes the “Imitation Game,” which involves three participants, such as a **human acting as a judge**, **another human**, and **a computer** that is attempting to convince the judge that it is human.



- The judge would type into a terminal program to “talk” to the other two participants. Both the **human** and the **computer** would respond, and the judge would decide which response came from the computer.

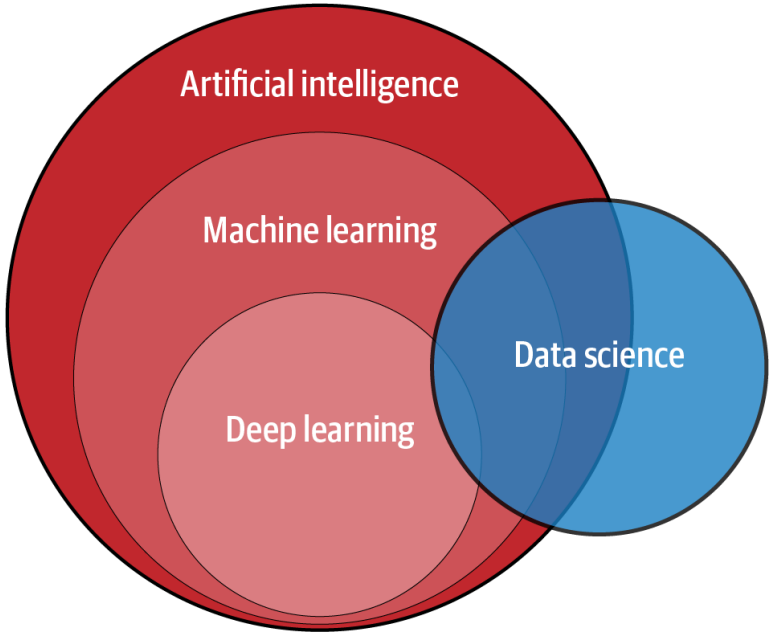
- If the judge couldn’t consistently tell the difference between the human and computer responses, then the computer won the game. The test continued in the form of the **Loebner Prize**, until 2019.

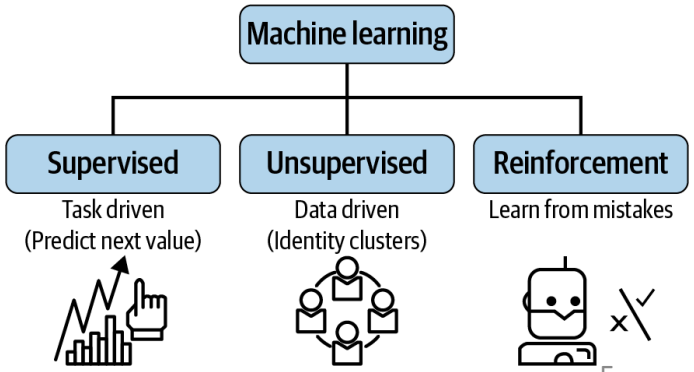
- The aim is simple enough: convince the judges that they are chatting to a human instead of a computer chat bot program.

Introduction to Machine Learning

cct | College Dublin
Computing • IT • Business

- We can observe from the figure that the relationships between **AI, machine learning, deep learning** and **data science** overlapping.
- **Machine learning** is a subset of AI that consists of techniques enabling computers to identify patterns in data and to deliver AI applications. Deep learning is a subset of machine learning that enables computers to solve more complex problems.
- Data science isn't exactly a subset of machine learning, but it uses machine learning, deep learning, and AI to analyze data and reach actionable conclusions.
- It combines machine learning, deep learning and AI with other disciplines, such as big data analytics and cloud computing.



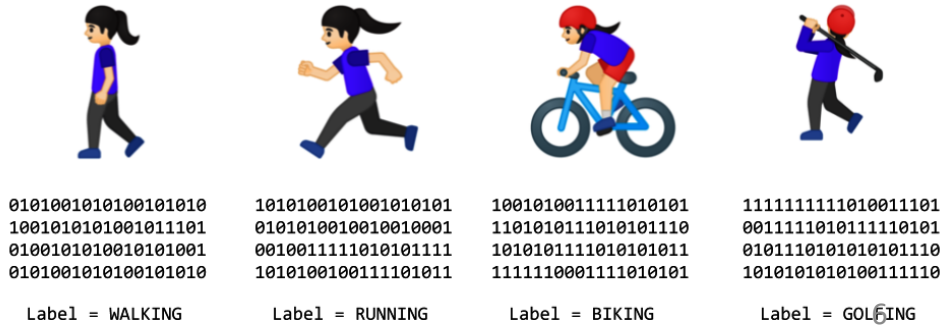


TP and ML

- Traditional programming (TP) approach is shown in the Figure. We have rules that act on data and give us answers.
- The rules and data are provided to the programme, and you can get an output based on the data structure.
- We get lots of data about our scenario, we label that data, and the computer can figure out what the rules are that make one piece of data match a particular label and another piece of data match a different label.
- Suppose we collect a lot of instances of this data while they're doing different activities.
- We end up with a scenario of having data that says "This is what walking looks like," "This is what running looks like," and so on

