# MACHINE LEARNING, DATA SCIENCE, BIG DATA MINING

# MACHINE LEARNING, DATA SCIENCE, BIG DATA MINING

## Machine Learning for Data Mining Applications in Big Data

**Prof. Dr. Dewan Md. Farid**
Department of Computer Science & Engineering
United International University


**https://cse.uiu.ac.bd/profiles/dewanfarid/**

# ABOUT THE AUTHOR

PROF. DR. DEWAN MD. FARID is a Professor of Computer Science and Engineering at United International University, and IEEE Senior Member. Prof. Farid worked as a Postdoctoral Fellow/Staff at the following research labs/groups: (1) Computational Intelligence Group (CIG), Department of Computer Science and Digital Technology, University of Northumbria at Newcastle, UK in 2013, (2) Computational Modelling Lab (CoMo) and Artificial Intelligence Research Group, Department of Computer Science, Vrije Universiteit Brussel, Belgium in 2015-2016, and (3) Decision and Information Systems for Production systems (DISP) Laboratory, IUT Lumière – Université Lyon 2, France in 2020. Prof. Farid was a Visiting Faculty at the Faculty of Engineering, University of Porto, Portugal in June 2016. He holds a PhD in Computer Science and Engineering from Jahangirnagar University, Bangladesh in 2012. Part of his PhD research has been done at ERIC Laboratory, University Lumière Lyon 2, France by Erasmus-Mundus ECW eLink PhD Exchange Program. His PhD was fully funded by Ministry of Science & Information and Communication Technology, Government of the People's Republic of Bangladesh and European Union (EU) eLink project. Prof. Farid has published 109 peer-reviewed scientific articles, including 30 highly esteemed journals like Expert Systems with Applications, Journal of Theoretical Biology, Journal of Neuroscience Methods, Bioinformatics, Scientific Reports (Nature), Proteins and so on in the field of Machine Learning, Data Mining and Big Data. Prof. Farid re-

ceived the following awards: (1) Dr. Fatema Rashid Best Paper Award (2nd Position) for the paper titled "KNNTree: A new method to ameliorate k-nearest neighbour classification using decision tree" in 3rd International Conference on Electrical Computer and Communication Engineering (ECCE 2023), CUET, Chittagong, Bangladesh, (2) JuliaCon 2019 Travel Award for attending Julia Conference at the University of Maryland, Baltimore, USA, and (3) United Group Research Award 2016 in the field of Science and Engineering. He received the following research funds as Principal Investigator: (1) a2i Innovation Fund of Innov-A-Thon 2018 (Ideabank ID No.: 12502) from a2i-Access to Information Program – II, Information and Communication Technology (ICT) Division, Government of the People's Republic of Bangladesh, and (2) Project Code: UIU/IAR/01/2021/SE/23 received from Institute for Advanced Research (IAR), United International University. Prof. Farid received the following Erasmus Mundus scholarships: (1) LEADERS (Leading mobility between Europe and Asia in Developing Engineering Education and Research) to undertake a staff level mobility at the Faculty of Engineering, University of Porto, Portugal in 2015, (2) cLink (Centre of excellence for Learning, Innovation, Networking and Knowledge) for pursuing Postdoc at University of Northumbria at Newcastle, UK in 2013, and (3) eLink (east west Link for Innovation, Networking and Knowledge exchange) for pursuing Ph.D. at University Lumière Lyon 2, France in 2009. Prof. Farid also received Senior Fellowship I and II awards by National Science & Information and Communication Technology (NSICT), Ministry of Science & Information and Communication Technology, Government of the People's Republic of Bangladesh respectively in 2008 and 2011 for pursuing Ph.D. at Jahangirnagar University. He visited 18 countries for attending international conferences, research and higher education. Prof. Farid delivered several invited/keynote talks including an invited research talk at Data to AI Group (DAI), Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA.

