LUT University
School of Engineering Science
Software Engineering
Master's Programme in Computer Science (Software Engineering)

Hafiz Muhammad Shahzad Sikandar

THE INTEGRATION OF MACHINE LEARNING INTO AN EXISTING SYSTEM

MASTER'S THESIS

Examiners: Professor Kari Smolander
Associate Professor Jussi Kasurinen

Supervisors: Professor Kari Smolander

ABSTRACT

LUT University
School of Engineering Science
Software Engineering
Master's Programme in Computer Science (Software Engineering)

Hafiz Muhammad Shahzad Sikandar

The Integration of Machine Learning into An Existing System

Master’s Thesis

66 pages, 23 figures, 12 tables, 0 Appendix

Examiners: Professor Kari Smolander
Associate Professor Jussi Kasurinen

Keywords: Machine Learning integration, Existing system, Artificial Intelligence, Machine Learning Algorithm, Clustering, Case Studies, ML architecture

We have been living in the era, where development of first version of applications is quite easy and affordable. As the application grow the managing data becomes the problem for many systems. That’s develop attraction for mostly systems to go forward with integration of ML.

This thesis explores the various aspects related to integration of ML, starting from introduction about the ML methods, algorithms, and models. Furthermore, the detailed explanation about the process of integration the ML, modeling the architectures for ML has been followed up. The process for integrating the ML itself includes six main steps to follow, start with the study of conventional or existing system, selection ML method, selecting the process model, selecting tool for ML integration (depending upon the platform),

implementation and deployment. The last and general part of integrating is to test the outcome and make continuous improvement.

Later, this thesis includes to explore the case studies. The selected two case studies emphasized the steps and factors needs to introduce the ML in the conventional systems. First case study, data parsing, text extraction and keywords extraction while clustering label ML technique has been implemented in second case study. The factors like stability, performance, flow of data, architecture, features, flexibility transparency and speed always influence the functionalities of the ML integration outcomes. Along with positive, there is always some risk and challenges while integrating ML like protecting data in term of security and privacy, getting the relevant data, maintaining the speed of ML system to increase the productivity of the system.

ACKNOWLEDGEMENTS

Accomplishing this master's thesis was the final and the important step towards the end of master studies with the title of MSc. in Computer Science with the major in Software Engineering. Hence, I want to express my gratitude to my supervisor, Professor Kari Smolander, who has guided me and supported me during this project. I also want to thank Mr. Umar Draz and Mr. M. Ahsan who supported me during my visit to Lappeenranta for meeting with supervisor. Along with this, I would also like to mention the co-operation of my Ex Chief Technology Officer (CTO) who has helped me to collect the data about the Discount-Based System (BDS) and explained in the second case study.

Although, I have been in Helsinki, but the time in the city of Lappeenranta and LUT University taught me a lot about life and helped me to find my own path in other aspects of life including studies, work, and other areas of life. I am really privileged to have best teachers not only in LUT University but also the previous schools and university studies. I am feeling really blessed for having had a chance to study in one of the best and prestigious university in the world and had an experience to study in the world the best education systems. Here, I would also like to thank all my dear friends as well as the professors, teachers and other staff of LUT, who contributed to the success of my studies and overall life during the years. Finally, my heartfelt thank you go to my parents, siblings, uncle, and friends, and my beloved elder brother Mr. Zubair Ahmad who has always supported me ever during all my studies and life.

Now it is time for me to focus more efficiently on my professional career. Even though the future excites me, and I am eager to seize new opportunities, a piece of me will always remain on the beautiful lake of Saimaa.

In Helsinki, Sunday, 29th March 2020
Hafiz Muhammad Shahzad Sikandar

Table of Contents

APPENDIX

(Arabic style page numbering starts on this page)

LIST OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
CTO	Chief Technology Officer
DBS	Discount Based System
DR	Dimensionality Reduction
ML	Machine Learning
NN	Neural Network
PM	Probabilistic Model
SL	Supervised Learning
SVM	Support Vector Machine
UNSL	Unsupervised Learning

1 INTRODUCTION

1.1 Background

We have been living in the era, where development of first version of applications is quite easy and affordable. As the application become bigger and bigger, the managing application become complicated and huge challenges. Due to handling the issue of increase in complexity of application as it grows, many systems introduced artificial intelligence (AI) and Machine Learning (ML) in their conventional systems. AI and Machine learning and two most hot topics in buzzwords and having been using in many applications connectively. Currently, ML is the most common application of AI which make a system intelligence enough to make decision by itself learning.

Over the years, topics related to artificial intelligence has been become a promising topic for public or private sectors. The classical artificial intelligence techniques mainly based on logic, representation of knowledge, planning and reasoning and as results software related understand languages and some robotics come into exist. Later, many researchers tried to focus on other approaches related to design of systems based on their learning and that leads to new and expanding area of machine learning [1].

Machine learning is mainly linked with implementation of Artificial Intelligence (AI). Machine learning is the only methodology commonly being integrated for AI system [2]. The various systems like image processing, image recognition, speech recognition, data analysis, web search etc. have been integrated to increase the efficiency in term of processing power. There are abundant of machine learning application in being used in various stages. Nowadays, a lot of organizations are using it and many prestigious organizations are planning to integrate it in their existing system. This research will help to study the different artefacts link with integration so that it could be useful for those organizations.

1.2 Goals and delimitations

The main purpose of this research is to study the process involve with the machine learning

integration with existing systems, explore the ML integration using different case studies and experiences and find out the answers for the research questions which is mentioned below.

- What kind of methods and techniques can be used to integrate machine learning in existing systems?
- What kind of architecture is being used in different existing systems and study one or two case studies about it?
- What is the simple process for integrating the machine in the existing system?
- What kind of risks and challenges are involved in integration machine learning in existing systems?

To identify, evaluate and interpret, a study needs to carry out in this research by using available literature, recent past practice and application which is relevant to topic. The focus of this research will be exploring the previously ML integration process and review it which is known as secondary study [3].

1.3 Structure of the thesis

The structure of the thesis is involved in four main sections, literature study, case studies, ML integration and discussion & conclusion. The Literature studies includes with the study of AI, ML and its methods & techniques while case studies part help to study previously ML integrated system and to compare with conventional systems. In ML integration section, the main studies must make after getting familiar with case studies and summarize the ML integration process.

Figure 1 illustrates the overall structure this study intends to follow. The first two components are belonged to Chapter 2, where firstly, need to finalize some learning materials which include the AI & ML algorithms. It starts with collecting required data or information to make better understanding of the topic and focus which leads to literature review later part. In literature review chapter, focus is on techniques and algorithms currently used in various existing system and its effectiveness. After knowing about the main concepts of ML, two case studies need to explore more one is from Wikipedia and one is from a private company in Finland (name is confidential) where ML integration has been made in conventional systems. Next component in this structure is the core component where

integration process, tools and architecture, challenges need to be discussed and then in discussion chapter, we need to summarize our finding throughout this research report. Finally, the conclusion needs to write in the last Chapter.

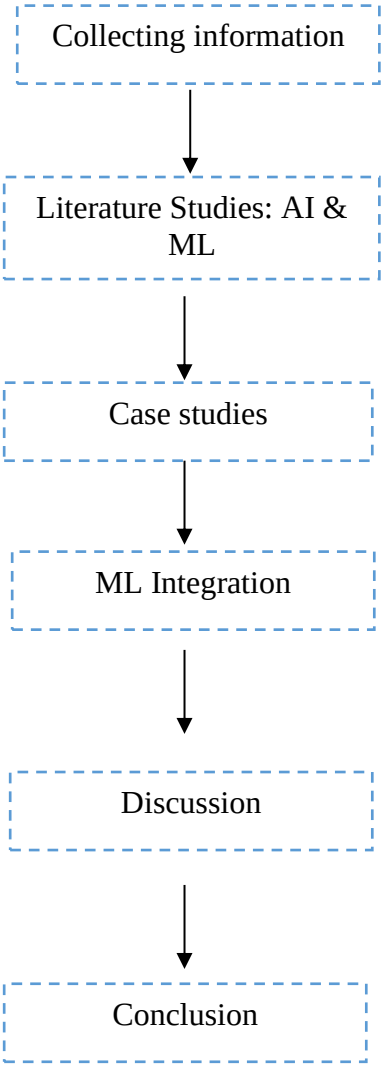


Figure 1: The structure of the thesis

2 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

In this chapter, the literature review has been carried out to study the machine background and focusing mainly of primary studies. During primary studies, several papers or literatures have been studied to develop the clear understanding about the chosen topics. As the topic is related to machine learning (ML) and it comes under the domain of artificial intelligence (AI), thus the primary studies in this chapter will start from digging in AI. This Chapter has divided 2 parts first one is AI and second ML.

2.1 Artificial Intelligence (AI)

AI is the main area where the selected topic came into being. It starts from defining it and then historical background and later explore it in detail. According to Oxford dictionary the term artificial intelligence in computer science can be defined as:

"The discipline which is related to study and development of computer systems capable of performing tasks normally requiring human intelligence, such as, **cognitive reasoning**, **visual perception**, **speech recognition**, **decision-making**, and **translation between languages**[1]."

According to Patrick from Massachusetts Institute of Technology:

"The AI is the study of computation to make a system that emphasize on **perceptions**, **reasoning** and **actions**" [4].

2.1.1 History of AI

It has been observed that current AI boom start almost seven years ago [5]. A serious of many ups and down have been seen which is divided into two main part "AI summer" and "AI winter" with growing interest in AI. A short history of AI can be seen in Table 1 given below.

[1] Oxford Definition of AI: https://www.oxfordreference.com/view/10.1093/acref/9780198609810.001.0001/acref-9780198609810-e-423

Table 1: History of AI. [5]

Timeline	Description
1956	At the Dartmouth conference, the AI termed has been used and introduced it as an academic discipline.
1956 – 1974	During these golden years, AI got government funding in promoting, and utilize it for various approaches like logic-base problem solving.
1974 – 1980	A high expectation linked with the limitations in term of capacity of AI projects moved to first 'AI winter', result in reducing funding and interest in AI research.
1980 – 1987	New successes and focus of researches have brought to introduce and increase in knowledge based expert systems and funding toward AI has increase to next level.
1987 – 1993	In 1987, during second 'AI winter', specialized hardware industry faced a collapse. Later, the negative perceptions have been developed by government and investors as expert systems showed some limitations and high cost to make it update and maintenance.
1993 – 2011	Initially, due to increase new success to optimism about AI, with the help of increasing computing power AI moved toward data driven. World champion Kasparov was beaten by IBM's DeepBlue in chess in 1997 and later, in 2002, Amazon used first time automated systems for recommendation module. Lastly, in 2011, two human well known champions at the TV quiz Jeopardy were beaten by Two apple release Siri IBM Watson.
2012 – Today	Several factors including freely availability of data, combining and calculation of computational power helps to make a breakthrough in machine learning and opening the new funding doors about AI potential. Firstly in 2012, Google driverless automated cars navigate while in 2016, a very complicated board game Go has been arranged where Google AlphaGo beaten board game champion.

2.1.2 Artificial Intelligence (AI) Techniques

After exploring the historical timelines, next question comes in mind is related to AI techniques, which helps studies to take forward to next level. There are several techniques in AI, few of those will be discussed in this section are given below:

1. Fuzzy Logic
2. Logic Programming
3. Probabilistic Reasoning
4. Ontology Engineering

2.1.2.1 Fuzzy Logic

Fuzzy logic includes the basic principle of approximate reasoning. Fuzzy logic focusses on approximate then exact solutions and model of reasoning. The main central concept in fuzzy logic is to unlike the conventional logical systems, fuzzy logic targets the modeling of imprecise way of reasoning which help to take part in human activities for making rational decisions in uncertain and imprecise situations [6].

2.1.2.2 Logic Programming

Mainly logic programming era began in 1970's as it was earlier work provided by some automatic theorems and AI. The main goal to achieve the AI by building automated deduction systems. In 1972, it has been led to fundamental idea which were convincing to use logic as programming language. As a result, in 1972, programming in logic (PROLOG) concept has been achieved as the first interpreter was developed in ALGOL-W (a language) by Roussel [7].

2.1.2.3 Probabilistic Reasoning

In uncertainty environment, probabilistic reasoning (PR) belongs to a fully accessible way for calculating the main theoretical and computational models that underlie the plausible reasoning [9]. By using Bayesian networks and Markov networks semantic and graphical model can be made for presenting PR.

There are few situations when PR can be preferably used in AI:

- Situation of predicates is unsure or unclear.
- Domain of predicates become larger and not easy to list down.
- When its certain that during the experiment a clear error have been found.

2.1.2.4 Ontology Engineering

To make understanding for ontology engineering, firstly need to make clear about ontology. An ontology can be defined as: An ontology is equivalent to a knowledge base for Description Logic [10]. The mentioned definition emphasizes the representing login in another logic languages including OBO format, common logic etc. Another main point include in the definition is the important part of knowledge base in descriptive logic-based system and which make it a simple lightweight ontology and a conceptual data model like Enhanced Entity Relationship (EER) or Unified Modeling Language (UML) which could be converted into OWL to make it application ontology or operational ontology by using the formalized data in OWL with a small difference. So, in the area of computer science, Information technology and system engineering is the field of study of models and methodologies for building ontology by using core concepts of formal representations within a domain and relationships between the concepts.