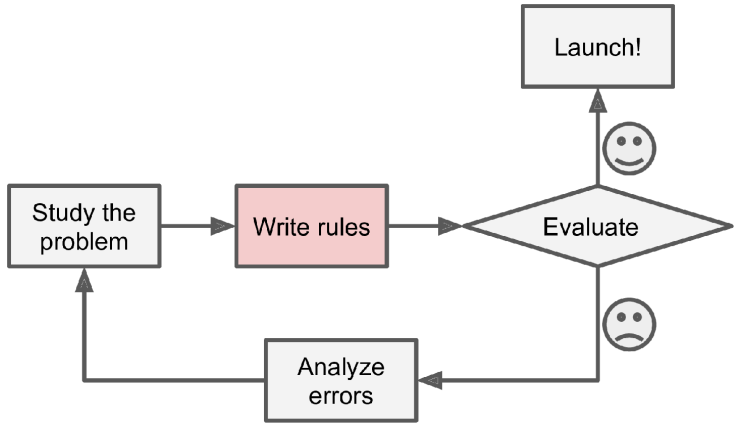
The traditional programming flow

Changing the axes to get machine learning

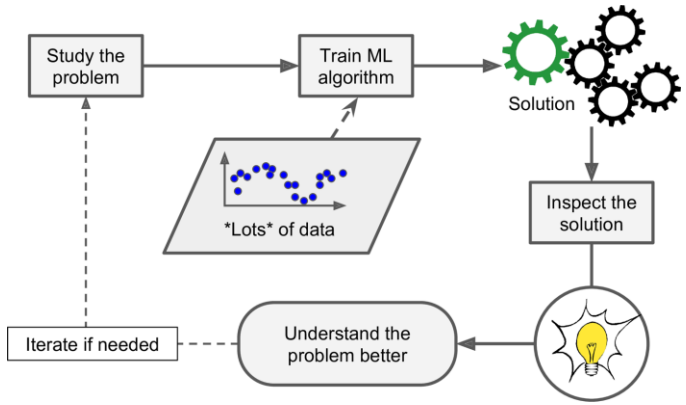


TP and ML

- We can develop a spam filter using traditional programming techniques using the following steps as
- First we would consider what spam typically looks like. We might notice that some words or phrases (such as “win,” “credit card,” “free,” and “amazing”) tend to come up a lot in the subject line.
    1. We would notice a few other patterns in the sender’s name, the email’s body, and other parts of the email.
    2. We would write a detection algorithm for each of the patterns that we noticed, and your program would flag emails as spam if a number of these patterns were detected.
- We would test the program and repeat steps 1 and 2 until it was good enough to launch.



The traditional approach

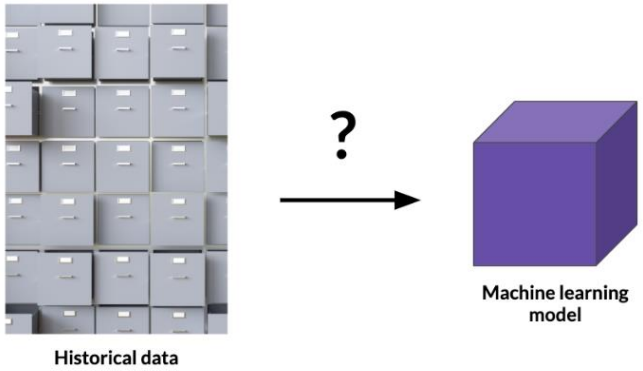


Machine Learning can help humans learn

What is Machine Learning?

cct | College Dublin
Computing • IT • Business

- Machine Learning is the science (and art) of programming computers so they can *learn from data*.
- **Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, 1959**
- **A set of tools for making inferences and predictions from data.**
- **Predict** future events
  - **Will it rain tomorrow?**
    - Yes (70% probability)
- **Infer** the causes of events and behaviors
  - **Why does it rain?**
    - Time of the year, humidity levels, temperature, location etc.
- **Infer** patterns
  - **What are the different types of weather conditions?**
    - Rain, sunny, overcast, fog, etc.

