A DATA MINING APPROACH ON OCCUPATIONAL COMPETENCY ASSESSMENT: THE CASE OF ADDIS ABABA

UMER MOHAMMED UMER

A Thesis submitted to the Department of Computer Science and Engineering, School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

July, 2019

A DATA MINING APPROACH ON OCCUPATIONAL COMPETENCY ASSESSMENT: THE CASE OF ADDIS ABABA

UMER MOHAMMED UMER

ADVISOR: DR. TILAHUN MELAK SITOTE

A Thesis submitted to the Department of Computer Science and Engineering, School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

July, 2019

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by Umer Mohammed Umerhave read and evaluated his/her thesis entitled ***"A data mining approach on occupational competency assessment: The case of Addis Ababa"***and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Information Systems Engineering.

_____	_____	_____
Student	***Signature***	***Date***
_____	_____	_____
Advisor	***Signature***	***Date***
_____	_____	_____
Chair Person	***Signature***	***Date***
_____	_____	_____
Internal Examiner	***Signature***	***Date***
_____	_____	_____
External Examiner	***Signature***	***Date***

DECLARATION

I hereby declare that this MSc Thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name: Umer Mohammed Umer

Signature: _____

This MSc Thesis has been submitted for examination with my approval as thesis advisor.

Name: Dr.Tilahun Melak Stote

Signature: _____

Date of submission: __July, 2019____

ACKNOWLEDGMENTS

First and foremost, I want to offer this endeavor to our Lord, the Almighty for the wisdom he bestowed upon me, the strength, peace of my mind, and good health in order to finish this research.

I would like to express my honest gratitude to my advisor Dr. TilahunMelakSitote for the continuous support of my study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of writing this thesis. My sincere thanks also goes to staff of Addis Ababa Occupational Competency Assessment and Certification Center (OCACC), who have dedicated their time for supplying information when it was needed. I also wish to express my gratitude to the information and certification staff of OCACC branch members for providing a collection of data. I also want to express my earnest gratitude to my wife and my all family.

My thanks extend to Mr. EngidaAsfaw, Mr. Mignote W/yesus and Mr. HabtamuTamrat for their kind sharing to get particular information on the thesis work and discussion. Finally I also like to thank W/roFetyaMohammed Ahmed and Mr. DanelZemeneMequanint for their continuous support and constructive comments on my report.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ARM	Association Rule Mining
DM	Data Mining
DMT	Data Mining Technique
EDM	Educational Data Mining
GTP	Government Transformation Program
IGC	International Growth Center
ML	Machine Learning
NTQF	National TVET Qualifications Framework
OCACC	Occupational Competency Assessment and Certification Center
OS	Occupational Standard
OC	Occupational Competency
OCA	Occupational Competency Assessment
OCAC	Occupational Competency Assessment and Certification
SCMS	Smart Candidate Management System
TVET	Technical and Vocational Education and Training
UNC	Unit of Competency
UNESCO	United Nations Educational scientific and Cultural Organization

ABSTRACT

There are many factors affect the process of occupational competency assessment such as language, regulations prepared by the government, and economic matter. The primary objective of this study was to identify patterns in the available data that could be useful for analyzing factors affecting a candidate's performance using data mining techniques. The dataset for this study has been obtained from Addis Ababa Occupational Competency Assessment and Certification Center (OCACC) and it contains 324, 236 records. From the available dataset, 3000 records have not been either assessed or not scheduled for the assessment so that they have been removed. Hence, a dataset of 321,236 records has been used for the final experiment. WEKA software for association and classification has been used as major tools. On top of that by computing the efficiency of experimented algorithms, it has been shown that the Apriori algorithm has better accuracy. As it has been indicated in the ARM analysis result if a candidate is young, single, unemployed with no work experience, he/she likely are not competent. Additionally, if the candidates also young, single, mode of training regular and applied for qualification based assessment, he/she also be incompetent. Likewise, if the candidates those who are young, mode of training regularly and have a year gap to assess their competency after accomplish their study, she or he will not also be incompetent. However, the study selects the feasible rules which have a possible role to realize a candidate's performance. Based on the analysis result, occupational standards (OS) have a higher influence on candidate's performance and also the majority of candidates have no awareness about occupational competency's (OC) which is derived from OS. On the other hand, from the resulted analysis, candidates have no full of physiological readiness for the assessment since there is no work environment exposure and also they are not assessed in the industries. Finally, the study recommended broadcasting online all Occupational Competencies (OC) which is generated from OS for the assessment to have a better awareness. And also training centers should work in collaboration with industries to make candidates a psychological preparation. Similarly, candidates should be assigned on the apparent ship to have work exposure in all types of practical activities. Finally, the study indicated that OCACC has a responsibility to assess each candidate in the industry. Since the candidate assumes the assessment as a job.

Keywords: Occupation, Competency, Assessment, Candidate, Data mining, ARM.

CHAPTER ONE: INTRODUCTION

1.1. Background of the study

A competent citizen has a significant role to play in bringing about universal change and development for a country. The development of a country can be determined by whether its citizens have a good education or not [1]. According to that, to create a competent citizen, who brings a universal change, Technical and Vocational Education and Training (TVET) institutes have an important role [2].

