Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

The Enhancement of Crowdsourced Software Engineering Based on Machine learning

"A Case Study Iraqi National Card Website"

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of the University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology – Software

By

Muataz Abd Al-Mohsen Khudair Bader

Supervised by

Assist. Prof. Dr. Ahmed Saleem Abbas Jassim

2022 A.D. **1443 A.H.**

سورة البقرة

أية (٣٢)

Declaration

I hereby declare that this dissertation entitled **"The Enhancement of Crowdsourced Software Engineering Based on Machine learning "A Case Study Iraqi National Card Website","** submitted to University of Babylon in partial fulfillment of requirements for the degree of Master in Information Technology \ Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source are appropriately cited in the references.

Signature:

Name: **Muataz Abd Al-Mohsen Khudair**

Date: / /2022

Supervisor Certification

I certify that the thesis entitled (**The Enhancement of Crowdsourced Software Engineering Based on Machine learning "A Case Study Iraqi National Card Website**) was prepared under my supervision at the department of Software/ College of Information Technology/ University of Babylon as partial fulfillment of the requirements of the degree of Master in Information Technology-Software.

Signature:

Supervisor Name: **Assist. Prof. Dr. Ahmed Saleem Abbas**

Date: / /2022

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled **"The Enhancement of Crowdsourced Software Engineering Based on Machine learning "A Case Study Iraqi National Card Website"** for debate by the examination committee.

Signature:

Assist. Prof. Dr. Ahmed Saleem Abbas

Head of Software Department

Date: / /2022

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**The Enhancement of Crowdsourced Software Engineering Based on Machine learning "A Case Study Iraqi National Card Website**) presented by the student (**Muataz Abd Al-Mohsen Khudair)** and examined him/her in its content and what is related to it, and that, in our opinion, it is adequate with (Viva Result) standing as a thesis for the degree of Master in Information Technology-Software.

Signature:
Name: Prof Dr. Zaki Saeed Tawfik
Title: Professor
Date: / / 2022
(**Chairman**)

Signature:
Name: Assist. Prof. Dr. Eng. Alharith A. Abdullah
Title: Assistance Professor
Date: / / 2022
(Member)

Signature:
Name: Dr. Mohannad M.Jassim Al-Yasiry
Title: Senior Lecturer
Date: / / 2022
(**Member**)

Signature:
Name: Assist. Prof. Dr. Ahmed Saleem Abbas
Title: Assistance Professor
Date: / / 2022
(**Member and Supervisor**)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name: Dr. Hussein Atiya Lafta
Title: Professor
Date: / / 2022
(Dean of Collage of Information Technology)

Dedication

I dedicate this simple effort to my:

Wonderful Father and mother (who passed away too early)

Brother and Sisters

My dear wife

Thanks to all my friends for their moral support and encouragement

Thanks for everything

Acknowledgement

Foremost, I am highly grateful to God for His unlimited blessings that continue to flow into my life, and because of You, I made this through against all odds.

To my supervisor, **Assist. Prof. Dr. Ahmed Saleem Abbas**: I feel highly indebted to you. I am deeply grateful for your suggestions on this topic, and without your support, comments, and guidance, it would be difficult to finish this work.

To my wonderful Parents (May God have mercy on them): Words cannot express my appreciation to you, the best father and mother. I wish to give you all thanks and love for your guidance, advice, invitations, and endless support.

Thank you to my brother and sisters for the endless support and I find in my heart nothing but gratitude for what you have given me throughout my study.

Furthermore, many thanks to my wonderful wife; you are supportive and helpful at every stage of this thesis.

My dear children: I am sorry for being busy and did not spend much longer time with you, especially in the last days of this work.

To my family and my friends: I give you all thanks for your belief in me, your constant support, encouragement, and cooperation at all times.

Thanks to the Ministry of Interior especially the Directorate of Communications and Informatics also the Directorate of Civil Status. Passport and Residence in particular for the great help that they have offered me. In addition, I thank the University of Babylon, especially the College of Information Technology, and all the teachers, especially those who contributed to getting me to this stage.

Last but not least, I would like to thank all the kind, helpful and lovely people who helped me directly or indirectly to complete this work and apologize to them for not being able to mention them by name here, but they are in my heart.

Muataz Abd Al-Mohsen Khudair

Abstract

If you had a big job, where would you start?

You may have contacted some friends and colleagues to find out how to handle the problem and use their collective suggestions to guide your actions. This is basically how crowdsourcing works. Instead of asking three friends, what would happen if you asked 300 people?

The needs to deal with large-scale stakeholders to ensure the correctness of software requirements make the crowdsourcing technique useful and helpful to improve the optimal level of requirements quality in terms and saves the development cost and time.

Crowdsourcing can be defined as exploiting the collective intelligence of the audience to get the job done. And since it depends on a lot of people, the right kind of crowdsourcing can often come up with a better solution than you would have if you combined only the thinking power of a few people, as crowdsourcing is done by gathering the largest number of Information, opinions or experiences through the Internet, social media and smart phone applications, and this process may be voluntary, free or paid.

Crowdsourcing is an emerging distributed model problem solving based on the combination of human and machine computation

The connection with the technique has always been a topic of interest how to translate these requirements into a product or service that people use. Crowdsourcing is used to support software engineering activities.

This thesis provides researchers in how crowdsourcing techniques improve the benefits of software engineering.

Accordingly, this message aims to take advantage of crowdsourcing in the unified card system for the purpose of providing various ideas that can solve difficult and unexpected problems faster, by collect data that Focus on the answers obtained to from the questions (18 questions) from the beneficiaries , which helps to establish a

joint development of the unified card department, which contributes to the development of Its products, services and systems, which helps to find ideal solutions to difficult problems in record time. That use

1- Decision Tree

2- Random Forest

3- K-Nearest Neighbor

4- Logistic Regression

5- Stochastic Gradient Descent

6- Likert Scale

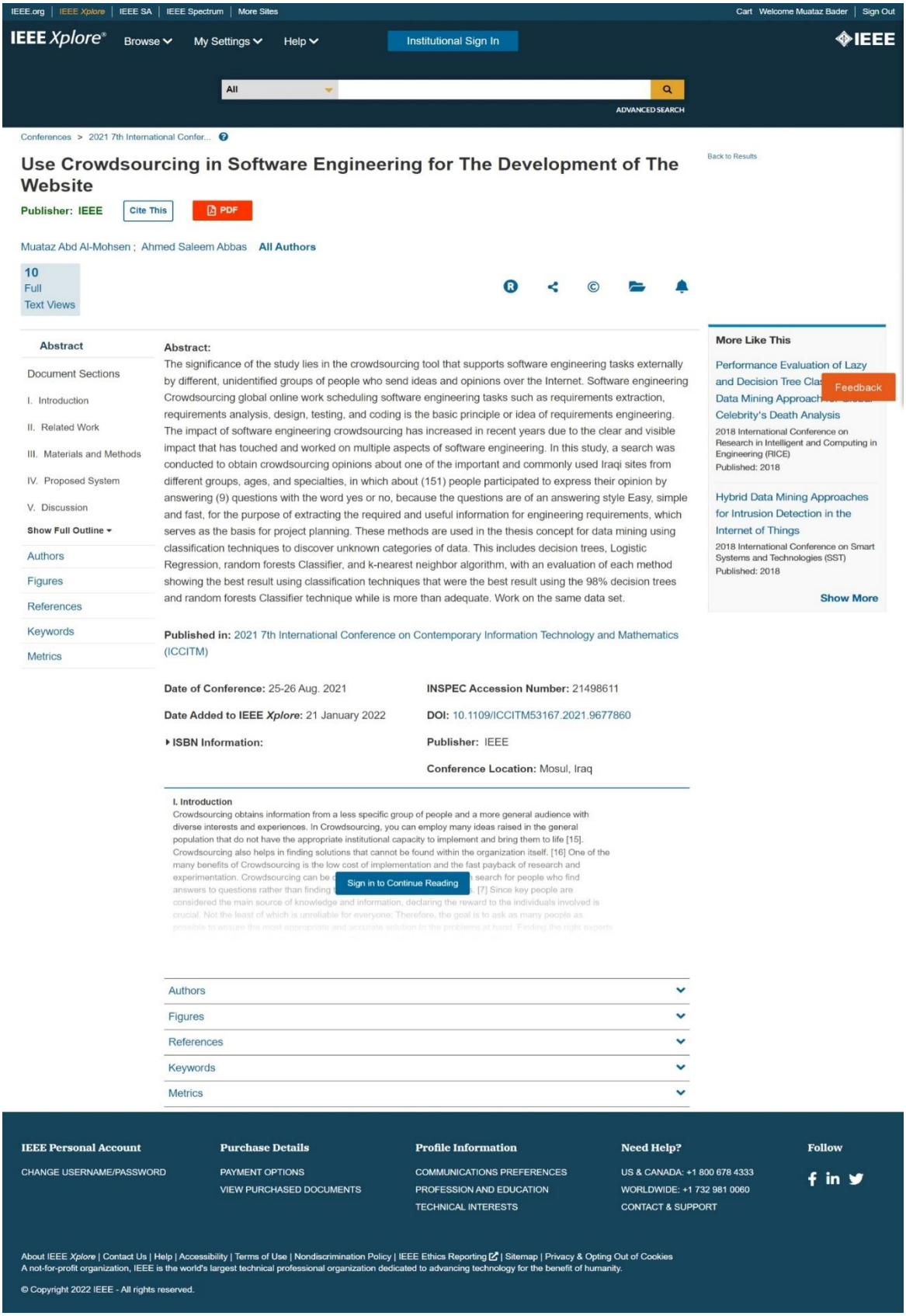
Moreover, the findings of the classification algorithms were evaluated based on performance metrics using accuracy, precision, recall, and f1-score measurements.

The results showed that the highest accuracy was achieved by applying Logistic Regression and Decision Tree on the two datasets that were used. The Accuracy, Precision, Recall, and F1-score are (1.00%,1.00%,1.00%,1.00%) in first dataset while the second dataset is (0.93%,0.92%,0.92%,0.92%), respectively.

In addition to extracting the results using the Likert scale with using Mc call model, the degree of acceptance in the first data set was (52.62%), While the second dataset results ranged between (0.94% - 0.57%).

Finally, the crowdsource could possibly be used to create adequate training datasets in order to improve the Software engineering task related to software testing and maintenance.

Declaration Associated with this Thesis



IEEE.org | IEEE Xplore | IEEE SA | IEEE Spectrum | More Sites

Cart Welcome Muataz Bader | Sign Out

IEEE Xplore® Browse My Settings Help Institutional Sign In IEEE

All

ADVANCED SEARCH

Conferences > 2021 7th International Confer...

Back to Results

Use Crowdsourcing in Software Engineering for The Development of The Website

Publisher: IEEE Cite This PDF

Muataz Abd Al-Mohsen ; Ahmed Saleem Abbas All Authors

10 Full Text Views

Abstract

Document Sections

I. Introduction

II. Related Work

III. Materials and Methods

IV. Proposed System

V. Discussion

Show Full Outline

Authors

Figures

References

Keywords

Metrics

Abstract:

The significance of the study lies in the crowdsourcing tool that supports software engineering tasks externally by different, unidentified groups of people who send ideas and opinions over the Internet. Software engineering Crowdsourcing global online work scheduling software engineering tasks such as requirements extraction, requirements analysis, design, testing, and coding is the basic principle or idea of requirements engineering. The impact of software engineering crowdsourcing has increased in recent years due to the clear and visible impact that has touched and worked on multiple aspects of software engineering. In this study, a search was conducted to obtain crowdsourcing opinions about one of the important and commonly used Iraqi sites from different groups, ages, and specialties, in which about (151) people participated to express their opinion by answering (9) questions with the word yes or no, because the questions are of an answering style Easy, simple and fast, for the purpose of extracting the required and useful information for engineering requirements, which serves as the basis for project planning. These methods are used in the thesis concept for data mining using classification techniques to discover unknown categories of data. This includes decision trees, Logistic Regression, random forests Classifier, and k-nearest neighbor algorithm, with an evaluation of each method showing the best result using classification techniques that were the best result using the 98% decision trees and random forests Classifier technique while is more than adequate. Work on the same data set.

More Like This

Performance Evaluation of Lazy and Decision Tree Cla... Data Mining Approach... Celebrity's Death Analysis

2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)

Published: 2018

Hybrid Data Mining Approaches for Intrusion Detection in the Internet of Things

2018 International Conference on Smart Systems and Technologies (SST)

Published: 2018

Show More

Feedback

Published in: 2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM)

Date of Conference: 25-26 Aug. 2021

Date Added to IEEE *Xplore*: 21 January 2022

ISBN Information:

INSPEC Accession Number: 21498611

DOI: 10.1109/ICCITM53167.2021.9677860

Publisher: IEEE

Conference Location: Mosul, Iraq

I. Introduction

Crowdsourcing obtains information from a less specific group of people and a more general audience with diverse interests and experiences. In Crowdsourcing, you can employ many ideas raised in the general population that do not have the appropriate institutional capacity to implement and bring them to life [15]. Crowdsourcing also helps in finding solutions that cannot be found within the organization itself. [16] One of the many benefits of Crowdsourcing is the low cost of implementation and the fast payback of research and experimentation. Crowdsourcing can be ... search for people who find answers to questions rather than finding ... [7] Since key people are considered the main source of knowledge and information, declaring the reward to the individuals involved is crucial. Not the least of which is unreliable for everyone; Therefore, the goal is to ask as many people as

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | Privacy & Opting Out of Cookies

A not-for-profit organization, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

Abbreviation	Description
SQ	Software Quality
IT	Information Technology
DM	Data Mining
NLP	Natural Language Processing
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
DT	Decision Tree
RF	Random Forest
KNN	K-Nearest Neighbor
LR	Logistic Regression
SGD	Stochastic Gradient Descent
ML	Machine Learning
P	Precision
R	Recall
F1	F1-Score

CHAPTER ONE

INTRODUCTION

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Over the years, crowdsourcing has become a practice in the area of software engineering. Even though the use of traditional development is still employed in executing tasks, there is an increase in the use of outsourcing and contract development in software projects. They are normally used for crowdsourcing so as to meet different needs like software testing, fixing of bugs, or the collection of a wide range of designs so as to provide a variety for a new user interface. The role of crowdsourcing in reshaping the contributions of developers to software projects can be seen through some mechanisms like experience sharing, sharing sites, online job marketplaces, competition sites, and bug bonuses.

The popularity of crowdsourcing is at its infancy stage, particularly, when it comes to the choice of the most suitable crowdsourcing program. Crowdsourcing is regarded as a model that is used to find solutions to problems by gathering information from human and machine sources. In the work done by Howe and Robinson in the year 2006, the concept of crowdsourcing was defined [2]. However, there is a globally accepted definition, which defines crowdsourcing as the process through which organizations outsource some of its tasks to numerous people by inviting them to participate in the work. Normally, Crowd-sourced Software Engineering is derived from Crowdsourcing.

One of the strategies used in this process includes call open format whereby, world-class software engineers are hired through the internet to carry out a wide variety of software engineering tasks like designing, extraction of requirements, coding, and testing. It is believed that through the use of this strategy, parallelism is achieved, thereby leading to a reduction of time spent in the marketplace [5]. More so, it is also believed that this approach is capable of minimizing costs and

defect rates with flexibility in development capacity. There have been successful implementations of crowdsourcing software engineering technologies on many crowdsourcing platforms such as TestFlight, TopCoder, Mob4Hire, AppStori, uTest. [3]

The crowdsourcing model have been applied in a wide range of design-based creative activities [10]. Over the years, Software engineering crowdsourcing has increasingly attracted the interests of professionals in the industry and the academia.

Recently, there have been works done in the area of resource mobilization, but it has been reported that resource mobilization in software crowdsourcing has received lesser attention from both the industry and academia [11]. However, some authors have argued that this is a misconception because there is a rapid growth of the field as it touches on several critical aspects of software engineering. Consequently, there is a growing literature on the various aspects and fields of software engineering applications.

The graphic representation of the growth trend in publications is shown in Figure (1.1) below. The graph shows a distribution of research works done in this area including journal papers, technical reports, conferences proceedings, while, Masters and Doctorate theses.

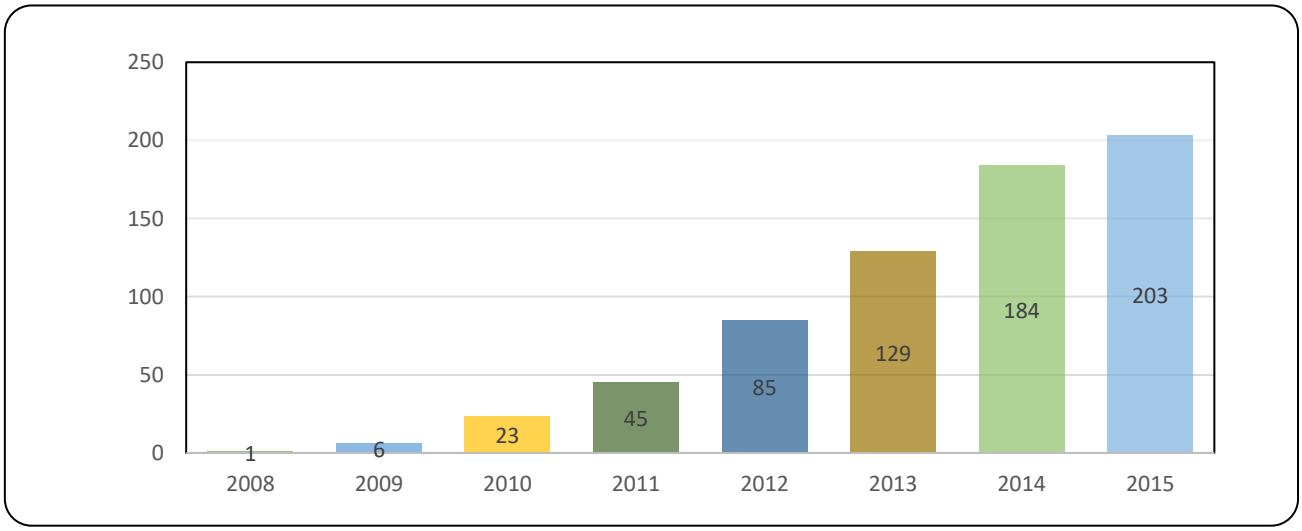


Figure (1.1): Published Studies in Area of Crowdsourced Software Engineering (2008- April 2015) [41]

The growth trend for these is shown in Figure (1.2). It can be seen that there is a commendable increase in publications in the area of crowdsourcing software engineering [41].

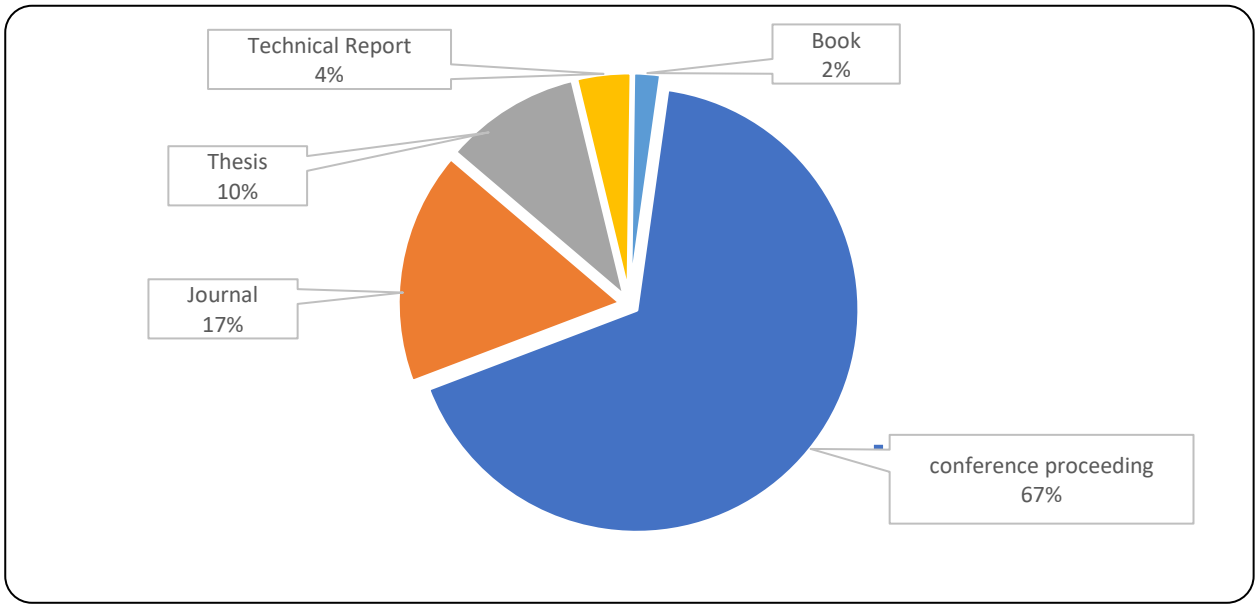


Figure (1.2): Publication Kind for Papers [41]

Since the term crowdsourcing was coined, different aspects of crowdsourcing have been researched, with researchers analyzing the economics of crowdsourcing competitions. In such studies, models for pricing strategies have been proposed by the authors. The proposals for these models have been made by analyzing reputation and earning a reward.

In a work done by the authors in [52], the relationship between the effectiveness of crowdsourcing and attention was investigated through an analysis of a dataset from YouTube site the results revealed that the productivity exhibited in crowdsourcing manifest a strong positive dependence on attention. To achieve the objective of their study, the number of downloads was measured [52].

Even though this is an area of software engineering that is yet to gain full maturity, it is fast growing and increasingly gaining popularity. The literature on crowdsourcing shows that software engineering may involve the direct outsourcing of software engineering tasks to the general public, indirect reuse of extant crowdsourced knowledge, or the proposal of a framework that can support

the enhancement or realization of Crowdsourced Software Engineering. Simply put, the concept of 'Crowdsourced Software Engineering' refers to the use of different strategies and techniques of crowdsourcing to enable the development of software [41].

It entails other activities that support software development, and not necessarily the actual production of the software itself. Crowdsourcing is carried out for reason including, eliciting for requirements, project management, to increase software security, software testing, or refinement of software.

Authors in [59] noted that only a few studies have highlighted the importance and uses of crowdsourcing in the development of software.

In (2009) working was represented as a sample of crowdsourcing scenario within the software development domain as a way of deriving the requirements for the delivery of an all-purpose crowdsourcing service in the cloud [49]. More so, the authors in an approach aimed at dividing programming tasks into micro tasks were proposed. The proposed approach is to be applied in crowdsourced software development [50].

In 2014, a case study of crowdsourcing software development was presented at a multinational corporation, where the challenges associated with the application of crowdsourcing software engineering were also presented [51].

There are different benefits that can be derived from Crowdsourced Software Engineering against the conventional methods of software development. One of the greatest benefits that can be derived from the use of crowdsourcing in software development is the reduction of costs for internal employees by integrating flexible external human resources, and the exploitation of a distributed model of production that is capable of hastening the process of software development [41].

1.2 Crowdsourcing

For the first time in June 2006, the concept of 'Crowdsourcing' was defined in a publication in Journal Wired article. The article which was written by Jeff Howe, was titled 'The Rise of Crowdsourcing,' [2]. Crowdsourcing is the process through which companies or institutions outsource tasks originally performed by their employees to a large network of people through open calls. Based on the existing definitions of crowdsourcing, the two key important elements of crowdsourcing are open call and an undefined large network force.

Crowdsourcing has been defined by different authors, thereby resulting in slight variations in the definitions. For example, in an article published by Brabham in 2008 [25], crowdsourcing was defined as an online model that supports problem-solving and distribution of tasks. From the period of 2006 to 2011, the concept of crowdsourcing has appeared in 32 published articles, with 40 definitions of the term [26]. Based on these definitions, the most relevant features of crowdsourcing have been highlighted as follows:

- Workforce flexibility
- Open-access production
- Voluntary participation.
- Mutual benefits among stakeholders.

There are several publications that point at the most important concepts and associated with crowdsourcing. These publications show that one of the main activities of crowdsourcing is the recruitment of a large workforce that has the right and required skills for a particular task.

Crowdsourcing is not limited to open call for the recruitment of the workforce.

There are so many crowdsourcing activities that can be seen online.

In 2001, Jack Hughes launched TopCoder, which was a marketplace that uses crowdsourcing in software development for the purpose of facilitating software development activities distributed over the Internet.

In addition to the introduction of the TopCoder system, a software development method was proposed by the author [28].

In March 2015, this emerged as the world's largest crowdsourcing platform for software engineering. This platform was awarded financial rewards when the number of software engineers on the platform reached 750000.

Crowdsourcing offers the following benefits:

- Easy access to a wide range of workers
- Various solutions.
- Minimal time-to-market.
- Reduced cost of labor.

The use of crowdsourcing has been extensively employed in numerous disciplines, such as protein structure prediction [29] information retrieval [30], transportation planning [8] weather forecasting [9] drug discovery [7], and software engineering [32].

1.2.1 A glance at History

Crowdsourcing began well before computers and the Internet age.

1- **18TH** Century – Longitudes

The emergence of crowdsourcing can be traced to as far back as the 18th century A.D, when in 1714, the British Government employed the use of crowdsourcing strategy to find someone that could develop a reliable method through which a ship's longitude while on the sea can be calculated. Consequently, John Harrison who emerged as the developer was awarded the "Longitude Prize." The awardee was a watchmaker, and as such, he was able to calculate the longitude with the aid of highly accurate clocks.[55]

2- **19TH** Century – Production of Oxford English Dictionary

In the 19th Century, there was a call from different brilliant individuals for the re-examination of the full English language.