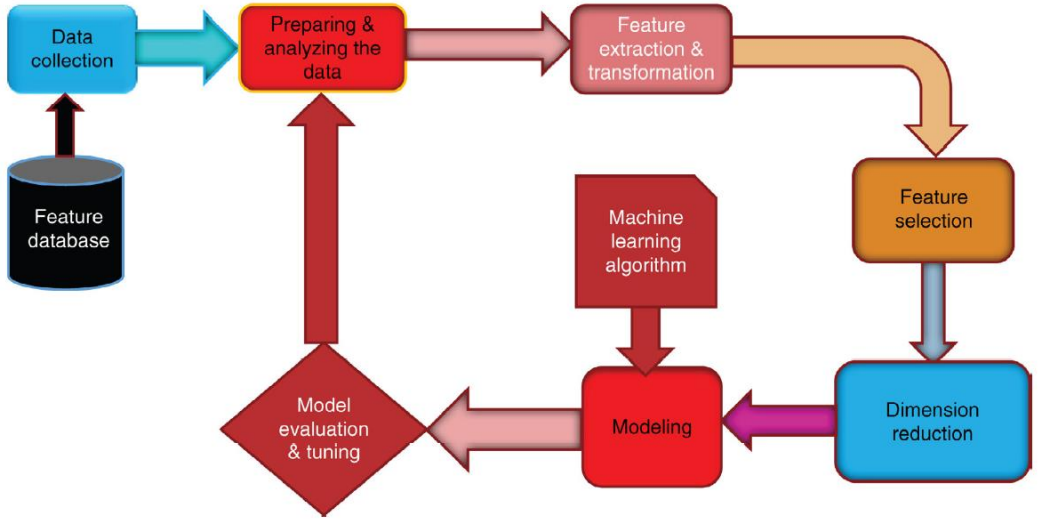


Machine learning workflow

Machine Learning Framework

cct | College Dublin
Computing • IT • Business

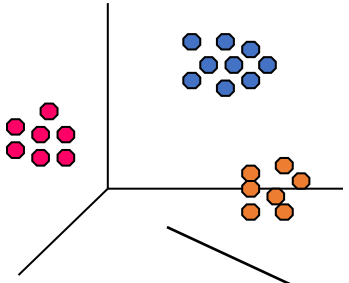
- Machine learning is a **unified algorithmic framework** designed to identify computational models that accurately describe empirical data and the phenomena underlying it, with little or no human involvement.
- A typical machine learning framework is presented as shown in Figure that shows the main stages highlighted in their blocks.
- To understand and develop an application utilizing machine learning framework, we use the procedures that are mentioned in the blocks.



A typical machine learning framework.

- Machine Learning can be used to perform a wide array of useful tasks including automatic detection of objects in images, speech recognition, knowledge discovery in the medical sciences, and predictive Analytics and many others.

Application Areas of ML

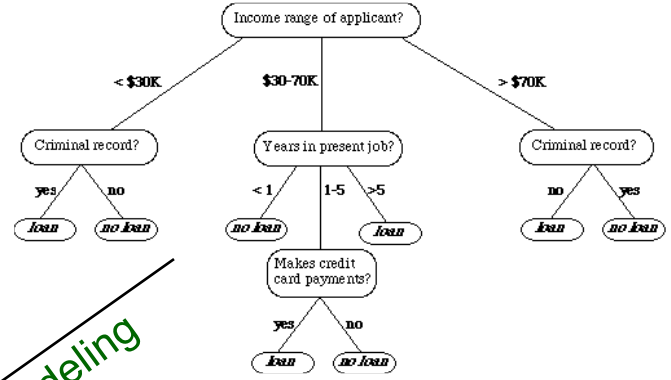


Clustering

Data

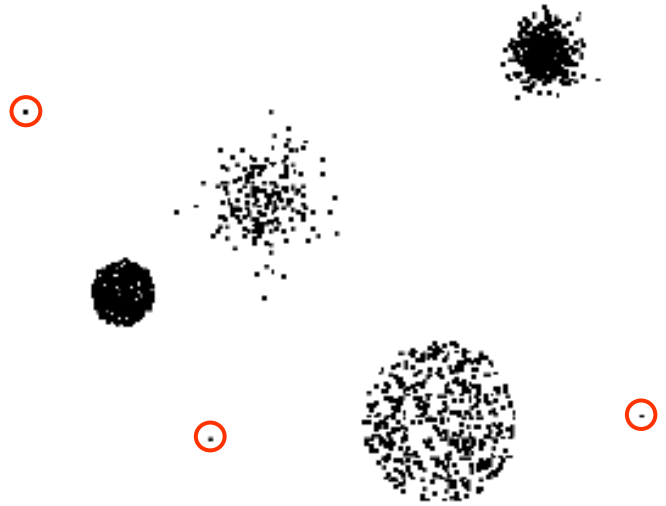
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Predictive Modeling



Association Rules

Anomaly Detection



Branches of Machine Learning

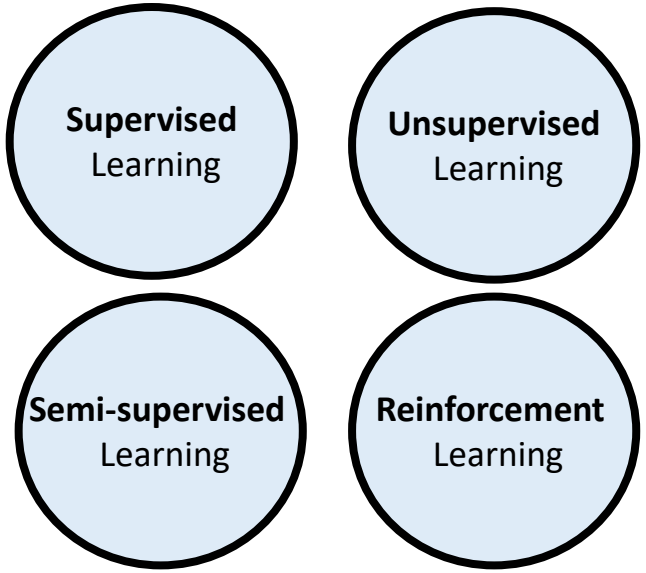
cct | College Dublin
Computing • IT • Business

- **Machine learning algorithms** fall into two broad categories

- ***Supervised Learning Algorithms*** are trained with labeled data. In other words, data composed of examples of the desired answers. For instance, a model that identifies fraudulent credit card use would be trained from a dataset with labeled data points of known fraudulent and valid charges. Most machine learning is supervised.

- ***Unsupervised Learning Algorithms*** are used on data with no labels, and the goal is to find relationships in the data. For instance, we might want to find groupings of customer demographics with similar buying habits.