According to [3] and [4],TVET refers to the educational method in addition to common education, the study of sciences, technologies, gathering of practical skills, attitudes, and knowledge relating to occupation in different sectors of economic and social life. And it also emphasizes skills, knowledge, and attitudes required for employment in a particular occupation or cluster of related occupations in any field of social and economic activity including agriculture, industry, commerce, the hospitality industry, public and private sectors. Particularly A key purpose of TVET is to provide special role on young people and adults with the knowledge, skills, and competencies toward to improve quality of life by enhancing employability for the unemployed, and facilitate transfer to new occupations for those currently employed [2]. Based on that, Skills development and TVET are now becoming increasingly important on the international and national policy agenda. For example, the United Nations Educational Scientific and Cultural Organization (UNESCO) advocates TVET, claiming that technical and vocational education that is driven by market demand is more effective in maintaining employment and income for the disadvantaged. Generally, there is a probability that TVET facilitates economic growth and poverty alleviation by serving as a mechanism to prepare people for occupational fields and by enhancing their effective participation in the world of work [5].

According to [6], Growth and Transformation Plan (GTP II), greater shares of economic production will come from industry and manufacturing with the consequent demands for middle- and higher-level skilled manpower to be supplied by the educational system. The business environment is changing at a fast pace due to the undeniable influence of rapidly emerging and

changing technologies in a fast globalizing world. This has increased the demand for variety in human skills necessary to respond to such drastic changes in a competitive manner for economic progress and business survival [7]. Meanwhile, the government believes that the present low factor productivity is due to the skill gap, and the industry provides less training. Therefore, the government focuses on vocational education is the means to close this skill gap to improve productivity and increasing professional's competitiveness in the global market [8].

According to [9]the driving goal of the national TEVT strategy of Ethiopia 2008 is to strengthen the culture of self-employment and support job creation in the economy. Based on that, Ethiopia mainly adopted its current TVET curriculum experiences from countries such as Australia and the Philippines. According to the trends of these countries, the new Ethiopian TVET strategy has decentralized the preparation of curricular materials to the institutions that deliver training. The difficulty may limit the current competency-based TVET curriculum in Ethiopia is a lack of experience to develop the curriculum at the local level in this decentralized responsibility to develop the curriculum at TVET institutions. In addition to the difficulty of decentralization, the continuous change made in the occupational standards is another dispute in the effective implementation of the reformed TVET approach. While TVET institutions have set themselves and started to provide training in certain occupational standards disseminated, the Ministry of Education in the meantime updates or replaces those occupational standards with the new ones. This has created resource wastage and grievance at institutions, management, instructors and students [10]. According to the ministry of education, the main objective of the new national TVET strategy is to create a competent, motivated, adaptable and innovative workforce that plays crucial roles in the poverty reduction and socio-economic development efforts of the country. Instead of that these efforts justified based on facilitating outcome-based system. This means that identifying competencies of the labor market and the outcome of training delivered in the system is measured through a process of verification which is known as occupational competence assessment (OCA). And the verification assessment is not only a TVET graduate's competence that is to be measured but also that of anyone who wants his/her competences be recognized [11].In order to meet the vital demand for trained and competent manpower, the issue of occupational competency assessment and certification (OCAC) has paramount importance to promote the social, economic, and political development of a nation. This reveals that to create a competent, motivated and innovative workforce the system of OCAC plays a tremendous role

[12]. OCAC process is geared towards assuring that professionals from various occupations possess the necessary integrated set of knowledge, skills, and attitudes that the industry and/or work environment requires. Hence, the city government of Addis Ababa has been engaged in such a process since its establishment having the vision to establish well organized and sound Occupational Competency Assessment and Certification Center (OCACC) which is led by the industry and produce globally competent workforce by the year 2020 [13].OCACC is one of certification center in Addis Ababa city and it was established in 2000 EC. The center provides competency assessment services for the country to ensure customer or professional's competency. Regarding this, OCACC provides occupational competency assessment for all professionals at various times, among this, there are industry experts, students, experienced practitioners and graduated students. According to that, professionals assessed to assure their competence [13]. The following diagram elaborates Addis Ababa OCACC assessment process.