This led to the birth of the Oxford English Dictionary. In 1879, the philosopher James Murray had a good idea which was in form of crowdsourcing. In his project, readers of his spoken English were asked to send him references of common and unusual words. He got largely positive responses to his request. Thus, in its truest meaning, that was the first time the concept of crowdsourcing was applied to produce the Oxford English Dictionary through outsourcing. [53]

3- **20TH** century – Toyota company

In 1936, Toyota Company, which is a popular Japanese automobile company, outsourced the redesigning of its logo to the general public. This project took the form of a competition for the designing of the logo, and as a result of the competition, 27,000 logos were received by the company, but the winning logo was that which had the three Japanese katakana letters for "Toyoda" in a circle [6]. Afterwards, the name of the company was changed to "Toyota".[54]

4- **21TH** century – Wikipedia

In the era of Web 2.0, the production of Wikipedia was the first attempt at crowdsourcing. Wikipedia was the first crowdsourcing project of Web 2.0, and as a result of the crowdsourcing project, while the online encyclopedia, popularly known as Wikipedia emerged.

Wikipedia is likely based on the idea of Rick Gates, who is the pioneer of the Internet. This innovator, in 1993, proposed the development of an encyclopedia in the World Wide Web to a Usenet newsgroup. The creation of articles was conducted as usual by the Authors. Their texts went through a peer-review process, and there was a chief editor. Toward the end of 2000, while the beginning of 2001 allowed people to read this website, unusual at the time, and make changes in the website to directly in the browser. For the first time on the 15th of January 2001, the online encyclopedia was launched on its own wikipedia.com. This is considered the birth of Wikipedia, which was initially a “fun side-project”. Subsequently, the use of crowdsourcing increased and became more prominent.[55]

1.2.2 Crowdsourcing Users

Crowdsourcing can be used as an efficient organizational strategy for the distribution of work to internet users. It allows organizations and businesses to access a huge number of on-demand workers that work on a project, and then their results are incorporated into the processes and systems of businesses.
(See figure 1.3) explain for Applications of Crowdsourcing [33].

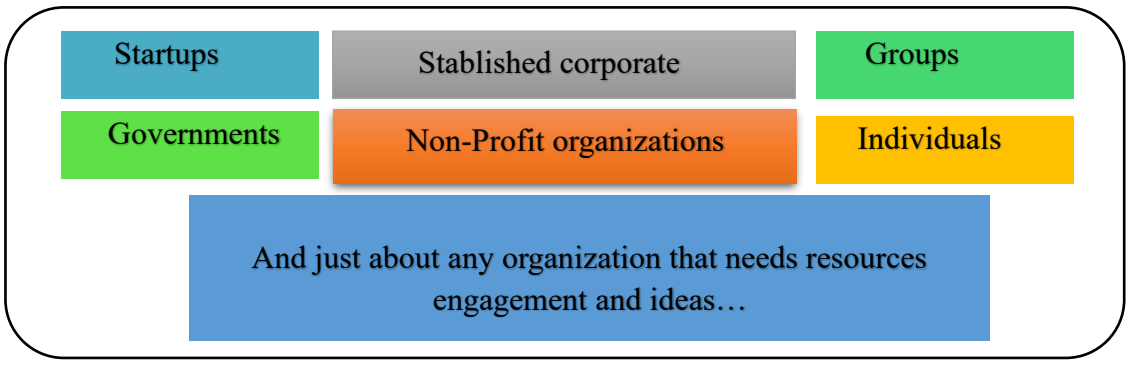


Figure (1.3): Applications of Crowdsourcing [55]

1.2.3 Steps Involved in Crowdsourcing

There are six steps involved when crowdsourcing is used. However, the three main steps involve preparation of the project, sharing the idea with the crowd, then a collection of final work products. Several of these steps are supported by crowdsourcing platforms that guide based on steps. Organizations can gain easy to crowds through these platforms, which also provide the crowds with the required compensation. Below is an illustration of the six steps needed for efficient crowdsourcing:[33]

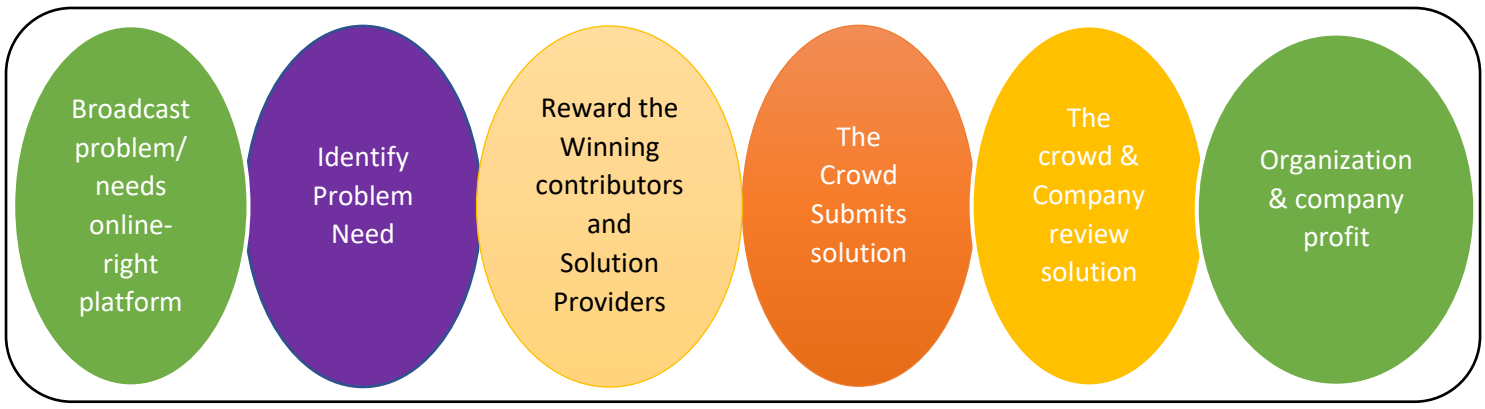


Figure (1.4): Steps to Initiate Crowdsourcing [33]

1.2.4 Advantages of Crowdsourcing

Crowdsourcing offers great advantages including the following: [41]

1. Lower costs: as compared to the traditional method of employing people as permanent employees to perform a task, crowdsourcing is more rewarding as it helps businesses and organizations to save cost by only hiring a group of people to carry out a task when the need arises, rather than permanently retaining them and paying them on a regular basis. In other words, it allows pay as you go, thereby saving cost and time.
2. Greater speed: because a large number of people is employed to perform a task at a time, the time needed to solve a problem is reduced, especially when it involves completing several minor tasks in real time.
3. More variety: using this approach allows companies the luxury of variety and choice, this is because some companies may be empty of ideas, especially when mini tasks are involved. However, crowdsourcing affords them the opportunity to derive new ideas from the knowledge backgrounds and life experiences of others.
4. Marketing and media coverage: Crowdsourcing offers excellent media coverage, which makes it a good source of cost value effective marketing.

1.2.5 Disadvantages of Crowdsourcing

there are also various disadvantages associated with crowdsourcing [40]

1. Breaching of confidentiality sometimes arises, when an individual or group of people are not loyal to the work, and its terms and conditions.
2. Collaborators may have difficulty in communication, especially when they come from different ethnic backgrounds, with different languages. More so, different time zones may slow down the pace of the work.
3. Difficulty in test coverage: Ensuring test coverage can be extra difficult, which will lead to increased requirement management so that errors can be accurately identified.

4. Due to the fact payments are made according to the number of errors found, instead of the magnitude of the errors, testers may only focus on finding many minor errors, instead of a few greater ones. This is to enable them earn more since the payment is based on the quantity of errors. [63]

1.2.6 Types of Crowdsourcing

There are three kinds of collective intelligence that can be used in the field of human-computer interaction, and in this subsection, the characteristics, benefits, and challenges of the three collective intelligences are discussed.[37]

1. Directed Crowdsourcing: In this case, workers are directed to pursue a given goal, and for that reason, when there is a conflict between the interests of the crowd workers and employer/requester, conflict emerges. In an event that such a conflict arises, then the crowd may demand for more incentives before completing the tasks. Even though monetary rewards help in increasing quantity, they do not enhance the quality of the task carried out. Thus, when directed crowdsourcing is used, the employer may need to put in place a mechanism for ensuring quality control. This way, companies can reap more meaningful outcomes. In addition, it has been argued that this type of crowdsourcing is the most suitable for tasks that workers need to demonstrate good performance with little training. The quality of work may be compromised by the crowd if it perceives the task as difficult [56].
2. Collaborative Crowdsourcing: this kind of crowdsourcing is not dependent on directed forms of contracting the job, but depends on the intrinsic motivation of the hired workers. This type of crowdsourcing is referred to as collaborative crowdsourcing and has proven to be the driving force behind the creation of Wikipedia. In this kind of crowdsourcing, workers are required to have a level of emotional connection to the work which is to be done, so that effective participation can be achieved.

As the case may be, stimulating adequate intrinsic motivation among users collaborating to work on similar projects for businesses can be very challenging when there is no emotional connection. In this type of crowdsourcing, the issue of internal conflict may emerge due to the absence of a well-established leadership structure. However, apart from impairing the process, it offers some benefits like triggering debate and discussions among the crowd, which in turn produces better insights or new perspectives. Nevertheless, gain new insights depends completely on the level of knowledge possessed by the crowd in question. Particularly, this kind of crowdsourcing is dependent on the knowledge which the crowd already possess, and this possess a major challenge [56].

3. Passive Crowdsourcing: in this type of crowdsourcing, an indirect relationship may exist between the requester and the crowd. Due to the normal behavior of the crowd, the useful output is produced. So, rather than the crowd being directed by the requester, their performance is monitored passively by the requester. In this case, the crowd must be provided with the infrastructures that are specifically dedicated to their work, so that interference can be prevented while valid and usable results are obtained. Interference must be avoided in this kind of crowdsourcing because quality results can only be obtained when ordinary behavior is well-observed [56].

Irrespective of the kind of crowdsourcing, the definition of crowdsourcing given by Howe (2006) [1] implies that all the tasks to be performed are aimed at reaching a common goal. More so, this author noted that crowdsourcing activities can be carried out by an unspecified and large group of users. Crowdsourcing activities can be flexible given the fact that they are carried out by a large unspecified group of people.

1.2.7 Different Types of Crowdsourcing Activities

Crowdsourcing is used to obtain knowledge and information on a particular topic.

1. Crowdfunding: involves fund raising for new innovations and initiatives by organizations or individuals. This can be carried out individually by individuals or corporately by organizations with this type of crowdsourcing, the chances of riskier activities acquiring their desired funding are higher, because the risk of investment is spread among the whole crowd.
2. Crowd Labor: to large complex tasks like text translation or image classification.
3. Crowd Research: this activity entails obtaining insights from intended audiences; it is usually used for marketing research or gathering of public opinions on an issue. This is achieved using surveys.
4. Creative crowdsourcing/Crowd casting: here, a crowd is motivated to find solutions to problems that require creativity such as design contests for the purpose of developing products, for rebranding, or designing logos. Normally, this type of crowdsourcing comes in form of rebranding. Normally, they are presented in form of a contest, whereby, a task or problem is suggested, and a reward promise is made for a better and faster solution.
5. Crowd collaboration: in this kind of crowdsourcing, a crowd is mobilized and directed to work collaboratively to provide a variety of solutions for a case presented by the director. For example, the crowd can provide customer support to fellow customers.
6. Crowd content: in this type of crowdsourcing, a crowd is directed to proffer solutions to problems in a manner that is non-competitive. The crowd is expected to achieve this by using their labor and knowledge. Most times, the use of this form of crowdsourcing is employed when the crowd is required to produce usable input or to analyze and provide feedback from large chunks of work. [56]

Apart from the numerous possibilities that abound in harnessing the power of the crowd, there are several other benefits that can be derived from the use of crowdsourcing. One of them is the different ways through which crowdsourcing can be used to gather input in an affordable and quick manner [56]. Crowdsourcing also creates new opportunities for organizations.

Secondly, it allows a large number of people to participate in problem-solving, thereby increasing the chances of getting better results and outcomes [56].

However, it is only when crowdsourcing methods and strategies are correctly applied that the benefits can be reaped. The quality of results that emerge from crowdsourcing can be negatively influenced by incorrect crowdsourcing practices, and this has been observed to be a trend that cuts across all models of crowdsourcing, types of the crowd, and crowd size, pointing to challenges in terms of quality assurance [56].

In addition, crowdsourcing is regarded as a traditional and effective way of solving problems with little expertise [37]. However, one of the challenges confronting crowdsourcing is ensuring a high level of quality when more complex tasks are being performed. Thus, it becomes important to gain insights into the dynamics between possible tasks to be performed and the crowd within the context of crowdsourcing. This will help in designing an appropriate method that is compatible with the goal of this thesis.

1.3 Crowdsourced Software Engineering

The term 'Crowdsourced Software Engineering' refers to the use of crowdsourcing strategies to support the development of software. In some previous works, it is referred to as ‘Software Crowdsourcing’, 'Crowdsourced Software Development,' and 'Crowdsourcing Software Development' [38]. However, the term 'Crowdsourced Software Engineering' is preferred given the fact that it emphasizes any software engineering activity, implying the inclusion

of every other activity that does not in itself yield software, for example, refinement of a test case, eliciting for requirements, and project planning.[39]
However, the definition be inclusive of all software engineering activities
(See Figure 1.5).
Crowdsourcing could be used to support any application or research that involves human subjects because there are certain crowdsourcing techniques that can be used to identify and recruit suitable human subjects.

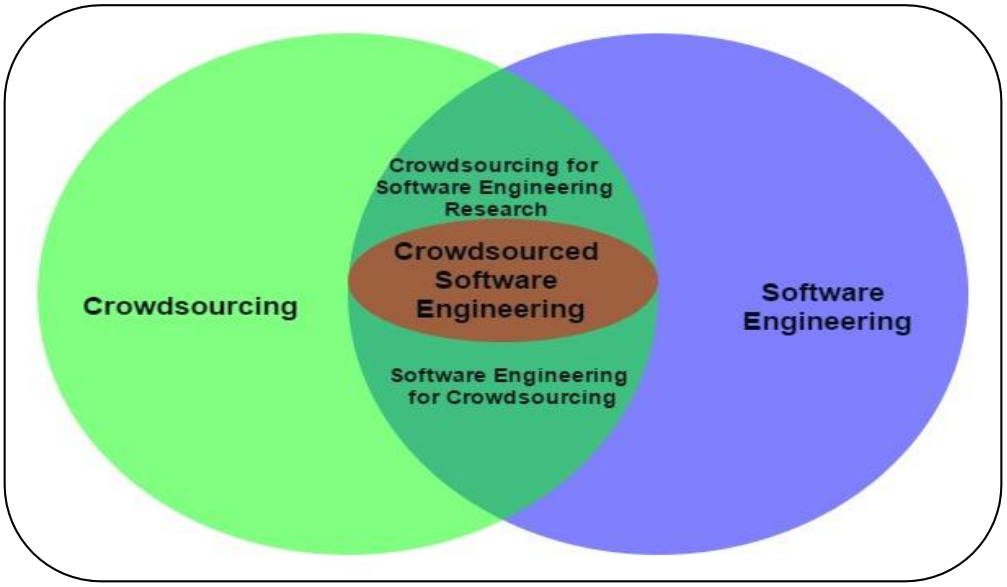


Figure (1.5): Crowdsourcing and Software Engineering [41]

Even though crowdsourcing is widely used in different software engineering activities, there is rarely a comprehensive definition of crowdsourced Software Engineering. Howe' definition still remains the most widely used definition which was earlier provided.
The Dagstuhl seminar report's definition [40] was formalized as a Wikipedia page on software Crowdsourcing. In this definition, the tasks involved in the development of software are well-defined, and the definition also notes that there is no restriction to who can be part of the labor force. However, the characteristic of a large potential workforce was not captured in the definition.

Crowdsourced Software Engineering generally involves three major stakeholders:

1. Actors (stakeholders)
2. Employers (requesters), they usually have software development work that needs to be done.
3. Workers; they participate in developing software and Platforms.

That provides an online marketplace during which requesters and workers can meeting. (See Figure 1. 6) shows a brief description of these three types of actors and their general activities in Crowdsourced Software Engineering.

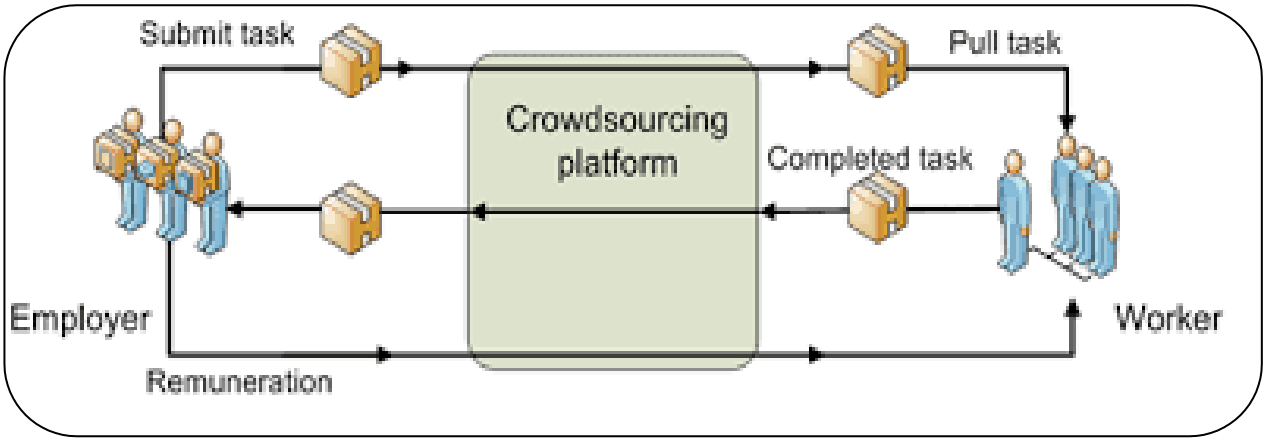


Figure (1.6): Actors in Crowdsourced Software Engineering [41]

1.3.1 Crowdsourcing Applications in Software Engineering

A very critical and widely accepted step which has an effect on the successful completion of software crowdsourcing projects for Software Design is Software Requirements Analysis. This can be accessed through existing commercial crowdsourcing marketplaces.

There are numerous platforms that support software interface design and software coding. There are three sub-areas for the application of crowdsourcing software coding: [41]

- Integration of development environment enhancement: here the knowledge of the crowd is used in supporting coding activities in integrated development. [61]

- Environments crowd programming environment: Instead of adopting crowd knowledge over the Internet to help in conventional coding activities, more attention is paid to the provision of systems that support crowd-based coding work.[5]
- Program optimization: here, the use of crowdsourcing is employed in compilation optimization.[41]

1.3.2 Challenges of Crowdsource Software Engineer

Crowdsourcing is accompanied by some challenges that cannot be ignored. Thus, in the work done in [62], the authors highlighted some of those challenges and how they can be overcome.

1. Sourcing for the Right Crowd: one of the ways through which one or more groups can be created is through the use of social media like Facebook. Creating one or more groups on Facebook, which allows the page owner to build a crowd in the form of followers that can generate new ideas and offer a wide range of suggestions through comments.
2. Emergence: Crowdsourcing can be used to derive a variety of perspectives regarding issues, but don't allow companies or organizations to get sidetracked. Given this challenge, the author suggested the development of tools that can help in achieving the clearly set goal in accordance to the mission and vision of the company.
3. Crowd Management: it is important to have a platform that has a combination of the right audience with the tools required for the management of resources.
4. Measure success of the challenge, there are several specific metrics that differ according to what the organization or company focuses on.
5. The Focus ideas for a final solution only or focus on the marketing side in order to reach a specific audience.

1.4 Problem Statement

Crowdsourcing has emerged as a novel model for problem-solving which makes it easier for computers to solve complex problems, such as the analysis of feelings and entities. By soliciting for answers through crowdsourcing, problems that were not visible or ambiguous, as well as the solutions, can be identified.

So far, the full potentials of crowdsourcing are yet to be reached. Thus, this work explains how the benefits of crowdsourcing can be derived at an optimal level. provides researchers in how crowdsourcing techniques improve the benefits of software engineering. which helps to find ideal solutions to difficult problems in record time and low cost.

1.5 Aim of the work

This study is aimed at proposing a model that can support the recruitment of the largest possible number of crowdsourcing so that new ideas can be obtained through crowdsourcing. The best solutions can emerge from the ideas obtained through crowdsourcing. the aims to take advantage of crowdsourcing in the unified card system for the purpose of providing various ideas that can solve difficult and unexpected problems faster, which contributes to the development of Its products, services and systems.

1.6 Related Work

1- Latoza et al. (2014). [50] Showed that when the process of software design is crowdsourced it allows designers to carry out a peer review of the designers presented by their peers. The researchers also revealed that the use of crowdsourcing resulted in top quality designs, where all the made use of each other's feedback on their initial plans and implementation for the improvement of the designs.

2- K. Mao, L. Capra (2017). [62] Argued that crowdsourcing is increasingly gaining popularity in software research engineering. They also suggested that

work on crowdsourcing software testing for framing the test design problem should be outsourced to the crowd.

3- Martijn van Vliet Thesis (2019). [56] In the work, a summary of a crowdsourcing approach was given. The approach allows the user to analyze large datasets as a possible solution for such types of problems. When the workload required for human tagging is reduced and the required effort divided among the crowd, organizations may be able to obtain usable requirements from this kind of user feedback. Possibly, an organization may wish to use a crowd to create enough training datasets so that the results from extant Natural Language Processing (NLP) techniques can be improved. Either of these approaches is in line with emerging themes in requirement engineer (RE) which focus on exploiting both artificial intelligence and human intelligence for a complementary approach for the elicitation of requirements.

4- Abel Menkveld Master Thesis (2019). [63] The idea of crowd-centric software development formed the basis for the thesis. The connection with the technique has always been a topic of interest: translating these requirements into a business or service that people love to use. The author primarily focused on combining IT and business, while using information scientist as the connecting bridge between the two divides.

5- Hani Barjes Salmeh Al-Bloush,jun Thesis (2020). [68] In this study, the author noted that when Crowdsourced Software Engineering is employed, the power of using a huge crowd can produce huge results such as reliability, flexibility, and quality with less cost and time. However, these benefits are accompanied by issues associated with Property rights in terms of the identification of Intellectual property ownership as well as the privacy of crowdsourcing tasks, originality of contents generated by the crowd, and acquisition quantity.

The researcher focused on safeguarding Intellectual property rights by developing a guideline for crowdsourcing platforms that are adopted in Software Engineering projects. It is expected that the guideline can help in

safeguarding Intellectual property through proper management and control of stakeholders.

6-Nour Jamal Absi Master Thesis (2021). [88] According to this author, the idea of crowdsourcing is adopted by companies in order to find novel ideas from employees, the public, and vendors that; these new ideas can be applied in their product growth working processes. In this study, it was also noted that through the process of crowdsourcing, a large amount of data can be derived, however, the data must be sorted out by experts so as to ensure that the right and the relevant idea is filtered from the large volume of available data.

These experts usually filter the ideas using computational algorithms combined with their human intuitions, since computers do not possess human intuition and are unable to make human judgments. The authors carried out the process of filtering the ideas using Random Forest Classifier teaching.

1.7 Gap

Crowdsourcing, as discussed related works, is performed by multiple individuals in different locations. It's far more scalable than outsourcing, Crowdsourcing is particularly good for enhancing smaller software quality (SQ) teams, as it brings in a more diverse range of people to give a fresh perspective and provide user feedback. Teams with irregular testing needs can use crowdsourcing to expand their testing capacity only when needed.

In addition to the flexible possibilities of the harnessed power of a crowd, additional benefits for the utilization of crowdsourcing exist. Firstly, crowdsourcing offers several ways through which input can be quickly gathered in an affordable manner that open new opportunities for organizations.

Secondly, it appears to broaden participation as it opens up the possibility for additional people to participate in problem-solving, which is linked to better outcomes and improved results.

Nevertheless, these benefits can only be reaped when the crowdsourcing activities and methods are performed and applied in the correct way because inappropriate practices negatively affect the quality of the results generated; which appears to be a traversing trend between all crowdsourcing models crowd types and sizes showing challenges in quality assurance.

More the idea of crowdsourcing has traditionally proven to be the most effective approach for solving problems that require little expertise while assuring an adequate level of quality for more complex tasks remains a challenge.

It is, therefore, crucial to further understand the dynamics between possible tasks to be performed and the crowd in a crowdsourcing setting, to ensure the construction of a correct method compatible with the goal project.

1.8 Thesis Organization

This thesis is split into five chapters. Every chapter begins with a short overview that offers a general description of the chapter. The essential contents of other chapters are as follow:

- Chapter Two: entitled "Theoretical Background." This chapter provides an overview of the pre-processing and data mining concepts, SGD, RF, DT, LR, KNN algorithms, and Likert Scale, additional evaluation of classification.
- Chapter Three: entitled "The Proposed System." It covers Propose Collect data and build a dataset together with the proposed system and its algorithms used.
- Chapter Four: entitled "Results and Discussion." This chapter demonstrates the results of the proposed system and the research experiments. It also discusses the evaluation of the system's performance.
- Chapter Five: entitled "Conclusions and Future Works." This chapter presents the research conclusion and the possible future research directions to improve work.