2.2 Machine Learning (ML)

We can define machine using Tom Mitchell definition [13]

"A computer program is said to be a **learn** when user experience E with respect to class of tasks T, a performance measure P is used to measure the performance at specific task T which help to make improvement in experience E".

In this section, main understanding about the machine learning (ML) need to develop here by describing different methods and techniques, types of ML and applications of ML which will help later to make decision about selecting modern data processing.

Normally, whatever instructions have been given to computer in the form programming software tells the computer what it need to do. ML is a way of making a systems good learner for different computer tasks by providing the learning material [11]. There are number of

questions come in my related to types of learning and what algorithms are famous in ML. Let's discuss the ML basic learning methods more deeply.

It is misconception that both are same. The term AI and ML are term in computer science. AI is a big domain where machine uses the intelligence property in the leaning system which has been shown in figure below.

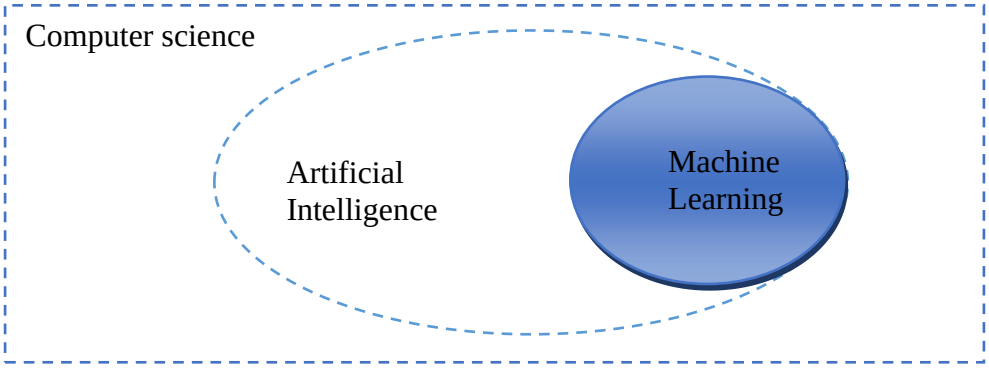


Figure 2: Computer Science, AI & ML

From the Computer Science Figure 2, it has been cleared that ML is a branch or sub part of AI system which use allow system to learn new things from data unlike the AI which is linked decision making. One big difference between AI and ML is success and accuracy level. ML more focuses on accuracy achieve regardless the success factor unlike AI which focus mainly of success of the decisions [5].

2.2.1 Type of Learning

As the ML system include the learning part of the system, so exploring some learning type can play an important part to develop the understanding about its methods which need to be implemented in future. There are three main types of learning which involve in ML, in Figure 3. To explain each learning, one or two commonly used methods involve in learning will be explored to have overview these learning types.

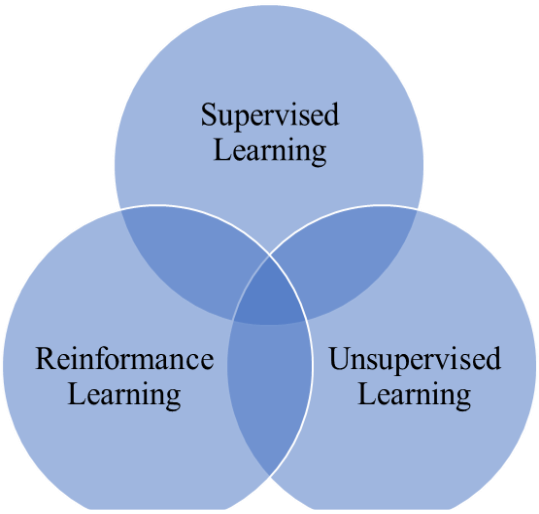


Figure 3: Types of Learning

2.2.1.1 Supervised Learning (SL)

The supervised learning (SL) mainly focusses on formalizing the idea by learning from various examples data. In SL, the learner (that is a computer program), takes the data which is divided into training data set and test data set [12]. The main objectives for the learner are to utilize the labelled example provided in training data sets and identify the unlabeled examples in test data set and get maximum accuracy level as much as possible. One of the goals of learner is to develop a program, set or rules or procedure to make a classification of new example in the test data set and analyze it using already class label we have got [12].

It would be clearer and more understandable with example. Let's take example, we have training data sets which is consisting on different images different vegetables where learner can get the identity of each vegetable image. And test data sets which consist on unidentified part of each vegetable image, from same classes. Then the goal comes for learner to make a program or rules or procedure to identify the elements in the test data sets.

SL can further be divided into two main problems type:

Classification: A problem when we use classification and with its help, we categorized the output in different categories. For example: during diagnosing the cells, we divide it effective or ineffective cells and cell with diseases or with no diseases.

Regression: A SL problem can be regression when we deal with real output value. For example, calculating the weights we get out in real value like 30kg or 30g.
Let's study different popular methods around the SL which describe it more deeply.

2.2.1.1.1 Neural Network

During the various data operations, it has been seen in many situations the given data is not in linear format. In this case, linear algorithm (linear regression) no more help and does not give the good results. For that situations, we must use another algorithm which is called Neural Network (NN), and sometimes its process looks like other SL algorithms. Basically, NN is divided into two layers, input layer and output layer just like logical classification. Graphically, the situation for NN can be explained, from the outside it looks same but different inside in term of operations and algorithms.

Let's explore the Rojas's NN "black box" concept as seen from outside [18]. There are number of inputs $x_1, x_2, x_3 \ldots x_n$ related to input layer while $y_1, y, y_1 \ldots y_n$ (output layer) Figure 4.

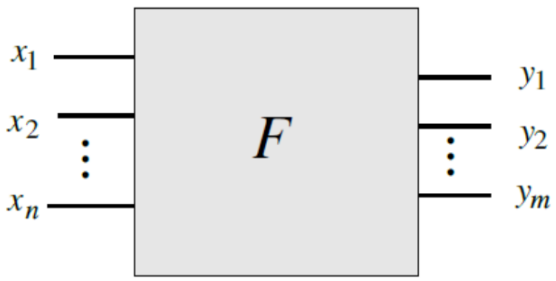


Figure 4: A Neural Network "black box"

NN basically combine the artificial intelligence to make artificial neural network and its operation tried to copy the functions of human brain. As a human brain have ability to do learning, reasoning and decision making and these neurons take part the main role in functionalities and improved over the year with the help of past experiences and history [18].

There is a collection of nodes to support the tasks for NN and these nodes can be seen in the Figure 5.

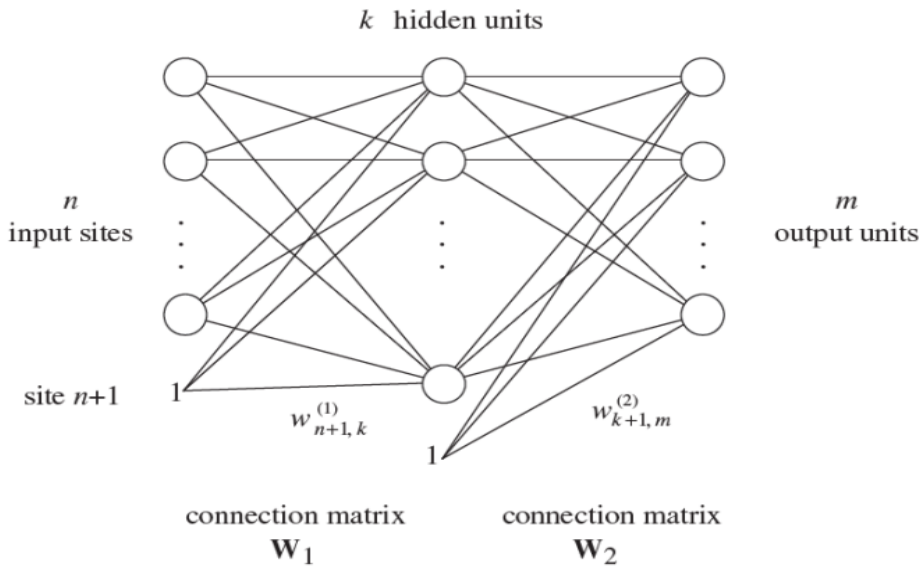


Figure 5: A three-layer neural network [18, p. 165]

These on themselves are primitive functions [17, p. 23]. As a collection though these nodes can do a lot. NNs are being used widely in pattern recognition solutions [17, p. 225-227].

2.2.1.1.2 Feed Forward Network

A direct graph is a good representation of a collections of neurons connected to each other in network. The neurons are the main building block in NN represented by Nodes and arrow shows the relationship between them and the direction can be assessed by arrowhead [19]. Each node has been assigned a number and link between them make them a pair of number, e.g. node 1 and 4 which are connecting to each other can be shown as (1,4).

During the studies, it has been seen that neurons are connected to each other with the help "synaptic weight" or simple "weight". By using these neurons synaptic connections, each neuron receives the weighted information in the network and an output has been produced which use an internal function known as activation function by processing those input weighted information. Here it needs to mention that inputs are those input signal which either comes from external environment or might be output information from some other neurons.

There are two main types of networks in term of architectures depending upon the connections between the neurons like "feed-forward neural network" and "recurrent neural network". A neural network where there is no cycle (without feedback loop), called feed-

forward neural network. In other cases, if a feedback loop exists then the network is called a “recurrent neural network”. Normally, neural networks are categorized based on “layers”. Feed-forward neural networks fall into two categories depending on the number of the layers, either “single layer” or “multi-layer” [19].

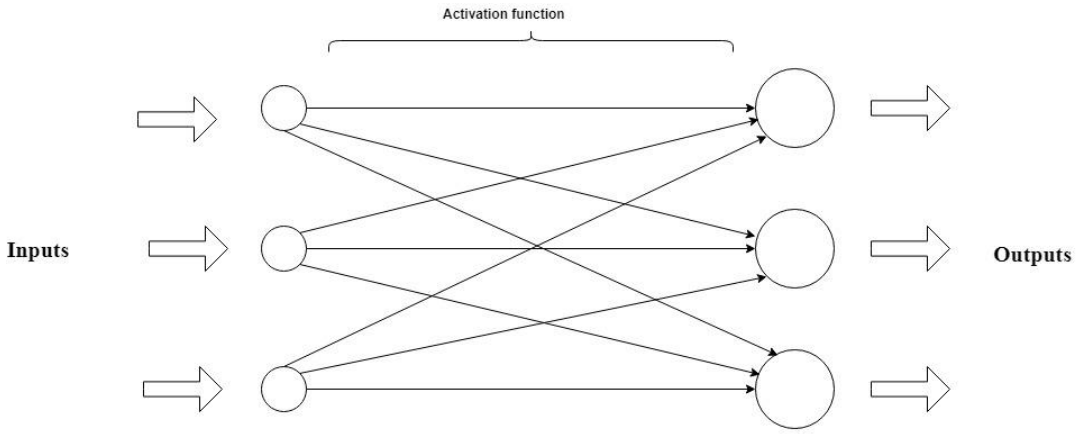


Figure 6: single layer fully connected feed-forward network

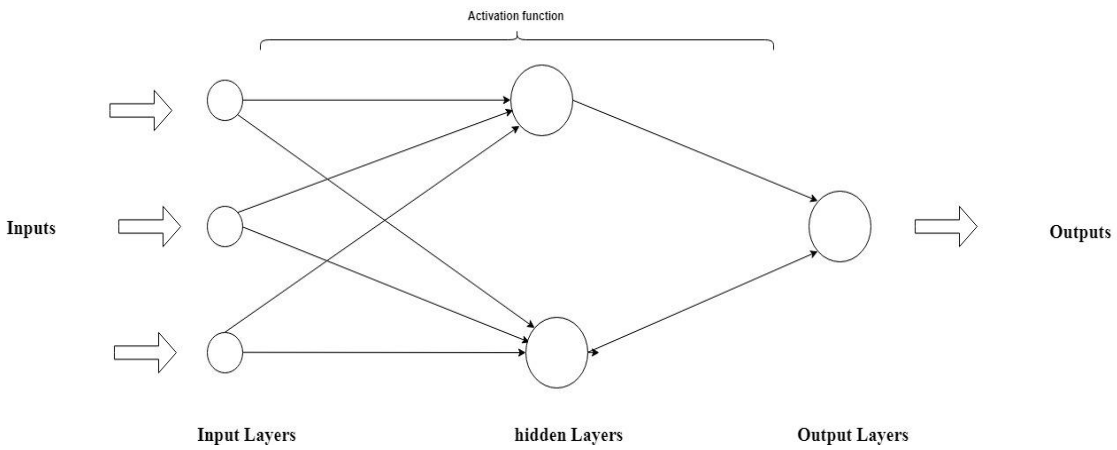


Figure 7: Multi layers fully connected feed-forward network

In Figure 6, a simple example of fully connected feed-forward network is single layers because there are no other layers in the network except input and output layers. In between the layers, an activation function is present which participate in processing, reasoning and transmitting information. If we introduce one or more layers in between input layers to output layers (hidden layers), then example is said to be a multi layers fully connected feed-forward network in Figure 7. In multi layers case, activation function is divided into different steps depending upon the hidden layer’s objectives [19].