# CONTENTS

# SYMBOLS

$\mathbb{X}$ The set of training instances/ training data
$N$ Size of $\mathbb{X}$
$(x^{(i)}, y^{(i)})$ The $i$-th instance pair in $\mathbb{X}$ (supervised learning)
$x^{(i)}$ The input (features) of $i$-th training instance in $\mathbb{X}$ (unsupervised learning)
$x_j^{(i)}$ The value of feature $j$ in $i$-th training instance
$D$ Dimension of an instance $x^{(i)}$
$K$ Dimension of a label $y^{(i)}$
$X \in \mathbb{R}^{N \times D}$ Design matrix, where $X_i$,: denotes $x^{(i)}$
$X_i$ A feature in training data
$\mathbb{F}$ Hypothesis space of functions to be learnt, i.e., a model
$C[f]$ A cost function of $f \in \mathbb{F}$
$C[\theta]$ A cost function of $\theta$ parameterising $f \in \mathbb{F}$
$(x', y')$ A testing pair
$\hat{y}$ Label predicted by a function $f$, i.e., $\hat{y} = f(x')$ (supervised learning)

$P(x, y)$ A data generating distribution

**PART I**

# LEARNING FROM DATA

# CHAPTER 1

# INTRODUCTION

Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.

—Albert Einstein, Physicist of the 20th century

Intelligence is the ability to adapt to change.

—Stephen William Hawking, Theoretical Physicist

## 1.1 Computational Intelligence

Computational Intelligence (CI) also known as Soft Computing, is a concept for advanced information processing/knowledge mining that usually refers to the ability of a computer/machine to learn a specific task from data/big data or experimental observation. It's a multidisciplinary field including advance artificial intelligence, machine learning, data mining, big data, pattern recognition, knowledge-based systems, decision support systems, high-performance computing, and market intelligence.

Figure 1.1: Machines are learning.

### 1.1.1 What is Machine Learning?

Machine learning (ML) is the field of computer science/intelligent computing or a branch of advanced artificial intelligence (AI) that involves the development of self learning algorithms to automatically extract rules/knowledge from data/Big Data in order to make predictions/decisions. The data is basically historical records/past data points. Instead of requiring domain experts/humans to manually extract rules and build models from analysing Big Data, machine learning offers a more efficient alternative for gaining the knowledge in data to gradually improve the performance of predictive models and make data-driven decisions. It ables the machine/computer to learn from experience without being specifically programmed. ML builds models based on historical data, known as *training data* for future prediction/decision making. ML is turning things (Big Data) into numbers and finding patterns in those numbers. ML is used in many real-life applications, e.g. internet search engines, email filters to sort out spam, detecting intrusions/unusual transactions, image classification etc.

**Arthur Samuel (1959)** Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

**Tom Mitchell (1998)** Well-posed Learning Problem: A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.

**Dewan Farid (2021)** Machine Learning is the process of extracting rules from Big Data for knowledge mining.

### 1.1.2 What is Data Mining?

Data mining is also known as *Knowledge Discovery from Data*, or *KDD* for short, which turns big data into knowledge. It's the process of extracting knowledge and uncovering hidden patterns from Big Data to understand the current and future trends. The main object of data mining is to extract patterns (known and unknown) from Big Data. The following definitions are found from different literatures.

1. It's the process of analysing data from different perspectives and summarising it into useful information.

2. It's the process of finding hidden information and patterns in a huge database.

3. It's the extraction of implicit, previously unknown, and potentially useful information from data.

### 1.1.3 Why Machine Learning and Data Mining?

Can we think of all the rules in complex real-life Big Data problems with long lists of rules? Also, environments are continually changing in real-life situations that's why we need machine learning and data mining. ML can adapt ("learn") to new scenarios when the concepts are changing over the time. The terms machine learning and data mining are commonly confused, as they often employ the same methods and overlap significantly. They can be roughly defined as follows:

- Machine learning focuses on prediction, based on known properties learned from the training data.

- Data mining focuses on the discovery of unknown properties in the data.

**Artificial Intelligence (AI)** The effort to automate intellectual tasks that are normally performed by humans. It enables computers to mimic human behaviour. The basic idea is to create intelligent machines that work and react like a human brain.

**Deep Learning** Subset of machine learning in which multilayered artificial neural networks (ANN) learn from Big Data.

**Big Data** Large volume of data with variety and velocity.

**Data Science** Data Science is the field of study that combines knowledge of artificial intelligence (AI), machine learning (ML), data mining (DM) to extract knowledge and insights from structured and unstructured data.