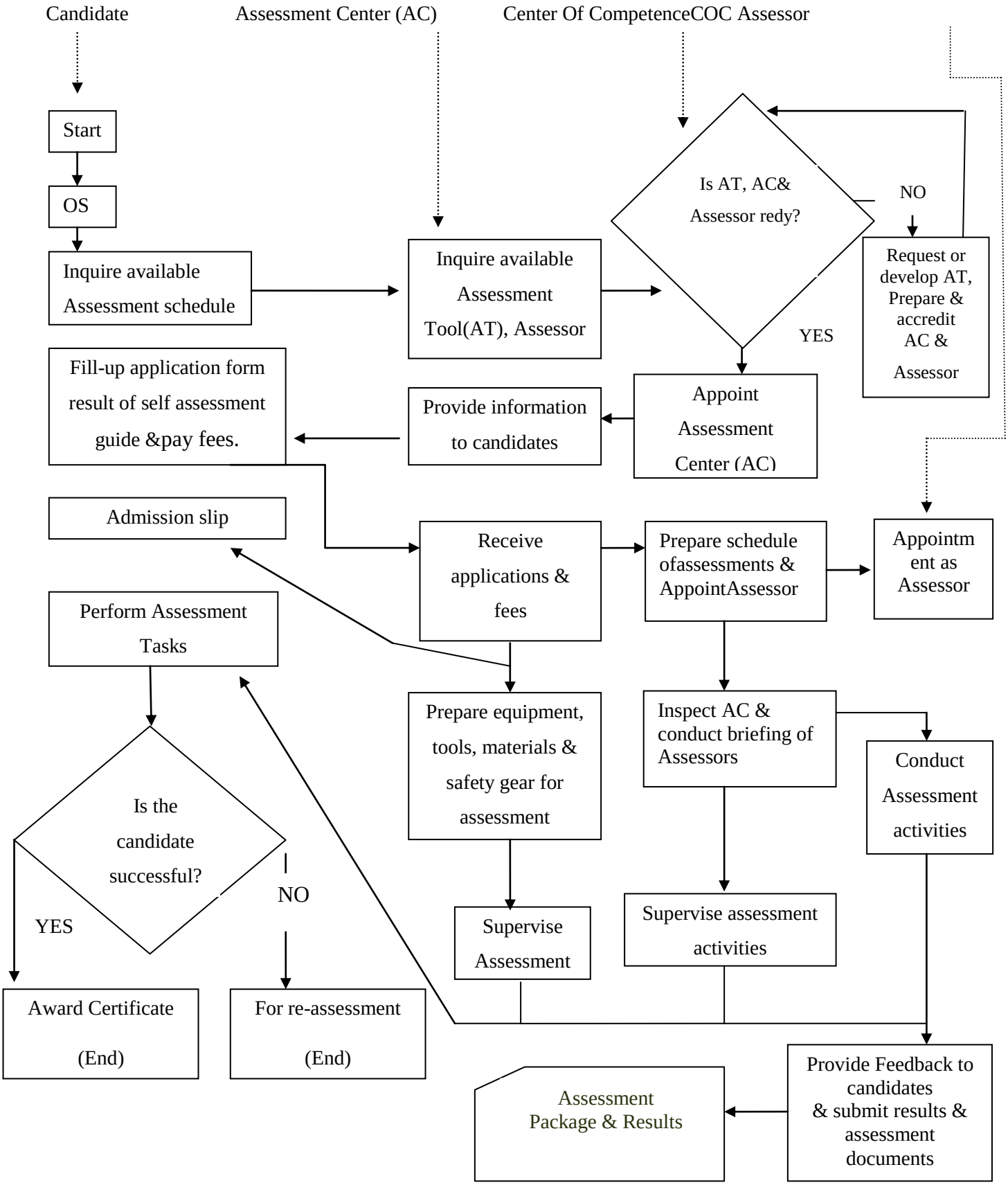


Figure 1.1: OCACC Assessment Process Flow Chart[14]

According to [15] define competency as the ability of an individual to perform a job properly and mannerly. And the authors also mentioned that competency is a set of distinct features that offers a structured guide enabling the identification, evaluation, and development of the behaviors in individuals. Likewise, Competency defined as the professionals skill to demonstrate knowledge, skills and attitudes in order to perform tasks and duties successfully and which can be measured against well-accepted standards (competency levels) required in employment as well as assessed against provided evidences at work location [16].Occupational competency assessment is a new phenomenon to the Ethiopian TVET system that it has become one of the main concerns of the TVET system reform process and, that requires the fulfillment of many aspects to administering the system effectively. Incompetency assessment the knowledge, skills, and attitudes of candidates are assessed with the standards expected in the workplace as it is expressed in relevant competence standards [17].

According to [18]the overall idea of competency-based assessment is the process of collecting evidence about one's knowledge, skill, and attitude to make judgments whether an individual is competent or not. The purpose of assessment is to confirm the fact that an individual can perform to the standard expected in the workplace as expressed in the relevant occupational standards. To guide the preparation of occupational standards, the Ethiopian Occupational Standards Development Guideline was developed first in July 2009 and then upgraded in January 2012. As of September 2012, and based on the needs of the world of work, occupational standards have been developed to 388 occupational titles including fields in agriculture; culture, sports and tourism; economic infrastructure; health; industry development; and labor affairs and social service [17].Occupational standards define the competencies or the abilityof workers according to requirements in the labor market. It describes the candidate's achievement to be qualified in a field of study.Competence includes skills, attitudes, and knowledge for performing a job. Identification and clustering of occupations are made in close cooperation with the Ministry of Labor and Social Affairs and the Civil Service Agency as well as other concerned bodies to ensure that the TVET occupational standards take into account the defined occupational titles from the National Occupational Classification indicated in the National TVET [12]. The Ethiopian TVET system is reorganized into an outcome-based system. This means that identified competencies of the labor market that are described in the occupational standards are the final benchmarks not only for training and learning activities but also for the assessment of

competencies and certification as well. Moreover, building an outcome-based TVET system creates access for equal recognition of competences acquired through whatever the means and ways of being competent.The overall frame and structure of the outcome-based TVET system is described in the National TVET Qualifications Framework (NTQF). The NTQF rationalizes all TVET provisions into a single nationally recognized qualification. It defines the different occupational qualification levels to be awarded. The levels detail the scope and composition of qualifications and degree of responsibility a qualified person would assume in the workplace [11].