CHAPTER TWO
THEORETICAL BACKGROUD

CHAPTER TWO
THEORETICAL BACKGROUND

2.1 Introduction

Due to the fact that crowdsourcing has emerged as a robust paradigm for solving large-scale problems, especially problems that are difficult for computers to solve, but easy for humans, the use of crowdsourcing is employed in the identification and extraction of the knowledge required. A combination of machine and human efforts can enable researchers to solve large-scale complex problems in an efficient manner. More so, higher accuracy can be achieved by leveraging human input through crowdsourcing. However, if a great number is crowdsourced, there will be a tremendous rise in the cost and time required for processing. Thus, combining the accuracy of a cost-effective and fast computer algorithm and human input is the best and most natural option that will yield better accuracy.

There is a wide range of evaluation strategies that can be used to ascertain if an idea is winning or not, and some of such strategies include, voting by participants; participants can vote in favor of or against an idea. Also, the participants can be asked to rate an idea. However, applying such strategies to a large collection of ideas is challenging. [36] To this end, this study is solution-oriented technical research which focuses on validating an artifact through simulation. In this chapter, the theoretical background of the study is discussed, while presenting an overview of the basic concepts of data mining (DM), a detailed description of the steps involved in preprocessing, and concepts associated with machine learning methods as well as a supervised classification method in general. Lastly, this chapter addresses the theoretical background of the performance benchmarks used for evaluating
the performance of classification algorithms. Such performance evaluation metrics for system evaluation include Accuracy, Recall, Precision, and F1-measure.

2.2.1 General Crowdsourcing Process

Based on the definition of crowdsourcing by Howe, there are three keys stakeholders involved in the process of crowdsourcing. These stakeholders include crowdsource (who is also referred to as Client or Employer); this stakeholder is responsible for executing software engineering tasks, and sourcing for the crowd online. The second group of stakeholders is the crowd, which is made up of contributors, workers, or participants; these individuals participate in software development tasks and crowdsourcing platforms that are referred to as Service Providers, Facilitators. The facilitators are responsible for providing an online infrastructure that supports the meeting of crowds and crowdsources. The roles that each stakeholder needs to play and the interactions that exist between each of them acts as a mediator between the crowdsources and crowd participants, where a Crowdsourced Software Engineering task is being accomplished. Regardless of the interactions that exist between the stakeholders, their identity is unknown to each other. They are usually not bound by similar rules, neither are they expected to carry out similar duties in terms of accomplishing their assigned tasks, which are also not known beforehand. [41]

Figure (2.1) next page shows the three keys' stakeholders involved in the process of crowdsourcing. More so, crowdsourcing platforms play a critical role in enabling the interaction between crowd participants and crowdsources. The figure also shows the different roles played by each of the stakeholders.

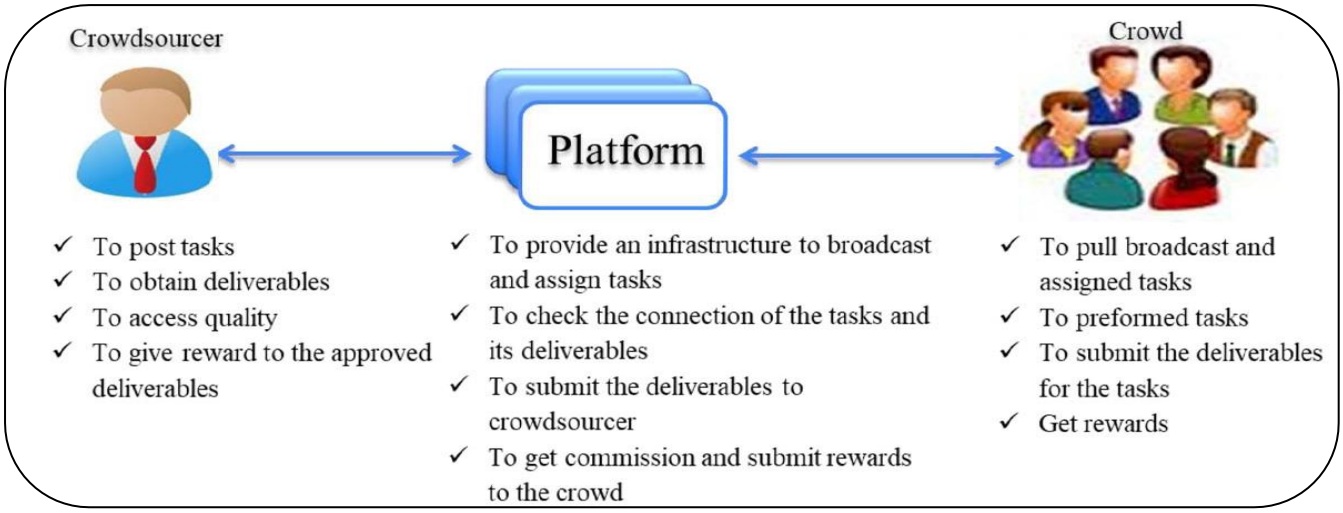


Figure (2.1): Actors in Crowdsourced Software Engineering [41]

The interaction between the stakeholders occurs through different processes and mechanisms, including enrolment mechanism, authentication mechanism, crowdsource task broadcast mechanism, task assignment mechanism, assistance mechanism, crowd skill declaration mechanism, time negotiation mechanism, price negotiation mechanism, result submission mechanism, result from verification mechanism, coordinate crowd mechanism, supervise crowd mechanism and feedback loops mechanism.

To summarize, the process of crowdsourcing allows organizations to access a crowd workforce online in a cost-effective manner, while creativity is harnessed. Extant literature on crowdsourcing shows that crowdsourcing is characterized by individuals or a group of experts or software engineers that are involved in a project that is aimed at finding a solution to a problem; the services they render are usually paid for in the form of incentives and monetary rewards. [56]

2.2.2 Task-Oriented Crowdsourcing

Crowdsourcing has emerged as a common paradigm for solving problems in different organizations and institutions because it allows them to leverage on the knowledge, intelligence, and experience of the crowd, which in turn allows the organization to have a variety of solutions and ideas to select from.

There has been extensive use of crowdsourcing in different tasks, including the collection of data, recognition, and categorization [24], etc. There is a strong connection between crowdsourcing and human computation, which allows humans to perform tasks that should be executed by machines due to the better performance demonstrated by humans. In other words, crowdsourcing allows humans that carry out certain computational tasks rather than using machines, because humans perform the tasks better than a machine. In most cases, crowdsourcing is a better option as compared to the use of machines, because it is a form of collective intelligence that is overlapping of human computation. It is task-oriented involving the aggregation of responses, designing of a task, and quality control. [68]

2.2.3 Answers Aggregation

Regardless of the benefits that can be derived from crowdsourcing, it is accompanied by some challenges such as the aggregation of the responses obtained from the crowd; this is one of the major challenges associated with crowdsourcing. Another challenge associated with crowdsourcing is the diversity in background and expertise of the crowd, which may result in disparities and uncertainties. More so, human error can occur as a result of inadequate expertise, carelessness, or the difficulty humans experience when questioning themselves. In addition, spammers or malicious workers can submit random responses with the aim of achieving financial benefits or rewards.[68]

2.2 Machine Learning

The process of machine learning entails training a computer program in a way that it is able to incrementally improve a given task. In terms of research, machine learning can be viewed from the perspectives of mathematical and theoretical modelling of the workings of a given process. On the other hand, in practical application, machine learning is the field of study that focuses on understanding how applications are created, showing iterative improvement.

Machine learning involves programming to enable the optimization of a computer's performance through the use of data. The fitness of a model can be determined using certain parameters, and the parameters of the model can be optimized using training data. Such a model is predictive in the sense that it can be used to predict the future and to gain insight from a given set of data. In machine learning, mathematical models are built using statistics because the main aim is to make inferences. The purpose of computer science is twofold: first, it is used in training efficient algorithms, which are required for solving problems to an optimal level, and for storing and processing a large amount of data. Second, once a model is learned, its representation and algorithmic solution for inference must also be efficient. [81] In some applications, the efficiency of an algorithm's inference can be determined by its time and space, predictive accuracy, and time complexity [44].

- This idea can be described using different concepts, but the three widely used concepts are given in Machine Learning below: [64]
- Supervised Learning.
- Unsupervised Learning.
- Semi-supervised Learning

Given the emergence and prevalence of machine learning and artificial intelligence, the various kinds of machine learning must be defined and understood. The most popular kinds of machine learning are shown in Figure (2.2) below.

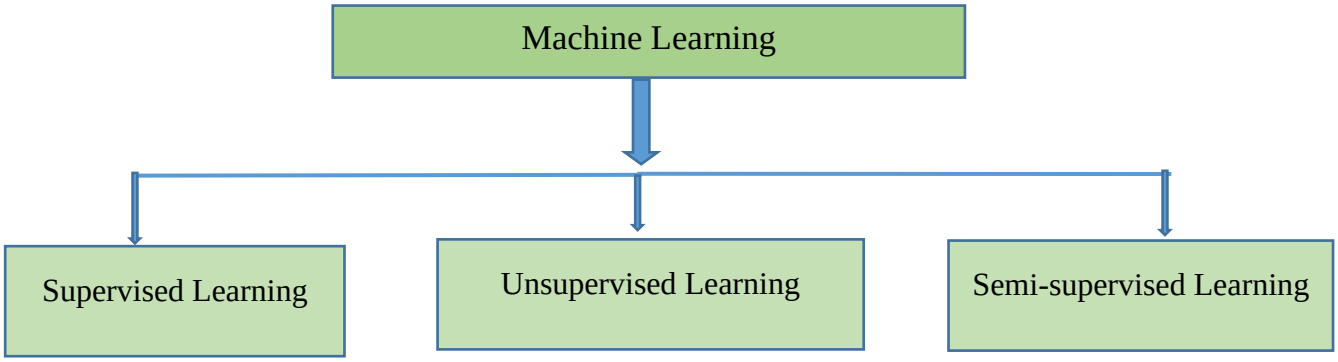


Figure (2.2): Major Kinds of Machine Learning [64]

2.2.1 Supervised Learning

One of the key functions of supervised classification algorithms is to assign class labels to data objects based on the correlations among the objects with predefined class labels [65]. The supervised classification algorithms are those algorithms that need to be augmented, where the input dataset is made up of two datasets (training dataset and testing dataset). The first dataset is characterized by an output that can be modified and categorized. The learning process of the whole model is implemented using the training dataset, and then they are tested using the testing dataset [66].

It includes different classification problems like Random Forest Classification and Logistic Regression.[67] The workflow of supervised machine learning algorithms is presented in Figure (2.3).

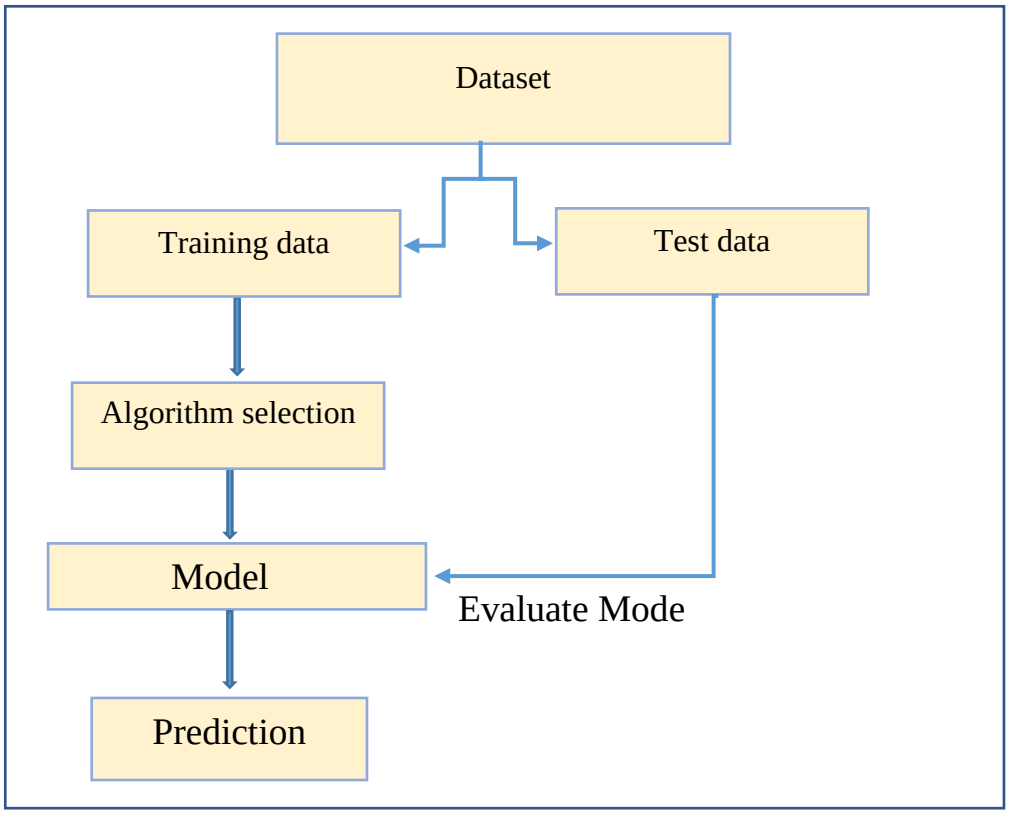


Figure (2.3): Workflow of Supervised Machine Learning Algorithm [46]

2.2.2 Unsupervised Learning

This algorithm is concerned with learning a few features from a given data. When the model is fed with new data, it makes use of characteristics that have been learnt previously to classify the data. The basic functions of the unsupervised learning algorithm are feature reduction and clustering [66]. However, the most important is clustering, which entails the discovery of a structure from a collection of uncategorized data.

In clustering, elements that are similar are categorized together in the same group, while the objects that are dissimilar are categorized together in one cluster [69].

There are many methods that are used for clustering, but the most popular one is the K-NN (k Nearest Neighbors). Using the clustering algorithms, the collection of natural occurring groups can be determined [70].

2.2.3 Semi-supervised Learning

The semi-supervised learning models are derived by combining supervised and unsupervised so as to leverage the power of both. It has great potentials in the areas of data mining and machine learning, where unrecognized data exits, thereby making the extraction of labeled data an uninteresting process [48].

2.2.4 Machine Learning with Crowdsourcing

The benefits and opportunities that abound in crowdsourcing have been recognized in the field of machine learning, thereby leading to the development of different techniques to address the issues of inaccuracy, uncertainty, and randomness that may occur as a result of using crowdsourcing in learning.

These issues are addressed by the techniques by focusing on two key issues, which are the quality of the prediction model and data quality [89]. The machine-learning community has recognized the importance of such systems in terms of creating opportunities for machine learning research [14].

The application of crowdsourcing in machine learning is relatively a new area of learning, which is concerned with human-in-the-loop learning activities and contributes significantly in the present era of big data.

Recently, there has been remarkable progress in the field of learning which combines crowdsourcing with machine learning, and for this reason, the machine learning community have introduced different techniques that can be used in dealing with the issues of randomness, inaccuracy, and uncertainty that may arise when crowdsourcing is applied in machine learning [54].

2.3 Preprocessing Method

The stage of pre-processing is usually implemented to ensure the quality of data, and data mining has emerged as one of the most important procedures in data preprocessing. In the preprocessing stage, the raw data is transformed into a format that is clear and understandable, eliminating uncertainties and ambiguities.

The aim of this step is to prepare the text or document to be used in the systems.

In the real world, data is normally inconsistent, incomplete, sometimes contains errors, and lacking in particular tendencies.

Data preprocessing has proven to be a good technique for solving such problems. Use of data preprocessing is often employed in database-driven applications like rule-based applications and customer relationship management.

Out of the many processes of Machine Learning, the most critical process is data processing as it allows the dataset to be encoded in a format that can be understood and analyzed by the algorithm. Preprocessing of data enhances the efficiency and accuracy of mining algorithms that involve distance measurements. Some of the tasks performed in the preprocessing step include cleaning of data, transforming data, and data reduction [71].

In order to make data usable, there are some preprocessing steps that must be implemented. The different steps of data preprocessing [57] are contained in Figure (2.4).

1- Data Cleaning: the cleaning of data involves activities such as the compensation of missing values, deletion of rows that contain missing data, reconciliation of disturbing data, and reconciliation of data variances. The elimination of noise from data is particularly crucial for machine learning datasets due to the fact that machines are not able to utilize any data that they cannot analyze and interpret.

   So, in order to make the data understandable for the machine, the data is cleaned by equally dividing to segments that are smoothed.

This is achieved by fitting the data into a multiple or linear regression function, or even by grouping the data into clusters with similar data; this process is referred to as clustering.

2- Data Integration: this process involves grouping data that have different representations so as to resolve conflicts that are present in data [57].

3- Data Reduction: databases can be slowed down by the use of a large amount of data, which also increases the cost of accessing the databases while increasing the difficulty of properly storing the data.

For the aforementioned reasons, the data reduction is implemented with the aim of presenting a reduced representation of the data in a data warehouse.[57]

4- Data Transformation: This process involves the normalization and generalization of data. The process of normalization entails ensuring that ever redundant data is eliminated, data is stored in a single place, and that all dependencies are logical. The aim of implementing this step is to ensure that the data is transformed into the most suitable forms for the process of mining.

The transformation of data is carried out so that the data values can be scaled within a given range (-1.0 to 1.0 or 0.0 to 1.0).

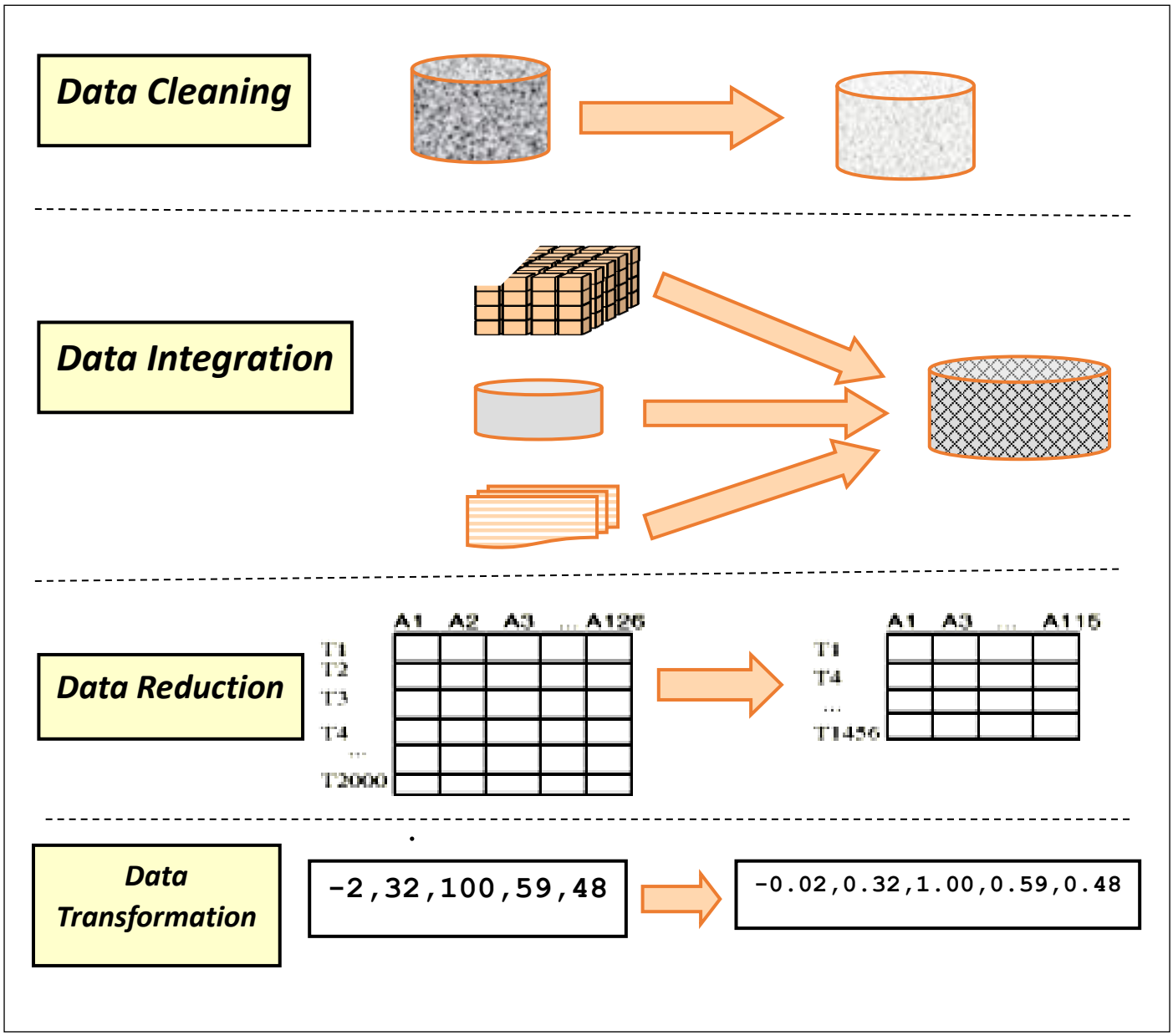


Figure (2.4): Forms of Data Preprocessing [57]

2.4 Implementing Data Preprocessing in Machine Learning

The data preprocessing procedure in Figure (2.5) shows the flow of the steps implemented at the data processing stage; all the stages are followed before the learning algorithms are applied to a given data mining task. In the first step, datasets are obtained from a source, and afterwards, the data is cleaned using some preprocessing mechanisms; the aim of this is to make the data usable. In the next stage, the data are fed into the learning algorithms as the input, and lastly, they are used in solving a particular problem such as regression, classification, clustering, pattern recognition, etc. Some or all of the data preprocessing steps may be important in the algorithms presented in Figure (2.5). It is important to note that the preprocessing step is a step that is critical to data normalization or discretization [58].

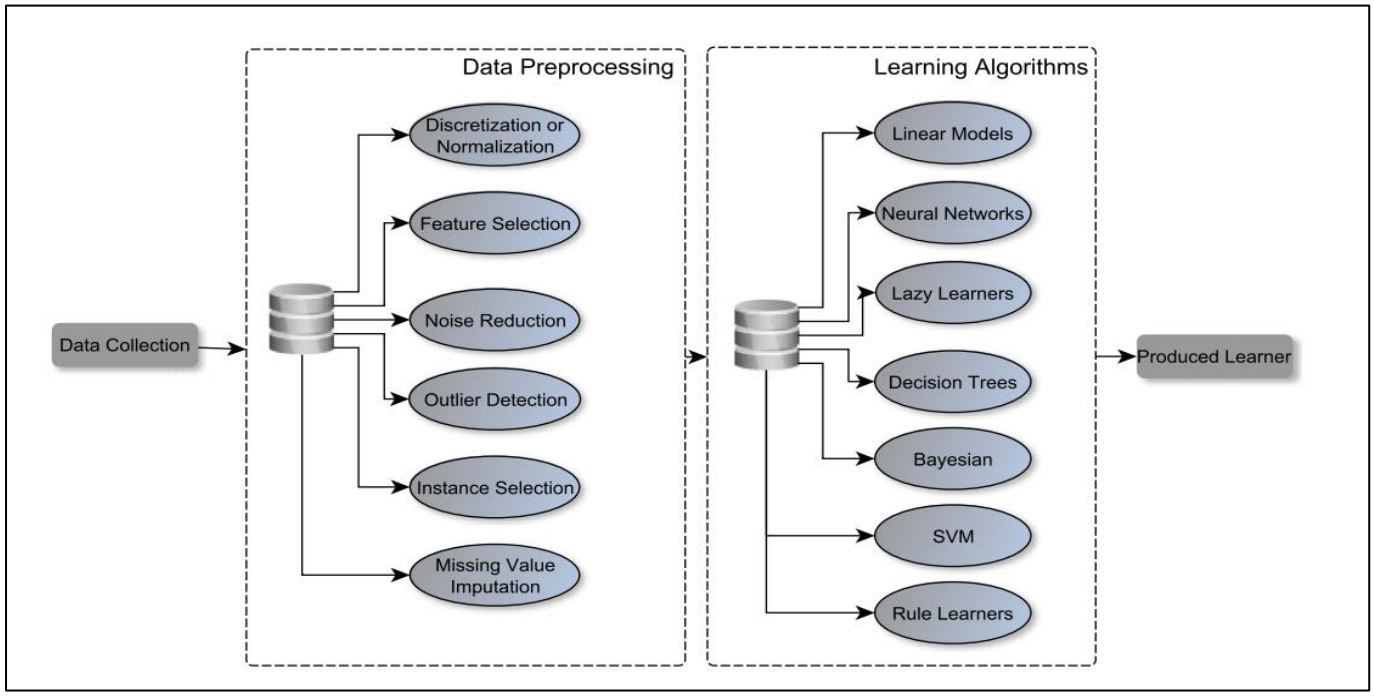


Figure (2.5): Predictive Modeling Process [58]

1. Getting the datasets- In machine learning, the datasets are the most critical component that can be manually collected, or large datasets can be obtained from websites. [58]
2. Feature scaling- This process involves the standardization of a dataset's independent variables in a given range.[72]
3. Splitting the dataset- This stage involves splitting the dataset for the machine learning model into two parts.
4. Separate Training Set from Testing Set. The first set, which is the training set is used in training the machine learning model, while the testing set is used to test the performance of the machine learning model.[72]

2.5 Data Mining Concepts

Data mining is the process of extracting beneficial knowledge or information from massive amounts of data saved either in databases, data warehouses, or other information depots [43]. The structure of a common data mining system may have the basic elements as in Figure (2.6).