- ***Reinforcement Learning (RL)*** is a popular and promising branch of AI that involves making smarter models and agents that can automatically determine ideal behavior based on changing requirements.



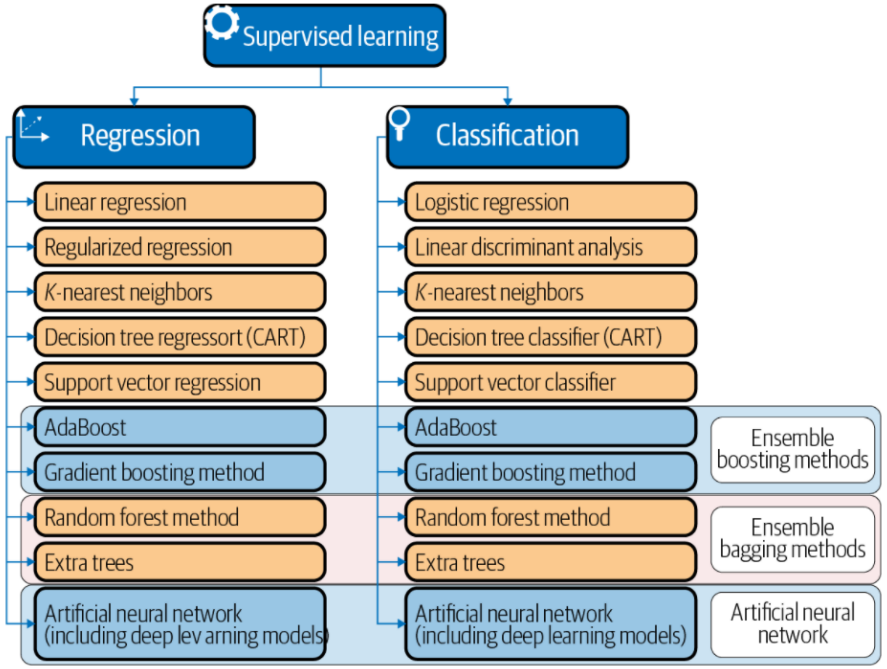
- ***Semi-supervised Learning*** is a machine learning branch that tries to solve problems with both labeled and unlabeled data with an approach that employs concepts belonging to clustering and classification methods.

Machine Learning Modelling

Regression

College Dublin
Computing • IT • Business

- Predict a value of a given **continuous valued variable** based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics and neural network fields.
- **Examples:**
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.



Models for regression and classification

- **Bootstrap aggregation** (Bagging) is an ensemble learning method that is commonly used to reduce variance within a noisy dataset.
- **Boosting** refers to the process of sequentially training weak learners to build a model.

Machine Learning Modelling

Classification

- Find a model for class attribute as a function of the values of other attributes.

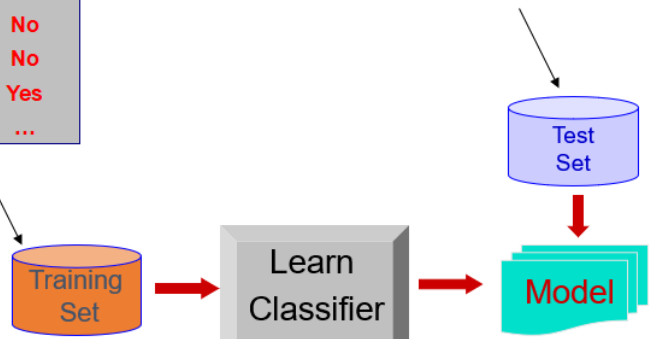
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Class

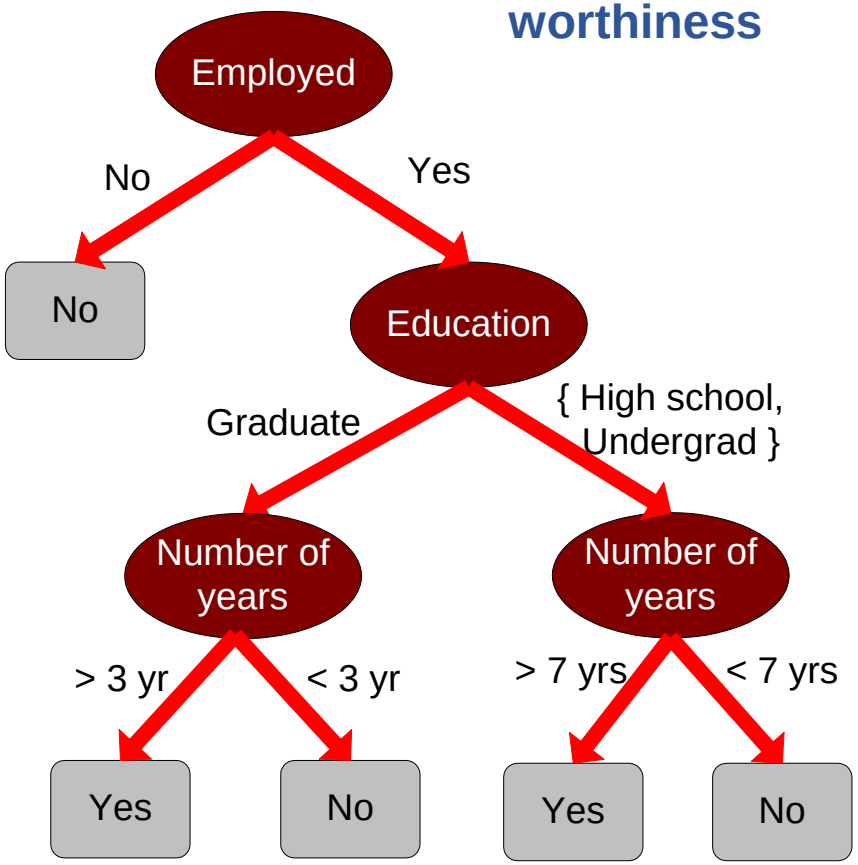
categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