**Data Warehouse** Data Warehousing (DW) is the process of compiling, managing and organising data from several sources into one common database, whereas data mining refers to the process of extracting useful data from the databases.

### 1.1.4 Data, Information, Knowledge

**Data** It's a collection of text, numbers, symbols, sound, picture or any recorded facts in raw or unorganised form that can be processed by a computer. Data can be collected from different sources, e.g. scientific data, medical data, demographic data, financial data, and marketing data. It's really important to have good quality data for Machine Learning research and modelling. Engendering data is a challenging and costly process, which is the most important part of all Data Analytics, Data Mining, and Big Data research.

**Information** Processed data is called information. Data that has been processed, e.g. grouped, normally by a computer, to give it meaning and make it interpretable. The patterns, associations or relationships in data can provide information.

**Knowledge** Processed information is called knowledge that is the understanding of information such as how to solve problems. Information can be converted into knowledge about historical patterns and future trends.

### 1.1.5 Types of Data

Data can be classified into two categories: **Unstructured Data** and **Structured Data**. Unstructured data have no rigid structure, e.g. images, video, natural language text, and speech etc. Structured data is like a table, which can be classified as follows:

1. Nominal/Categorical (e.g. marital status, eye colour, political party etc)
2. Numerical (discrete: number of children, defects per hour; and continuous: weight, sales, sensor stream data)
3. Ordinal data, which has order but the distance between values is unknown. E.g. how would you rate your health from 1-5? where, 1 being poor, 5 being healthy.
4. Time-series Data (Data across time), e.g. the historical scale values of gold from 2012-2018.

Real world data are generally incomplete, noisy, and inconsistent. ***Low quality data will lead to low quality mining results***. Generally, big data contains errors, missing attribute values, and lacking certain attributes of interest. Measure for data quality is a multidimensional view, which follows:

1. Accuracy: correct or wrong, accurate or not.
2. Completeness: not recorded, unavailable.
3. Consistency: some modified but some not.
4. Timeliness: timely update?

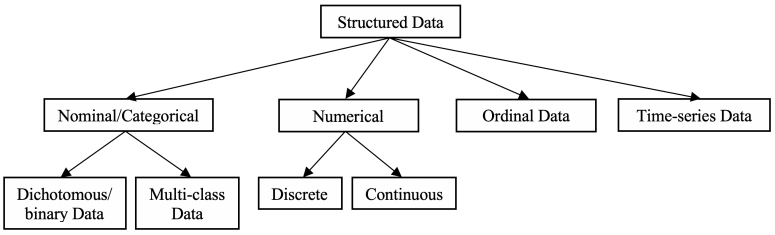


Figure 1.2: Types of structured data.

5. Believability: how trustable the data is?
6. interpretability: how easily the data can be understood?

## 1.2 Different Types of Learning in Machine Learning

The general idea of machine learning is to develop concepts from historical data for acquiring knowledge through experience. There are different types/problems of learning in the field of machine learning that are listed below.