The process of data mining involves the use of a wide range of techniques that are specifically designed for such purpose. There are two kinds of data mining tasks, which are description and prediction [35]. In the prediction tasks, the value of a given attribute is predicted using supervised learning techniques based on the values of other known attributes. Two tasks that are involved in predictive modelling are regression and classification [36].

Description tasks involve clustering, mining associations, sequence discovery, and summarization. These methods entail the discovery of clear patterns in data through the use of unsupervised learning techniques.

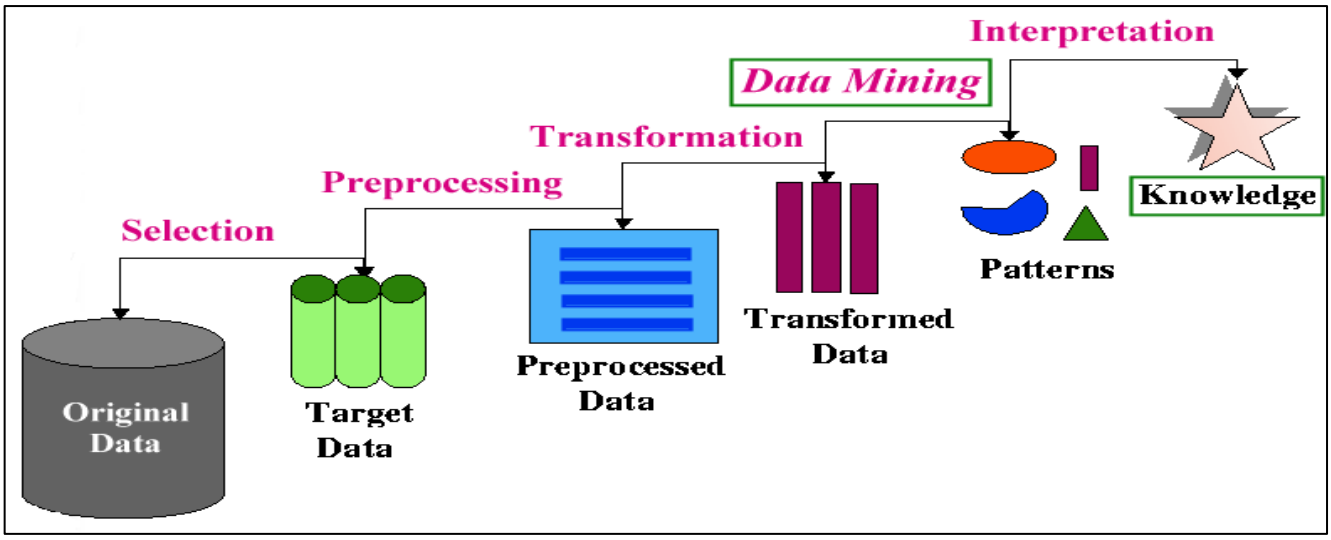


Figure (2.6): Data Mining System Structure [57]

In this study, classes of unknown data are discovered using classification techniques such as Decision Trees, Naïve Bayes, Random Forest, Support Vector Machine, and Neural Networks. The basic concepts of data mining are depicted in Figure (2.7).

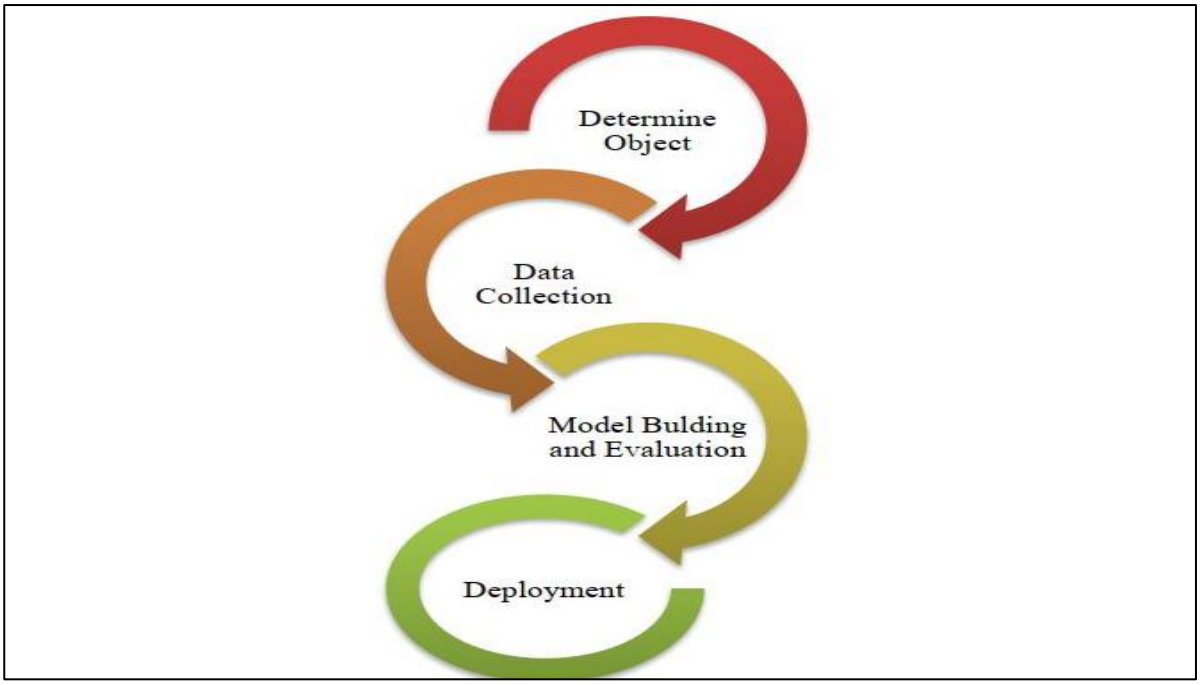


Figure (2.7): Data Mining Concept [37]

2.5.1 Data Mining Challenges [73]

A new trend of problems associated with data mining research and development is emerging as a result of the paradigm shift towards naturally distributed complex problem-solving environments. There are two broad categorizations of such problems, and they are described as follows:

1. Distributed Data: This entails the storage of the data to be mined in distributed computing environments on various platforms. For both organizational and technical reasons, all the data cannot be stored in a central place. This has necessitated the creation of tools, algorithms, and services that can support the mining of distributed data.
2. Distributed operations: The future will experience increased data mining algorithms and operations on the grid, and such, there is a need for the development of new tools, algorithms, and services that can support the smooth integration of these resources into distributed mining systems for solving complex problems.
3. Massive data: The increasing presence of massive data makes it necessary to develop algorithms that can be used in the mining of such large datasets.
4. Complex data types: with the increase in complex data structures, sources, and types (e.g., multi-relational, natural language text, and object data types), there is a need to develop new algorithms, tools, methods, and service that can enable grid-enabled mining of the data.
5. Data privacy, security, and governance: With regards to the security, privacy, and governance of data, there are grave consequences that are associated with automated data mining within distributed environments. To this end, there is a need to put in place data mining technologies that can address such issues.
6. User-friendliness: one of the factors that influences the use of a system is user-friendliness, and as such, the technical complexity of a system must be hidden from an end-user.

This can be achieved by developing novel easy-to-use infrastructure, software, and tools that can be used in the areas of resource identification, grid-supported workflow management, resource allocation, user interfaces, and scheduling.

2.5.2 Crowdsourcing for Search and Data Mining

With the emergence of crowdsourcing, new ways of designing, training, and evaluation are being sought by professionals in academia and in the industry.

Crowdsourcing has made available the opportunity to combine automated systems with human computational abilities so that search results can be validated through quality control and cost control by Crowd-powered for Data Mining Classification Algorithm Machine Learning [83].

2.6 Classification Algorithms

Classification is a process whereby a specific set of data is categorized into classes, and this can be done for both structured and structured data. The process begins with a prediction of the class of categorizing a given set of data into classes, it can be performed on both structured and unstructured data. These classes are normally known as label, target, or categories. In the field of machine learning, the process of classification is a concept of supervised learning which involves the categorization of a set of data into classes. This process is accompanied by some problems such as face detection, speech recognition, classification of the document, handwriting recognition, etc. There are two major categories which these problems can fall under, including multi-class problems or binary classification problems. In the field of machine learning, there is a wide range of algorithms that exist for the purpose of classification. The prediction of outcomes by the algorithms is made possible through the identification of correlations between traits.

Afterwards, the previously unseen set of data is fed to the algorithm as the prediction set, including the same set of attributes, which are not the same as the prediction attribute yet to be defined.

One of the most effective ways to evaluate the performance or efficiency of an algorithm is through its ability to predict accurately.

In other words, prediction accuracy is one of the metrics used in determining the efficiency of an algorithm [44].

The process of classification is critical to different post-processing operations. Here, the data is analyzed using machine learning techniques that automate the construction of the analytical model. It is perceived that the system is able to learn from data, identify patterns, and make decisions with very minimal human intervention [61].

In the current study, in this thesis, the use algorithms have been employed including, Random Forest Classifier (RF), Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Stochastic Gradient Descent (SGD). Additionally, the Likert Scale, In the subsequent sections, the algorithms are explained in detail.

2.6.1 Decision Tree (DT)

The Decision Tree is one of the most efficient ways through which models can be generated in the form of a tree structure [77]. With the Decision Tree, a dataset can be divided into smaller subsets, thereby resulting in the formation of an incremental decision tree. In other words, this technique produces a tree with leaf nodes and decisions. Both categorical and numerical data can be handled by decision trees. The central idea behind the decision tree algorithm is demonstrated in Figure (2.8), while the rule for the decision tree is presented in Equation (2.1) [78].

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t)\, log2\, p(i|t) \quad \ldots\ldots\ldots\ldots (2.1)$$

Where (c) denotes the number of classes,

(i/t) represents the portion of records that belong to class (i) at a given node (t).

The manner in which the decision to split the data is made by a decision tree is controlled by the entropy. In reality, it has an effect on the way a Decision Tree draws its boundaries.

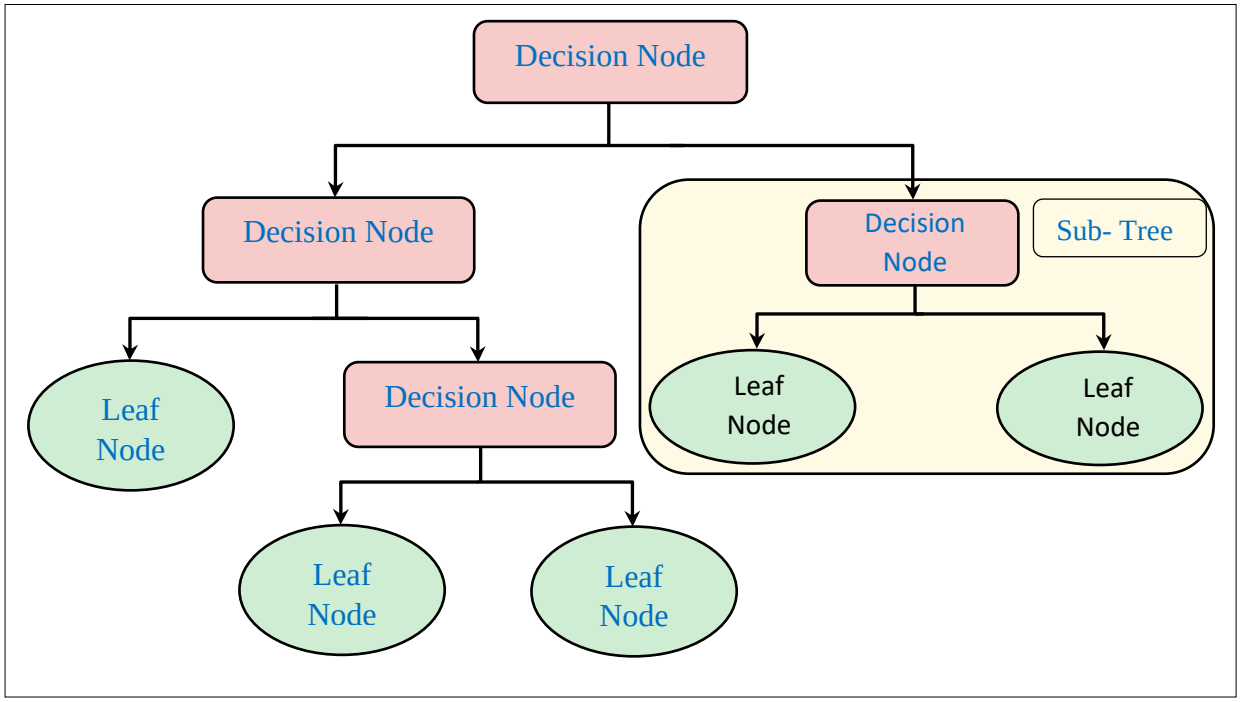


Figure (2.8): The Main Idea of Decision Tree [78]

The construction of the tree occurs in a top-down recursive dived and conquer fashion. A decision node will have two or more branches and a leaf representing a decision or classification. On the decision tree, the topmost node that corresponds to the best predictor is referred to as the root node. The most interesting aspect of the decision tree is its ability to handle both numerical and categorical data. when Pseudocode is explained in chapter three.

A decision tree induction algorithm is shown in Algorithm (3.1) The input to this algorithm consists of the training records (E) and the attribute set (F).

The algorithm works by recursively selecting the best attribute to split the data (Step 7) and expanding the leaf nodes of the tree (Steps 11 and 12) until the stopping criterion is met (Step 1). [13]

The details of this algorithm are explained below:

```
Treegrowth (E, F)
1: if stopping _cond (E, F) = true then
2:    leaf = createNode ().
3:    leaf. label = Classify
4:    return leaf.
5: else
6:     root = createNode ().
7:     root.test_cond=fin_best–split (E, F)
8:     let v = {v/v is a possible outcome of root.test_cond}
9:     for each v € V do
10:       Ev = {e/ root.test_cond € = v and e € E}
11:       child = TreeGrowth (Ev, F)
12:       add child as descendent of root and label the edge (root → child) as v
13:   end for
14: end if
15: return root
```

A- Advantages and Disadvantages [78]

- The major advantage of a decision tree is that is simple and easy to understand as well as visualize. Due to its simplicity, not much data preparation is needed.
- The disadvantage which accompanies the decision tree is its creation of complex trees that may not by efficiently categorized. They are characterized by minimal instability, and for that reason, even the slightest change made to the tree can be a hindrance to the entire structure of the decision tree.

B- Applications

- When there is a need to explore data.
- For the recognition of the pattern.
- Option pricing in finances
- For the identification of disease and risk threats.

2.6.2 Random Forest Classifier (RF)

Random forest (RF) refers to a technique that is used for classification and fits a large number of decision tree classifiers on various sub-samples of a dataset. RF is built on the foundation of the decision tree [69]. Here, the creation of a large number of trees is done using a training set, and after the trees are created, validation is carried out so that future observation can be predicted. The output that can be obtained from the use of this classifier can be both continuous and categorical value output. A wide range of benefits can be derived from the use of the Random Forest algorithm, including the following:

(1) It can be used to perform both classification and regression tasks.

(2) It can handle missing values and maintain accuracy for missing data.

(3) It has the ability to deal with a large dataset of a higher dimensionality [75].

The central idea behind the algorithm is presented in Figure (2.9).

when Pseudocode is explained in chapter three.

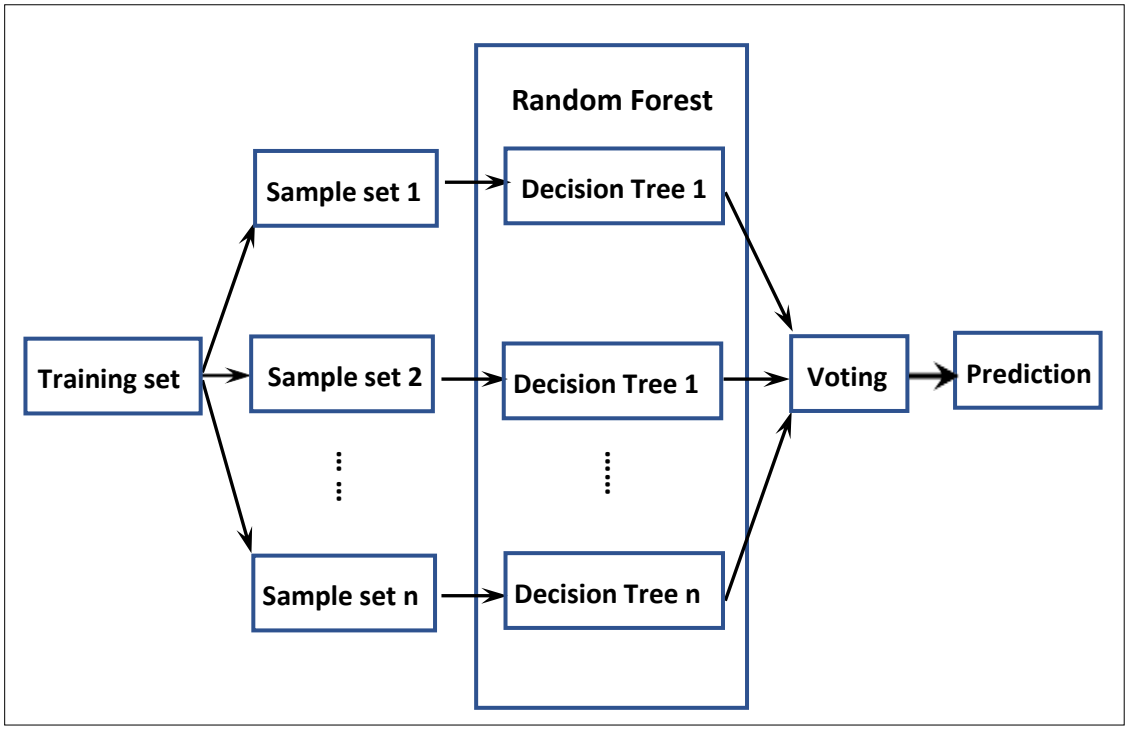


Figure (2.9): The Main Notion of Random Forest [74]

The Random Forest classification is given in Algorithm (3.2). The algorithm selects random D data points from the training set. and building the decision trees associated with the selected data points (Subsets). finally for new data points.[53]
The details of this algorithm are explained below:

To generate c classifiers:
1: **For** i = 1 to c **do**
2: Randomly sample the training data D with replacement to produce D_i
3: Creat root node Ni containing D_i
4: Call BuildTree (N_i)
5: **end for**
6: **Buildtree** (N):
7: **if** N contains instances of only one class **then**
8: **return**
9: **else**
10: Randomly select x% of the possible splitting features in N
11: Select the feature F with the highest information gain to split on
12: Great f child nodes of N, N1, …, Nf, where F has f possible values (F_1, …, F_f)
13: **for** i = 1 to f **do**
14: Set the contents of Ni to Di is all instances in N that match Fi
15: Call BuildTree (N_i)
16: **end for**
17: **end if**

Algorithm (3.2): Random Forest Classification Algorithm

Advantages and Disadvantages

- In comparison to decision trees, the random forest is more advantageous because of its high accuracy due to the reduction in over-fitting.
- It is also accompanied by some disadvantages, out of which the most significant one is its complexity in terms of implementation and it is sometimes slow in real-time prediction. [75]

B- Applications

- Industrial applications in terms of determining if a loan applicant is high-risk or low-risk.
- For prediction of mechanical failure in automobile engines.
- Prediction of social media share scores
- Performance scores.

2.6.3 K-Nearest Neighbor Algorithm (KNN)

This classification algorithm is one of the most basic and simplest methods of classification. Due to this, it is recommended for use as the first choice for a classification study, especially when the researcher has little or no prior knowledge about data distribution.

Out of all the machine learning algorithms, KNN is one of the simplest algorithms that can be used in performing supervised learning tasks. It can be employed for both regression and classification purposes. The major requirement for this is to define a distance between two given samples. The basic function of KNN is to classify a given input element by assigning the most common label amongst its training set based on the distance [82].

The score of similarity of every one of the nearest neighbors. Equation (2.10) shows the Euclidean distance which is used because it functions like humans' normal way of comprehending the real world.

$$\textbf{Euclidean}\ (x, y) = \sqrt{\sum}\ (xi - yi)^2 \quad \ldots\ldots\ldots\ldots\ (2.2)$$

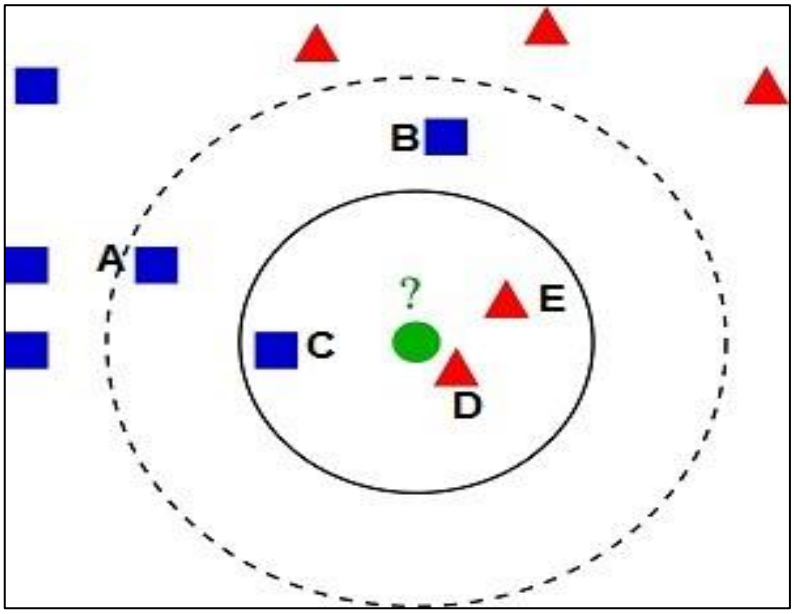


Figure (2.10): The Main Idea of KNN [82]

Where m denotes the number of unique samples in the set of documents,

(***xi***) is a weight of the term (***i***) in data (***x***),

(***yi***) is a weight of the term (***i***) in data (***y***)**.**

when Pseudocode is explained in chapter three.

The K- Nearest Neighbor classification method is given in Algorithm (3.3). The algorithm computes the distance between each test example, (x', y') and all the training examples (x, y) ∈ D to determine its nearest-neighbor list, D".[13]

The details of this algorithm are explained below:

1: Let K be the number of nearest neighbors and D be the set of training examples
2: for each test example z = (x',y') do
3: Compute d(x',x), the distance between z and every example (x,y) € D
4: Select Dz ≤ D, the set of K closest training example to z
5: y' = argmax ∑(x',y') € D I (v= yi)
6: **end for**

Algorithm (3.3): k-Nearest Neighbor Classification Algorithm

A- Advantages and Disadvantages

- From an advantageous point of view, the KNN algorithm offers the benefit of simple implementation and robustness training data that contains noise. This algorithm is efficient regardless of how large the data is.
- The disadvantage that is associated with the use of the KNN algorithm is that there is no need for the K value to be defined, and it is also accompanied by a relatively high cost of computation as compared to other algorithms.

B- Applications

- Industrial applications when there is a need to find a similar task as compared to others.
- Applications that involve the detection of handwriting.
- Recognition of images.
- Video recognition
- Analysis of stock

2.6.4 Logistic Regression (LR)

Logistic Regression (LR) is a popular statistical algorithm that is capable of producing a probability model that can be used in numerous applications.

Use of logistic regression has been extensively employed in the retrieval of information; it has been studied in the field of machine learning. More so, it is characterized as a discriminative model that can be employed in probabilistic categorization,

where it produces the posterior probabilities for test examples that can be suitably engaged in other systems. In an event that the main goal is to classify,

then considering a test example x, logistic regression can directly estimate the conditional probability of assigning a class label y to the example seen in [79]:

$$Z = \frac{1}{1+\exp(-yw^T x)} \quad \ldots\ldots\ldots\ldots (2.3)$$

where $\boldsymbol{y} \in \{+1, -1\}$ is the class label, and $\boldsymbol{w} = (w1, w2, \ldots, wm)$

and $\boldsymbol{x} = (x1, x2, \ldots, xm)$ is the data vector 2,

($\boldsymbol{T}$) are the weight vector and intercept of the decision hyperplane respectively.

It is easy to generalize logistic regression to multiple classes by considering multi-class classification as several binary classification problems. In order to determine if the class should be assigned, the probability estimate can be compared with a threshold. Alternatively, this can be determined by computing the decision that produces optimal expected effectiveness. Accuracy in predictions for future inputs can only be achieved by a logistic model if the training data is not over-fitted. when Pseudocode is explained in chapter three.