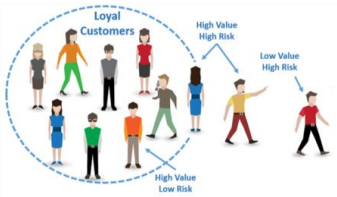


Model for predicting credit worthiness



Classification

Applications



cct College Dublin
Computing • IT • Business

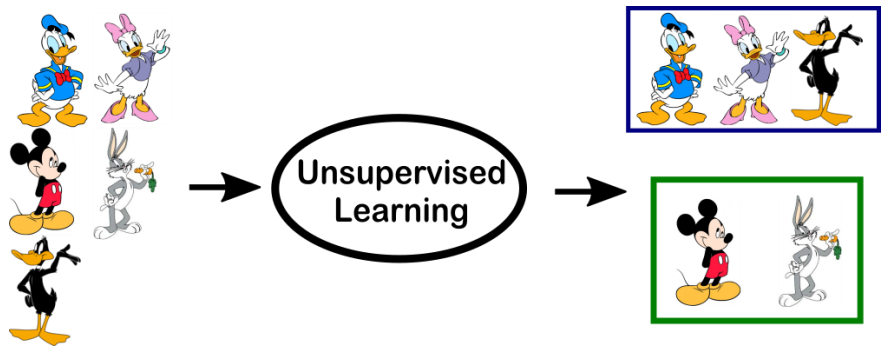
- **Fraud Detection**
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he/ she buy, how often he/ she pays on time, etc.
    - Label past transactions as fraud or fair transactions. This forms the **class** attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

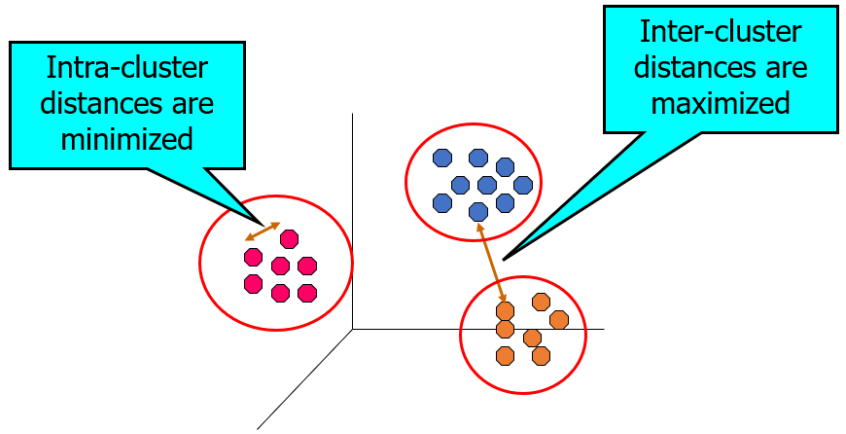
- **Churn prediction for telephone customers**
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he/ she calls, what time-of-the day he/ she calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Machine Learning Modelling

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- In unsupervised learning, as we might guess, the training data is unlabeled. The system tries to learn without a teacher.
- The most important unsupervised learning algorithms.
- **Clustering:** K-Means, DBSCAN and Hierarchical Cluster Analysis (HCA).
- **Anomaly Detection and Novelty Detection:** One-class SVM and Isolation Forest.
- **Visualization and Dimensionality Reduction:** Principal Component Analysis (PCA), Locally Linear Embedding (LLE) and t-Distributed Stochastic Neighbor Embedding (t-SNE).





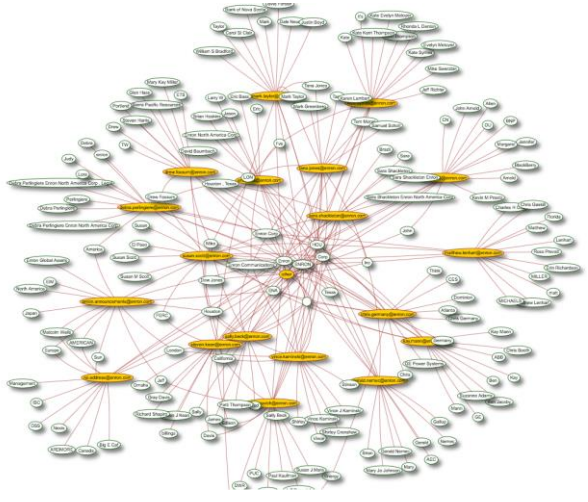
Clustering
Applications

cct College Dublin
Computing • IT • Business

- **Market Segmentation:**
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in the same cluster vs. those from different clusters.