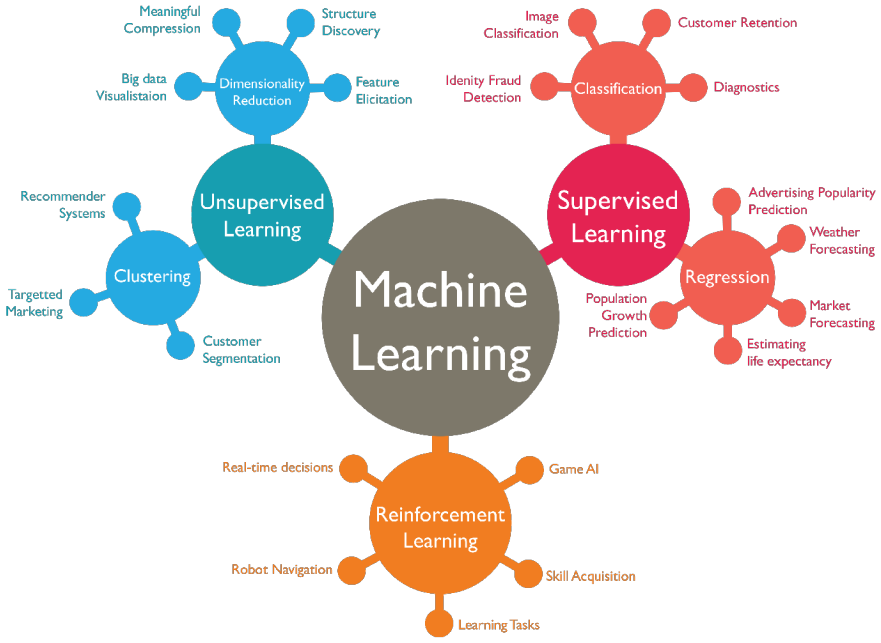


Figure 1.3: Types of Machine Learning tasks.

### 1.2.1 Supervised Learning

Supervised learning is the process of building models when input variables, $X$, called features and output variable, $y$ called label or a class employ algorithms to learn mapping function $y = f(X)$ from input/data to output/class. The goal is to map the function so well that the new input data/instance $x'$ can be classified correctly. In supervised learning, we have data and labels. The ML model tries to learn the relationship between data and labels. Supervised learning can be bifurcated into the following tasks:

***1.2.1.1 Regression*** It consists of mathematical methods that allow us to predict the numerical/real/continuous class-value/label of output variable $y$ based on the value of one or more predictor/input variables $X_i$. Linear regression is one of the most common forms of regression analysis that is used in prediction and forecasting.

***1.2.1.2 Classification*** It refers to the predictive modelling problems where the class-values/labels of output variable are categorical/nominal/discrete, i.g. classify if an email is spam or not. There are three types of classification problems:

1. Binary-class classification.
2. Multi-class classification.
3. Multi-label classification (e.g. what items does this photo contain? What topics is this YouTube video about?).

Table 1.1: Binary-class Classification Problem.

| | $X_1$ | $X_2$ | $X_3$ | $\cdots$ | $X_M$ | $y^{(i=0,1)}$ |
|---|---|---|---|---|---|---|
| $x^{(1)}$ | $x_1^{(1)}$ | $x_2^{(1)}$ | $x_3^{(1)}$ | $\cdots$ | $x_M^{(1)}$ | $y^{(i)}$ |
| $x^{(2)}$ | $x_1^{(2)}$ | $x_2^{(2)}$ | $x_3^{(2)}$ | $\cdots$ | $x_M^{(2)}$ | $y^{(i)}$ |
| $x^{(3)}$ | $x_1^{(3)}$ | $x_2^{(3)}$ | $x_3^{(3)}$ | $\cdots$ | $x_M^{(3)}$ | $y^{(i)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x^{(N)}$ | $x_1^{(N)}$ | $x_2^{(N)}$ | $x_3^{(N)}$ | $\cdots$ | $x_M^{(N)}$ | $y^{(i)}$ |

### 1.2.2 Unsupervised Learning

Unsupervised learning, also known as clustering, is used to group unlabelled data (when we only have input data, $\mathbb{X}$ and no corresponding class-values) for finding patterns/similarities and interesting structure in the data. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Therefore, in unsupervised learning we have only data but no labels. The ML model tries to find the pattern in data without something to reference on. Clustering is commonly used for dimensionality reduction, e.g. PCA (Principal Component Analysis).

Table 1.2: Multi-class Classification Problem.

| | $X_1$ | $X_2$ | $X_3$ | $\cdots$ | $X_M$ | $y^{(i=0,1,2,\ldots,P)}$ |
|---|---|---|---|---|---|---|
| $x^{(1)}$ | $x_1^{(1)}$ | $x_2^{(1)}$ | $x_3^{(1)}$ | $\cdots$ | $x_M^{(1)}$ | $y^{(i)}$ |
| $x^{(2)}$ | $x_1^{(2)}$ | $x_2^{(2)}$ | $x_3^{(2)}$ | $\cdots$ | $x_M^{(2)}$ | $y^{(i)}$ |
| $x^{(3)}$ | $x_1^{(3)}$ | $x_2^{(3)}$ | $x_3^{(3)}$ | $\cdots$ | $x_M^{(3)}$ | $y^{(i)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x^{(N)}$ | $x_1^{(N)}$ | $x_2^{(N)}$ | $x_3^{(N)}$ | $\cdots$ | $x_M^{(N)}$ | $y^{(i)}$ |