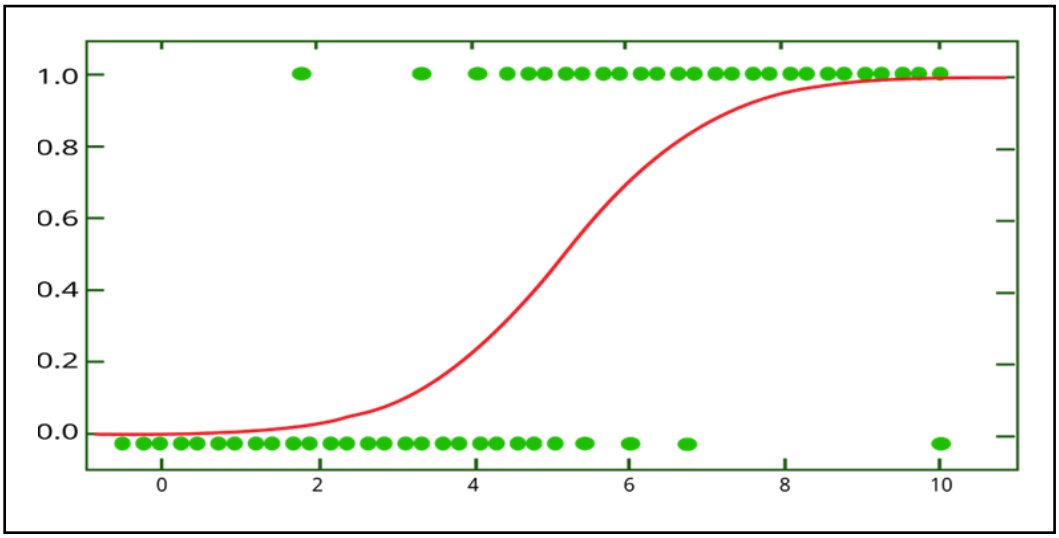


Figure (2.11): The Main Idea of LR [74]

The Logistic Regression in Algorithm (3.4). that use one or more independent variables to determine an outcome can be applied to both, and its output posterior probabilities can be conveniently processed.[85]

The details of this algorithm are explained below:

1: We want our classifier to output values between 0 and 1
2: When using liner regression we did $h\theta(x)=(\theta Tx)$
3: For classification hypothesis representation we do $= C(x)=g((\theta Tx))$
4: Where we define $g(z)$
5: z is a real number
6: $g(z)= 1/(1+e\text{-}z)$

Algorithm (3.4): Logistic Regression Algorithm

A- Advantages and Disadvantages

- Particularly, the logistic regression is designed to perform classification tasks, and it helps in providing insight on the effect of an independent variable on the outcome of the dependent variable.
- The major limitation of the logistic regression algorithm, which is considered as its disadvantage is the fact that it only works with binary predicted variables, assuming that the data does not contain missing values and that the predictors are not related to each other. [74]

B- Applications

- Identification of risk factors for diseases
- Classification of Words
- Forecasting of Weather
- Voting Applications

2.6.5 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is regarded as a simple but efficient optimization algorithm; the values of parameters of functions which reduce cost function can be found. More so, it is a method that is not categorized under a given family of machine learning models. Regardless of its simplicity, it can be efficiently used to fit linear classifiers and regressors under convex loss functions like Logistic

Regression and Support Vector Machines. This algorithm is one of the algorithms that has been in existence in the community of machine learning for a long time and has recorded successes in its application to large-scale datasets due to the fact that the coefficients are updated for each training instance instead of at the end of the instance. Successful and wide application of Stochastic Gradient Descent has been made to large scale and sparse machine learning problems that often arise in natural language processing and text classification.

The sparse nature of the data makes it easier for the classifiers in this module to scale through. [81]

$$Wt + 1 = wt - at. \nabla f_{it} (wt) \ldots\ldots\ldots\ldots (2.4)$$

Whereas (wt) is the value of the parameter vector at time t,

(at) is the learning rate or step size,

and (∇f_{it}) refers denotes the gradient of the loss function.

when Pseudocode is explained in chapter three.

The Stochastic Gradient Descent in Algorithm (3.5). introduce SGD algorithm The sequence of random variables {t}t≥0. introduce a key assumption that each realization ∇f (w; t) is a function.[29]

The details of this algorithm are explained below:

```
1: Initalize: w0
2: iterate:
3: for t = 0,1m2... do
4:     Choose a step size (i.e., learning rate) nt > 0.
5:     Generate a random variable εţ
6:     update the new iteratw wt+1= w1 – nt ∇ƒ (wi, εţ)
7: end for
```

Algorithm (3.5): Stochastic Gradient Descent Algorithm

A- Advantages and Disadvantages

- The only advantage associated with the use of this algorithm is the fact that it is efficient and easy to use.
- On the other hand, the disadvantage of using this algorithm is that it has the requirements of several hyper-parameters, and demonstrates sensitivity to feature scaling and the number of iterations.

B- Applications

- Internet of Things
- It can be used to update coefficients of linear regression or even parameters like weights in neural networks.

2.7 Likert Scale

The Likert scale is a popular rating scale and is effectively used to assess opinions, attitudes, or behaviors. Likert scales are popular in survey research for they allow to easily operationalize personality traits or perceptions.

To collect data, presented participants with Likert-type questions or statements and a continuum of possible responses, generally with 5, 7, 9 questions. Each question is given a numerical score so that the data can be analyzed quantitatively.

Likert items are used to measure respondents' attitudes to a particular question.

To analyze, the data is usually coded as follows.[10]

• 1 = Strongly disagree

• 2 = Disagree

• 3 = Neutral

• 4 = Agree

• 5 = Strongly agree

One must recall that Likert-type data is ordinal data. can only say this one score is higher than another, not the distance between points.

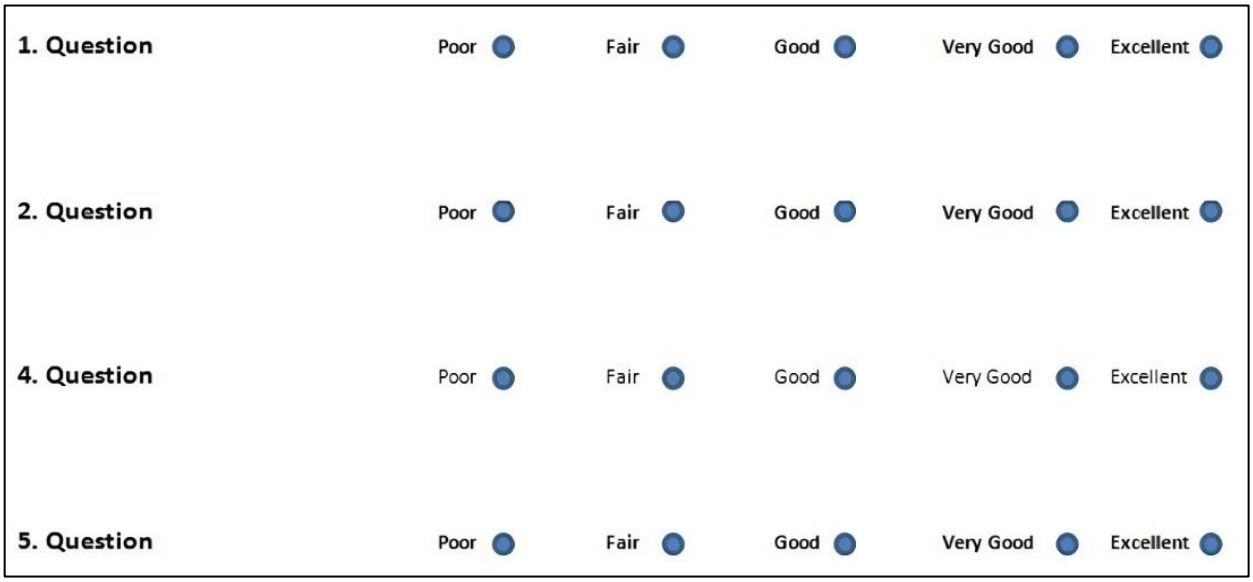


Figure (2.12): Likert Scale [10]

A- Advantages and Disadvantages

There are many advantages and disadvantages of using a Likert Scale:

- Ease of implementation: This scale can be easily understood and applied to various customer satisfaction.
- Quantifiable answer options: conduct statistical analysis on the received results.
- Analyze the rank of opinions: Likert scale offers a ranking of the views of these people surveyed.
- Simple to respond: Respondents can understand the intent of this scale and quickly answer the question.[26]
- Some respondents normally lie in the questionnaire due to either business or attempt to keep privacy hence leading to wrong conclusions.
- sometimes respondents may misunderstand the questions asked and give wrong feedback leading to skewed results.

B- Application

- Likert scale's type is suitable with studies in social and behavioral sciences that have to do with perceptions, attitudes, emotions, opinions, personalities, and descriptions of people's environment.
- use to enhancement the reliability of results.[28]

2.8 Performance Metric for Classification Algorithms

The efficiency or performance of a given classification algorithm can be assessed through the use of a variety of benchmarks, that are also referred to as metrics. Some of such metrics include Accuracy, Precision, F1-score, and Recall. The aforementioned metrics are calculated based on the computing confusion matrix, which is a matrix that is used in deriving the summary of the number of examples that have been correctly or wrongly predicted by a classification model as highlighted in Table (2.1) below. A description of the key values of this table is given below:

Table (2.1): Confusion matrix for classification [80]

		Predicted Class	
		Positive +	Negative -
Actual Class	Positive +	*f*++ (TP)	*f*+- (FN)
	Negative -	*f*-+ (FP)	*f*- - (TN)

2.8.1 Evaluation Metrics

With the use of the evaluation metrics, the performance of the classifier can be estimated. The use of evaluation metrics is employed at the evaluation step, where the input is the expected class labels from the categorization stage as well as the real related class labels. The estimation of the classifier's performance can be done based on the results obtained from the comparison between the expected class and the actual

class labels [80]. Before the classification metrics are defined, it is important to understand the meaning of the following abbreviations: TP, FP, FN, and TN.

1- **True positive (TP):** refers to the positive cases that are correctly classified by the classifier.

2- **False Negative (FN):** represents the positive cases that are wrongly classified by the classifier.

3- **False Positive (FP):** refers to the incorrect classifications of negative instances by the classifier.

4- **True negative (TN)**: denotes negative instances that are correctly classified [44].

In this situation, the fundamental statistical evaluation that can be used include recall, accuracy, F-measure, and precision.

A- Precision: this measures the correctness of the model as well as the positive predicted value. If the model achieves a higher precision score, then it means that the rate of false positive is lower as shown in Equation (2.5) below [80].

The precision of class referred to as (P), is calculated using the following equation:

$$\boldsymbol{P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \dots\dots\dots (2.5)$$

B- Recall: this is the metric used in evaluating the model in terms of sensitivity, and it measures the positive cases with correct categorization.

A higher recall score means a lower number of positive instances that were wrongly categorized as negative as presented in Equation (2.6) [80]

The recall of class referred to as (R), is calculated using the following equation:

$$\boldsymbol{R} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \dots\dots\dots (2.6)$$

C- F1-Score: This matrix refers to the weighted average of the recall and precision measures, and is usually used in determining the accuracy of imbalanced datasets.

A high F1 value means that the overall performance of the system is high. Equation (2.7) shows how F1 can be calculated [80].

The F1-measure of a class referred to as (F1), is calculated as follows:

$$\boldsymbol{F1 = 2 \times \frac{precision \times recall}{precision + recall}} \ldots\ldots\ldots (2.7)$$

D- Accuracy: this measure is used to determine the number of correct predictions which is divided by the total number of predictions.

Equation (2.8) can be used to compute the accuracy of a given system [80].

$$\boldsymbol{A = \frac{TP+TN}{TP+TN+FP+FN}} \ldots\ldots\ldots (2.8)$$

2.8.2 Evaluation of Software Quality

Normally, the fitness-of-purpose defines a product of a high quality or superior performance. This means that such a product meets the needs of users. In order to carry out quality software evaluation and to determine quality assurance, it is important to have models that provide descriptions of the quality of the software, and it can link with the process of development. There are two ways to ensure the quality of software. The first one focuses on a direct specification and evaluation of a software product. This means that specific metrics are used in evaluating the performance of the software.

The second approach focuses on quality assurance of the process of software development.

For each characteristic, a set of attributes that can be measured is determined. This way, the quality of software can be evaluated, no clue is provided to show the manner in which a high-quality software can be constructed. [14]

The aim of using these requirements in the process of software development is to ensure that customers are satisfied with the software product because the satisfaction

of the customers is regarded as the major index of measuring the quality of a software product.

Software engineering basically seeks to find approaches and techniques that can be used to develop software products of high quality at an affordable cost. The increased usage of computers and related applications in the industry necessitates the development of high-quality software that can facilitate business success as well as the safety of humans.

there is a good base for understanding software quality. can combine the quality aspects with other relevant factors to get a holistic view of software quality.

The software quality model defines four important indicators of software quality:

- Reliability
- Performance efficiency
- Security
- Maintainability

Software quality measures whether the software satisfies its requirements. Software requirements are classified as either functional (specify what the software should do) or non-functional (like portability, privacy, security, supportability, and usability).

2.8.3 Mc Call Model

The overall goal of any software management is "Quality built-in with cost and performance as prime consideration". This means that the software should be built with certain quality aspects that fulfill the needs of the user. Its performance is kept on the top priority. The performance is also based on the demands of the user and the developer's perspective.

Given the intangible and abstract nature of software, researchers and practitioners are bound to find ways to characterize software in order to make benefits and costs visible. Jim McCall produced the McCall software quality model for the US Air Force in 1977.

This is used to maintain harmony between the users and the developers. Successful software is developed that fulfills the user needs in consideration with the developer's point of view.

The Mc Call model established product quality through several features that have been categorized as follows:

Product Review (flexibility, maintenance, and testing),

Product Transition (portability, reusability, and interoperability), and

Product Operation (reliability, correctness, usability, and integrity).

The major contribution of the McCall method is to determine the correlation between metrics and quality. The use of the model is employed as a foundation for the creation of other quality models [84].

In other words, other quality models are constructed based on the model.

The Mc Call model is presented in Figure (2.13).

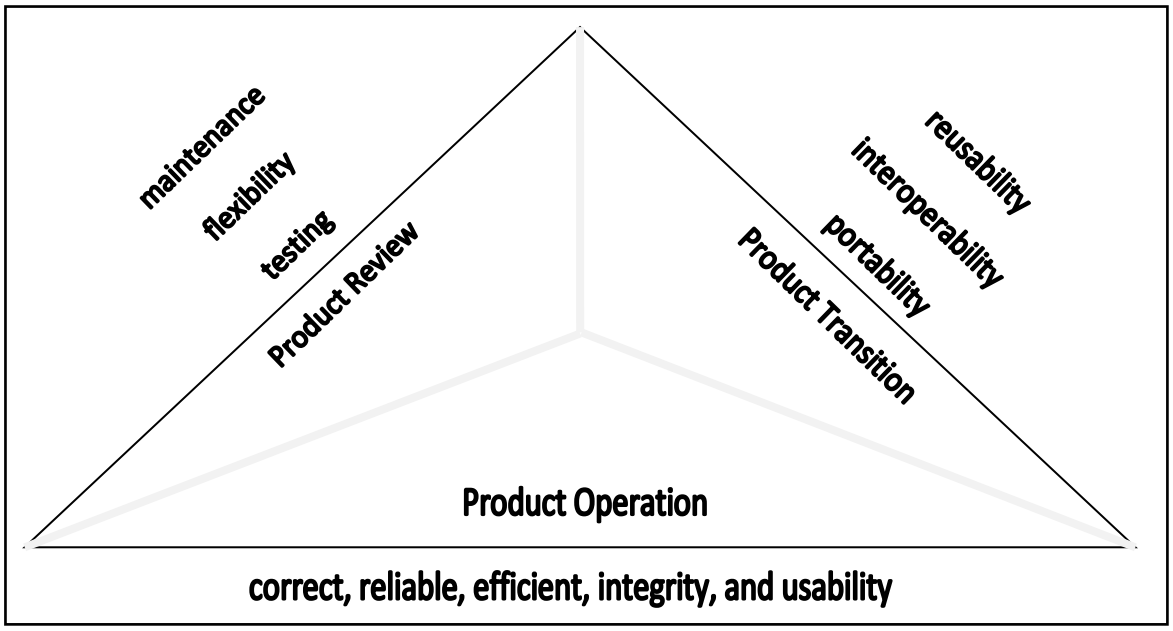


Figure (2.13): Mc Call Model [84]

CHAPTER THREE

PROPOSED SYSTEM AND IMPLEMENTAYION

CHAPTER THREE

PROPOSED SYSYTEM AND IMPLEMENTATION

3.1 Introduction

In this Chapter, the main steps of the practical part of this thesis are described. It first presents the architecture of the proposed system.

those contents two parts where it is the first create new set by collect data for crowdsourcing this data set and the secondly got data set used thesis that work in the same idea.

Then, the data collection procedure is discussed. This is followed by explaining the algorithms used, and it explains all implementation steps of this project procedure working in this thesis.

3.2 Software and Hardware

The system is implemented by using

- python language (IDLE 3.6).
- windows11, 64-bit Operating System.
- Google forms
- The location from Iraqi National Card Website: https://reg.nid-moi.gov.iq
- Microsoft Excel worksheet
- CPU Intel(R) Core (TM) i7-10510U CPU @ 1.80GHz 2.30 GHz.
- Memory: 8GB. RAM
- Hard 1 TiB

These specifications are more appropriate to implement the system.

3.3 Proposed System Architectural

1. presents the architecture and the practical side of the proposed system with use dataset (1). Figure (3.1) shows the whole stages of the proposed system.

2. presents the architecture and the practical part of the proposed system that use dataset (2). Figure (3.2) shows the whole stages of the proposed system.
3. initially, the first stage is collecting data the first dataset that was created by building on crowdsourcing feedback about an Iraqi National Card Website. while the second dataset that belongs to another thesis in the same domain.
4. The second stage is data will be pre-processing, which includes two sub-steps that use data transformation forms suitable for the mining process.

   This use normalization is done in order to scale the data values in a specified range. This is to obtain the necessary data that would be used in the proposed system.
5. The third stage is data mining which encompasses many processes to prepare the proposed system that contains a crowdsourcing dataset. where data will be split into training and testing datasets respectively in which training data utilized for building the model and its effectiveness will be checked on test data.

   with explain the algorithms used, and all implementation stages of the project procedure that used in this thesis.
6. The last stage, fourth is the Proposed System shows the result for using each algorithm in metric and explains the software quality factor.

work was done on two datasets where it is the firstly got dataset from the used thesis that work in the same idea. and secondly new collect data and build a new dataset from crowdsourcing.

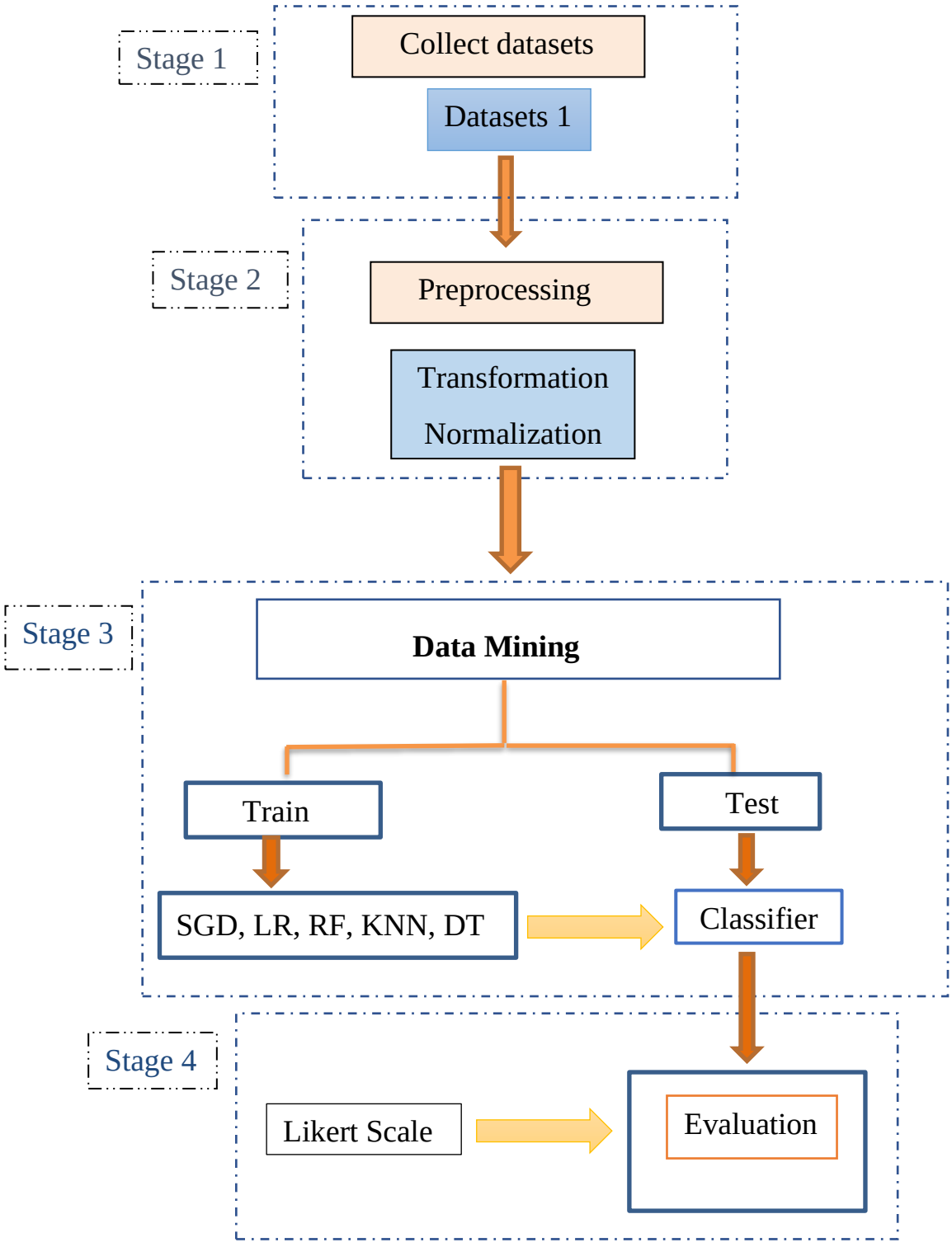


Figure (3.1): Proposed System Architecture (Datasets 1)

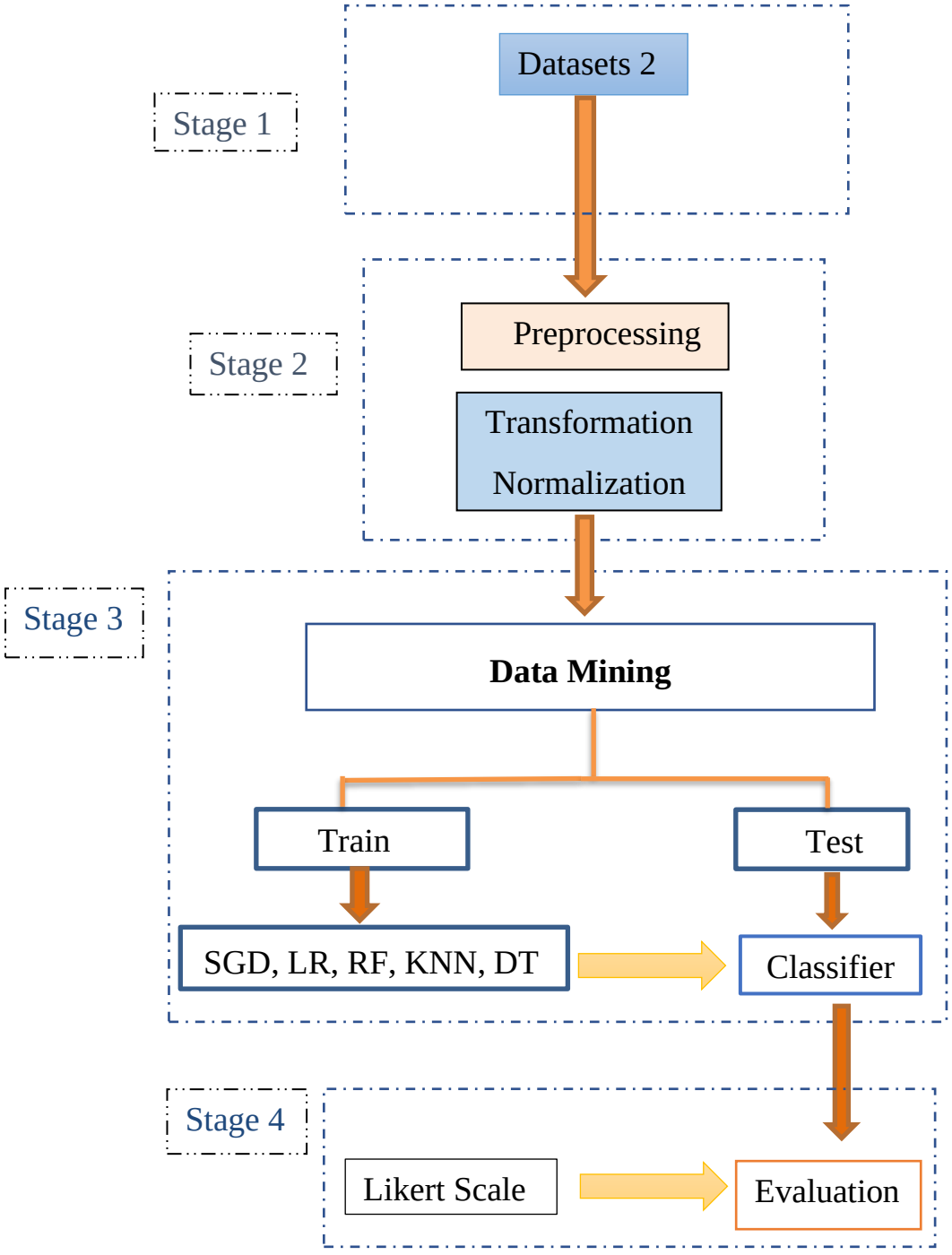


Figure (3.2): Proposed System Architecture (Datasets 2)

3.4 Propose Collect Data and Build a dataset (1)

1- The first stage Data requirements establish the process used to identify, prioritize, precisely formulate, and validate the data needed to achieve business objectives. The data requirements document is prepared a data collection effort by the user group is required to generate and maintain system data.
2- To gather data from the target group, 100% data gathering need access to the entire target group and getting information directly.
3- The second stage in the data acquisition process is the collection process of data.
4- The data are forms questionnaires they consist of a set of questions Conversion of data to other patterns introduces a certain amount of error that must be controlled. sources of data may introduce some degree of error in their collection processes.
5- The final stage, third Once the collected data is in electronic form, some "processing" is usually necessary to mitigate obvious errors, and some analysis is usually necessary to convert data into useful information.
6- data editing is the application of checks that identify missing, invalid, duplicate, inconsistent entries. Editing is a final inspection-correction method. Untreated, missing data can introduce serious errors into estimates.
7- It is data quality is better achieved through clarity of definitions, forms design, data collection procedures. while Coding is the process of adding codes to the data set as additional information or converting existing information into a more useful form.