2.2.1.2 Unsupervised Learning (UNSL)

In unsupervised learning (UNSL), a machine receives simple inputs (x1, x2, x3, ...) but unlike SL or reinforcement learning, no target outputs. It might be tough to imagine what machine could learn as it does not get proper feedback from its environment. Although, it is quite possible to develop a formal framework or model for UNSL with the help of notion including the goals to representation systems for input to assist it in decision making, prediction for future inputs, effective communication with another machine etc. Thus, UNSL can be thought of finding patterns in available data and that's would be considered a pure unstructured data [14]. The main challenge in UNSL is its subjective behavior in the absence of no goal for the analyses input such as predict response.

Two very simple and conventional example of techniques involve in UNSL are clustering and dimensionality reduction (DR). We will discuss those two one by one in this section.

2.2.1.2.1 Clustering

In clustering, data has been organized in a way that same labeling data come in same group and such group are called cluster. A cluster contain a collection of data having similarities in term of properties and other collections have dissimilarities with other cluster data. See the clustering Figure 8, part-a shows the initial data in rough random format while part-b shows the data in clustered format, red and blue. Each cluster have the similarities in data. There are few methods which can be used for clustering like K-means in details.

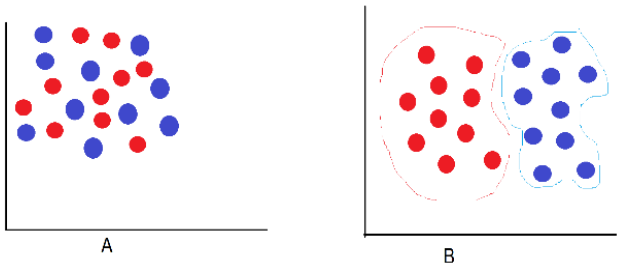


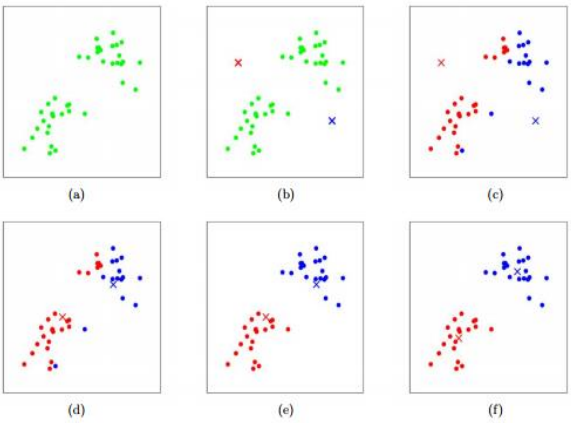
Figure 8: Clustering

K-Means

When provided data has a set of features but absence of labeling unlike the supervised machine learning having Support Vector Machine. We can't make it in supervised learning so UNSL need some other method like K-mean.

K-Means one of the most commonly used clustering algorithm. The idea behind this algorithm is to store the k centroid which can be used to define the clusters. The cluster belongness can be recognized by its distance closer to a specific cluster centroid [15]. K-means tries to find best centroid by changing the position of assigning data point to cluster by using current centroid and selecting cluster by using current data point of cluster.

A simple version for k-means algorithm, training data example is represented by dots while cross shows the centroid cluster. In Figure 9, it starts from 1 to 6 stages, (a) original data set, (b) initial randomly generated centroid cluster, (c to f) initial illustration of two clusters, run two iterations of k-means. Every iteration comes with assignment of training example closest to centroid cluster and painting them and result in a stage where blue color dots belong to blue cluster centroid other red dots belong red cluster [15].



1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

Figure 9: k-mean Clustering [15]

Figure 10: k-mean Clustering [15]

Let's talk about more about the pseudocode in Figure 10, start with initialization of centroid clusters $\mu_1, \mu_2, \mu_3, \ldots \mu_k$, where k belongs to real numbers and use the training example. In second stage, it runs the iterations unless it finds all convergence which is divided into two sets data i^{th} and k^{th} data which have been made in first part by finding the minimum distance to centroid cluster.

2.2.1.2.2 Dimensionality Reduction

In the modern application, expanding the amount data tremendously and raising the complexity level for the data as a result the dimensionalities of data has been increased.

Many big applications related to biology, geology, astronomy, robotics, latest mechanical engineering and well known field of science and technology which require to explore various new data analysis techniques including data analytics, data manipulations, dimensionality reductions and data visualizations produced high dimensional and huge datasets[16].

The main aim for DR is to translate the data from high dimensions to low dimensions in a way that same input objects should be mapped to nearest points on a manifold [16]. Two shortcomings have been seen while using DR, firstly, from the input to manifold could not get a function that can be applied to new points whose relationship to the training points is unknown. Secondly, many methods presuppose the existence of a meaningful (and computable) distance metric in the input space.

The main challenge in finding a function which can map high dimensional input dataset to low dimensional output dataset is the provided neighboring relationships between the sample input space. Graphically, neighborhood relationships information could be received from source data that may not be available for test points, such as prior knowledge, manual labeling, etc.[2]

2.2.1.3 Reinforcement Learning

Reinforcement learning (RL) takes us back to early studies of cybernetics, statistics, neurology, psychology, and computer science. During the last five years, it has been seen the interest in artificial intelligence and machine learning societies rapidly increased. It can simply describe that a way of programming agents by reward and punishment regardless of need to specify how the task is to be achieved. The main challenge in reinforcement learning is to make agent with the introducing learning behavior through trial and error interaction with a dynamic environment. RL handle the problems related to making learning what to do when a specific situation comes and what actions must take to maximize a numerical reward signal [23].

Reinforcement learning is different from unsupervised learning type of machine learning which mainly dealing with finding the structure in unlabeled datasets. Classification of

[2] Introduction to Dimensionality Reduction: https://www.geeksforgeeks.org/dimensionality-reduction/

machine learning in two not enough and do not classify properly. Although, one might think of reinforcement learning as a kind of unsupervised learning because it does not rely on correct behavior of given datasets, reinforcement learning tries to maximize a reward signal instead of finding structure hidden in example data [23]. There mostly, research categorized the reinforcement learning as third type of machine learning paradigm alongside of supervised learning, unsupervised learning, and perhaps other paradigms as well.

Examples:

To make understanding clear, let's talk about few examples in real life.

- In real time, a very simple example includes the adoptive controller which helps to control the parameters for petroleum refinery's operation. The main tasks for adoptive controller are to optimize the yield, cost, quality or trade off by using specific marginal costs without sticking strictly to the set points originally suggested by engineers.
- In a new room, a mobile robot decides to enter in room and find a more trash to collect and find a way to put it back to a charging station. To decide, it uses the current status of the battery (current state), try to assess the situation in quick response or actions and easy in find the recharger in the past.

2.2.2 Machine Learning Models

The ML models play an important role while integrating it in existing system due to compatibility check between models and systems functionalities. In general, ML models can be divided into three kind of models on the bases of methodologies and approaches have been used in them. These model categories are given below.

2.2.2.1 Logical Model

In logical model, we have logical expression which consist on two parts segments and selection of models. An expression which return true or false as a result is called logical expression. Initially, data need to group using logical expression based on same characteristics and then apply the different algorithms match with problem's situation. For example, during the solving of classification problems, all the instances are grouping belong to same class.

Generally, logical models are categorized into two ways, one is tree model and other one is rule model. Rule models focuses on a collection of implications like if happens this then happens this (if-then). In these models, if part defines the segmentation part and then part shows the behavior. In algorithm section, the explanation of tree models like decision tree is available.

2.2.2.2 Geometric Model

In previous part, we have read about the logical models, having decision tree models where data is being partitioned by logical expression. In this section, a short description about the second ML model, geometric model which focus on considering the geometry of the instance spaces. In geometric model, the different features of the problem can be described in different dimensional like 2D or 3D having (x or y-axis) and (x, y and z-axis) respectively. Sometimes, the given features are not in geometric form, then we need to model the features to patch up with intrinsically geometric. For example, we calculate the temperature which is not in geometric form so it can be modeled in two axes like temperature against axes. Linear model is one of the examples of geometric model that is discuss later.

2.2.2.3 Probabilistic Model

After reading about above two ML model's family, there comes the third type models include in ML which is probabilistic model (PM). Probabilistic model uses the probability functions for the classification of new entities. In PM, each feature and target are considered and imagined as random variable. By using these variables, level of uncertainty needs to find out. Predictive and generative are two main types of probabilistic model. Predictive probability uses the concepts of conditional probability like one variable can be predict from another variable. On the other hand, generative model uses the joint distribution to estimate the target, joint distribution uses the conditional or marginal distribution which play roles in identical variables.

2.2.3 Machine Learning Algorithms

In this section, we will discuss some algorithms use in ML. There are a huge number of algorithms, but our focus will be only on famous ones.

2.2.3.1 Support vector machine

In 1992, Guyon and Vapnik introduced Support vector machine (SVM) for the first time. The main success of SVM is due to its ability to perfect in recognition of handwritten digits. With the help of kernel methods, SVM is actively playing an important role in ML research[3].

SVM is a ML algorithm which focuses on analyzing the data for classifications and regression analysis. The algorithm divides the given data sets and sort it in two different categories and as a result it delivers the output map of sorted data. SVM can be used in solving different problems like text categorization, image data classification, handwritten recognition.

Let's go through the SVM pseudo code:

- we are given a set S of points xi ∈ R where {i = 1, 2, 3, ... n}
- data is divided into two categories
- Each xi belong to either class
- given a label $y_i \in \{-1,1\}$.
- Set S leaving all the points on the same side having same class.
- Finally, an equation of hyperplane establishes
- To performs classification, an N-dimensional hyperplane has been constructed that optimally separates the data into two categories.

Classification of data with SVM:

Let's see how classification work here using the Figure 11. On the left side, the objects are non-classified or illusion form. It is cleared from figure; objects belong to two different classes. A 2-dimentional hyperplane as separating line has been used in right side figure which is a decision plan and divides the object into two subsets such that in each subset all elements are similar.

[3] Support vector machines http://cs.joensuu.fi/pages/whamalai/expert/svm.pdf

Figure 11: Support vector machine (svm) simple classification

Advantages:

- SVMs deliver a unique solution.
- SVMs gain flexibility in the choice of the form of the threshold

Disadvantages:

- Due to non-parametric technique, in SVM, there is a lack of transparency in results.

2.2.3.2 Decision tree

A decision tree is a visual and explicit way of representing decision and decision makings. A decision tree acts as classifier to express recursive partitioning of the instance space. In decision tree, there several nodes, the main nodes which don't have any incoming edges called 'root'. So, the root node acts as main starting node and all other nodes have incoming edges. The node with outgoing edge is called test node or internal node and rest of the nodes are called leave node which can also call decision node.

According to a discrete function, each internal node is divided into two or more child nodes. Simply, each test validates a specific attribute so that instance is partitioning on the basis on attributes values[4].

[4] Decision Trees: http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

```
TreeGrowing (S,A,y)
Where:
S - Training Set
A - Input Feature Set
y - Target Feature
Create a new tree T with a single root node.
IF One of the Stopping Criteria is fulfilled THEN
    Mark the root node in T as a leaf with the most
    common value of y in S as a label.
ELSE
    Find a discrete function f(A) of the input
        attributes values such that splitting S
        according to f(A)'s outcomes (v1,...,vn) gains
        the best splitting metric.
    IF best splitting metric > treshold THEN
        Label t with f(A)
        FOR each outcome vi of f(A):
            Set Subtreei= TreeGrowing (σf(A)=vi S,A,y).
            Connect the root node of tT to Subtreei with
                    an edge that is labelled as vi
        END FOR
    ELSE
        Mark the root node in T as a leaf with the most
            common value of y in S as a label.
    END IF
END IF
RETURN T
TreePruning (S,T,y)
Where:
S - Training Set
y - Target Feature
T - The tree to be pruned
DO
    Select a node t in T such that pruning it
        maximally improve some evaluation criteria
    IF t≠Ø THEN T=pruned(T,t)
UNTIL t=Ø
RETURN T
```

Figure 12: Top-down algorithm for decision tree [25]

A top down simple decision tree induction algorithm has been illustrated in Figure 12. It focuses three main aspects or parameter Training Set S, Target feature y and an input feature to create a decision tree T. Unlike the creating tree algorithm, tree pruning absence input feature attribute which has been show in algorithm.

In Figure 13, you will see the decision tree for direct mailing system having restriction of age greater than or equal to 25 and customer gender must be a male. Here, the main root node would age where decision tree is going to start.