According to [12], the competency level utilizing the following four occupational standards listed below:

- **Level 1(Foundation)**: indicates an initial stage under the usual standard for work and competent in the performance of a range of diverse occupation activities, most of which may be usual and expected.

- **Level 2(Intermediate)**: People who work under supervision or who work in teams are considered as 'Intermediate' or level 2and competent in a significant range of varied work activities, performed in a variety of contexts. Some of the activities are difficult or non-routine, and there is some individual responsibility or autonomy.

- **Level 3 (Advanced)**:People at level 3 are employees who do not have the responsibility of managing or organizing people, but do not work under supervision and have the freedom to move about at work and compete in a broad range of varied work activities performed in a wide variety of contexts, most of which are complex and non-routine.

- **Level 4 (Management)**: people who are responsible for organizing and managing people and production and competent in a broad range of difficult, technical or professional work activities performed in a wide range of context and with a generous level of personal duty and autonomy.

In particular, if there is a professional citizen who trained based on occupational standards (OS) and took occupational competency assessment “OCA”, it is easy to understand the value of the

country. However, when we see the ten years OCACC report, there are 732,023 professionals took occupational competency assessment (OCA) in various areas from 2000 EC up to 2010 EC in OCACC. However, 383,247 that means52% can be able to pass the assessment in the given year [19]. Instead of that, it is necessary to identify the factors which affect the candidate's performance using data mining techniques.

As far as my investigation, there is no research, documented related to occupational assessment quality and it factors in the use of data mining technique (DMT) and machine learning (ML). In the existing system, the organization which is OCACC has a research department. And there are different types of statistical studies and reports up there. Instead of that, we can analyze the existing process using SWOT analysis.

SWOT analysis is the phases of the strategic management process, which is external and internal analysis. By conducting an external analysis, an organization identifies the critical threats and opportunities in its competitive environment. It also examines how competition in this environment is likely to evolve and what implications that evolution has for the threats and opportunities an organization is facing. While external analysis focuses on the environmental threats and opportunities facing an organization, internal analysis helps an organization identify its organizational strengths and weaknesses. It also helps an organization understand which of its resources and capabilities are likely to be sources of competitive advantage and which are less likely to be sources of such advantages [20].

However, organizational SWOT Analysis summarized as follows.

Strength of the organization

1. The organization has research department and try to fill the assessment process gap using traditional studies.
2. Adequate and reliable ICT infrastructure.
3. Adequate and well-resourced office accommodation.

Weakness of the organization

1. The existing Occupational Competency Assessment (OCA) researches handle using traditional statistical methods.
2. Lack of ICT as well as data science specialized manpower in the research department.

Opportunities for the organization

1. Data mining gives an opportunity to the organization to analyses or to make different analysis easily.
2. Data mining also brings to the organization quality analysis results for quality assessment process.

Threats

1. Bureaucratic structure of the organization to maintain data mining techniques.
2. Failed to accept new technological advancement.

However, this study aimed to associate factors that affect a candidate's performance through the Occupational Competency Assessment (OCA) Process using data mining techniques and consequently to learn and suggest some possible solutions for the problems of the issues under consideration.

1.2. Motivation

First of all Learning and improving organizational issues using data mining techniques is a very interesting and inspiring issue.

Now a day's there are a number of young people with various skills in Ethiopia. Since there are many professionals apply for occupational assessment in the past ten years. In every year huge amounts of professionals data is recorded in OCACC database. However, as an information technology expert when I saw the candidate data, there are several candidates are failed in different fields, therefore understanding and evaluating the problem domain is an interesting issue. According to that a proper attention is needed in assessment process since there is enough information for better analysis and decision making. Meanwhile, data mining discovers hidden

information from candidate database, and also this information meaningful for the organization. Regarding this, the motivation behind this study is, therefore, to implement data mining techniques to discover hidden patterns or interesting knowledge from the existing database that enables to summarize accurate factors that release candidates performance from the expected outcome.

1.3. Statement of the Problem

Nowadays, Ethiopia is committed to making great efforts to produce a competent workforce through occupational competency assessment and certification system that can fulfill the minimum standards listed in the national occupational Standards. This is because highly competent human resource is now a key phrase at the heart of academicians and development authorities. It is also one of the most concerns of modern society to keep sustained competitive advantage for a nation that can last much longer. In this effort, the role of stakeholders (assessors, supervisors, assessment centers, training institutions, Center of competency, industries for cooperative training, and other necessary inputs) assumes a central position to cultivate market-oriented competent citizens that can play an incalculable role in materializing the aforementioned national goals [13].

According to [12] one of the focuses of the policy strategy is on producing entrepreneurial and competent citizens that can create their own business and self-employed both quantitatively and qualitatively. For that matter, competency-based assessment is very important to improve quality and increase competitiveness. Regarding this, Addis Ababa City Government OCACC provides occupational competency assessment for all professionals in various occupations, however, the required manpower was not found based on federal TVET quality standards. Hence, the issue of factors on the candidate performance is not clear so far, while it could be possible to discover the feasible factors that affect the candidate's performance and realize the quality of the assessment process using data mining techniques.

Data mining is widely used in educational field to find out the problem in educational activity on educational environments. Student performance is of great concern in the educational institutes where several factors may affect the performance of student [21]. According to that, several researches conducts data mining techniques to analyze factors which is release students

performance and also they can be predict students' performance, attitude and interest. However, from the perspective of data mining, there is no research documented that analyze factors that affects candidates or professionals performance in vocational educational environment using data mining techniques. Thus, to discover the problems from the database that can be hidden and affects candidate's performance the following basic research questions are proposed.