Table 1.3: Multi-label Classification Problem.

| | $X_1$ | $X_2$ | $X_3$ | $\cdots$ | $X_M$ | $y^{(i=0,1,2,\ldots,P)}$ | $y^{(j=0,1,2,\ldots,P)}$ |
|---|---|---|---|---|---|---|---|
| $x^{(1)}$ | $x_1^{(1)}$ | $x_2^{(1)}$ | $x_3^{(1)}$ | $\cdots$ | $x_M^{(1)}$ | $y^{(i)}$ | $y^{(j)}$ |
| $x^{(2)}$ | $x_1^{(2)}$ | $x_2^{(2)}$ | $x_3^{(2)}$ | $\cdots$ | $x_M^{(2)}$ | $y^{(i)}$ | $y^{(j)}$ |
| $x^{(3)}$ | $x_1^{(3)}$ | $x_2^{(3)}$ | $x_3^{(3)}$ | $\cdots$ | $x_M^{(3)}$ | $y^{(i)}$ | $y^{(j)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x^{(N)}$ | $x_1^{(N)}$ | $x_2^{(N)}$ | $x_3^{(N)}$ | $\cdots$ | $x_M^{(N)}$ | $y^{(i)}$ | $y^{(j)}$ |

### 1.2.3 Semi-supervised learning

Semi-supervised learning is the process of learning models from semi-supervised data where we have both labeled and unlabelled instances.

### 1.2.4 Reinforcement Learning

Reinforcement learning is an agent (algorithm), which performs actions in an environment and is rewarded or penalised based on the whether the actions were favourable or not, Fig 1.4. In reinforcement learning, there's no training data and the algorithm works on a rewards-based system. The agent selects an action in an environment that will lead to rewards, e.g. creating a game.

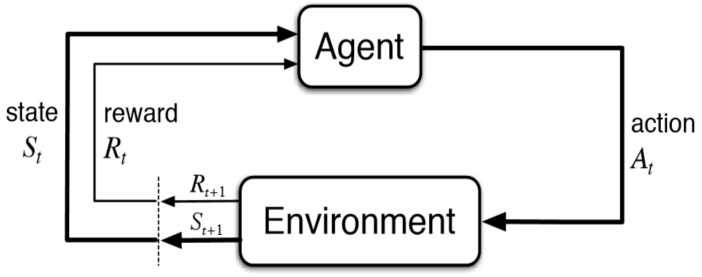


Figure 1.4: Reinforcement Learning.

### 1.2.5 Transfer Learning

Transfer learning takes knowledge from one/existing model and uses it in your own/other model.

### 1.2.6 Active Learning

Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an instance, which may be from a set of unlabelled instances.

### 1.2.7 Sequence to Sequence

Sequence to sequence (Seq2seq) is one of the machine learning techniques used for language processing. Applications include language translation, image captioning, conversational models and text summarisation, e.g. given a sequence of English text, translate it into French.

## 1.3 Illustration of Supervised Learning

To illustrate supervised learning, we consider a small dataset in Table 1.4 described by four features/attributes namely Outlook, Temperature, Humidity, and Wind, which represent the weather condition of 14 days. Each feature has several unique values. The Play column in Table 1.4 represents the decision/class category of each day. It indicates whether a particular weather condition is suitable or not for playing tennis.