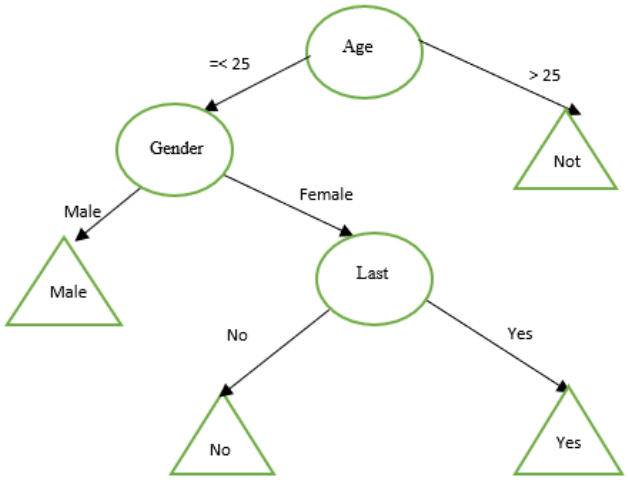


Figure 13: Decision Tree to illustrate Response to Direct Mailing

Advantages:

- Scaling of data is not required.
- Normalization of data is not required.
- Due to easy and simple process, briefing to technical persons and stakeholders never an issue.
- DT requires less effort for data processing.

Disadvantages:

- Need higher time to train model.
- Inadequate for applying regression and predicting continuous values.

2.2.3.3 K-Nearest Neighbour (KNN)

K-nearest neighbour algorithm (KNN) is quite a simple, non-parametric and straightforward algorithm which uses a classifier belong to closest neighboring class. Commonly, there are more than one neighbors participate, so knn can become a good choice where k refers to numbers of nearest neighboring classes [26]. If the training examples are based on run-time then run-time needs to be allocated, called as memory-based classification. Due to directly link with training examples, also known as example-based classification.

The Figure 14 illustrates the Nearest Neighbour Classifier, here k value is 3 so its 3-NN and applicable on two class in 2-dimentional feature space. The q_1 and q_2, two classify examples

have shown in the figure, q_1 is quite straightforward having 3 nearest neighbors same class O while q_2 is complicated having mixed neighbors (one from O class and two from X class). In q_2 case, complexity can be resolved by using majority voting system and calculate the weight using distance functions like Euclidean distance/Manhattan distance/Correlation distance[5].

Thus, knn algorithm has been divided into two steps, first one for storing all the training data while second step further has two stages. In second step, finding the nearest neighbors and class which the neighbour belongs to [26].

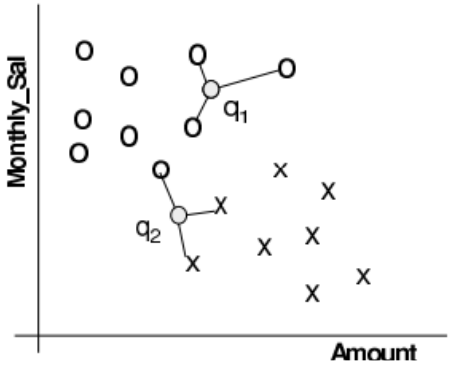


Figure 14: Decision Tree to illustrate Response to Direct Mailing

Advantages:

- Easy to implement
- No assumptions – means there is nothing to assume because of non-parametric nature.
- It supports classification and regression problem so it can be used for both.
- Cost of learning in knn is almost zero[3].

Disadvantages:

- K-NN is very slow algorithm.
- As K-NN works well in small input variables, so it becomes worst as dimensionality increase.
- K-NN need to calculate the parameter K (nearest neighbour).

[5] https://www.cs.upc.edu/~bejar/apren/docum/trans/03d-algind-knn-eng.pdf

2.3 Summary

To implement the ML in existing systems, its better to understand first the basic of machine learning and exploring its models and methods. In this Chapter, study has been started with AI (as ML is a type of AI). By summarizing this chapter, it has been divided into two part. Firstly, a short description about AI helps to understand background of this by studying the history and techniques of AI. In the later part, ML (the key focus of our study) has been explored in terms of types of learning, models and algorithms. The overall structure of this chapter can be seen in Figure 15.

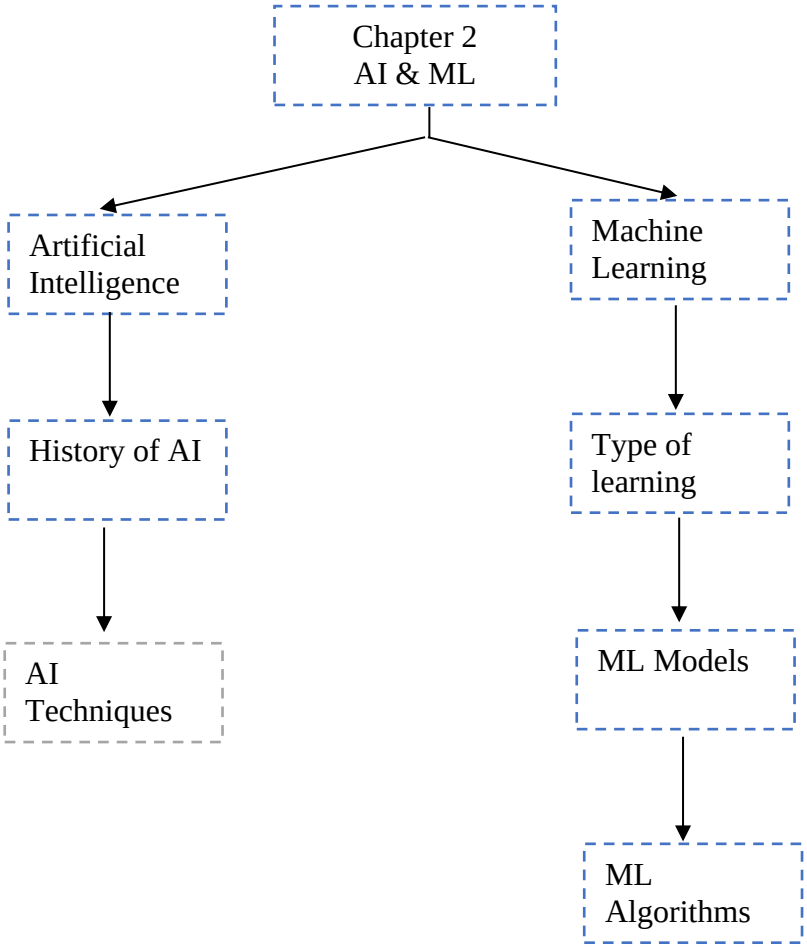


Figure 15: Summery of this Chapter

3 MACHINE LEARNING INTEGRARION

In this Chapter, main studies will explore the ML architectures and models. After studying the ML models in Chapter 2 and architectures in this chapter, we will try to combine the knowledge and move toward machine learning models and then studies the tool used for integrations and integrate the machine.

3.1 ML integration architecture

In this section, our focus would be to understand and trying to explore each major component in designing the architecture for ML. There are different ways of designing ML architectures and only depends upon the system need to integrate with ML. In every ML architectures, few components are common and used in various systems.

The focus of a ML system includes the processing, manipulating, and reasoning about data and applying different operations using different algorithms. According to Gartner, there is a specific process to execute to basic architecture. To understand the initial map for architecture, have a look on the Figure 16 below describe the stages for the ML process:

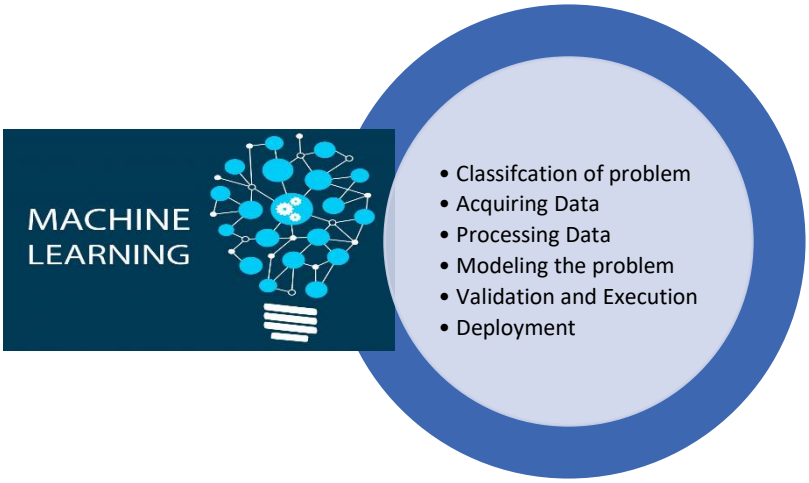


Figure 16: Stages of ML integration Process

Gartner's ML process includes many elements related to infrastructure and being used in many cloud-based systems. The cloud is most relevant and excellent combination for many ML application due to its fast processing and handling of huge data. The Table 2 below shows a short description of the stages in ML process.

Table 2: Stages of ML integration Process

Stages	Description
Classification of problem	The main focus on developing the problem taxonomy by keeping in mind the problem or problem statement.
Acquiring Data	Try to acquire the data which supports the problem and it can come through various ML process or device like IoT.
Processing Data	Include processing the data operation using various data algorithm to make it ready for execution later on.
Modeling the problem	Modeling the acquired data and apply different ML algorithm to design the solution.
Validation and execution	This stage includes the validating the result and run the ML basic routine and refining the results.
Deployment	At the end, results need to deploy the problematic place to solve it and see the real-life results.

After going through the case studies, it has been realized that architecture and ML methods are the most essential artifacts in integration of ML in the existing. The various ML methods have already been described in previous chapter, so only architecture will be discussed now with respect to ML integration.

When we talk about the architecture related to integration of ML, the Gartner's architecture weighted high in term of ML. The fig shows how the above-mentioned components of integrated ML works in the existing system. These are the common components that integrated ML system have to perform the various operations.

It starts with data acquisition which retrieve the data from various sources like ERP databases, IoT devices or web servers or etc. which connect to batch data in data warehouses. Then these batched data are being sent for further processing where various feature of these feature need to extract. For example, in various ML integrating System, where initial data is being sent for the operations like parsing data where different feature are being extracted.

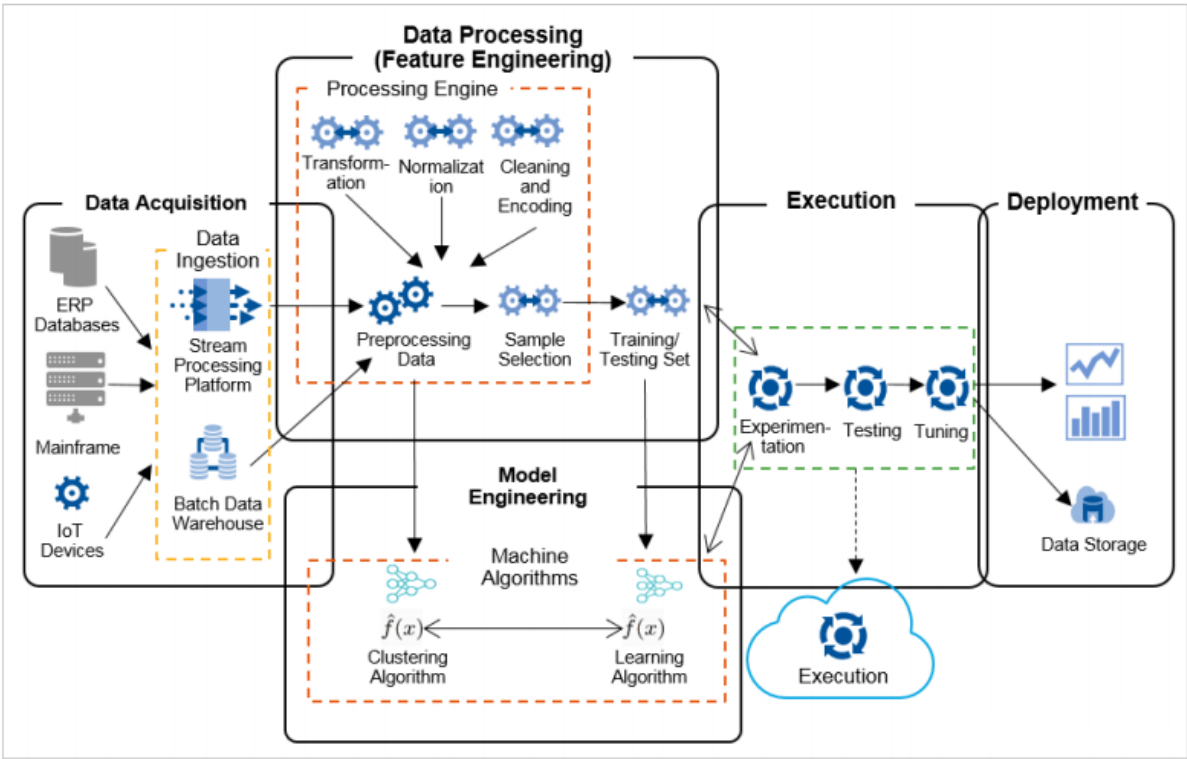


Figure 17: The ML integration architecture (with components) [30]

After processing the data, the data has been sent for modeling data using various ML algorithms it might clustering or anything other based on nature of requirements. After applying the algorithms on learner, it need to send for execution which need to perform experiments or testing of algorithms and if does not follow the guidelines then it must go back to process the algorithms again to achieve the basic objective of the system. If we see again the case study two, we can see the after applying some sort of algorithm on parsing the data, it needs to pass the tests (unit testing or regression testing or etc.) otherwise it will process again to get the green signal from the testing or experiments. After passing the testing, finally its ready to deploy on data server or storage and back to the case study, its ready for data analysis purpose.