3.5 Data Collection

Data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. performing research for academic purposes, data collection allows

you to gain first-hand knowledge and original insights into a research problem to make assumptions

about future probabilities and trends. Figure (3.3) shows Propose Collect data and build questions by google forms to get the answers for questions.

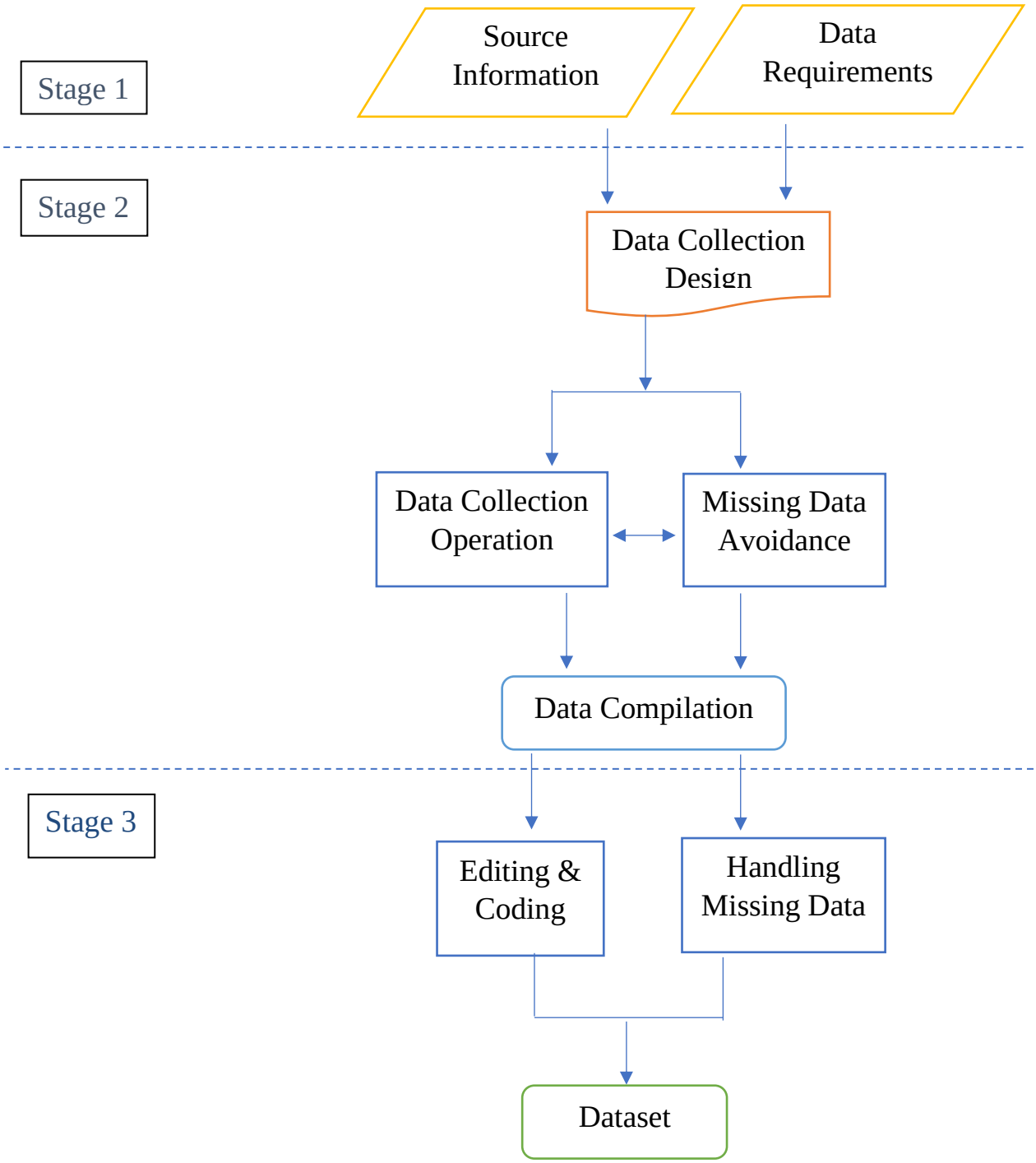


Figure (3.3): Propose Collect Data and Build a Dataset (1)

A- First Case Study

- Using crowdsourcing which is done in Iraq/Baghdad. From the Ministry of Interior / Directorate of Civil Status. Passport and Residence/National Card.
- all answers to the questions were at five levels, starting with a score of (0) to (4) degree which is a (0) strongly agree, (1) agree, (2) neutral, (3) disagree, and ending with a (4) strongly disagree.
- Three questions whose answers were definite (Academic achievement, Marital status, work nature)
- People who were asked the question (10001) to (18) questions independently so that their total number is (180018) decisions in the below explaining the question.
- The questions directed to the crowd are derived from the idea of Mc Call. (See in section 2.8.3) These questions have been grouped into three points of view:
  - PRODUCT OPERATION
    - Correctness [1-The content of the site is clear?]
    - Efficiency [2- Academic achievement? 3- Marital status? 4-work nature?]
    - Integrity [5- Is the site fast loading and uploading? 6-The site is protected?]
    - Reliability [7- Reliability on site?]
    - Usability [8- Create other websites? 9- Save time and effort? 10- Easy access to the site?]
  - PRODUCT REVISION
    - Maintainability [11- Did you like the design of the site?]
    - Flexibility [12- Is the website easy to use? 13- It is easy to get the work done?]
    - Testability [14- Are you satisfied with the site? 15- You need help?]
  - PRODUCT TRANSITION
    - Portability [16- User of websites?]
    - Reusability [17- Accept the questionnaire in the future?]
    - Interoperability [18- It is recommended to use it?]

B- Second Case Study

- Data collected, used, and/or analyzed for the Business Informatics Master Thesis "Applying the Requirements Engineering for Software Architecture model in software products: a case study working crowdsourcing," conducted at Utrecht University in the Netherlands.
- Each User Story (n = 56) has been tested against eight criteria.
- Find it at the following location dataset from the website: https://data.mendeley.com/datasets/7r9j67wxzb/1.
- That publish in Year: 2019 that DOI identifier: 10.17632/7r9j67wxzb.1.
- Introducing the company where we perform the case study: Tournify. This Dutch startup provides an online tournament for sports organizers, delivered as a service.
- Each User Story is evaluated on its quality manually by three experts individually. The experts use the description of the criteria to analyze the User Stories. The eight criteria and descriptions Are shown in Table (3.1).[63]

Table (3.1): The Eight Criteria to Used Assess User Stories [63]

Criteria	Description
Well-formed	A user story includes at least a means and a role
Atomic	A user story expresses a demand for exactly one feature
Minimal	A user story contains nothing more than means, role, and ends
Conceptually sound	The means express a feature, and the ends express a rationale
Problem-oriented	A user story only specifies **the** problem has not the solution
Unambiguous	A user story avoids terms or abstractions that leading to multiple interpretations
Full-sentence	A user story is a well-formed full sentence
Estimable	A story does not indicate a coarse-grained requirement that is difficult to plan and prioritize

3.6 Quality of the Crowdsourced User Stories

Inter-rater reliability assesses the level of agreement between independent raters on some sort of performance or outcome. [63] With inter-rater reliability, there must be a standardized and objective operational definition by which performance is assessed across the spectrum of "agreement" ratings can be made at a categorical (yes / no), ordinal (Likert-type scale), The number of ratings taken and the number of independent raters also play a significant role in choosing the correct test for inter-rater reliability.[87]

[Likert-type scale is a rating scale used to assess opinions, attitudes, or behaviors. Likert scales are popular in survey research because they allow you to easily operationalize personality traits or perceptions. To collect data, you present participants with Likert-type questions or statements and a continuum of possible responses. Each question is given a numerical score so that the data can be analyzed quantitatively].[86]

Tests for inter-rater reliability show that the average pairwise percent agreement between the three judgments varies from (65.5% to 91.7%) for each criterion.

The results of the analysis are shown in Table (3.2).[63]

Table (3.2): Results of The Analysis [63]

Criteria	Average pairwise
Unambiguous	65.5
Problem-oriented	73.8
Minimal	79.8
Conceptually	79.8
Full-sentence	81.0
Atomic	84.5
Well-formed	90.5
Estimable	91.7

3.7 Data Pre-Processing

The first stage of the proposed system is pre-processing. a data mining technique that involves transforming raw data into an understandable format. that is a very important step which is done for a better-quality input the established steps need to go through to make sure data is successfully preprocessed.
to use data transformation, with data cleaning already begun to modify data, but data transformation will begin the process of turning the data into the proper format need for analysis and other downstream processes that performing by
Normalization is a process that ensures that no data is redundant, it is all stored in a single place, and all the dependencies are logical. It is done in order to scale the data values in a specified range (1 or 0).

3.8 Classification

It is a significant stage; the main goal of the classification algorithm is to identify the category of a given dataset; these algorithms are mainly used to predict the output for the data.

In the present study, five approaches have been utilized for classifying the data, The following algorithms are used for data and evaluation.
which are

1- Decision Tree (DT)

Step-1: Begin the tree with the root node, says S, contains the complete dataset.
Step-2: Find the best attribute in the dataset.
Step-3: Divide the S into subsets that contains possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

2- Random Forest (RF)

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

3- K-Nearest Neighbor (KNN)

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of (**x**, **y**) number of neighbors.

Euclidean (**x**, **y**) = $\sqrt{\sum}$ (**xi - yi**)**2**

Step-3: Take the (**x**, **y**) nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these (**x**, **y**) neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

4- Logistic Regression (LR)

Step-1: For every attribute value, make rule

Step-2: Count the appearance of the class

Step-3: Assign class to attribute value

Step-4: Compute error rate for each rule: $g(z) = 1/(1+e\text{-}z)$

Step-5: Select the minimum error rule.

5- Stochastic Gradient Descent (SGD)

Step-1: The first vector of parameters is randomly selected.

Step-2: Continue the phases below till a minimum is reached approximately.

Step-3: Examples are mixed in the training set randomly.

Step-4: For i = 1 to n.

Step-4: Wt + 1= wt - at. ∇f it (wt)

Step-5: Here, Q (w) refers to the experimental risk while Qi (w) refers to the value of the loss function.

3.9 Likert scale

As the Likert scale is categorical data and has labelled, at the beginning, you structure dataset, marketing, and build a model of ML. The Likert scale is a multiclass classification. turned dataset into five classes (0,1, 2, 3, 4).

CHAPTER FOUR
RESULT AND DISCUSSION

CHAPTER FOUR
RESUILT AND DISCUSSION

4.1 Introduction

The presented chapter presents the implementation related to the suggested system on a set of data for the purpose of testing the effectiveness related to the system and evaluating the degree that the required aims.

4.2 System Specifications

Hardware, as well as software tools, were the main requirements for the suggested system. Accordingly, sub-section 3.2. was provided software specifications, while providing hardware specifications.

4.3 Results of Data Collection

About ten thousand crowdsourcing were gathered for about a month. These crowdsourcing were collected in google form format and saved in an excel sheet where the output of this stage was the input to the next stage, namely pre-processing. while another data set was ready on the other thesis in section (3.4.1 - B).

4.4 Results of Classification

Five classification algorithms were used in the proposed system. The first algorithm was a Logistic regression the second was Stochastic gradient descent, the third Decision tree, the four Random Forest classifier, and finally K-nearest neighbour. The five algorithms that have calculated Accuracy, Precision, Recall, and F1- measure shows the result case study for the two data sets.

4.4.1 First Case Study

A - The proposed system has experimented on the total number of the crowdsourcing was (10003) which were divided into (7002) as a training set (70%) and (3001) as a testing set (30%).

These data were used to construct our classification where the model's accuracy was enhanced after increasing the number of training data.

Five classification algorithms were used in the proposed system. These algorithms are used the same inputs that are used in the system. Table (4.1) and figure (4.1) show the performance measures of a system and explain the outcomes of the algorithms for the dataset inputs that are used.

The five algorithms have calculated Precision, Recall, F1-measure and Accuracy show the result case study for the first data set.

Table (4.1): Classification Result

	Precision	Recall	F1-score	Accuracy
LG	1.00	1.00	1.00	1.00
SGD	0.95	0.78	0.82	0.78
DT	1.00	1.00	1.00	1.00
RF	1.00	1.00	1.00	1.00
KNN	0.85	0.85	0.85	0.85

The results accuracy of the classified data set was the high result percentage (100%) by using logistic regression, decision tree, and random forest classifier. because the primary work of these algorithms is a classification model, prediction, and tree structure to find a best-fitting relationship between the dependent variable and a set of independent variables.

while other algorithms the Stochastic gradient descent is a result percentage (78%) because this algorithm calculates the derivative from each training data instance and calculating the update immediately. And is sensitive to feature scaling.

finally, K-Nearest Neighbor algorithms were low, percentage (85%). because Classification is computed from a simple majority vote of the nearest neighbors of each point.

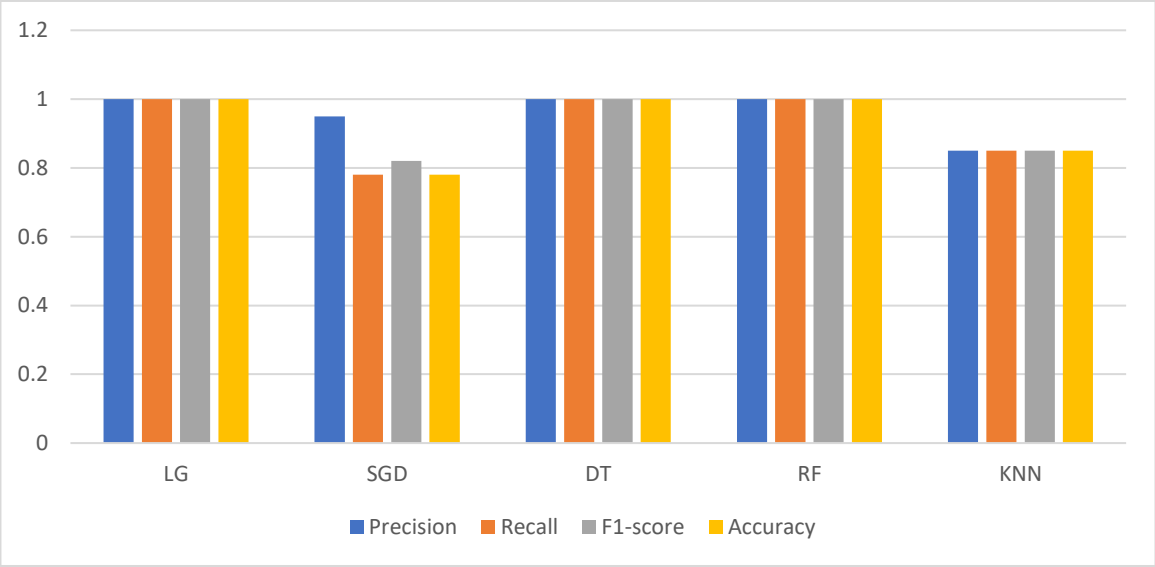


Figure (4.1) Flowchart Classification Result

B- on the second hand used the Likert scale used the same inputs used in the system's algorithm. Table (4.2) and Figure (4.2) show the results of the data set inputs used. By calculating the results of the questions and indicating the extent of high acceptability and rejection and in between.

that collect agrees the highest percentage (52.62%) and collects disagree itis (31.57%) while lowest percentage itis natural percentage (15.78).

This means that there is acceptance of the system by crowdsourcing and support of such systems in the future.

Table (4.2): Likert Scale Result

strongly agree	agree	neutral	disagree	strongly disagree
42.10	**10.52**	**15.78**	**26.31**	**5.26**

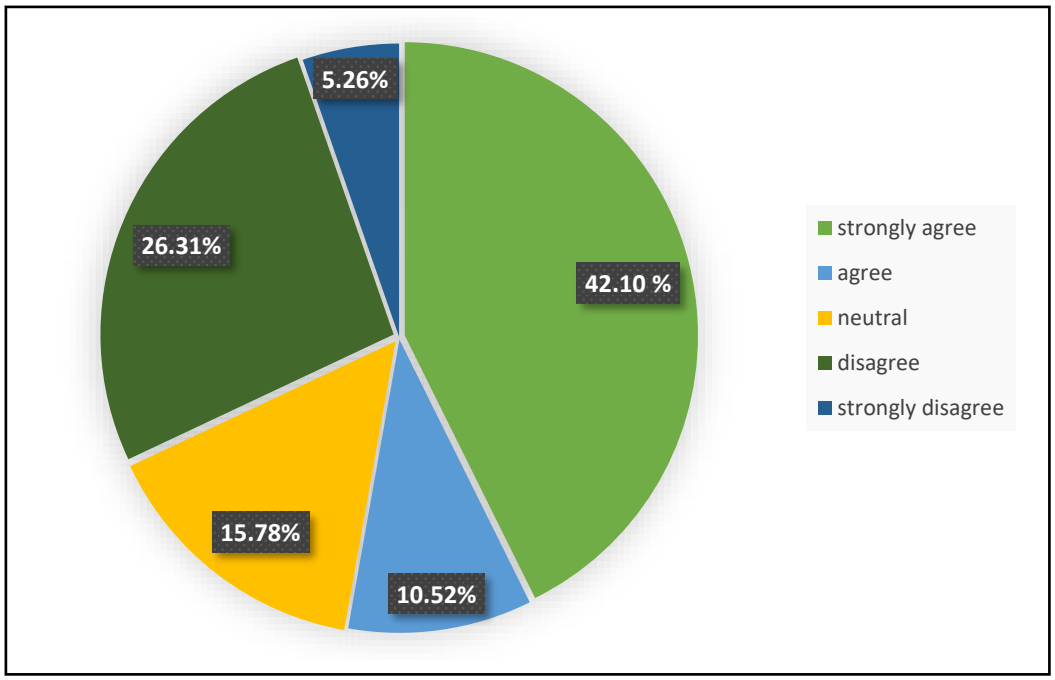


Figure (4.2) Flowchart Likert Total

To get more results from using crowdsource was used Likert Scale with the Mc Call model, as the details of the results are in follow Table (4.3) and figure (4.3).

Table (4.3): Likert Scale Likert Mc Call Model Result

Features	strongly agree	agree	neutral	disagree	strongly disagree
Operation	50.0	16.6	33.3	0	0
Revision	80.0	20.0	0	0	0
Transition	80.0	20.0	0	0	0

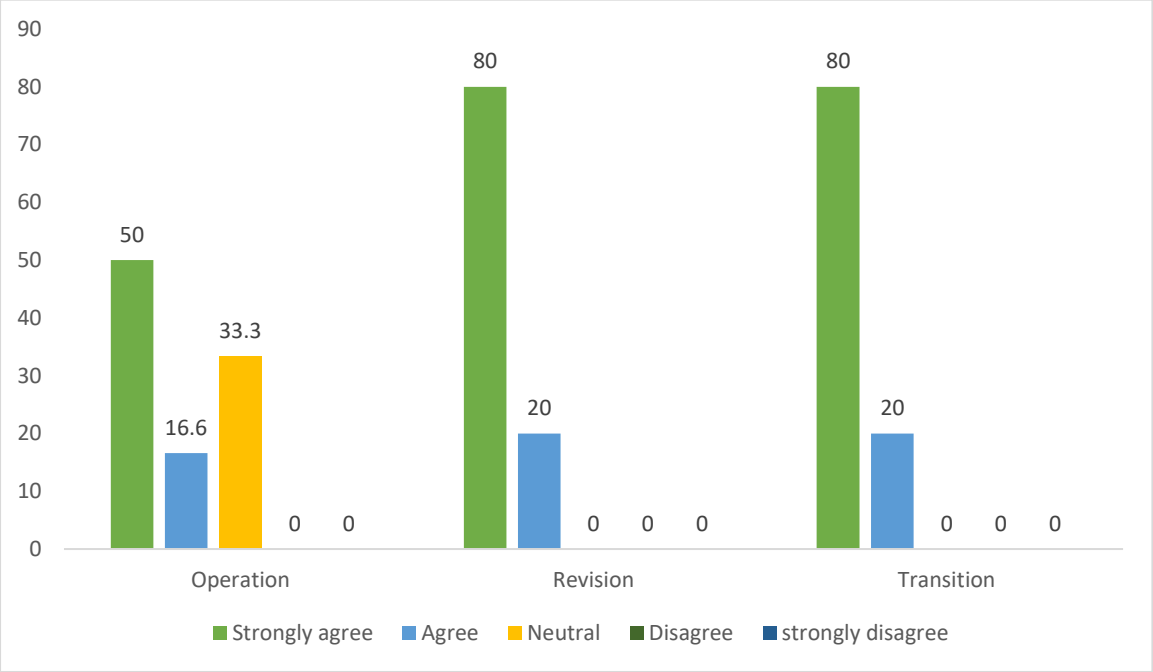


Figure (4.3) Flowchart Likert Mc Call Model Result

Through the table and the chart, we note that crowdsourcing has good acceptability in product operation with a percentage of (66%), while product revision and product transition had a high acceptability rate (100%) and we note the full support for the system and the absence of any percentage of refusal all feature when dealing with a Mac Call model (See in section 2.8.3).

This means that the system needs some revisions for the purpose of making improvements to the system, especially in product operation.

4.4.2 Second Case Study

A- The proposed system was a worker on the total number of the data was (56) which were divided into (39) as a training set (70%) and (16) as a testing set (30%).

The same five classification algorithms were used in the proposed system.

These algorithms are used the same inputs that are used in the system. Table (4.3) and figure (4.3) show the performance measures of a system and explain the outcomes of the algorithms for the dataset inputs that are used.

The five algorithms that have studied precision, Recall, F1-measure, and Accuracy show the result case study for the second data set.

Table (4.4): Classification Result

	Precision	Recall	F1-score	Accuracy
LG	0.93	0.92	0.92	0.92
SGD	0.86	0.62	0.64	0.62
DT	0.93	0.92	0.92	0.92
RF	0.86	0.77	0.77	0.77
KNN	0.93	0.92	0.92	0.92

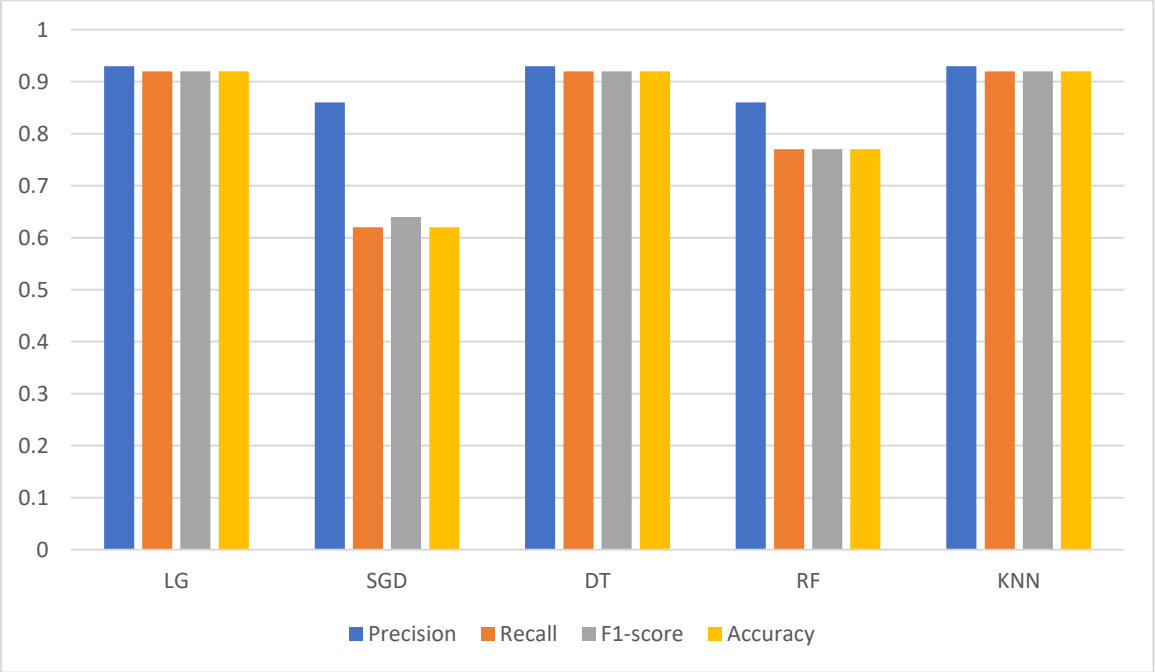


Figure (4.4): Flowchart Classification Result

The results accuracy of the classified data set was the high result percentage (92%) by using logistic regression, decision tree, and K-nearest neighbour. because the primary work of these algorithms is a classification model, predict, and the outcome is measured with a dichotomous variable, rules are learned sequentially using the training data one at a time, it works on storing instances of training data.

while other algorithms the random forest classifier is a result percentage (77%) because the outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees.

finally, Stochastic gradient descent was low, percentage (65%). Because this algorithm works continuously and non-stops when updating from a certain point.

the algorithms results were best the result from the Likert scale in the same dataset used the same thesis.