1.4. Research questions

- RQ1. What are the patterns that are frequently associate each other in the candidate database?
- RQ2. Are the rules generated from frequent patterns are interesting and have influence on candidates' performance?
- RQ3. Has the discovered knowledge negative implication on the assessment process?

1.5. Research Objective

1.5.1. General objective

In light of the statement of the problem, the general objective of this study is to point out the dominant factors affecting the candidate's Performance during the assessment process using a data mining techniques.

1.5.2. Specific objectives

- Determine the dominant factors affecting the candidate's Performance during the assessment process using Association rule mining (ARM).
- Determine the dominant factors affecting the candidate's Performance during the assessment process using classification techniques.
- To justify the results of ARM based on classification technique

1.6. Significance of the Study

The purpose of this study is to find out and assess the major factors and magnifying the problems associated with the performance of candidates on the occupational competency assessment process and also the study has the following significances.

- **For Researchers**

The study points out new problems that don't get researchers attention and for further investigation. Likewise, other researchers can discover knowledge from the resulting data.

- **For OCACC**

The study is important to have awareness about the problems of occupational competency assessment process to take action such as re-evaluate their assessment process and also the outcome of the study allow closer follow up and more attention to quality during the process of assessment.

- **For Training Institute**

The study helps to aware of the Competency level of training institutes so that they can take -

some corrective measures in curbing the problems.

- ✓ To help to take measures on their training methods and their weaknesses.
- ✓ To help their institution is an excellence center.

- **For Government**

- ✓ Used as an indicator for the quality of training up to changing the Occupational Standard (OS) with the curriculum.
- ✓ The study gives evidence for governmental and non-governmental organizations that work in the area of TVET institutes.
- ✓ To make industries in a position to consume the knowledge, skilled and productive manpower for the country's development.
- ✓ To help to enhance the competence and competitive of TVET trainers and assessors.

- ✓ To help to reduce unemployment.
- ✓ To help to introduce some control mechanisms that ensuresfairness/ethics.

Moreover, the study is also promising in terms of filling the gap of assessment process quality. More importantly, the findings of this study can also lead to new problems for further investigation. As indicated before this study is a data mining approach, to find out factors that affect a candidate's performance, according to that by using different mining techniques and algorithms, these studies tend to be more useful for more investigation for future works.

1.7. Delimitation of the Study

The study is delimited on candidates who took occupational competency assessment at Addis Ababa OCACC.

1.8. Scope and Limitation of the Study

Occupational assessment service is given throughout Ethiopia except sum areas like "woreda". However, due to lack of infrastructure such as central database, the main focus of the proposed study is restricted on the City Government of Addis Ababa. Furthermore, the data for this research gathered a five years retrospective data only. Since before five years there is only head office which offers the assessment services. And also the numbers of fields or occupations are added in the given resent year. Likewise, the data collected from three branches which have a routine assessment from five branches. In terms of variables, the study is limited to point out the dominant factors affecting a candidate's competency performance.

As a limitation, the data stored in various places and there is no appropriate related work on this area of occupational assessment process using data mining techniques.

1.9. Operational Definition

- **Assessment**: is the means of determining if a candidate possesses the required competencies of an occupational qualification as stated in the Occupational Standard (OS). It is a process of collecting a piece of evidences and making a judgment on whether

competence has been achieved. It does not discriminate whether one acquires the competencies inside or outside the TVET institutions [11].

- **Candidate**: An individual seeking recognition of his/her competences to acquire a certification from OCACC to make themselves competitive in the labor market at a national level; Also it helps to increase their confidence rather than others, level of employment and self-employment will be increased [11].
- **Assessor**: An individual who meets the required qualifications to be authorized by the Center of Competence to assess whether a candidate possesses certain Competences or all the competencies defined by an occupational qualification level [11].
- **Competent (C)**:an assessment result is proven to have knowledge and skills for specific fields of study [11].
- **Not yet competent (NYC)**: An assessment result is given to a candidate by an assessor when the examiners believe that the candidate has not proven the possession and application of knowledge, skills and proper attitude to the standard of performance in the workplace [11].
- **Competence**: The possession and application of knowledge, skills and proper attitude to the standard of performance in the workplace [11].
- **Occupational Standard**: A standard defined by experts of the world of work indicating the competencies that a person must possess to be able to perform up to the expected level and be productive in the world of work [11].

1.10. Organization of the Thesis

The remaining of this thesis report is organized as follows. Chapter 2 presents a literature review and related works. This chapter discusses researches that have been conducted on educational data mining, data mining methods, algorithms, techniques. The methodology of the thesis work, data description, and process model also presented in chapter three. The experiment results, evaluation, and discussion are described in Chapter 4. Lastly, in Chapter 5 conclusions, recommendation and future works are pointed out.

CHAPTER TWO: LITRACHER RIVIEW

2.1. Machine Learning and Data Mining

Machine learning (ML) and Data mining (DM) methods are research areas of computational science whose rapid growth due to the advances in information analysis research, growth in the database industry and the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores [22].