Nowadays, the ML architectures are increasingly integrated with different cloud-based system where various micro-services are being used along with ML. DevOps infrastructure is highly used along with ML to optimize the software development strategies in many to detect the failure or pass in new release for different software delivery life cycle. The DevOps provide the opportunities for integrating custom ML and AI in existing systems, as

well as opportunities to leverage ML and AI as a tool for improving operational efficiencies within the IT organization.

3.2 Developing the Machine Learning (ML) model

When we go deep in ML integration, the model plays an important role to describe basic flow of the system. To integrate the ML in system, we must look at generic model which we have been realized after completing the case studies. One of the key points is to understand the existing system's component to integrate especially where performance need to improve by introducing the ML in system.

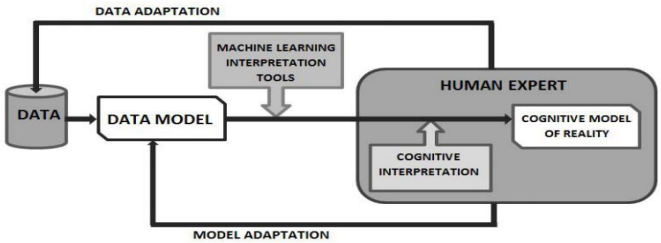


Figure 18: The graphical representation of ML Model [30]

The graphical representation of a ML model which we can integrate in any existing system can be seen in the Figure 18. The suggested of ML model having different components which perform different functions based on system requirements. ML system most importantly deals with data operations and data is the most important attribute in the system.

The communication between Data, Data model and interaction way with human expert system represented in the figure. Data model plays the role of bridge between input data and output data while ML interpretation tools. The interpretation tools need to be discussed in next section, while human expert component need to be adapted with provided data model.

3.3 Machine Learning Integration Tools

There are the numbers of tools or opportunities which various organizations are used to integrate the ML in the existing system. The tools selection depends upon the platform of the existing system where you want to introduce the ML. The tools which are commonly used in integrations of ML is given below in the Table 3. For example, if the existing system

is being developed java, then the tools to implement ML must also compatible and Weka[6] sounds a best tool for implementing ML algorithms.

Table 3: ML integration support tools/library/platform

Name	System Platform	Description	Features
Scikit-learn	Python	Scikit-learn is used for machine learning development if the existing system support Python. This tool provides the library for the python programing language to integrate the ML functionalities in existing system	• It helps to data processing, data analysis and data manipulation. • It helps to integrate many ML algorithms like classification, Dimensional reduction, clustering and regression. • Mostly features are freely available.
TensorFlow	JavaScript	It is very useful when existing system developed in JavaScript. It helps to provide library for developing ML system for trained and build the data model.	• It is helpful when ML system need to integrate the neural network. • By using TensorFlow library in the system, data model can be converted easily ML model. • Overall, very helpful for training and building data model according to existing system. • Good programming skill is required to implement this in ML because it's difficult to learn.

[6] https://www.cs.waikato.ac.nz/ml/weka/

Weka	Java	A data mining software using the existing system is developed in ML.	It supports several ML algorithms including clustering, classification, regression, data visualization and other data mining.
KNIME	Multi-language platform	Knime is knows as a tool for integrating the data analytics, reporting, fixing and integration platform. By using the data pipelining concept, it helps to combine component of ML and data mining.	• It helps to integrate the ML algorithms written in various programming languages like C, C++, R, Python, Java, JavaScript etc. • It is very easy to learn. • Easy to implement in data mining operations.
Colab	Python	Colab supported by google is a cloud service which supports Python. It helps the system to build the machine learning applications using the other supported libraries of PyTorch, Keras, TensorFlow, and OpenCV	• It is useful for learning purpose system for ML education • It assists in machine learning research system.

There are several ML integration tools but, in this section, we have tried to highlight the mainly and famous tools.

3.4 Machine learning Integration

The ML integration in existing system can be made by combining the previously discussed components together. The previous components like ML algorithms, ML process model, ML integration architectures need to be decided before starting the integration ML in conventional system. While deciding these components, implementer need to find the answers of following point:

- Study the conventional system feasibility.
- Selection of algorithms.

- Selection of architecture or process model.
- Selection of tools according to platform.
- Implementation.
- Deployment.

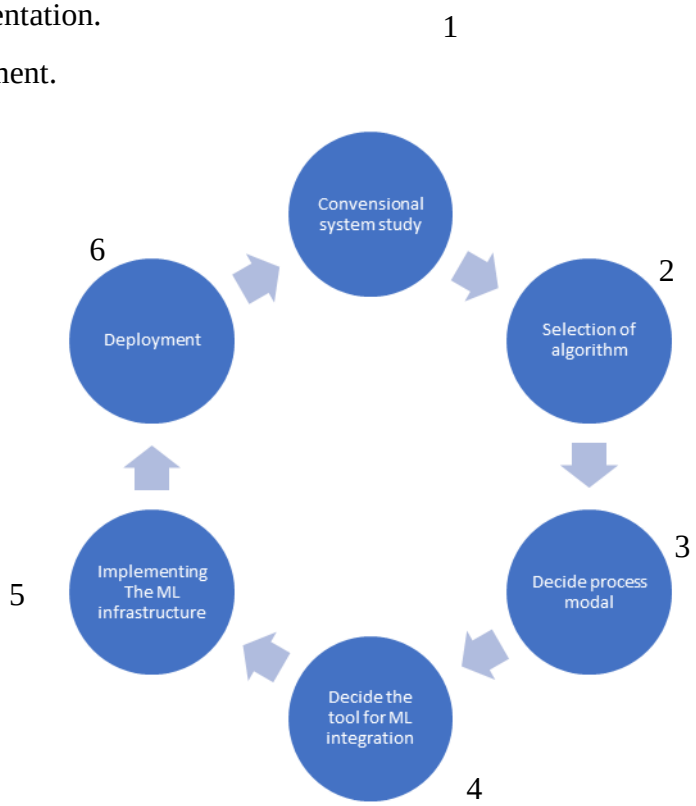


Figure 19: ML integration based on components

For instance, ML integration in conventional system initiates with the existing system studies and move next to algorithms selection like which algorithm is more suitable clustering or regression or etc. depending upon the system requirement. Then next one is to finalize the process model which might include the follows the steps or data model and then take the decision about the tools by keeping in mind the selected component's compatibility. Finally, implement the selected design infrastructure and deploy the ML integrated system.

In Figure 19 shows the basic ML integration component which need to follow while replacing it in existing system. While studying the components of ML integration, three components are the key and plays important roles in term of functionalities of the existing system. It is necessary to explore more those components of the system.

First one is selection of algorithms because incompatible algorithm selection can mislead the systems capability. To overcome this highlighted problem, it's keen to study the

algorithms and read the pros and cons of the algorithms. The algorithms always have some limitations to operate in conditions (specific or general). If algorithms match with the system required functions, then it need to be selected. Secondly, selecting the ML integration tools also key component by keeping in mind the system platform as well. For example, while choosing the tool in case study one, the crucial task was to choose the tool because most of the convention platform has been developed in JavaScript programming language. Lastly, the implementing the ML integration infrastructure in existing systems because at this stage make or break situation happens.

3.5 Summary

In this chapter, main study includes:

- Developing the ML integration Architecture by analyzing the Gartner's studies [30].
- Explore the ML integration Model and process.
- Listing the key ML integration tools based on their compatibility with different platforms.
- Exploring the key ML integration process important steps (extra).

4 MACHINE LEARNING INTEGRATION - CASE STUDIES

In this chapter, focus would be the studies of two case studies where integration has been used in already running system.

4.1 Case study one

In first case study, we are going to make studies about a well reputed product which follow the basic systems by using a machine learning architecture to perform various operations including processing and manipulation of data. For writing this case studies, I have personal interaction with CTO in that company (hiding the name for copyright or personal data laws) to discuss about the systems perspective operations.

In term of use cases that integration need to address, it is important to understand the flaws or bugs in conventional system. Firstly, the conventional system has a problem in parsing input files, managing the expanding data and problem in processing (absent of cloud-based system). Thus, these three are the main cases which need to focus during the implementation of ML in existing system.

4.1.1 Background

Nowadays, its IA and ML era, in many systems, ML methods have been playing a major role in growing the productivity and performance like campaigning for brands and finding the right target audience [1]. The selected system can be named as discount-based system (DBS). The DBS provides the basic access to consumer for finding the all sales products which are on discount labeling by different fashion brands. The DBS acts as platform for both seller and buyer to provide to combine all the products are for sales. The basic objective this system is to facilitate the users in buying cheapest products without visiting the different branch specific websites by keeping the users' preferences and provide best experiences for the users. Later, we will discuss various aspects of the integration in this system.

4.1.2 How it works?

The DBS starts with the integration different data in various format depends upon the brand or store specific systems. After integration the provided data, the products are available on DBS platform for various users in available markets. The users or buyer can pass through the selected stores and buy their products easily and securely. System also supports other internal tasks including managing various part of the system.

4.1.3 Architecture of conventional system

Let's study the architecture studies involve in this existing system design before the integration of ML. A simple graphical higher-level diagram can be seen below Figure 20 to represent conventional system. The figure shows the architecture of conventional system where input files need to process and store the data which make the speed of data flow is lower across the conventional system. Server part of the old system contained some small integrated tools, but no machine learning methods were being used.

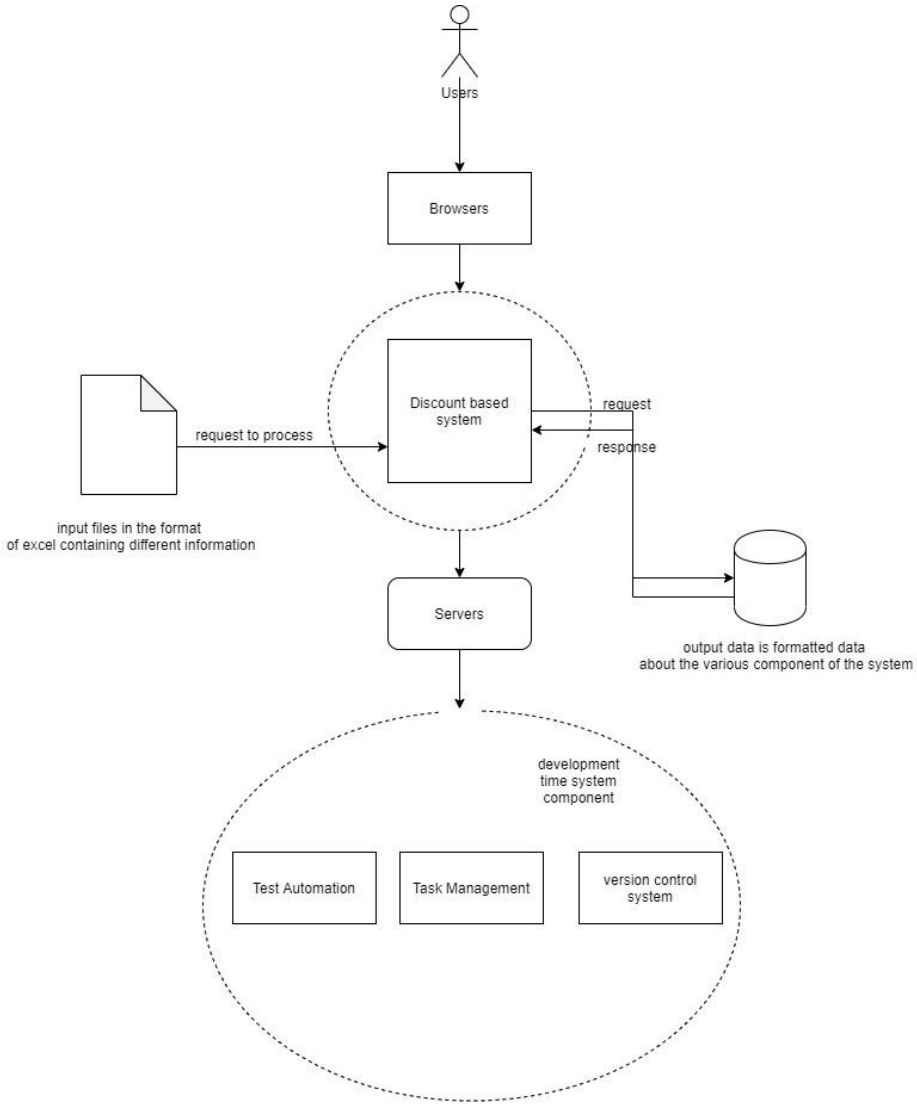


Figure 20: The architecture of conventional system

There are some development time components which is also play an important role of the existing system given in Table 4.