Table 1.4: Weather Data: A Binary Classification Problem.

| Day | Outlook | Temperature | Humidity | Wind | Play |
|---|---|---|---|---|---|
| $Day_1$ | Sunny | Hot | High | Weak | No |
| $Day_2$ | Sunny | Hot | High | Strong | No |
| $Day_3$ | Overcast | Hot | High | Weak | Yes |
| $Day_4$ | Rain | Mild | High | Weak | Yes |
| $Day_5$ | Rain | Cool | Normal | Weak | Yes |
| $Day_6$ | Rain | Cool | Normal | Strong | No |
| $Day_7$ | Overcast | Cool | Normal | Strong | Yes |
| $Day_8$ | Sunny | Mild | High | Weak | No |
| $Day_9$ | Sunny | Cool | Normal | Weak | Yes |
| $Day_{10}$ | Rain | Mild | Normal | Weak | Yes |
| $Day_{11}$ | Sunny | Mild | Normal | Strong | Yes |
| $Day_{12}$ | Overcast | Mild | High | Strong | Yes |
| $Day_{13}$ | Overcast | Hot | Normal | Weak | Yes |
| $Day_{14}$ | Rain | Mild | High | Strong | No |

A set of rules learned from the Table 1.4.

1. If Outlook = Sunny and Humidity = High then Play = No
2. If Outlook = Sunny and Humidity = Normal then Play = Yes
3. If Outlook = Overcast then Play = Yes
4. If Outlook = Rain and Wind = Strong then Play = No
5. If Outlook = Rain and Wind = Weak then Play = Yes

In the slightly more complex form shown in Table 1.5, two of the attributes - temperature and humidity have numeric values.

Table 1.5: Weather Data with Some Numeric Attributes.

| **Outlook** | **Temperature** | **Humidity** | **Wind** | **Play** |
|---|---|---|---|---|
| Sunny | 85 | 85 | Weak | No |
| Sunny | 80 | 90 | Strong | No |
| Overcast | 83 | 86 | Weak | Yes |
| Rain | 70 | 96 | Weak | Yes |
| Rain | 68 | 80 | Weak | Yes |
| Rain | 65 | 70 | Strong | No |
| Overcast | 64 | 65 | Strong | Yes |
| Sunny | 72 | 95 | Weak | No |
| Sunny | 69 | 70 | Weak | Yes |
| Rain | 75 | 80 | Weak | Yes |
| Sunny | 75 | 70 | Strong | Yes |
| Overcast | 72 | 90 | Strong | Yes |
| Overcast | 81 | 75 | Weak | Yes |
| Rain | 71 | 91 | Strong | No |

A set of rules learned from the Table 1.5.

1. If Outlook = Sunny and Humidity $>$ 75 then Play = No
2. If Outlook = Sunny and Humidity $\leq$ 75 then Play = Yes
3. If Outlook = Overcast then Play = Yes
4. If Outlook = Rain and Wind = Strong then Play = No
5. If Outlook = Rain and Wind = Weak then Play = Yes

### 1.3.1 Input: Concepts, Instances, and Attribute

The input takes the form of *concepts*, *instances*, and *attributes*. We call the thing that is to be learned a *concept description*. Each instance is characterised by the values of attributes that measure different aspects of the instance. There are many different

types of attributes, although typical data mining schemes deal only with numeric and nominal, or categorical attributes.

**Concept** is the thing to be learned.

**Concept description** is the output produced by a learning scheme or classifier.

**Instances** are the things that are to be classified or associated or clustered. Each dataset is represented as a matrix of instances versus attributes, which in database terms is a single relation, or a *flat file*.

**Attribute** is a data field, representing a characteristic or feature of a data object. The nouns *attribute, dimension, feature*, and *variable* are often used interchangeably in the literature. The value of an attribute for a particular instance is a measurement of the quantity to which the attribute refers.

## 1.4 Machine Learning Process

Learning is the concept of estimating the model parameters so the predictions are consistent with true labels. There are six steps in the machine learning process: (1) data collection, (2) data preparation, (3) training model, (4) testing model, (5) deploying model, and (6) retrain model that's shown in Fig. 1.5.

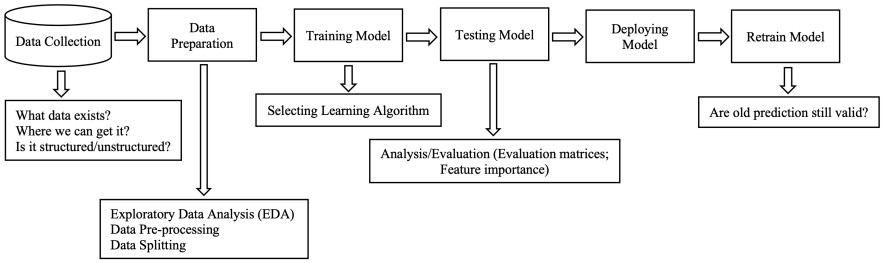


Figure 1.5: Machine Learning Process.