B- on the second method used the Likert scale used the same inputs used in the system's algorithm. Table (4.4) and Figure (4.4) show the results of the data set inputs used in the thesis.

The output results were calculated based on the output used in another thesis.

the highest Unambiguous percentage is (0.94%) and Minimal itis the lowest percentage itis (0.57%).

In general, the results were higher than the results of the other thesis because use the Likert scale in a new idea and a different method from another thesis.

were the extracted results it is following Unambiguous, Estimable (0.94), Problem-oriented Conceptually (0.91), Full-sentence (0.85), Atomic (0.83), and Well-formed (0.66), finally Minimal (0.57).

Table (4.5): Likert Scale Result

Criteria	Average Pairwise
Unambiguous	0.94
Estimable	0.94
Problem-oriented	0.91
Conceptually	0.91
Full-sentence	0.85
Atomic	0.83
Well-formed	0.66
Minimal	0.57

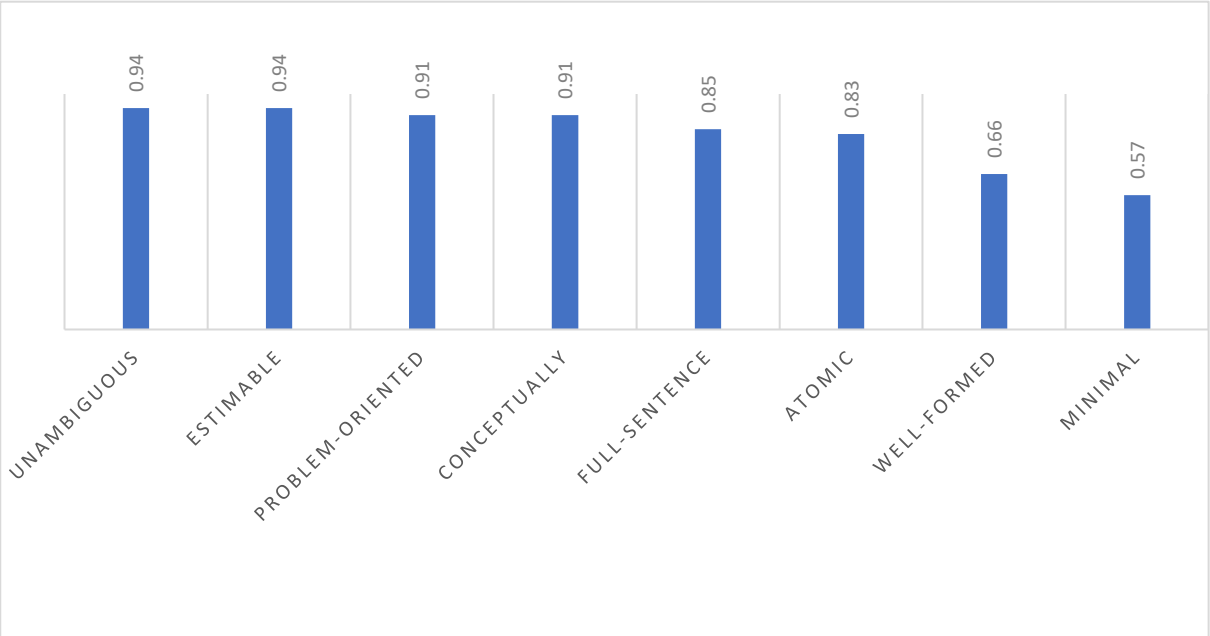


Figure (4.5) Flowchart Likert Scale Result

4.5 Summarization

This chapter included experimental environment, practical explanation of data pre-processing and feature extraction as well as results of all algorithms in the proposed system (the first Logistic regression, the second was Stochastic gradient descent, the third Decision tree, the four Random Forest classifier, and finally the fifth K-nearest neighbour) and used the Likert scale of the two datasets.

CHAPTER FIVE
CONCLUSION AND FUTURE WORKS

CHAPTER FIVE

CONCLUSION AND FUTURE WORKS

5.1 Thesis Summary

The data was collected for a period of nearly a month in 2021 by crowdsourcing contribution. and preprocessing techniques were applied on the two datasets before passing the research data to the proposed classification model.

by testing, all algorithms showed Decision Tree (DT), and Logistic regression (LR) Classifiers are found to be the best algorithms used. In addition to using the Likert scale to find respondents to choose from a linear set of responses that increase or decrease in intensity or strength.

5.2 Conclusions

The machine learning community has noticed the great opportunities offered by crowdsourcing and has developed a large number of techniques to deal with the problems of inaccuracy, randomness, and uncertainty when learning with crowdsourcing. By focusing on two primary issues: data quality and forecast model quality. And to set proposing model to support to be use crowdsource

Crowdsourcing is a good idea that should be practiced with great care. One must take into account the experience of those he uses to ensure that he actually gets what he needs at the end of the project.

In this thesis, we built a crowdsourced dataset based on the Mc Call model, data mining was used to divide the data into two groups testing and training through the use of algorithms and metrics different to get the best result from the algorithms used. while this thesis aim at the strengths and weaknesses of the system were identified and the improvements and development it needed in a particular aspect. Through crowdsourcing and the various questions directed to them.

5.3 Future Work

The results from this thesis can function as the foundation of future work.

1- The focus of the future work can be laid upon further evolving the used methods.

2- The method could be potentially improved, be it an improved training method or the application of more strict quality controls.

3- Further researches could affect the quality of the results either positively or negatively.

4- Such selective aggregation of the results could result in new insights while reusing the same data gathered during the tests of research.

5- More extensive analysis could be applied to the generated results from this project to better explore the performance of the method.

6- Further research into the performance of the different types of contributors could also be conducted, either focused on their opinions and ideas.

REFERENCES

References

[1] J. Howe, crowdsourcing why the power of the crowd is driving the future of business, June 2006.

[2] "The rise of crowdsourcing," Wired magazine, vol. 14, no. 6, pp. 1–4, 2006

[3] K. R. Lakhani, D. A. Garvin, and E. Lonstein, "TopCoder(A): Developing software through crowdsourcing," Harvard Business School Case, 610-032, January 2010

[4] K.-J. Stol and B. Fitzgerald, "Two's company, three's a crowd: A case study of crowdsourcing software development," in Proceedings of the 36th International Conference on Software Engineering, 2014, pp. 187–198.

[5] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, ´ et al., "Predicting protein structures with a multiplayer online game," Nature, vol. 466, no. 7307, pp. 756–760, 2010.

[6] T. C. Norman, C. Bountra, A. M. Edwards, K. R. Yamamoto, and S. H. Friend, "Leveraging crowdsourcing to facilitate the discovery of new medicines," Science Translational Medicine, vol. 3, no. 88mr1, 2011.

[7] D. C. Brabham, T. W. Sanchez, and K. Bartholomew, "Crowdsourcing public participation in transit planning: preliminary results from the next stop design case," Transportation Research Board, 2009.

[8] A. T. Chatfield and U. Brajawidagda, "Crowdsourcing hazardous weather reports from citizens via twittersphere under the short warning lead times of EF5 intensity tornado conditions," in Proceedings of the 47th Hawaii International Conference on System Sciences.IEEE, 2014, pp. 2231–2241.

[9] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," in ACM SigIR Forum, vol. 42, no. 2. ACM, 2008, pp. 9–15.

[10] Ankur Josh and Dinesh Kumar Pal, Likert Scale: Explored and Explained Article in Current Journal of Applied Science and Technology · January 2015 DOI: 10.9734/BJAST/2015/14975

[11] S. Zogaj, U. Bretschneider, and J. M. Leimeister, "Managing crowdsourced software testing: A case study-based insight on the challenges of a crowdsourcing intermediary," Journal of Business Economics, vol. 84, no. 3, pp. 375–405, 2014.

[12] B. Kogut and A. Metiu, "Open-source software development and distributed innovation," Oxford Review of Economic Policy, vol. 17, no. 2, pp. 248–264, 2001.

[13] Introduction to data mining, VI PIN KUMAR University of Minnesota, and Army High Performance Computing Research Center.

[14] lease, M. 2011. On quality control and machine learning in crowdsourcing. In The 3rd Human Computation Workshop (HCOMP) at AAAI, 97–102

[15] J. M. Hughes, "Systems and methods for software development," August 2010, US Patent 7778866 B2.

[16] D. C. Brabham, "Crowdsourcing as a model for problem solving an introduction and cases," Convergence: the international journal of research into new media technologies, vol. 14, no. 1, pp. 75–90, 2008.

[17] E. Estelles-Arolas and F. Gonz ´ alez-Ladr ´ on-De-Guevara, "Towards an integrated crowdsourcing definition," ´ Journal of Information Science, vol. 38, no. 2, pp. 189–200, Apr. 2012.

[18] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popovic, ´ et al., "Crystal structure of a monomeric retroviral protease solved by protein folding game players," Nature Structural and Molecular Biology, vol. 18, no. 10, pp. 1175–1177, 2011.

[19] R. Johnson, "Natural products: Crowdsourcing drug discovery," Nature chemistry, vol. 6, no. 2, pp. 87–87, 2014.

[20] C. Muller, L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh, "Crowdsourcing for climate and atmospheric sciences: current status and future potential," International Journal of Climatology, 2015.

[21] T. D. Breaux and F. Schaub, "Scaling requirements extraction to the crowd: Experiments with privacy policies," in Proceedings of the 22nd IEEE International Requirements Engineering Conference, Aug. 2014, pp. 163–172

[22] K. T. Stolee and S. Elbaum, "Exploring the use of crowdsourcing to support empirical studies in software engineering," in Proceedings of the 4th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 2010, pp. 1–4.

[23] Y. Usui and S. Morisaki, "An Approach for Crowdsourcing Software Development," Proceedings of the Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement, pp.32–33, 2011.

[24] R. Prikladnicki, L. Machado, E. Carmel, and C. R. B. de Souza, "Brazil software crowdsourcing: A first step in a multi-year study," in Proceedings of the 1st International Workshop on CrowdSourcing in Software Engineering, June 2014, pp. 1–4.

[25] D. C. Brabham, “Crowdsourcing as a model for problem solving an introduction and cases,” Convergence: the international journal of research into new media technologies, vol. 14, no. 1, pp. 75–90, 2008

[26] Michael R. Hyman, NMSU Jeremy J. Sierra, Texas State University, Likert Scales: Design Issues May 2010

[27] D. Sobel, Longitude: The true story of a lone genius who solved the greatest scientific problem of his time. Macmillan, 2005.

[28] Adetayo Olaniyi Adeniran Federal University of Technology, Application of Likert Scale Type and Cronbach's Alpha Analysis in an Airport Perception Study. Published Date: 22 March 2019

[29] Lam M. Nguyen, New Convergence Aspects of Stochastic Gradient Algorithms, Journal of Machine Learning Research 20 (2019) 1-49.

[30] M. Lease and E. Yilmaz, “Crowdsourcing for information retrieval,” in ACM SIGIR Forum, vol. 45, no. 2. ACM, 2012, pp. 66–75

[31] T. D. Breaux and F. Schaub, “Scaling requirements extraction to the crowd: Experiments with privacy policies,” in Proceedings of the 22nd IEEE International Requirements Engineering Conference, Aug. 2014, pp. 163–172

[32] K. T. Stolee and S. Elbaum, “Exploring the use of crowdsourcing to support empirical studies in software engineering,” in Proceedings of the 4th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 2010, pp. 1–4.

[33] Pavan kumar Bengaluru Area, India, The power of crowd sourcing SlideShare Apr. 17, 2013.

[34] Y. Usui and S. Morisaki, “An Approach for Crowdsourcing Software Development,” Proceedings of the Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement, pp.32–33, 2011.

[35] R. Prikladnicki, L. Machado, E. Carmel, and C. R. B. de Souza, “Brazil software crowdsourcing: A first step in a multi-year study,” in Proceedings of the 1st International Workshop on CrowdSourcing in Software Engineering, June 2014, pp. 1–4

[36] M. N. Huhns, W. Li, and W.-T. Tsai, “Cloud-based software crowdsourcing (Dagstuhl seminar 13362),” Dagstuhl Reports, vol. 3, no. 9, pp. 34–58, 2013.

[37] Jefrey P.Bigham Carnegie Mellon University MICHAEL S. BERNSTEIN Stanford University and EYTAN ADAR University of Michigan. Human-Computer Interaction and Collective Intelligence 2015.

[38] H. Xue, “Using redundancy to improve security and testing,” Ph.D. dissertation, the University of Illinois at Urbana-Champaign, 2013.

[39] J. Lin, “Understanding and capturing people’s mobile app privacy preferences,” Ph.D. dissertation, Carnegie Mellon University, 2013.

[40] R. Snijders, “Crowd-centric requirements engineering: A method based on crowdsourcing and gamification,” Master’s thesis, Utrecht University, 2015.

[41] Ke Mao↤, Licia Capra, Mark Harman, Yue Jia. A Survey of the Use of Crowdsourcing in Software Engineering. Department of Computer Science, University College London, Malet Place, London, WC1E 6BT, UK

[42] Ethem Alpaydın, "Introduction to Machine Learning", the MIT Press Cambridge, Massachusetts London, England, 2010.

[43] N. Dutta, S. Umashankar, V. K. A. Shankar, S. Padmanaban, Z. Leonowicz, and P. Wheeler, “Centrifugal Pump Cavitation Detection Using Machine Learning Algorithm Technique,” in 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2018, pp. 1–6.

[44] S. Menaka and N. Radha, “Text classification using keyword extraction technique,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 12, 2013.

[45] M. Fan, “Term Paper: How and Why to Use Stochastic Gradient Descent?”

[46] A. Dey, “Machine learning algorithms: a review,” Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 1174–1179, 2016.

[47] F S. Vijayarani, M. J. Ilamathi, and M. Nithya, “Preprocessing techniques for text mining-an overview,” Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2015.

[48] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” Synth. Lect. Artif. Intell. Mach. Learn., vol. 3, no. 1, pp. 1–130, 2009.

[49] Vukovic, M. (2009). Crowdsourcing for Enterprises Maja Vukovi. Proceedings of Congess on Services-I, 686-692.

[50] Latoza, T. D., Ben Towne, W., Adriano, C. M., & Van Der Hock, A. (2014). Microtask Programming: Building Software with a Crowd. User Interface Software and Technology Symposium, 43-54. 10.1145/2642918.2647349.

[51] Stol, K., & Fitzgerald, B. (2014). Two’s Company, Three’s a Crowd: A Case Study of Crowdsourcing Software Development. Proceedings of ICSE, 2014, 187–198.

[52] Huberman, B. A., Romero, D. M., & Wu, F. (2009). Crowdsourcing – Attention and Productivity. Information Science,35(6), 758–765. doi:10.1177/0165551509346786.

[53] Hoang Nguyen Hanoi University of Mining and Geology, Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach, August 2019.DOI: 10.1016/j.resourpol.2019.101474.

[54] Victor Sheng, Jing Zhang "Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions.July 2019. DOI:10.1609/aaai.v33i01.33019837

[55] Fangwen Yuan, Jun Liang, Zhaokun Xue Crowdsourcing: Today and Tomorrow An Interactive Qualifying Project Submitted to the Faculty of the Worcester polytechnic polytechnic insttute.

[56] Martijn van Vliet (4171934) (m.vanvliet@uu.nl) Department of Information and Computing Sciences, Utrecht University the Netherlands a Crowdsourcing Technique for the Requirements Elicitation from Online Reviews August 1, 2019

[57] The Knowledge Discovery Process Data Preparation & Preprocessing.ppt. Bamshad Mobasher DePaul University

[58] Stamatios-Aggelos n. Alexandropulos, Sotiris b.Kotsiantis and Michael n.Vrahtis , Data preprocessing in predictive data mining. The Knowledge Engineering Review, Vol. 34, e1, 1–33. © Cambridge University Press, 2019.

[59] Mao, K., Capra, L., Harman, M., Jia, Y. "A survey of the use of crowdsourcing in software engineering. Journal of Systems and Software", (126), pp. 57-84. 2017.

[60] Crowdsourcing for Software Engineering Article in IEEE Software · March 2017 DOI: 10.1109/MS.2017.52

[61] Xie, T., Bishop, J., Horspool, R.N., Tillmann, N., de Halleux, J., 2015. Crowdsourcing code and process via Code Hunt, in Proc. 2nd International Work- shop on CrowdSourcing in Software Engineering.

[62] K. Mao, L. Capra, M. Harman, and Y. Jia, "A survey of the use of crowdsourcing in software engineering," Journal of Systems and Software, vol. 126, pp. 57 – 84, 2017.

[63] Abel Menkveld "Applying the Requirements Engineering for Software Architecture Model in Software Products: a case study using crowdsourcing" Master Thesis. May 2019.

[64] N. Dutta, S. Umashankar, V. K. A. Shankar, S. Padmanaban, Z. Leonowicz, and P. Wheeler, "Centrifugal Pump Cavitation Detection Using Machine Learning Algorithm Technique," in 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2018, pp. 1–6.

[65] E. E. N. Vollset and O. E. E. Folkestad, "Automatic classification of bank transactions." NTNU, 2017.

[66] Zhaoan Dong · Jiaheng Lu · Tok Wang Ling · Ju Fan · Yueguo Chen Using Hybrid Algorithmic-Crowdsourcing Methods for Academic Knowledge Acquisition, 2017

[67] Luz, N., Silva, N., Novais, P.: A survey of task-oriented crowdsourcing. Artificial Intelligence Review pp. 1- 27 (2014). DOI 10.1007/s10462-014-9423-5

[68] Hani Barjes Salmeh Al-Bloush, Jun 2020 Guideline for Safeguarding intellectual Property Rights of Crowdsourced software engineering activities

[69] T. Soni Madhulatha, "An Overview on Clustering Methods", IOSR Journal of Engineering, Vol. 2(4), pp. 719-725, Apr 2012.

[70] Unnati R. Raval, Chaita Jani, "Implementing and Improvisation of K-means Clustering ", International Journal of Computer Science and Mobile Computing, Vol.4, No.11, pp. 72-76, 2015.

[71] M. Y. Al musaddar, "Improving Arabic Light Stemming in Information Retrieval Systems," Diss. MSC Thesis. Computer Engineering Department, Faculty of Engineering, Research and Postgraduate Affairs, Islamic University, Gaza, Palestine, vol. 1, 2014.

[72] k-nearest neighbor Article in Scholarpedia January 2009
DOI:10.4249/scholarpedia.1883 · Source: DBLP.

[73] Bhoj Raj Sharmaa*, Daljeet Kaura and Manjub aDepartment of Computer Science, Eternal University, Baru Sahib, Sirmour (H.P) Computer Science Department, BMJ Group of Colleges, Bathinda, (PB) Accepted 20 June 2013, Available online 25 June 2013, Vol.3, No.2 (June 2013) A Review on Data Mining: Its Challenges, Issues, and Applications.

[74] Y. Liu, Q. Pan, and Z. Zhou, "Improved Feature Selection Algorithm for Prognosis Prediction of Primary Liver Cancer," in Third IFIP TC 12 International Conference, ICIS 2018 Beijing, China, November 2–5, 2018, pp. 422–430.

[75] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 2, no. 6, pp. 493–507, 2012.

[76] B. Luo, Q. Zhang, and S. D. Mohanty, "Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment," arXiv Prepr. arXiv1805.01586, pp. 1–7, 2018.

[77] X. Wu and V. Kumar, The top ten algorithms in data mining. United States of America: CRC press, 2009.

[78] S. Agarwal, G. N. Pandey, and M. D. Tiwari, "Data mining in education: data classification and decision tree approach," Int. J. e-Education, e-Business, eManagement e-Learning, vol. 2, no. 2, p. 140, 2012.

[79] M. M. Al-Tahrawi, “Arabic text categorization using logistic regression,” Int. J. Intell. Syst. Appl., vol. 7, no. 6, p. 71, 2015.

[80] Hassan Mohammad Dawoud, " Combining Different Approaches to Improve Arabic Text Documents Classification", MSC thesis in Computer Engineering, Islamic University – Gaza Palestine, 2013.

[81] Lecture 5: Stochastic Gradient Descent CS4787 — Principles of Large-Scale Machine Learning Systems.

[82] K-NN 15.097 MIT, Spring 2012, Cynthia Rudin Credit: Seyda Ertekin.

[83] Crowdsourcing for Search and Data Mining, Matthew Lease School of Information University of Texas Austin, TX

[84] Jose P. Miguel, A Review of Software Quality Models for the Evaluation of Software Products · Universidad Peruana Cayetano Heredia, DOI: 10.5121/ijsea.2014.5603, November 2014.

[85] Jian Zhang, Modified Logistic Regression: An Approximation to SVM and Its Applications in Large-Scale Text Categorization School of Computer Science, Carnegie Mellon University.

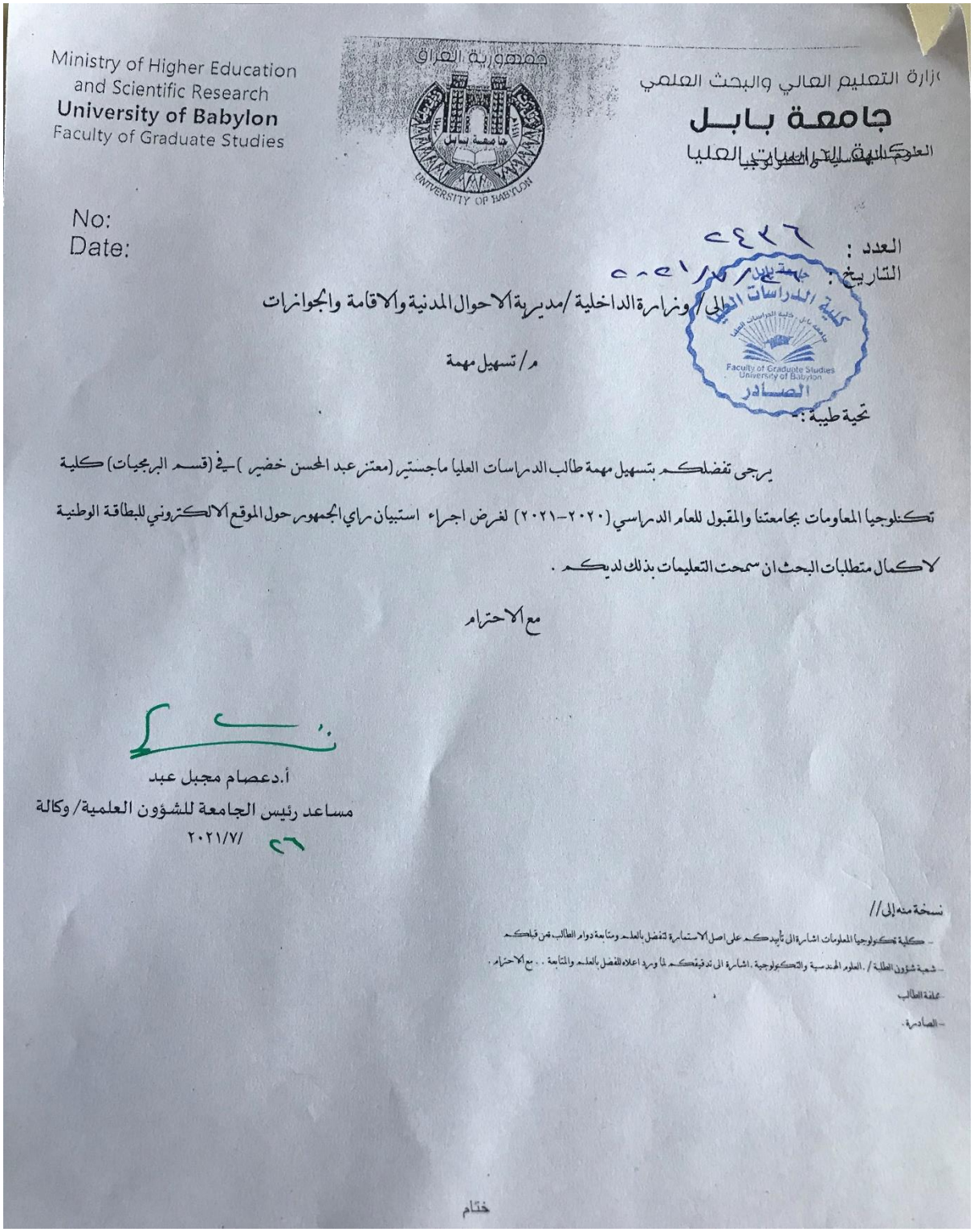
[86] Tyler Rinker, On the Treatment of Likert Data, University at Buffalo, May 2014.

[87] Brendan Eagan, Testing the Reliability of Inter-Rater Reliability, Educational Psychology University of Wisconsin – Madison Wisconsin USA

[88] Nour Jamal Absi Halabi, Discovering the most promising in a crowdsourcing platform for product development. American University of Beirut, Lebanon February 2021.

Appendix A

Formal Document (1)



Ministry of Higher Education and Scientific Research
University of Babylon
Faculty of Graduate Studies

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
العلوم للدراسات العليا

No:
Date:

العدد : ٤٤٦
التاريخ : ٢٠٢١/٧/٢٦

جامعة بابل
كلية الدراسات العليا
Faculty of Graduate Studies
University of Babylon
الصادر

الى / وزارة الداخلية /مديرية الاحوال المدنية والاقامة والجوازات

م/ تسهيل مهمة

تحية طيبة :-

يرجى تفضلكم بتسهيل مهمة طالب الدراسات العليا ماجستير (معتز عبد المحسن خضير) في (قسم البرمجيات) كلية تكنلوجيا المعلومات بجامعتنا والمقبول للعام الدراسي (٢٠٢٠-٢٠٢١) لغرض اجراء استبيان راي الجمهور حول الموقع الالكتروني للبطاقة الوطنية لاكمال متطلبات البحث ان سمحت التعليمات بذلك لديكم .