Table 4: Development time component of the existing system

Components	Description
Task Management	The task management includes the various functions like integration of managing tool. JIRA management tool has been used to managing the task during the development and support of this DBS. Jira is world top class software development tool for the team follow the agile methodology. The JIRA support the agile functionalities like task management for team and agile project management related tasks. Another communication tool is involved in the operation is to integration the of SLACK which very useful for communication and collaboration tool within the team.
Version Control	A version control is said to be a component of software configuration management who involve in managing different sort of changes to documents, source code of software that might be collections of various information related to system. Currently GitHub is been used as version control system. The various changes in the information in term of source code is being identified by using version naming in ascending order like version1.0, version1.1 and soon.
Test automation	Test automation part consist on different kind of code-based testing including testing library like PhpUnit etc. The code-based testing to run these tests for every line including source code and documents of the systems. The main issues in the current pipeline is to add some sort of automated system which will help the validation and testing it server-end.

4.1.4 Challenges

While studying the conventional system, a list of few challenges has been found which can be addressed by integration ML part in the existing system.

Main challenges are to find the answers of following questions:

- What challenges the system is facing right?
- Why we want to integration the ML in the system?
- What is architecture modification need to make?
- How could we integrate ML operations or function in the system?

When we study the system architecture provided above, system need some sort of modification in term of architecture and handling data flow and improve it. During the discussion with related company personal, it has been realized that managing the increasing data day by day creating a huge challenge for the system. That why there need some major changes in the system to enhance the productivity and speed in term of operations. The architecture wise, it needs to add some sort of data parsing in the beginning and data processing including DevOps methodology and finally integration some statistics services that will discuss in next section.

4.1.5 Approach

The main approach in this case studies has been included with the introduction of parsing method while processing data. The parsing has been divided into three components which performs three different kinds of function. There are the main parts of the parsing which is given in the Table 5:

Table 5: ML Data parsing Techniques using in Latest DBS

Method	Description
Feature extraction	A ML technique which has been used normally in text analyzing for getting insights from data. It performs its functions by extracting pieces of data from the already existing given data or text, so if we wants to extract the important information available in trained data (using trained data model) like keywords or piece of data (like brands, price, tracking information and etc.). After organizing the data, it can be used in different supporting text analyzer tools.
Text Classification	Text classification (a.k.a. *text categorization* or *text tagging*) perform the tasks by performing the assignment of data set of predefined categories to free text. Text classifiers has been used

	to design the data, structure, and categorize pretty much anything. For example, new given rough data which can be organized by categories, discount can be organized by prices, brand mentions can be organized by sentiment, and so on. Here, the text classifier C can be organized by utilizing some general parsers where different categories are already developed in the system.
Keywords extraction	The given data consist on different important keywords which is more relevant to the terms. This technique helps to assigning the index to various keywords which need to be search and generate clouds tags support the Search Engine Optimization (SEO) related operations, clouds analyzing, marketing etc.

4.1.6 The ML integration general framework

The integration the ML in the existing system need some sort of pre studies and feasibility of the system by calculating relation between different component of the system. To understand the relationship among system's component, it is essential to make the flow of information clear to get clear understanding.

After the study of main challenges of the system main approach has been developed for the ML integration in the previous section. The general integrated framework for the existing system with inclusion of ML methods is given below in Figure 21.

The current framework includes and covers three major aspects of the new introduced system. Firstly, the parsing the input data files, extracting the terms by using the trained data which is consist on general parsers. Generic parsers used keywords extraction method and assigned the indexed to extracted terms. A simple classifier has been used here to group the data based on indexes and the organized the list which need to deliver next operations of the system. For example, the tags data are being used for the online marketing and SEO as well.

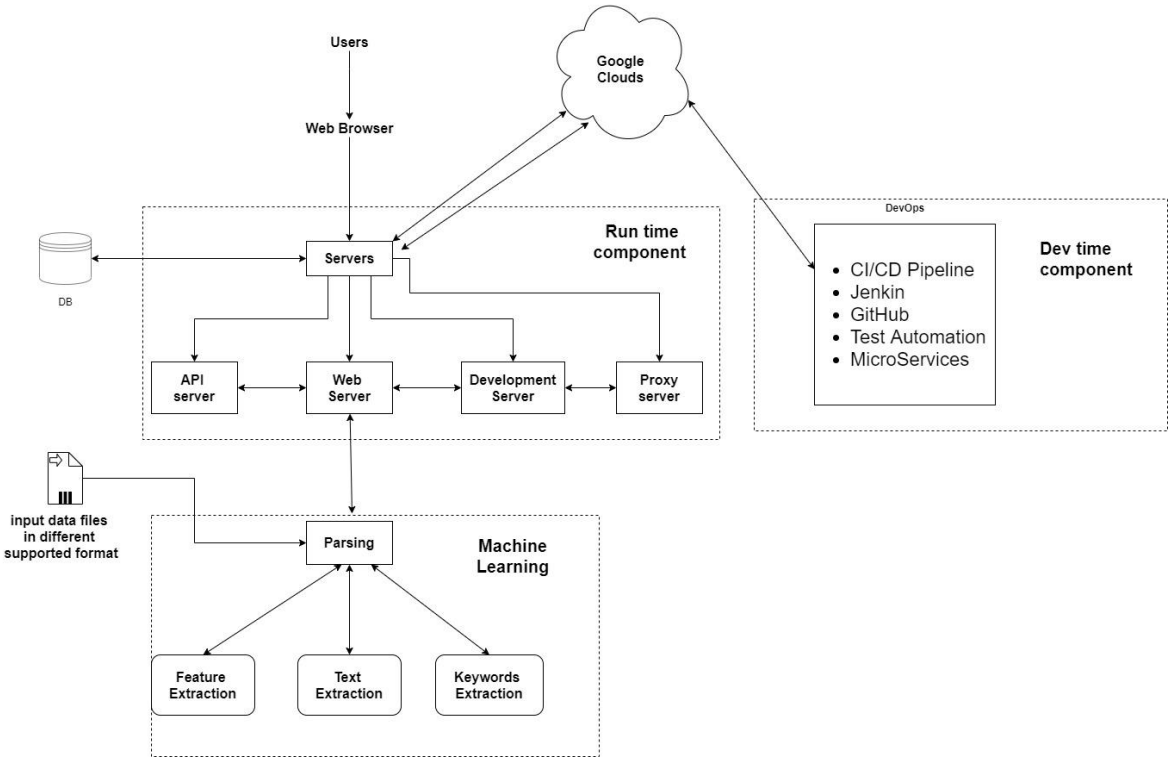


Figure 21: The general framework with ML integration in existing system

Secondly, the server's component is divided into few other small servers to share the load among servers like API server, Web Server, Development server, Proxy server and Database server. The main server (Web server) which directly interact with the user, access by the web browsers and deals with different request and get view response to the user.

Finally, the last component of the system which includes the cloud-based ML techniques. DevOps are the main services here to optimize the system and supported by various ML methods like keywords output in data parsing is used for traffic analyses.

4.1.7 Summary

By the summarizing the key finding during this case study, we have found that there is not always need to modify the data related operations. Sometimes introducing the new architecture in the system with related to ML in existing application resulted in increase the productivity of the system.

In Table 6, a comparison has been made with the system prior to integration of ML and after the ML integration. Thus, the comparison table highlights only the main factors involve in introducing ML. For instance, Figure 18 represents the existing system in the case study with the absence of ML methods while Figure 19 described the new system (after ML methods integration) architecture.

Table 6: Conventional system vs ML integrated system

Conventional System	ML integrated System
No ML methods	ML Methods are being used
Low system's speed, because system data is getting expand, and old conventional system was unsuccessful to speed up.	System processing speed increased because parsing power has been helped to increase as the trained data. Using that ML methods, parsing the given data, making data analysis tools.
Make flow of data tough and complex	Make flow of data very easy by diving the system on different server essential for specific functions by integrating cloud-based ML which helps the new system to make data flow smooth and simple and easy to manage.
Figure 18	Figure 19

During the discussion with responsible person, the system feedback is much positive and had made the user's life easy and increase the profit in term of sales as well. In future, it can be improved further to make easier the flow of information as system is growing rapidly and huge number of data is being received on daily visitors to DBS. In future, these visitor's data can be optimized more to improve online or digital marketing for the DBS.

4.2 Case study two

In this case, the focus to study the integration part in one of the most important components of the system and the focused system is the Wikipedia. We are going to study the system before and after the integration of ML in their specific component. The case studies involve in enhancing the cluster labeling in Wikipedia using a ML technique like clustering and labeling terms.

In term of use case involve in this case, the only use case to address is to improve feature extraction and keywords search by integrating ML method (clustering algorithm) and replace the conventional methodology with ML clustering methodology.

4.2.1 Background

As we know that the electronic information era is going rapidly increasing with the advancement many digital processing. As the result, with gaining of huge amount textual data which have come up with new and efficient data processing techniques to organize the data in organized forms.

In this scenario, the clustering algorithm seem to be most relevant for organizing textual data which allow and help to make the digital copies of each part of the work for personal use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. The prior specific permission or fee is required to copy or republish on various servers or redistribution lists.

While the studies, it has been studied in many clustering applications especially based on human interaction, human users directly interact with the available clusters. In this kind of application, we must label the cluster in a way that human users can easily get the understanding what is cluster about. A most common approach involving labeling clustering is to use statistical methods for feature selection. Normally, this has been carried out by identifying the key terms available in the text which represent the main cluster topic. But despite that many times keywords or phrases get failed to return the meaningful readable label for the set of documents.

Table 7: Lists of top-5 important terms extracted using the JSD selection method and top-5 labels extracted using Wikipedia for several ODP categories. [27]

ODP Category	Top-5 JSD important terms	Top-5 Labels Using Wikipedia Enhancement
Bowling	*bowl*, bowler, lane, bowl center, league	Bowls, *Bowling*, Bowling (cricket), Bowling organisations, Bowling competitions
Buddhism	buddhist, *buddhism*, buddha, zen, dharma	*Buddhism*, History of Buddhism, Buddhism by country, Tibetan Buddhism, Buddhists
Ice Hockey	hockey, nhl, hockey league, coach, head coach	*Ice hockey*, Ice hockey leagues, Hockey prospects, Canadian ice hockey coaches, National Hockey League
Electronics	voltage, high voltage, circuit, laser, power supply	*Electronics*, Power electronics, Diodes, Power supplies, Electronics terms
Tennis Players	wimbledon, tennis, defeat, match today, wta	*Tennis Players*, Tennis terminology, Tennis tournaments, 2002 in tennis, 2000 in tennis
Christianity	church, catholic, ministry, christ, grace	*Christianity*, Christian denominations, Non-denominational Christianity, Christian theology, Christianity in Singapore

While illustrating the concept, look at the table 7 shows the main important terms (5) that could be extracted for six Open Directory Project (ODP) topics with the help of JSD selection method. A collection of clusters of 100 web documents based of random sample from the corresponding ODP category represent each topic in the term. With the help of 100 ODP categories, the original label linked with the category has been discovered provided by human assessor which included 85% of the categories of the category's text. Although, label's text obtained by human are hardly identified significant by using feature selection methods.

4.2.2 How it works?

Overall, the conventional system takes the input in the form documents and process the text or terms in the documents and stores that text information in database (DB) server. After that initial information, the terms are sent to next components. Most of its internal working is protected one, not easily available online. Before the clustering methods, the information handling and text extraction methods are some random one which is not easy and compatible with existing system.

4.2.3 Challenges

The main challenges involve behind the integration of ML is to improve the system performance in term of processing and feature extraction. The main challenges are given below:

- Improve the feature extraction processing.
- Improve the flow of data without changes or modify the other component architectures.
- Speed of processing need to increase.
- Managing text involved in provided documents.

4.2.4 Architecture of conventional system

In term of architecture, it's not possible to study and access the overall architecture for this existing system. But one thing is sure, in term of architecture, that we have replace the methods for text extraction which follow the same architecture as before (see Figure 22). Only modification in architecture is to integrate the ML clustering methods.

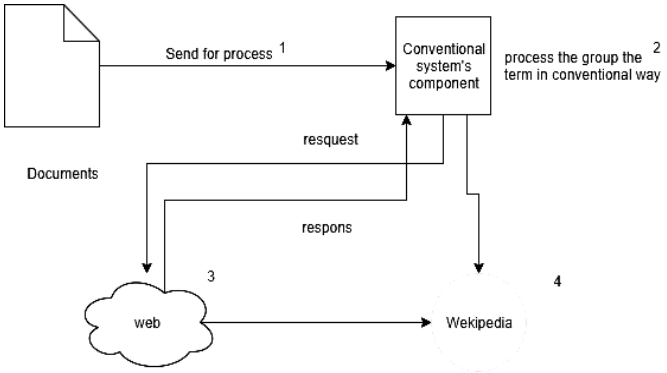


Figure 22: Architecture of conventional component [27]

4.2.5 Approach

In this section, we can discuss the approach which has been used in this case studies. During this case studies, a short investigation regarding the contribution of external knowledge based for labeling cluster it has seen with the help set of documents using Wikipedia, different related topics are being identified by the close work presented by Syed [28].