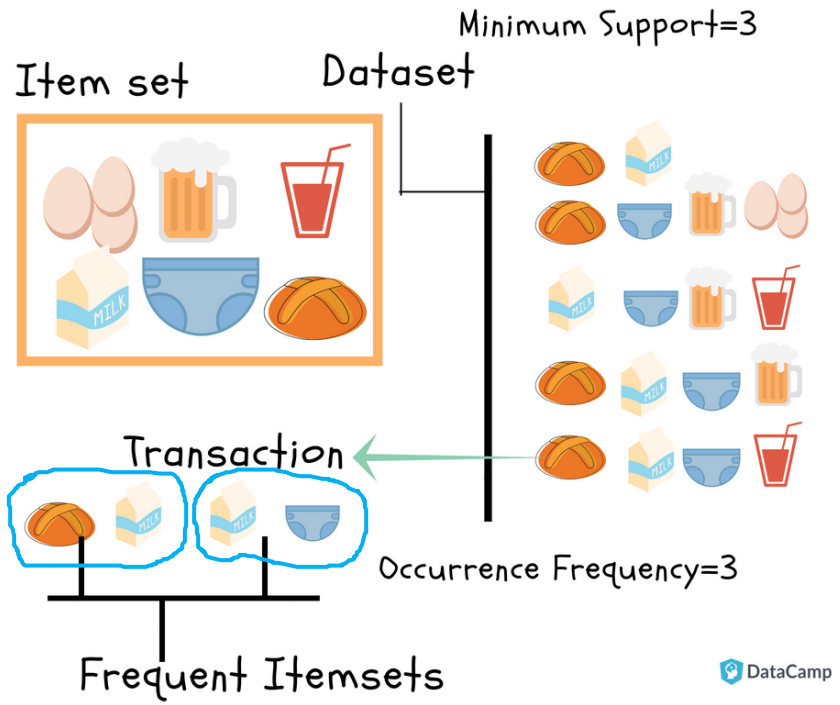
- **Document Clustering:**
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Association Rule Discovery

- **Given a set of records each of which contain some number of items from a given collection**
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.



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

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Analysis

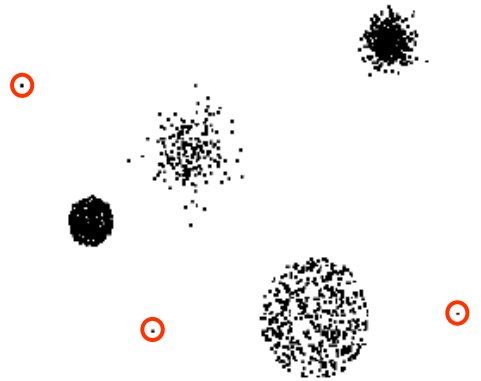
Applications

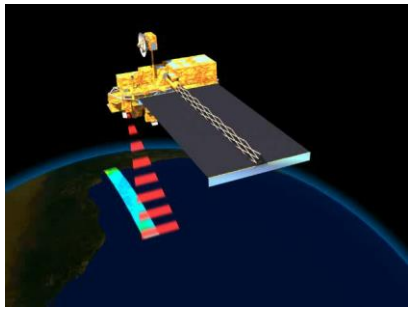
cct College Dublin
Computing • IT • Business

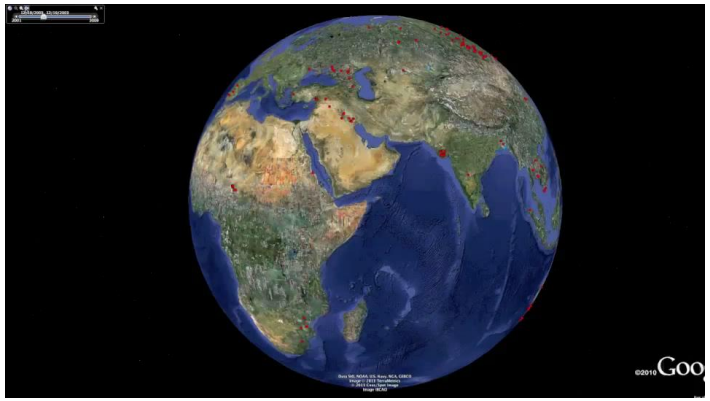
- **Market-basket analysis**
  - Rules are used for sales promotion, shelf management, and inventory management.
- **Telecommunication alarm diagnosis**
  - Rules are used to find combination of alarms that occur together frequently in the same time period.
- **Medical Informatics**
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases.

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- **Applications:**
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.