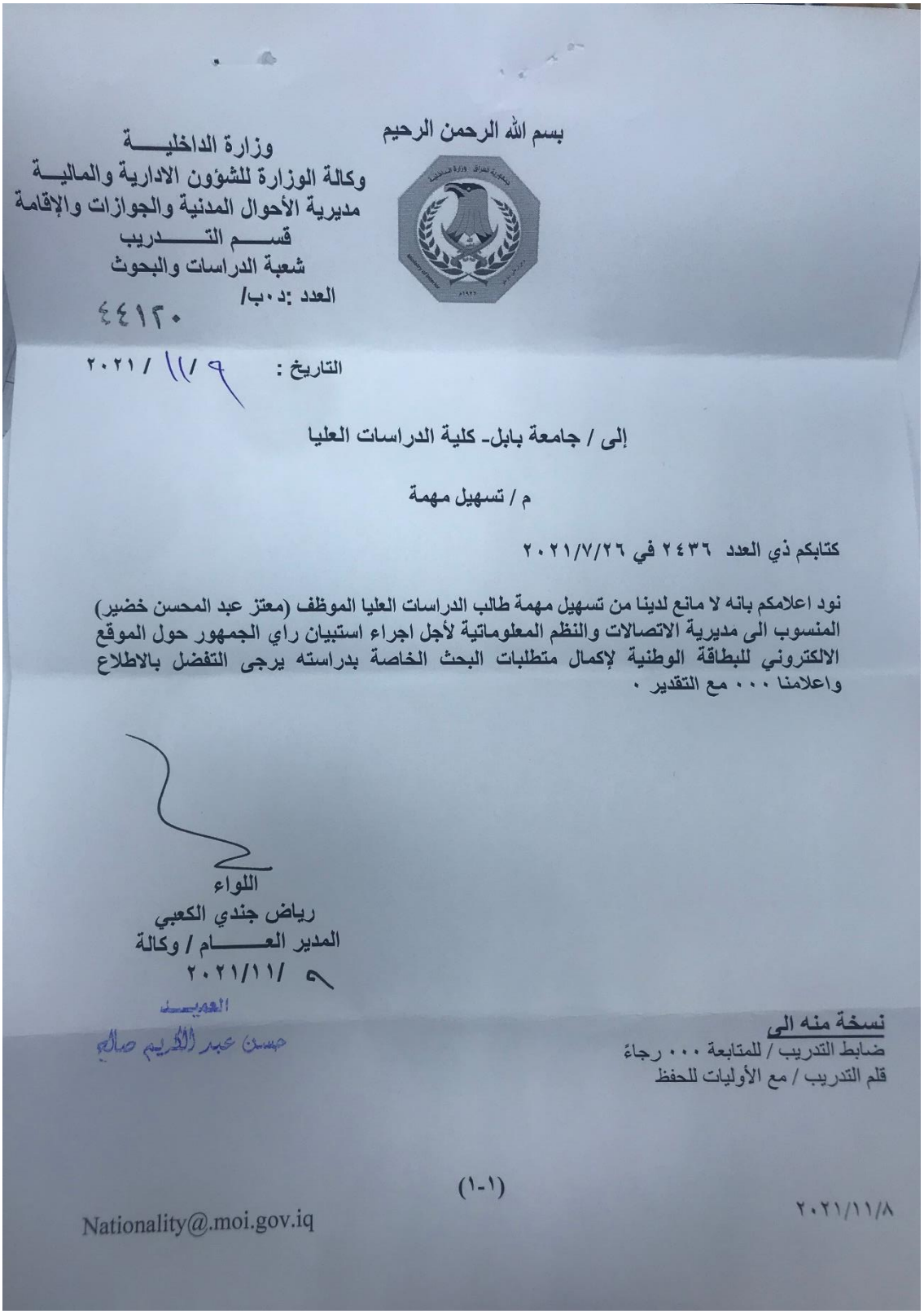
مع الاحترام

أ.دعصام مجبل عبد
مساعد رئيس الجامعة للشؤون العلمية/ وكالة
٢٠٢١/٧/

نسخة منه إلى//
- كلية تكنولوجيا المعلومات اشارة الى تأييدكم على اصل الاستمارة لتفضل بالعلم ومتابعة دوام الطالب فمن قبلكم
- شعبة شؤون الطلبة / . العلوم الهندسية والتكنولوجية . اشارة الى تدقيقكم لما ورد اعلاه للفضل بالعلم والمتابعة . . مع الاحترام .
- محافظ الطالب
- الصادرة .

ختام

Formal Document (2)



بسم الله الرحمن الرحيم

وزارة الداخليـــــة
وكالة الوزارة للشؤون الادارية والماليـــة
مديرية الأحوال المدنية والجوازات والإقامة
قســـم التـــــدريب
شعبة الدراسات والبحوث
العدد :د٠ب/
٤٤١٣٠

التاريخ : ٩ /١١/ ٢٠٢١

إلى / جامعة بابل- كلية الدراسات العليا

م / تسهيل مهمة

كتابكم ذي العدد ٢٤٣٦ في ٢٠٢١/٧/٢٦

نود اعلامكم بانه لا مانع لدينا من تسهيل مهمة طالب الدراسات العليا الموظف (معتز عبد المحسن خضير) المنسوب الى مديرية الاتصالات والنظم المعلوماتية لأجل اجراء استبيان راي الجمهور حول الموقع الالكتروني للبطاقة الوطنية لإكمال متطلبات البحث الخاصة بدراسته يرجى التفضل بالاطلاع واعلامنا ٠٠٠ مع التقدير ٠

اللواء
رياض جندي الكعبي
المدير العـــــام / وكالة
٩ /١١/ ٢٠٢١

العبيد
حسن عبد الكريم صالح

نسخة منه الى
ضابط التدريب / للمتابعة ٠٠٠ رجاءً
قلم التدريب / مع الأوليات للحفظ

(١-١)

Nationality@.moi.gov.iq

٢٠٢١/١١/٨

Appendix B

The First Dataset

الموقع متناسق	البيانات المطلوبة واضحة	الموقع لا يحتوي على تعقيد	الموقع مختصر	وجود قواسم مشتركة في البيانات	قابلية التنقل بين الحقول	المرونة في الموقع	سهولة استخدام الموقع
اوافق	اوافق بشدة	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق
اوافق	اوافق	مقبول	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	مقبول	اوافق	اوافق	اوافق	مقبول	مقبول	اوفق بشدة
اوافق	مقبول	اوافق	مقبول	اوافق	اوافق	اوافق	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	مقبول
اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوفق بشدة
اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	مقبول	اوافق بشدة	اوافق	مقبول
مقبول	اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق
اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق
اوافق بشدة	اوافق بشدة	اوافق بشدة	مقبول	اوافق بشدة	اوافق	اوافق	اوفق بشدة
اوافق	اوافق بشدة	اوافق	اوافق	مقبول	اوافق	اوافق بشدة	اوفق بشدة
اوافق	مقبول	اوافق	اوافق	اوافق بشدة	اوافق	اوافق بشدة	مقبول
مقبول	اوافق بشدة	اوافق بشدة	اوافق	اوافق	اوافق	اوافق بشدة	اوافق
مقبول	مقبول	مقبول	مقبول	مقبول	مقبول	مقبول	مقبول
اوافق بشدة	اوافق بشدة	مقبول	مقبول	اوافق بشدة	اوافق	اوافق	اوفق بشدة
اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق
مقبول	اوافق	اوافق	اوافق	مقبول	مقبول	مقبول	مقبول
مقبول	مقبول	مقبول	مقبول	اوافق بشدة	مقبول	مقبول	اوفق بشدة
مقبول	اوافق بشدة	اوافق بشدة	اوافق	اوافق	اوافق	اوافق بشدة	اوافق
مقبول	اوافق	اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق
اوافق بشدة	مقبول	مقبول	مقبول	اوافق بشدة	مقبول	مقبول	مقبول
اوافق بشدة	اوافق	اوافق	اوافق	اوافق	مقبول	اوافق	اوافق

راضي بالموقع	سريع التحميل والرفع	سهولة التعامل مع الموقع	أعجبك تصميم الموقع	محتوى الموقع واضح	هل أنت مستخدم أنترنت	يحتاج إلى تمرين	الموقع مستقل	قابلية العمل على الأنظمة	قابلية العمل على الأجهزة
اوافق	اوافق	اوافق	اوافق	اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	اوافق	اوافق	مقبول	اوافق بشدة	اوافق	مقبول	اوافق	اوافق	اوافق بشدة
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق بشدة	اوافق	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	اوافق
مقبول	مقبول	اوافق	مقبول	اوافق	اوافق	اوافق	اوافق	مقبول	مقبول
اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق
اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق بشدة	مقبول	اوافق	اوافق بشدة	اوافق بشدة	مقبول
اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق
اوافق	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق
اوافق	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	مقبول	اوافق بشدة	اوافق بشدة	اوافق
اوافق	اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق
اوافق بشدة	اوافق	اوافق بشدة	مقبول	مقبول	مقبول	اوافق	اوافق	اوافق بشدة	اوافق بشدة
اوافق	اوافق بشدة	اوافق	اوافق بشدة	مقبول	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة
مقبول	مقبول	مقبول	مقبول	اوافق	اوافق بشدة	مقبول	مقبول	مقبول	مقبول
اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق	اوافق
اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق بشدة	اوافق	اوافق	اوافق
مقبول	مقبول	مقبول	اوافق بشدة	اوافق بشدة	اوافق	ارفض	اوافق	مقبول	مقبول
اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	مقبول	اوافق بشدة	اوافق بشدة	مقبول
اوافق بشدة	اوافق	مقبول	اوافق بشدة	اوافق	اوافق بشدة	مقبول	اوافق بشدة	اوافق بشدة	اوافق بشدة
اوافق بشدة	اوافق بشدة	اوافق	مقبول	اوافق بشدة	اوافق	مقبول	اوافق	اوافق بشدة	اوافق بشدة
اوافق	اوافق بشدة	اوافق بشدة	اوافق	اوافق	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق بشدة	اوافق
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق بشدة	اوافق	اوافق	اوافق	اوافق
مقبول	مقبول	مقبول	اوافق بشدة	مقبول	اوافق	اوافق بشدة	مقبول	مقبول	اوافق بشدة
اوافق	اوافق	اوافق	اوافق	اوافق	اوافق	ارفض	اوافق	اوافق	مقبول

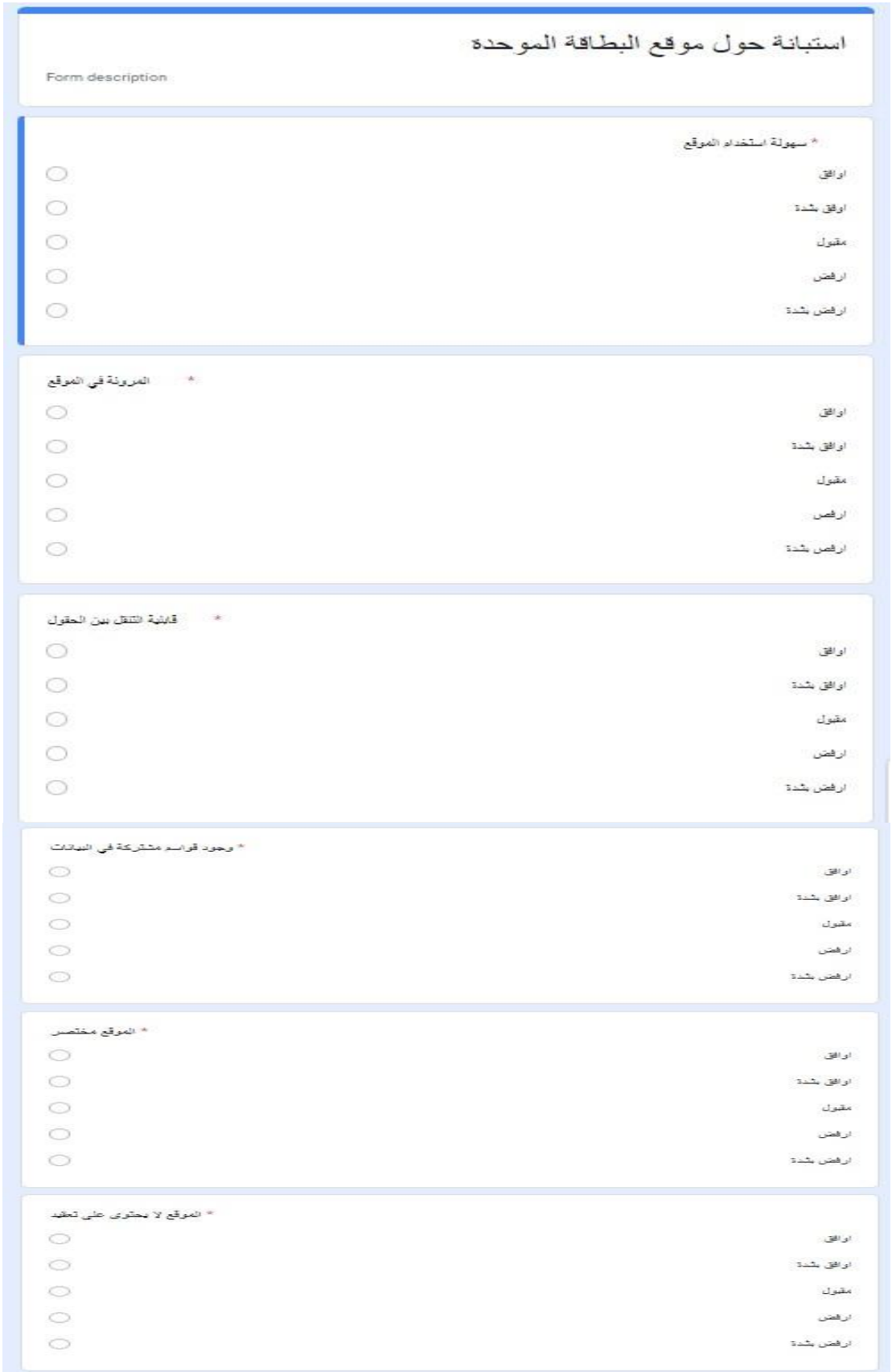
The Second Dataset

User Story	Well-Formed	Atomic	Minimal	Conceptually-sound	Problem-oriented	Umambigious	Full sentence	Estimatable
Als organisator wil ik niets. Maar wij organiseren nu een FIFA 2019 toernooi en denk dat dit heel veel gebeurt. Geen teams maar individuele spelers, mogelijk met een teamnaam, veel spelers meer dan in menig toernooi, kortere wedstrijden, geen velden maar consoles. Ik kan me voorstellen dat je Tournify daarvoor ook een template geeft of iets anders aanpast, zodat het makkelijk wordt ook e-Sports toernooien te organiseren. Denk dat het technisch nu al wel kan maar doe eens een check wat er beter kan.	0	0	0	1	0	1	1	0
Als organisator wil ik teamfoto's toevoegen, zodat het persoonlijker wordt.	1	1	1	1	1	1	1	1
Als organisator wil ik het account kunnen inperken om het te delen, zodat ik met meerderen een toernooi kan organiseren maar niet iedereen alle admin-rechten heeft.	1	1	1	1	1	1	1	1
Als organisator wil ik beschikbaarheid van scheidsrechters beperken tot een deel van een dag of weekeinde, zodat ik tegemoet kan komen aan wensen die men heeft voor de eigen beschikbaarheid.	1	1	1	1	1	1	1	1
Als organisator wil ik ook voordat het toernooischema gemaakt wordt, een inschrijvingssite voor een toernooi kunnen maken, zodat ik ook het aanmelden door andere teams voor het toernooi in dezelfde Tournify kan afhandelen.	1	1	1	1	1	1	1	1
Als organisator wil ik als toernooi organisator als er meerdere leeftijds categorien op hetzelde moment spelen de scheidrechters ook apart in kunnen plannen, dit kan nu niet alleen handmatig en niet automatisch, zodat dit sneller gebeurt is	1	1	0	1	1	1	0	1
Als organisator wil ik de zekerheid dat een e-mailadres dat bij inschrijving wordt doorgegeven ook echt valide is, zodat een bevestiging ook zeker weten aankomt. Hiervoor zouden jullie gebruik kunnen maken van een third party check , zoals https://www.mailgun.com/email-								

User Story	Well-Formed	Atomic	Minimal	Conceptually-sound	Problem-oriented	Umambigious	Full sentence	Estimatable
Als organisator wil ik dat de (losse) wedstrijden op het scherm getoond worden op tijd en niet op volgorde waarin de wedstrijden zijn gemaakt, zodat er geen wedstrijden op onlogische wijze getoond worden op de schermen zodra je achteraf een wijziging gaat doorvoeren in het wedstrijd schema.	1	1	1	1	1	1	1	1
Als organisator wil ik graag vaste tijdblokken met getoonde tijd en omschrijving kunnen toevoegen, zodat je ook lunch of andere activiteiten zichtbaar kunt maken voor de deelnemers in hun app.	1	1	1	1	1	1	1	1
Als organisator wil ik de ranking kunnen tonen op de schermen, zodat alle deelnemers eenvoudig kunnen zien welke plaats ze uiteindelijk behaald hebben.	1	1	1	1	1	1	1	1
Als organisator wil ik graag meer statistieken meteen zichtbaar hebben in de beheermodule in een hoofdscherm, zodat je niet continue hoeft door te klikken om betaalde aantal te achterhalen zoals aantal teams, aantal velden, aantal scheidsrechters, aantal wedstrijden etc.	1	1	1	1	1	1	1	1
Als organisator wil ik graag de volgorde van de velden in het beheerscherm kunnen wijzigen of door te slepen of door het veldnummer te wijzigen waardoor de volgorde wijzigt, zodat velden die alsnog worden toegevoegd op de juiste plaatst komen te staan. Of om zo tijdelijk even velden naast elkaar te zetten voor een beter overzicht.	1	1	1	1	0	1	0	1
Als organisator wil ik graag memovelden voor intern gebruik, zodat je belangrijke info kunt opslaan of delen met anderen die toegang hebben tot de beheermodule.	1	1	1	1	1	1	1	1
Als organisator wil ik graag het symbool voor verplaatsen van deelnemers naar andere teams laten vervangen door de tekst: verplaats ipv een pijltje die wijst naar de prullenbak, zodat duidelijker is dat dit een functie is van verplaatsen van deelnemers en het niet lijkt dat het is voor het verplaatsen naar de prullenbak.								

Appendix C

A Survey about the National Card website


استبانة حول موقع البطاقة الموحدة

Form description

* سهولة استخدام الموقع

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* المرونة في الموقع

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* قابلية التنقل بين الحقول

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* وجود قواسم مشتركة في البيانات

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* الموقع مختصر

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* الموقع لا يحتوى على تعقيد

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* البيانات المطلوبة واضحة

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* الموقع متناسق

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* قابلية العمل على الأجهزة

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* قابلية العمل على الأنظمة

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* الموقع مستقل

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* يحتاج الى تمرين

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* هل أنت مستخدم أنترنت

- اوافق
- اوافق بشدة
- مقبول
- ارفض
- ارفض بشدة

* محتوى الموقع واضح

- اوافق
- اوافق بشده
- مقبول
- ارفض
- ارفض بشده

* أعجبك تصميم الموقع

- اوافق
- اوافق بشده
- مقبول
- ارفض
- ارفض بشده

* سهولة التعامل مع الموقع

- اوافق
- اوافق بشده
- مقبول
- ارفض
- ارفض بشده

جمهورية العراق
وزارة التعليم والبحث العلمي
جامعة بابل
كلية تكنولوجيا المعلومات
قسم البرمجيات

تحسين التعهيد الجماعي في هندسة البرمجيات بالاعتماد على تعلم الآلة

" دراسة تطبيقية على الموقع الإلكتروني للبطاقة الوطنية الموحدة في العراق "

أطروحة

مقدمة الى مجلس كلية تكنولوجيا المعلومات في جامعة بابل والتي هي جزء من متطلبات نيل درجة الماجستير في تكنولوجيا المعلومات – البرمجيات

من قبل

معتز عبد المحسن خضير بدر

بأشراف

أ.م. د احمد سليم عباس جاسم

1443 هـ **2022 م**

الخـلاصـــة

اذا كان لديك عمل تجاري كبير ، فمن أين ستبدأ؟

ربما تكون قد اتصلت ببعض الأصدقاء والزملاء لمعرفة كيفية التعامل مع المشكلة واستخدام اقتراحاتهم الجماعية لتوجيه أفعالك ،هذه هي الطريقة التي يعمل بها التعهيد الجماعي، بدلاً من سؤال ثلاثة أصدقاء ماذا سيحدث إذا سألت 300 شخص؟

الحاجة إلى التعامل مع أصحاب المصلحة على نطاق واسع لضمان صحة متطلبات البرامج تجعل تقنية التعهيد الجماعي مفيدة لتحسين المستوى لجودة المتطلبات من حيث توفير تكلفة التطوير والوقت.

يمكن تعريف التعهيد الجماعي على أنه استغلال الذكاء الجماعي للجمهور لإنجاز المهمة. ونظرًا لأن ذلك يعتمد على الكثير من الأشخاص ، فإن النوع الصحيح من التعهيد الجماعي يمكن أن يأتي غالبًا بحل أفضل مما كنت ستحصل عليه إذا جمعت فقط قوة تفكير عدد قليل من الأشخاص ، حيث يتم التعهيد الجماعي من خلال جمع أكبر عدد من المعلومات أو آراء أو تجارب عبر الإنترنت ووسائل التواصل الاجتماعي وتطبيقات الهواتف الذكية ، وقد تكون هذه العملية تطوعية أو مجانية أو مدفوعة الأجر.

أذن التعهيد الجماعي هو حل مشكلة نموذجي موزع يعتمد على الجمع بين الحساب البشري والآلة ولطالما كان الارتباط بهذه التقنية موضوعًا مهمًا حول كيفية ترجمة هذه المتطلبات إلى منتج أو خدمة يستخدمها الأشخاص من حيث استخدم التعهيد الجماعي لدعم أنشطة هندسة البرمجيات.

توفر هذه الرسالة للباحثين كيفية تحسين تقنيات التعهيد الجماعي لفوائد هندسة البرمجيات حيث تهدف هذه الرسالة إلى الاستفادة من التعهيد الجماعي في نظام البطاقة الموحد بغرض تقديم أفكار متنوعة يمكنها حل المشكلات الصعبة وغير المتوقعة بشكل أسرع ، وذلك من خلال جمع البيانات التي تركز على الإجابات التي تم الحصول عليها من الأسئلة (ثمانية عشر سؤالاً) من المستفيدون وعددهم ما يقارب عشرة الاف شخص مما يساعد على تطوير لعمل البطاقة الموحدة ويساهم في تطوير منتجاتها وخدماتها وأنظمتها بالاضافة الى إيجاد الحلول المثالية للمشكلات الصعبة في وقت قياسي.

يتكون النظام المقترح من اربعة مراحل رئيسية.

المرحلة الأولى وهي جمع البيانات التي ركزت على إجابات المستفيدين من نظام البطاقة الموحدة.

المرحلة الثانية وهي المعالجة المسبقة للبيانات التي تم تطبيق النظام عليها.

المرحلة الثالثة استخدام طرق التصنيف المختلفة باستخدام عدة خوارزميات

والمرحلة الرابعة تتضمن النتائج المستخرجة من خلال استخدام الخوارزمياتوالتي كانت

1- Decision Tree

2- Random Forest

3- K-Nearest Neighbor

4- Logistic Regression

5- Stochastic Gradient Descent

6- Likert Scale

تم تنفيذ هذه التقنيات لألقاء الضوء على أدق تقنية لاستخدامها في تصنيف بيانات. علاوة على ذلك، تم تقييم نتائج خوارزميات التصنيف استنادا الى قياس الأداء باستخدام

(Recall, Accuracy, Precision and F1-Score)

أظهرت النتائج انه تم تحقيق اعلى دقة من خلال تطبيق خوارزمية التالية

(Logistic Regression, Decision Tree and Random Forest)

حيث بلغت مقياس الأداء (١٠٠٪) لهذه الخوارزميات الثلاثة بالإضافة الى استخدام خوارزمية

(Stochastic Gradient Descent) التي حققت نسبة (٧٨٪) وأخير تم استخدام خوارزمية الخامسة والأخيرة وهي (K-Nearest Neighbor) التي حققت نسبة (٨٥٪).

بالإضافة الى استخدام مقياس (Likert Scale) الذي أوضح قياس أراء المساهمين وبيان مدى رضاهم عن النظام المعمول به.

إذ حصلت موافقة بشدة على اعلى تقييم بنسبة (٤٢١٠) تليها غير موافقة حصلت على نسبة (١٥٧٥) ومن ثم طبيعي حصلت على نسبة (١٥٧٨) أما موافق فحصلت على نسبة (١٠٥٢) وأخيرا غير موافق بشدة حصلت على اقل نسبة (٥٢٦).

ويتضح مما يلي أن التعهيد الجماعي يسعى الى بعض التعديلات أو التغييرات لتناسب أراء وأفكار الجمهور الذي تم بيان رأيهم بالنظام المعمول عليه الإن.