In approach for this case studies, we can find the few main points involve in this:

- Firstly, find the Wikipedia those pages which are more relevant to cluster need to be labeled.
- Secondly, used the meta data of Wikipedia pages which includes many important aspects (like page title or categories) to support the main experiments.
- Look at the Table 1, for the set of ODP topics Wikipedia labels are being extracted by labeling systems and it is really satisfied with the provided human automated labels.

To evaluate this work in this case studies, a sample of the ODP collection and the 20-news group known as 20NG collection has been used. The evaluation framework is being followed presented in [29], a collection of uniform samples of 100 categories from ODP is extracted with each association of manual label. Our experiments show that for both benchmarks, our labelling framework can provide Match@5 $\geq$ 0.85. This means that for more than 85% of the categories, the manual label (or an inflection, or a synonym of it) appears in the top five recommended labels.

4.2.6 A general framework for integrated ML

To make a framework, a general proposed framework has been made for cluster labeling using some external resources. Figure 23 has illustrated the framework which consist on some main component or parts like indexing, clustering, term extraction, candidate label extraction and finally candidate evaluation.

In generally, the designed system can be described in a following way:

- Initially, the system gets the input which is in the form of set of textual documents.
- After that input receiving, the inputs are parsed, and index has been assigned to it and generates the inverted index.
- Next, with the help of initial index, new term has been extracted for other components which lead to clustering data for the cluster components.
- For each generated clustered data, some important terms are being extracted to estimate the best matched content of the component of the clusters.
- There are several candidate labels available in the clustered data which help to identify the important terms. Here the candidate labels can be chosen various set of important terms or external resources (from different web servers or Wikipedia).

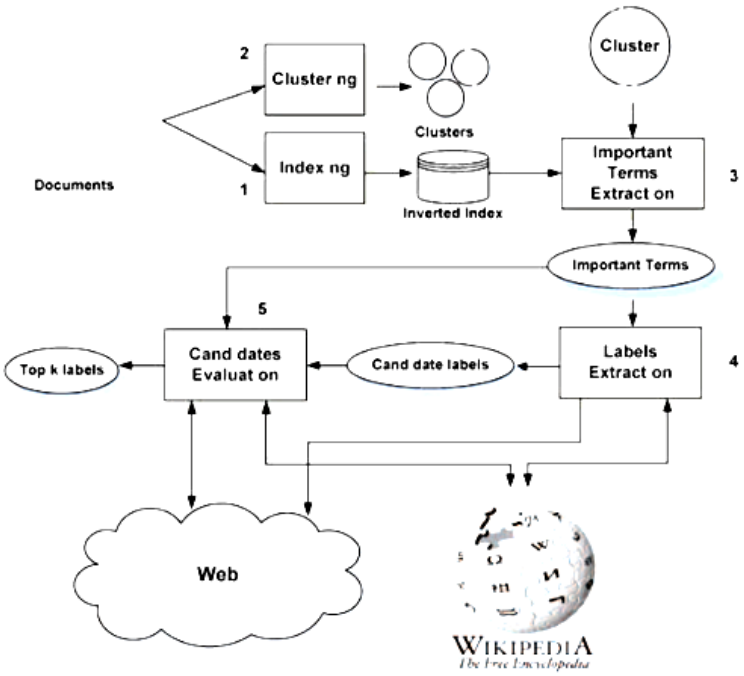


Figure 23: General framework for cluster labeling [27]

- At the final stage, a list of key suggested labels has been obtained by the system by evaluating the list of candidate labels.

Let's explain briefly the different components of the framework which is involved in various functions or operations. In the framework, there are five main steps involve the selected ML integration in the existing system's component and their functions in these steps.

4.2.6.1 Indexing

In the framework, the presented documents are parsed and assigned the token and represented in vector space using vector representation in system's vocabulary. The weight to the terms can be calculated using the weighted schemes by *tf-idf* a vector space model. The *Lucene* open source search system is being used to search index and assigned that index to documents. The inverted and inverse can be calculated using the *(tf(t, d))* and *(idf(t))* respectively where t is term and d represents documents in the entire collection.

4.2.6.2 Clustering

The clustering algorithm plays an important in this framework, the key objective of this algorithm includes the creation coherent clusters to help the clusters documents by sharing the same topics, for the representation of mutual topic belong to documents within the specific clusters by expecting the labels obtained by the system. During the clustering, the cluster can be represented by using the centroid of cluster's document while the weights of the terms include in the cluster's centroid are distributed among many cluster's documents by changing it to bias terms. As the result, the weight of the term t, document d and the cluster C can be given as:

$$w(t, C) = ctf(t, C) \cdot cdf(t, C) \cdot idf(t) \qquad [27]$$

There is no limitation to use the labeling framework to a certain clustering algorithm, but the coherency of the clusters identified by the system is expected to significantly affect the quality of labelling.

4.2.6.3 Important terms extraction

The given inputs are given in the form of cluster C ∈ C that containing the wished list of terms T (C) = (t_1, t_2,..., t_k), ordered by their estimated importance, which help to represent

the content of the cluster's documents. These terms have a list of single keywords and N-grams of various length.

The feature selection is tightly linked with important term extraction which is the process of selecting a subset of the terms for text representation, and it is frequently applied by text categorization and text clustering methods. To evaluate the feature selection the common approaches can be used according to their ability to distinguish the given text from the whole text. In this case studies, the main aim is to find a set of terms T (C) that best separates the cluster's documents from the all available collection.

The extraction of important term based on the method given by Carmel's method which was originally proposed in the context of the query difficulty model. We need to find a set of terms that maximizes the Jensen-Shannon Divergence (JSD) distance between the cluster C and the entire collection. A scoring is being assigned to each term according to its contribution to the JSD distance between the cluster and the collection. The term having highest scored need to be selected as cluster important terms.

4.2.6.4 Extracting labels

The important term T(C) is given, we need to extract candidate label for the cluster next. One of straight forward method or type involve in labeling is to extract it directly from the content of cluster's document. But it has been seen in many cases that it does not provide the suitable labels or not that much meaningful for the end users like in table 4.2. Therefore, the external sources as complimentary sources for this task are needed to use.

The candidate labeling process can be summarized in the following simple way:

- Need to generate index using ***Lucene search system***.
- Execute the query *q* on Wikipedia index with the help disjunction of key terms
- As the result query *q*, list of documents D(q) that need to sort based on score.
- For each document $d \in D(q)$, we must consider both the document's title and the set of document associated categories as potential candidate cluster labels (denoted $L(C)$).

4.2.6.5 Candidates evaluation

The several judges' methods have been used in this case studies for the evaluation of candidate labels. Here, limited number of judges would be mentioned:

- MI Judge: The mutual information judge simple works with scoring each candidate by average pointwise mutual information related to label of the set of cluster's important terms within a given external textual corpus.
- SP judge: The second type of judge methods used in this case studies, termed Score Propagation (SP) judge, scores each candidate label with respect to the scores of the documents in the result set related to that label. The SP judge is helpful to propagates documents' scores to candidates that are not directly associated with those documents but share the list of common keywords with other related labels.

4.2.7 Summary

By summarizing the findings, we have tried to arrange an investigation about the study related to cluster labeling algorithm to enhance the utilizing power of Wikipedia knowledge base. In this case study, we have tried to explain the general framework to perform the required functions or operations that help to extract the candidate labels from the available terms and Wikipedia and assigned the score to the terms then later on, the top scored candidates according to the evaluation of several independent judges. Later, an experiment has been carried out for the implementation perspective which in included in getting meta-data related different Wikipedia articles, which identical to the content of cluster's terms can give the best labeling clusters of the textual document. Our candidate extraction approach that has been used in this case studies is based on identifying Wikipedia articles that are having same characteristics to the cluster's content and then extracting titles and categories from those pages.

In comparison two system's component, conventional versus integrated ML system's component it can be seen in comparison Table 8 below:

Table 8: Conventional system vs ML integrated component

Old System's component	ML integrated System's component

Low speed of processing	Improve the speed because indexing is always increased the searching any component in the system. For example, in database search operation can be done if index has been used in Databases queries.
Complex data flow	Divide the different component of the text extraction using the cluster according to functionalities of the component. For example, indexing and validation perform the specific function that is easy to understand by function's names.
Figure 20	Figure 21

Overall, the Cluster labeling technique with Wikipedia is extremely successful, as highlighted by the results in this work, especially in collections of documents whose topics are covered well by Wikipedia concepts. For domain specific collections, with topics that are not completely covered by Wikipedia, the proposed candidates may disturb or fluctuate the system's performance due to their irrelevance to the documents' topics. For these kinds of collections, an intelligent decision should be made about the use of Wikipedia content or another external resource or content; alternatively, a choice could be made to focus only on inner terms for labeling. After analyzing the available collection with respect to Wikipedia the decision needs to make. In future, these collections specific decision making can be developed as part of the labeling framework is left for further research.

5 DISCUSSION AND CONCLUSIONS

This section unrolls the main results, themes and explaining detail with while linking with the literature review. Keeping in view of research objectives and questions main findings are explained. A new perspective and interesting points are highlighted which will help in making a conclusion and suggestion.

To address the main research objective, it was crucial to explore the main components or factors involve in integration in ML which has already described in previous Chapters.

5.1 The Basic Factors of ML integration

The data and architectures are two main factors which involve in integrating of ML in conventional systems (in term of functions), like making the system intelligent to learn by itself. In the previous chapters, its has already been briefly described the importance of architecture in ML. Thus, here we can highlight the importance of other important factor that is data while integrating ML.

Data is the core of the ML and many of most exciting models don't work without large data sets. If the architecture and other components help to develop machine, then data is known as the oil for machine learning[7]. Data is transformed into meaningful and usable information for helping either the processing the system's functions or human readings.

Studies shows that there must be a clear separation for the outlining the data architecture. During the performing data operations, commonly viewpoints for data are depending upon the data type like application data, technical data and business data is of utmost importance. Not all the data that has been used for training in ML integration model can be obtained from business processes. Sometimes these data can be originated from some other sources like photo collections, traffic data, weather data, financial data etc. For instance, first case study shows the parsing operation by using the feature extraction methods use the data which is produced by various other business process from parent systems. Those data collections

[7] ML Reference Architecture:
https://freeandopenmachinelearning.readthedocs.io/en/latest/architecture.html

might produce the data from sources including some IoT devices, web servers or business operations.

There are some other factors involve in integration of ML which has been briefly described in Table 9. According to Gartner [30] and case studies, these factors are also known as key elements of integrated ML system to validate the performance of the system.

Table 9: Key factors in integrated ML system

Name	Description
Stability	It helps to understand maturity of the systems, and get the answer how stable the selected framework is? While integrating ML, there must be a proper methods or plan to check and test the stability of the system so that it could improve on demand.
Performance	There must be some sort of benchmark for testing and analysis the integrated ML system which includes in training and production is highly recommended.
Features	There must be clear support for the recommended features of the systems, and it must perform the same features efficiently in newly integrated system than conventional system.
Flexibility	The flexibility is the most important factor involve in ML methods and it must not be compromise. For example, if the integrated ML systems are working precisely for one API, how easy to switch to another API?
Transparency	A lot of knowledge of engineers and programmers are required to develop the integrated ML system due to its tasks (training or production). The selected ML integration framework need to clear enough to describe the quality, working, maintenance process and its business dependencies.
Speed	Speeding up time consuming and recurrent development tasks.

5.2 Machine Learning Methods compatibility

To integrate the various methods in ML, it is key to keep in mind the basic functionalities of the conventional systems and best feasible condition for the ML methods to perform as good as possible.

During the exploration of case studies, it has been seen that ML methods has been wisely chosen by keeping in mind the needs, functions and requirement of either existing systems or existing component.

If we go through the case studies one, it has been clearly and briefly described the reasons behind selection of ML techniques during the integration of ML. Before integrated the ML methods in conventional system, a short research had been made team to study the different pros and cons of the various algorithms and methods. In that report, team members have reported some outcomes of their finding and then they have chosen the mentioned ML techniques mentioned in (first) case study. The team has given the commendation about those selected ML methods by keeping in mind the need and requirement of the conventional system. The selected three methods are best in business in data parsing because of simple implementation, easy to upscale due to its adaptation and improve the speed of the data parsers. For example, the keyword extraction techniques help to assigning the index to various keywords which need to be search and generate clouds tags support the Search Engine Optimization (SEO) related operations, clouds analyzing, marketing etc. These techniques not only just worked for the integrated system but also helped the promoting brands for marketing teams.

In case study two, we have seen the cluster labeling techniques has been wisely used which is best in business to achieve the system requirements. The ML integration results show that it helps to increase the resource availability like if first cluster fails once then it can pick another cluster and share the workload [27] in the integrated system. This efficiently prevents the loss of valuable time and information if a server fails and continue its working. Similarly, other algorithm components like indexing features also suits the labeling component in the conventional system by gaining higher processing power and speed.