Machine learning (ML) investigates how computers can learn (or improve their performance) based on data. Instead of that the main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions [23]. On the other hand, Data mining can be analyzing data from different perspectives and summarizing it into useful information. And it is an analytical tool for analyzing, categorize, and summarize the relationships among data.

Data mining satisfies its main goal by categorizing valid, potentially useful, and easily understandable correlations and patterns present in existing data. This goal of data mining can be fulfilled by modeling it as either Predictive or Descriptive nature. The Predictive model works by making a prediction about values of data, which uses known results found from different datasets which means it determines what might happen in the future. In order to that, this model desires larger data set expertise and toolset. In general, the tasks include the Predictive data mining model classification, prediction, regression, and analysis of time series. The Descriptive model generally identifies relationships in datasets. It serves as an easy way to explore the properties of the data examined earlier and not to predict new properties [24]. Generally descriptive model describes what happened in the past. These are generally pre-canned reports. Particularly Prediction is often referred to as supervised data mining while descriptive data mining includes the unsupervised and visualization aspects. The data mining model illustrated in the following Figure 2.1:

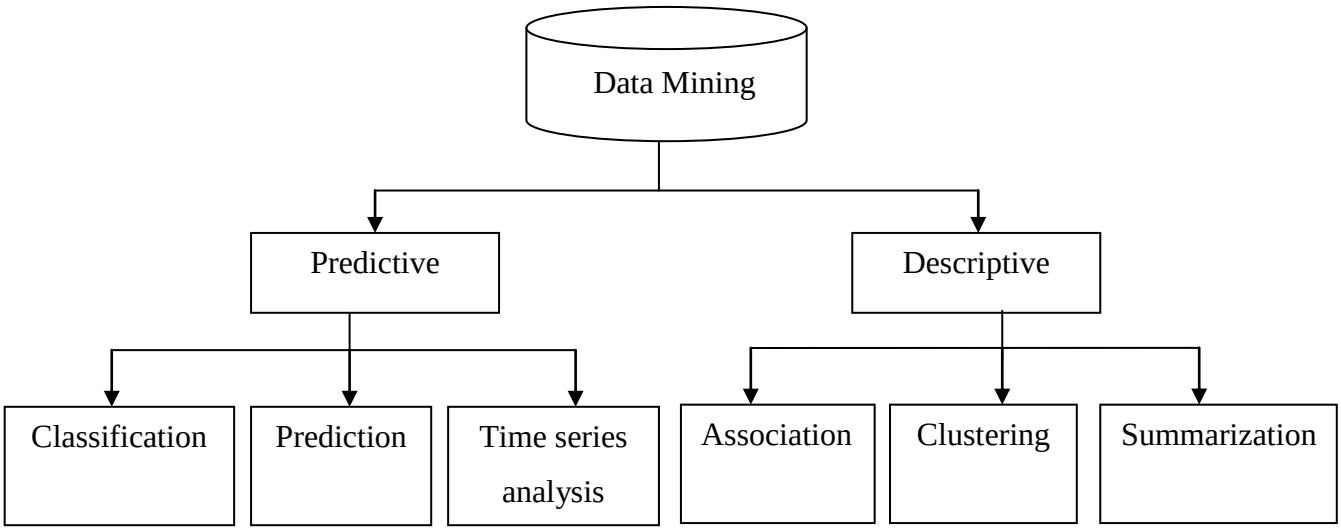


Figure 2.1: Data Mining Models

2.2. Knowledge Discovery from Data

As a whole, data mining is the process of extraction interesting non-trivial, hidden, previously unknown and potentially useful patterns or knowledge from a huge amount of data. And it can be expressed as Knowledge discovery from huge data, which means in a database there is knowledge extraction, data/pattern analysis, data archeology, information harvesting, etc. [25].

Data mining also a step in the knowledge discovery process, which means in knowledge discovery there is Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation which is the process of visualization and knowledge representation techniques are used to present the mined knowledge to the user. Based on that data mining is the fifth and essential process to discover or extract knowledge from huge data by using intelligent methods and techniques in order to extract data patterns [25] and [26].

Knowledge Discovery in Database (KDD), also refers to extracting or “mining" knowledge from large amounts of data and also it is an organized process of identifying valid, novel, useful and understandable patterns from large and complex datasets [27]. The KDD process can be decomposed into the following steps as illustrated in Figure 2.2:

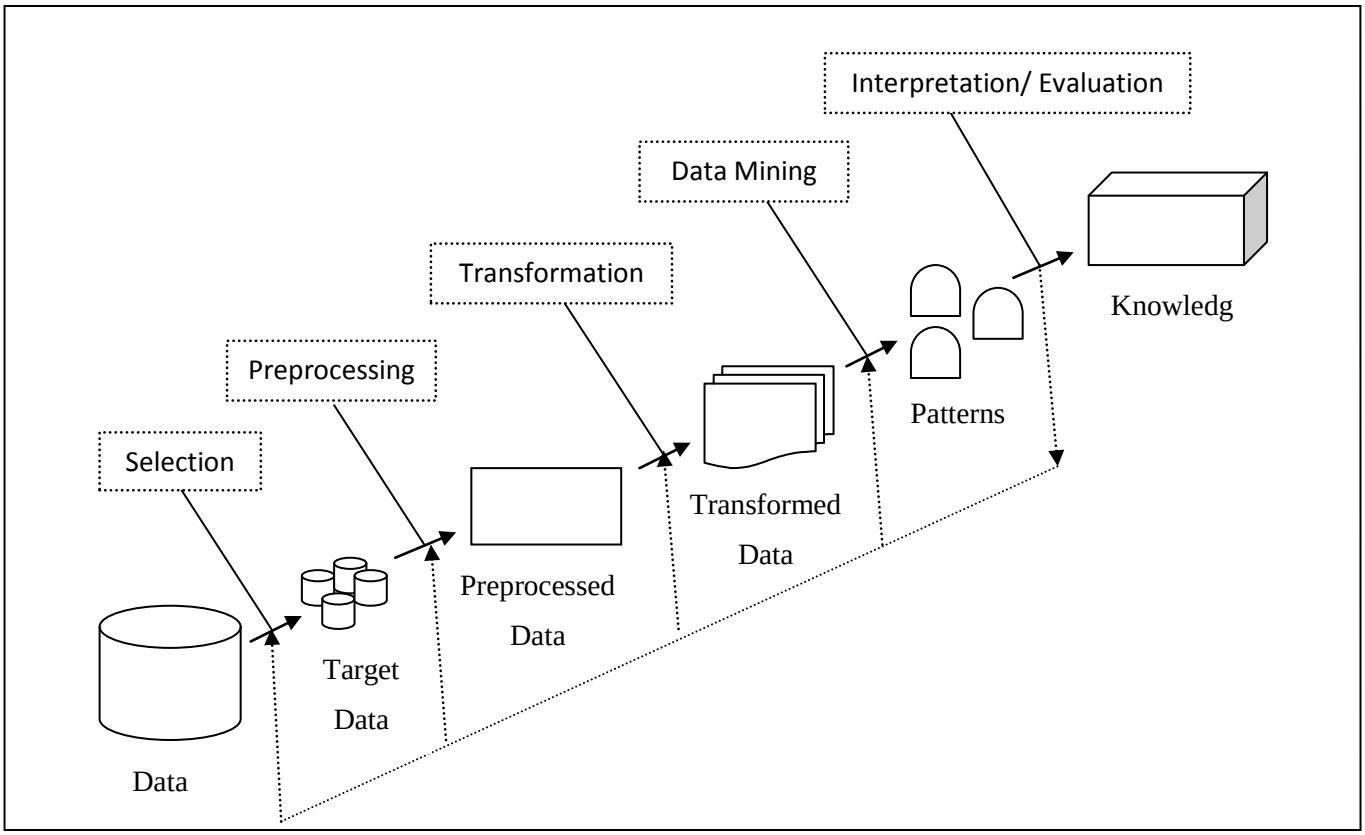


Figure2.2: Data mining as a step in the process of knowledge discovery [28]

- **Data Selection** - At this stage, a target dataset relevant to the domain of application is selected or created from many different and heterogeneous data sources.
- **Data cleaning/Preprocessing** - High-quality data is a requirement for the KDD process to produce reliable knowledge. Hence this stage includes tasks such as missing value imputation, removal of noise or outliers, mapping of feature values onto appropriate domains and attributes discretize to improve the quality of data contained in the dataset.
- **Data reduction** - In a high dimensional dataset, all attributes are not equally significant. The task of information reduction is to categorize prime attributes and take away non-relevant ones from the dataset by conserving the knowledge concerned within the dataset as much as possible to increases the quality of the mined results.
- **Mining** - It is a process to extract hidden information from real-world datasets. Depending on the goals of the data discovery method like prediction or description, this step applies algorithms to extract the knowledge from the transformed data.

- **Interpretation and evaluation** - The knowledge obtained as a result of performing a data mining task must be correctly interpreted and properly evaluated to ensure that the resulting information is meaningful and accurate. Also, it is to be presented to the users in an understandable manner. Visualization as a whole, it is the processor a technique used to show the complex results of a data mining task in a simplified manner.

The application of information mining technique has been broadly applied in several business areas such as health, education and finance for the purpose of data analysis and then to support and maximize the organizations' customer satisfaction in an effort to increase loyalty and retain customers' business over their lifetimes [22][23][27][29], and[30]. In general data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data and transactional data[23].

2.3. Data Mining Techniques, Algorithms and Tools

2.3.1. Data Mining Techniques

To identify valid, useful and easily understandable correlations and patterns presented in existing data, Data mining techniques have a great role[31]. There are several data mining techniques have been used including classification, clustering and association rule. In addition to that, there are a wide variety of applications in real life and various tools are available which supports different algorithms [31][32]. The following is a description of the main techniques, algorithms, and tools used in the area of EDM.

2.3.1.1. Classification

Classification is one of the data mining techniques which are useful for predicting group membership for data instances. This approach frequently employs a decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. Basically, there are two types of attributes available that are output or dependent attribute and input or the independent attribute. In general classification test data are used to

estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data rows[31][30] and [33].

2.3.1.2. Clustering

As usual, clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. In educational data mining, clustering has been used to group students according to their behavior, performance and activates. In general, the clustering technique defines the categories and puts objects in every category, while in the classification techniques, objects are assigned into predefined classes[32][9].Among the important and interesting parts of clustering techniques; it is document clustering and linguistic information acquisition[34].