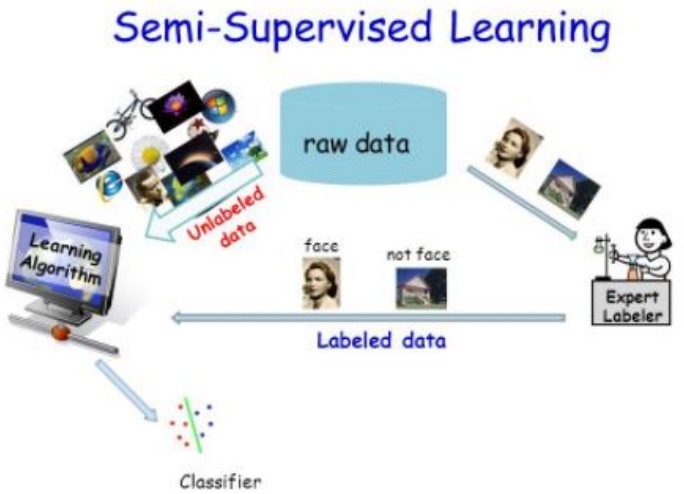


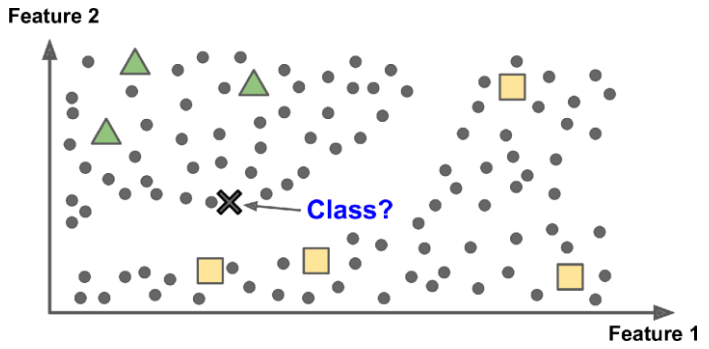


Semi-supervised Learning

cct | College Dublin
Computing • IT • Business

- Since labeling data is time-consuming and costly, we have plenty of unlabeled instances, and few labeled instances.
- Some algorithms can deal with data that's partially labeled. This is called as the Semi-supervised learning.
- Some photo-hosting services, such as Google Photos, are good examples of this.
- Once you upload all of your family photos to the service, it automatically recognizes that the same person **A** shows up in photos 1, 5, and 11, while another person **B** shows up in photos 2, 5, and 7.
- This is the unsupervised part of the algorithm (clustering). Now all the system needs is for you to tell it who these people are. **Add one label per person** and it is able to name everyone in every photo, which is useful for searching photos.





Semi-supervised learning with two classes (triangles and squares): the unlabeled examples (circles) help classify a new instance (the cross) into the triangle class rather than the square class, even though it is closer to the labeled squares

Reinforcement Learning

cct | College Dublin
Computing • IT • Business

- Consider we are teaching the dog to catch a ball, when we throw a ball, the dog catches a ball, we will give a cookie. If it fails to catch a dog, we will not give a cookie. So the dog will figure out what actions it does that made it receive a cookie and repeat that action.

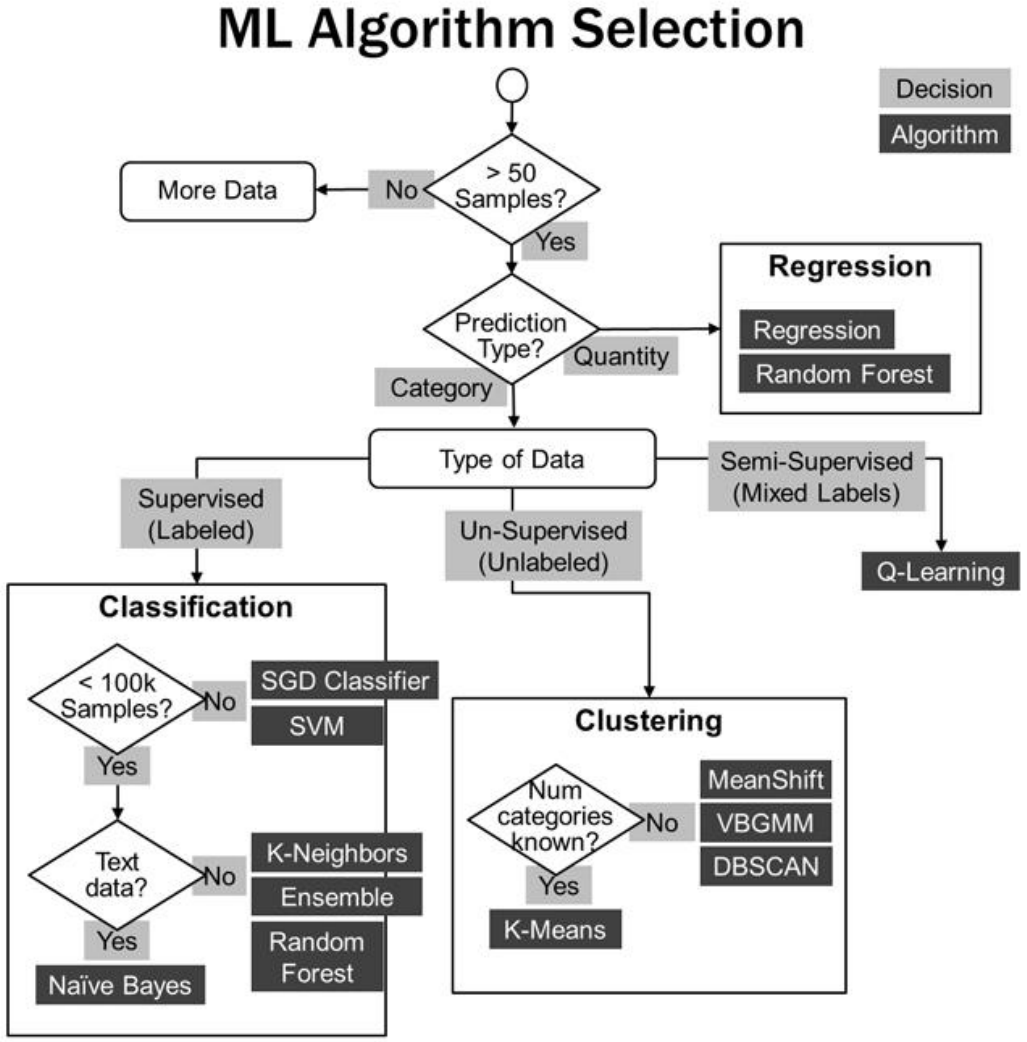
- Similarly, in an **RL environment**, we will not teach the **agent** what to do or how to do, but we will give feedback to the agent for each **action** it does. The feedback may be positive (**reward**) or negative (**punishment**).
- The learning system which receives the punishment will improve itself. It is a trial-and-error process. The reinforcement learning algorithm retains outputs that maximize the received reward over time.