The table below shows the general integrated ML techniques and its compatible environment (prerequisite) in conventional system to effectively implement it. For example, if we look

on keyword's extraction method in ML, (Table 6), to train the various ML components. In Table 10, there is list of few common ML methods and its prerequisites. Only two of the prerequisites would be mentioned in the table which have been used in case studies.

Table 10: ML integration methods vs prerequisite

Name	Prerequisite
Keywords extraction	• Classified the important words and expression • You need know the important words to train the data • You must know the picking of important word or expression from data sets. • Knowledge about its training data.
Clustering	• Knowledge about the required classification of data. • When we need clustering for words with similar definitions to optimize the search approaches. • When the available data is large, and it demands some cluster divisions.

5.3 Risks and challenges in integration

After studying the ML integration in already in existing systems in the form of case studies and ML methods, there are limitations in integration ML in various systems. While integrating the ML in existing system following limitation and risk can be kept in my mind.

In term of studying the risks and challenges involve in introducing ML in conventional system, it could be varied on the chosen systems design & architecture, integration process & models and assessment of functionality. Other than these risks, there are also number of other critical risk and challenges like data privacy & protection including the data poisoning attacks, selection of ML methods or techniques, speed of processing and upgradation old to latest ML techniques in the selected system.

The short description about the risk and challenges in the Table 11, not only it has been faced in selected case studies but also in Gartner Technical professional advices (general). For

example, Gartner's step for integration ML process also includes these challenges especially acquisition of relevant data and importantly data privacy. To find out the challenges, case studies implementation was also helpful in term of speed up the processing of functionalities because while discussing with CTO in first case study, "there is a clear improvement of speed up processing compare to prior to ML integration". Thus, speed of the ML systems also selected in this risk and challenges Table below.

Table 11: Risk and challenges in integrating ML in conventional systems

Name	Description
Acquisition of relevant Data	A most common challenge of ML application during the integration is the acquisition of relevant data. This is also a limitation as the quality, availability, and composition (e.g. are the meta-data included well? Is the data labeled well?) of the data collection at hand have a strong influence on the performance of ML algorithms.
Data privacy	While integration ML, it has been seen that mainly data privacy should never compromised. The main interaction of ML integrated system is the data interaction by performing various data operations. Missing any steps or component can lead to threaten the data privacy. During the implementation of ML in conventional systems, data privacy is known as one of biggest challenge to achieve. For example, nowadays, many conventional systems are replacing with ML integrated system where decision making has been performed by using various collections from other sub systems. This is critical and crucial to maintain both during decision making and integration ML which is one of key challenge.
Data protection	Similarly, many hackers or viruses attacks the data server to steel the data in many ML integrated systems. Sometimes, ML integrated system is being divided into separate data server to make the system simple and easy. Some systems use KISS (Keep It Simple and Stupid) principle, but in this case, it can lead to compromising on data protection. That's why, these two factors are closely linked with each

	other and lead to main challenge while introducing the ML in existing systems.
Speed of processing	One of the objectives for introducing ML in existing system is to increase the speed of process and the efficiency of the system and this led to next challenges to manage. To achieve the objectives of the system, it is key to select the right and suitable ML methods to handle this challenge. Another important factor which link with the speed of the ML system is the architecture which help the system for the flow of information and data.
Implementation	With the implementation perspective, its not an easy task for the developer to integrate ML in the conventional system. Another challenge related to implementation is to find the capable and skilled developer due to limited number of developers. Implementing the Data security, Data privacy and simplicity in integration at the same time is itself become a critical challenge for ML system.

5.4 Advantages of ML Integration

In machine learning Integration, there is a saying which fit on this "No Free Lunch" theorem. In a nutshell, Studies shows that there is not a single algorithm which can be integrated in existing system to work perfectly on neural network and decision tree at the same time. So, the strength and weakness of various ML architecture is much important and relevant to us to highlight here because it was super keen to study it previously integration ML in the selected case studies (case studies chapter).

In this section, the discussion mainly focuses on the few advantages to attract many enterprises for integrating ML techniques to improve the perform of the system compare to conventional system.

During the discussion with responsible person about the first case study, it has been seen the major improvement in the system in terms of performance, speed and time. These three advantages are the key factors while introducing the ML in the existing system. For example, before the introduction keywords extraction, conventional system was slow and because of

some old data fetching and parsing methods. Old extraction methods included traditional functional programming which need huge amount of time to complete it in the presence of increasing data. Data handling and tool formation is second most important advantages in that case study to increase the performance related to data. Another advantage includes in ML integration is related to introduction of cloud-based ML infrastructure in the existing system. That cloud component of the system also used ML methods where parsed data is being utilizing for run time development processes.

Similarly, in case second study, the clustering technique have the best advantage over other methods in the certain circumstances due to its effectiveness. Before integrating the selected methods, a short study has been carried to check the feasibility of the selected methods with others. After implementing the clustering method, the following advantages have been achieved:

- No more complexity: compare to previous methods to improve the keywords search in the system, clustering methods provides the simplicity in their functions combining with indexing, judges etc.

- Speed: Overall, the ML has improved the speed of the system by removing the old methods barriers (see in case study).

Other than these strengths about integrating ML in the conventional system, there are also some other advantages which could be achieved by integrating the ML in existing system which are following Table 12.

Table 12. General advantages for Integrating ML

Name	Description
Less human interaction	With the integration ML, we trained the data in a way that system process by itself which cause the less human interaction with the systems.
Continuous improvement	Whenever we integrate the algorithm in conventional system, the next time the algorithms get improved in accurately and

	efficiently with time. With the increment in amount of data, the functions and process get more tasks to accomplish that lead to continuous improvement.
Handling multi-dimensional data	The integrated ML algorithms perform best for handling data which might include in data manipulation, data optimization etc. while that data might be multi-dimensional or not.
Expanding Scope of the ML System	With the integration ML methods in system, the scope of the system goes expand in term of scalability and variety.

5.5 Summary

This Chapter covers the following points:

- Discussion about our previous studies and finds.
- A briefly description about factors link with ML integration.
- A short discussion about Machine Learning Methods compatibility.
- Highlights the risks and challenges involves ML integration.
- Advantages of the ML integration.

6 SUMMARY & CONCLUSION

By summarizing the report, the ML integration plays an important part to increase the productivity of the existing system. During the study about the ML methods and the integration of ML in existing systems by using the case studies and literature, the following outcomes have been found:

The ML methods knowledge plays an important and key role to start the integration ML in existing systems, that knowledge can be obtained in chapter two where focus is on AI and ML. A few methods have been discussed in that chapter to highlight the algorithms.

The chapter three has described the ML integration components including integration process, architecture, tools and models. These studies help to formalize the proper understanding and explore the various aspects of the systems including the utilizing the Gartner's research report [30].

Two case studies have been included, firstly, a Finnish company where they have moved the system from conventional to ML integrated systems known as discount-based system (DBS) by using the various data parsing methods (keywords extraction, feature extraction and etc.). Secondly, the Wikipedia introduced the ML clustering algorithm in one specific component related to text search to speed up process and simplified it.

Later part of the report consists on explanation and discussion to explore the various aspects of ML integration which includes the ML methodology compatibility, factors which effect the specific or overall performance of the ML integration systems. Finally, our study shows that the ML integration process vary from system to system depending upon the flow of information in the system, design and requirement of the conventional system, resources for integration and capability of the existing systems.

6.1 Conclusion

In conclusion, it has been a great and crucial task to integrate the ML in various existing systems. The studies show that during the integrating ML in conventional system, feasibility study, ML methods selection, designing architecture, selection of ML integration tools, and

implementing the selected ML methods and architecture are the most essential and relevant players. We can't ignore any of these mentioned important component or factors which directly or indirectly relate with the performance of selected systems.

While integrating the ML, the risk and challenges are needed to keep in mind because it can cause the system to lead to mismatch the system expected outcome. The risk and challenges handling most importantly link with the selection of ML methods. In Gartner studies, ML integration process are also played an important role in this regard [30]. In the previous sections, few advantages and disadvantages have been clearly highlighted (in Chapter 2). For example, in case studies chapter, selected ML methods are more compatible with the system's requirement.

Thus, the ML integration process vary from system to system depending upon the flow of information in the system, design and requirement of the conventional system, resources for integration and capability of the existing systems.

REFERENCES

1. Bottou, L., Curtis, F.E. and Nocedal, J., 2018. Optimization methods for large-scale machine learning. Siam Review, 60(2), pp.223-311. Available at: https://arxiv.org/pdf/1606.04838.pdf.
2. B. Anna ,14 November 2017, "Machine Learning – Existing Applications"https://www.apriorit.com/dev-blog/472-machine-learning-applications
3. Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26.
4. Patrick Henry Winston, "Artificial Intelligence, 3rd Edition", Massachusetts Institute of Technology.
5. "WIPO Technology Trends 2019 Artificial Intelligence", World Intellectual property Organization. Available at: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf.
6. Zadeh, L.A., 1988. Fuzzy logic. Computer, 21(4), pp.83-93. Available at: http://www.geocities.ws/hhvillav/00000053.pdf
7. Lloyd, J.W., 2012. Foundations of logic programming. Springer Science & Business Media. Available at: http://cgi.di.uoa.gr/~prondo/SEMANTICS/Lloyd.pdf
8. Margaret R., 2016 "expert system, Search Enterprise AI". Available at: https://searchenterpriseai.techtarget.com/definition/expert-system
9. Pearl, J., 2014. Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier. Available at: https://www.georg-buechner.org/book/1048085570/download-probabilistic-reasoning-in-intelligent-systems-by-judea-pearl.pdf
10. Keet C.M., 2018, "An Introduction to Ontology Engineering",. Available at: https://people.cs.uct.ac.za/~mkeet/files/OEbook.pdf
11. Kubat, M., 2017. An introduction to machine learning (Vol. 2). Cham, Switzerland: Springer International Publishing.

12. Learned-Miller, E.G., 2014. Introduction to supervised learning. I: Department of Computer Science, University of Massachusetts. Available at : https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf
13. Mitchell. T.M., 1997, "Machine Learning",. Available at : https://courses.edx.org/asset-v1:ColumbiaX+CSMM.101x+1T2017+type@asset+block@AI_edx_ml_5.1intro.pdf
14. Ghahramani, Z., 2003, February. Unsupervised learning. In Summer School on Machine Learning (pp. 72-112). Springer, Berlin, Heidelberg.. Available at : http://www.inf.ed.ac.uk/teaching/courses/pmr/docs/ul.pdf
15. Piech, C.,2013, "K Means",. Stanford University: Artificial Intelligence: Principles and Techniques, Available at: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
16. Pierson, E. and Yau, C., 2015. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome biology, 16(1), p.241. Available at: https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-015-0805-z
17. Bishop, C.M., 2006. Pattern recognition and machine learning. springer.
18. Rojas, R., 2013. Neural networks: a systematic introduction. Springer Science & Business Media.
19. Sazli, M.H., 2006. A brief review of feed-forward neural networks. Communications, Faculty of Science, University of Ankara, 50(1), pp.11-17. Available at: https://www.researchgate.net/publication/228394623_A_brief_review_of_feed-forward_neural_networks
20. Brownlee, J., 2017. A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Machine Learning Mastery, 19., Available at : https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/
21. Bengio, Y., Simard, P. and Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), pp.157-166. Available at:

http://www.comp.hkbu.edu.hk/~markus/teaching/comp7650/tnn-94-gradient.pdf

22. S. Akshay., Long Short-Term Memory' [lecture notes], CS 638 and CS 838 - Building Deep Neural Networks, University of Wisconsin System, uploaded 28 March 2017, http://pages.cs.wisc.edu/~shavlik/cs638/lectureNotes/Long%20Short-Term%20Memory%20Networks.pdf
23. Richard S. Sutton, Andrew G. Barto, ” Reinforcement Learning: An Introduction”, 2nd edition , 2015, Available at : https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf
24. Mohammed, M., Khan, M.B. and Bashier, E.B.M., 2016. Machine learning: algorithms and applications. Crc Press., Available at : https://www.researchgate.net/publication/303806260_Machine_Learning_Algorithms_and_Applications
25. Rokach, L. and Maimon, O., 2005. Decision trees. In Data mining and knowledge discovery handbook (pp. 165-192). Springer, Boston, MA., Available at : http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf
26. Cunningham, P. and Delany, S.J., 2007. k-Nearest neighbour classifiers. Multiple Classifier Systems, 34(8), pp.1-17.Available at: https://www.researchgate.net/profile/Sarah_Delany/publication/228686398_k-Nearest_neighbour_classifiers/links/0fcfd50d0c1d1f41ad000000/k-Nearest-neighbour-classifiers.pdf
27. Carmel, David, Haggai Roitman, and Naama Zwerdling. "Enhancing cluster labeling using wikipedia." Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009.
28. Z. S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In ICWSM ’08, 2008.
29. H. Toda and R. Kataoka. A clustering method for news articles retrieval system. In WWW ’05, pages 988–989, New York, NY, USA, 2005. ACM.
30. Sapp, Carlton E. "Preparing and architecting for machine learning." Gartner Technical Professional Advice (2017): 1-37.