## 1.5 Data Collection and Preparation

To collect the data we need to ask the following questions:

- What kind of problems are we trying to solve?
- What data sources already exist?
- What privacy concerns are there?

- Is the data public?
- Where should we store the data?

Data Preparation performs the following tasks:

- Exploratory Data Analysis (EDA), i.g. learning about the data, understand and know the data;
- Data Preprocessing (preparing data to be modelled)
- Data Splitting

### 1.5.1 Exploratory Data Analysis (EDA)

Learning about the data/understanding and knowing the data:

- What are the feature variables (input) and the target variable (output)?
- What's the kind of data? e.g. structured/unstructured
- Are there missing values?
- Where are the outliers? how many of them are there? why are they there?

### 1.5.2 Data Preparation

- Feature Imputation (filling missing values)
- Feature Encoding (turning values into numbers)
- Feature Normalisation (scaling) or Standardisation
- Feature Engineering: transform data into (potentially) more meaningful representation by adding in domain knowledge.
- Feature Selection (selecting most valuable features)
- Dealing with Imbalances

### 1.5.3 Data Splitting

- Training Set (usually 70-80%)
- Validation Set (typically 10-15%): Model hyper-parameters are tuned on this;
- Testing Set (10-15%)

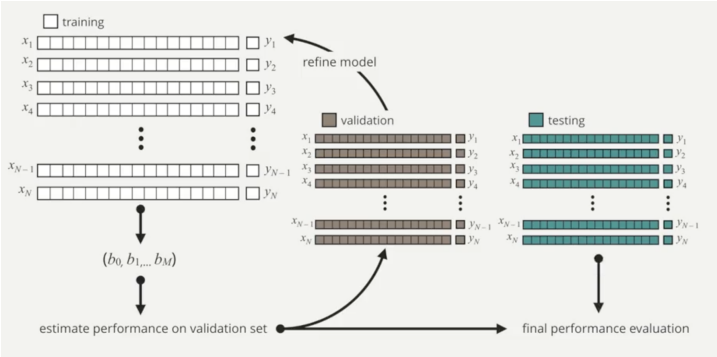


Figure 1.6: Data Splitting.

## 1.6 Why Data Preprocessing?

Data preprocessing transforms raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. There are several data preprocessing techniques: *Data cleaning* can be applied to remove noise and correct inconsistencies in data. *Data integration* merges data from multiple sources into a coherent data store such as a data warehouse. *Data reduction* can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering. *Data transformations* (e.g., normalisation) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. Data goes through a series of steps during preprocessing:

**Data cleaning** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

**Data integration** Data with different representations are put together and conflicts within the data are resolved, e.g., using multiple databases, data cubes, or files.

**Data transformation** Data is normalised, aggregated and generalised.

**Data reduction** This step aims to present a reduced representation of the data in a data warehouse.

**Data discretisation** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

### 1.6.1 Data Cleaning

Data in the real world is dirty. Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error. We can use domain expert

knowledge to correct inconsistent data. Table 1.6 shows some of the examples of noisy data.

Table 1.6: Some Examples of Noisy Data

| **Noise Type** | **Example** |
|---|---|
| Missing attribute value | Occupation = " " (missing data) |
| Containing errors | Salary = "-10" (an error) |
| Containing discrepancies in codes | Age = "42", Birthday = "01/01/2010" |
| Intentionally entering value | Jan. 1 as everyone's birthday |

Fill in missing attribute or class values:

1. Ignore the tuple: usually done when class label is missing.
2. Use the attribute mean (or majority nominal value) to fill in the missing value or for all samples belonging to the same class.
3. Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

### 1.6.2 Data Integration

Data integration combines data from multiple sources into a coherent store, e.g., integrate metadata (metadata is data about data or metadata is data that describes other data) from different sources. It also detects and resolves data value conflicts. Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

### 1.6.3 Data Transformation

Data transformation includes the followings:

1. Normalisation: Scaling attribute values to fall within a specified range.
2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalisation: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

### 1.6.4 Data reduction

Data reduction obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results. A

database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complex dataset.