- In the above analogy, the **dog** represents the **agent**, giving a **cookie** to the dog on catching a ball is a **reward** and not giving a cookie is **punishment**.

- It depends on out **policy** that we should give a reward after each step or after completion of some number of steps.
- A RL agent can explore for different actions which might give a good reward, or it can (exploit) use the previous action which resulted in a good reward. If the RL agent explores different actions, there is a great possibility to get a poor reward.

ML Algorithm Selection

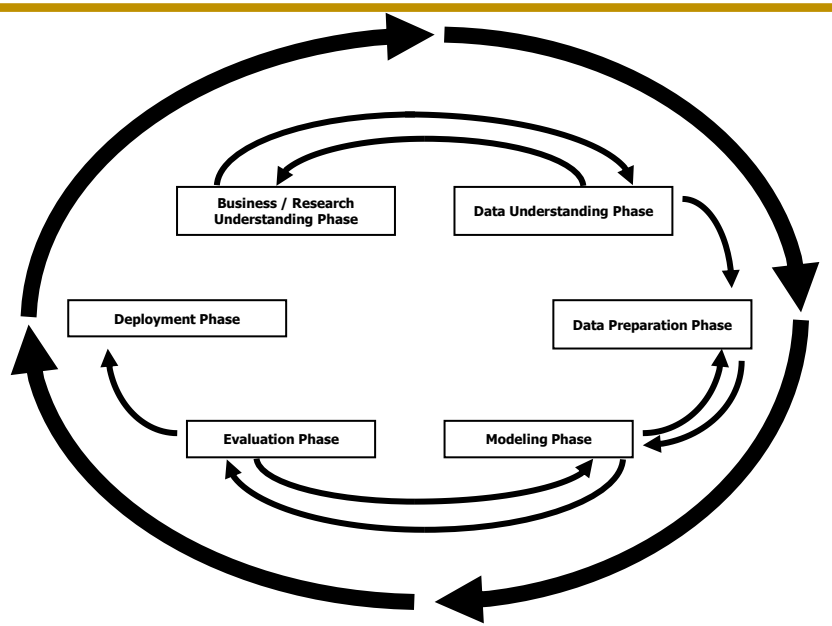
- The flow chart is used to show the selection of **Machine learning algorithm** based on the data.
- It is clear from the flow chart that the three major kinds of algorithms are considered as important in the machine learning.
- It is the responsibility of Data Scientist to select an appropriate **ML algorithm** to provide an accurate results for the relevant problem.



Cross Industry Standard Process
CRISP-DM

College Dublin
Computing • IT • Business

- **Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996**
  - Fits data mining into the general problem-solving strategy of business/ research unit
  - Industry, tool and application neutral
  - Data mining projects follow iterative, adaptive life cycle consisting of 6 phases



CRISP-DM Lifecycle

- Iterative **CRIP-DM** process is shown in outer circle
- Most significant dependencies between phases shown
- Next phase depends on results from preceding phase
- Returning to earlier phase possible before moving forward

Cross Industry Standard Process
CRISP-DM

cct College Dublin
Computing • IT • Business

1. Business/ Research Understanding Phase
- Define project requirements and objectives
- Translate objectives into data exploration problem definition
- Prepare preliminary strategy to meet objectives

2. Data Understanding Phase
- Collect data
- Perform exploratory data analysis (EDA)
- Assess data quality
- Optionally, select interesting subsets

3. Data Preparation Phase
- Prepares for modeling in subsequent phases
- Select cases and variables appropriate for analysis
- Cleanse and prepare data so it is ready for modeling tools
- Perform transformation of certain variables, if needed

4. Modeling Phase
- Select and apply one or more modeling techniques
- Calibrate model settings to optimize results

5. Evaluation Phase
- Evaluate one or more models for effectiveness
- Determine whether defined objectives achieved
- Make decision regarding data exploration results before deploying to field

6. Deployment Phase
- Make use of models created
- Simple deployment example: generate report
- Complex deployment example: implement parallel data exploration effort in another department
- In businesses, customer often carries out deployment based on your model

Resources/ References

- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, September 2019, ISBN: 9781492032649.
- Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media, Inc. October 2016.
- Data Mining And Machine Learning, Fundamental Concepts And Algorithms, MOHAMMED J. Zaki, Wagner Meira, Jr., Cambridge CB2 8BS, United Kingdom, 2020.
- Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.
- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/
- Some images are used from Google search repository (https://www.google.ie/search) to enhance the level of learning.