1. Reducing the number of attributes (removing irrelevant attributes e.g., information gain, principle component analysis).
2. Reducing the number of attribute values (reducing the number of attributes by grouping them into intervals).
3. Reducing the number of tuples

### 1.6.5 Discretisation and generating concept hierarchies

1. Unsupervised discretisation (class variable is not used): Equal-interval: split the whole range of numbers in intervals with equal size, and Equal-frequency: use intervals containing equal number of values.
2. Supervised discretisation (uses the values of the class variable): Using class boundaries.
3. Generating concept hierarchies: recursively applying partitioning or discretisation methods.

## 1.7 Places to Find Free Datasets

https://www.dataquest.io/blog/free-datasets-for-projects/

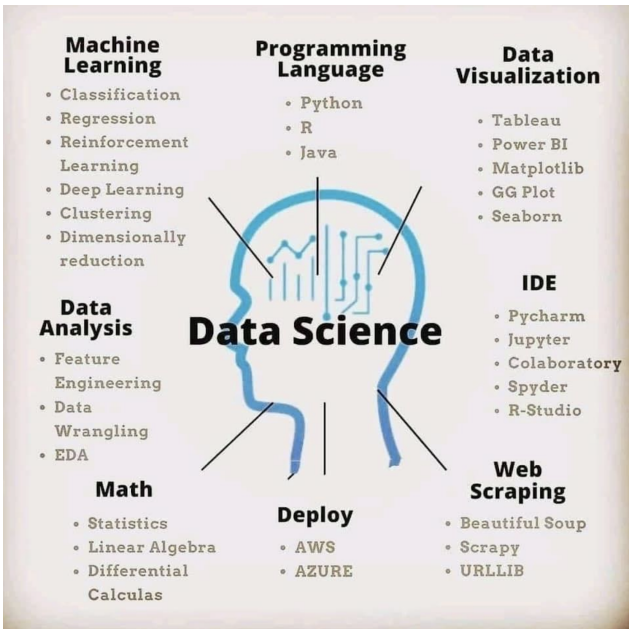


Figure 1.7: Data Science.

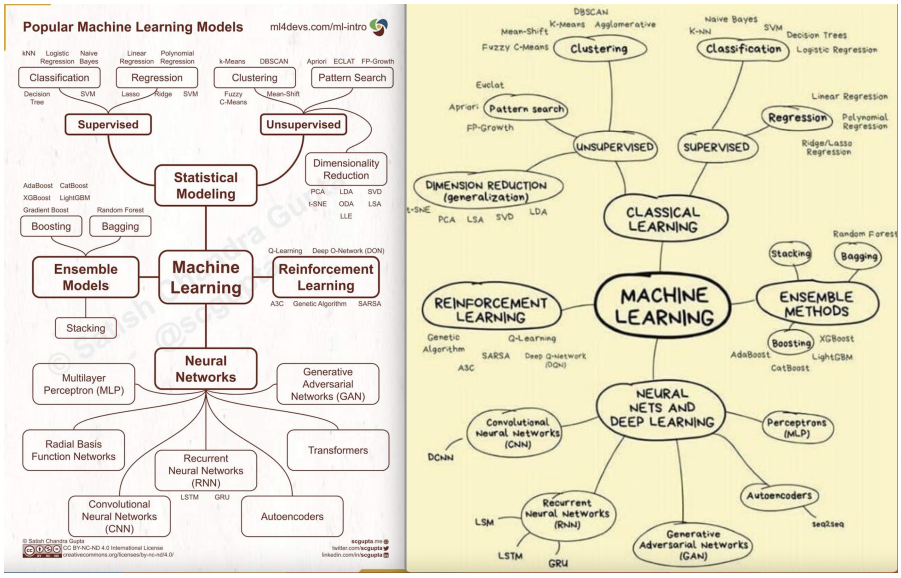


Figure 1.8: Machine learning models.