Document clustering:

- Improving precision and recall in information retrieval.
- Browsing a collection of documents for purposes of information retrieval (scatterand gatherer strategy)
- Organizing the results provided by the search engine
- Generation of document taxonomies (cf. YAHOO!)

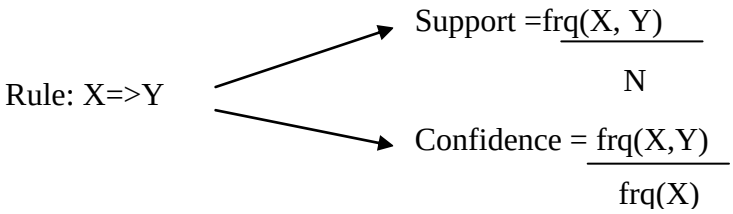
Linguistic Information Acquisition

- Induction of statistical language models
- Generation of sub categorization frame classes
- Generation of Ontology
- Collocation classification[34].

2.3.1.3. Association Rule

Association analysis is the discovery of association rules showing attribute-value conditions that occur oftentimes along in an exceedingly given set of information. Association rules area unit if/then statements that facilitate to uncover relationships between unrelated information in a database, relational database or alternative information repository.

On the other hand, association rules are usually needed to satisfy user-specified minimum support and user-specified minimum confidence at the same time [33][28] and [35].



Association rule mining and frequent item-set analysis invent in market basket analysis 20 years ago whereby trends and patterns in consumer purchasing activities could be identified and used to increase profit[36]. Association rules mining generally a process of finding all frequent item sets and then generates strong association rules from the frequent item-sets[36]. But the problem of finding association rule is also found in all frequent item-sets using minimum support and find association rules from frequent item-sets using minimum confidence[37]. The following diagram will show how to generate association rules.

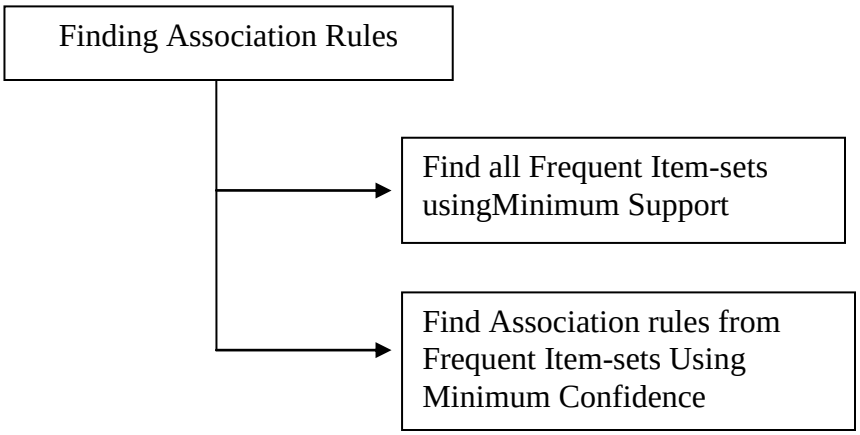


Figure2.3:Generating Association Rules [37]

2.3.2. Data Mining Algorithms

There are different algorithms presented over time for classification, clustering and for generating association rules. Among these, under the classification technique, there are ID3 Algorithm, Neural network, C4.5 Algorithm, Naïve Bayes Algorithm, SVM Algorithm, and ANN Algorithm. Under the clustering technique, there are major clustering algorithms, which

are Hierarchical clustering, Data partitioning and Data grouping [34]and under Association rule mining algorithm there are AIS, Partitioning algorithms;Apriori, Apriori-TID, Apriori Hybrid, Fp-growth, and TertiusApriori Algorithms are presented [37].

2.3.2.1. Classification Algorithm

ID3 Algorithm

The id3 calculation starts with the unique or original set as the root hub. On every cycle of the algorithm, it emphasizes through every unused attribute of the set and figures the entropy (or data pick up IG (A)) of that attribute. At that time chooses the attribute that has the minimum entropy (or biggest data gain) value. The set is "S" then break up by the selected attribute (e.g. marks < 50, marks < 100, marks >= 100) to produce subsets of the information. The algorithm proceeds to recourse on each and every item in the subset and considering only items never selected before. Features of the ID3 algorithm, it produces a more accurate result than the C4.5 algorithm and Detection rate is increase and space consumption is reduced [38].

C4.5 Algorithm

C4.5 is an algorithmic rule won't to produce a decision tree that is the associate expansion of previous ID3 calculation. It enhances the ID3 algorithmic rule by managing each continuous and discrete property, missing values and pruning trees once construction. The decision trees created by C4.5 can be used for grouping and often referred to as a statistical classifier. C4.5 creates decision trees from a set of training data the same way as the Id3 algorithm. As it is a supervised learning algorithm it requires a set of training examples which can be seen as a pair: input object and the desired output value (class).The algorithmic rule analyzes the training set and builds a classifier that has to have the capacity to accurately prepare each training and test cases.Among features of the C4.5 algorithm, the build Models can be easily interpreted, Easy to implement, Can use both discrete and continuous values and Deals with noise [39].

Neural Network

The field of Neural Networks has arisen from diverse sources ranging from understanding and emulating the human brain to broader issues of copying human abilities such as speech and can

be used in various fields such as banking, legal, medical, news, in classification program to categories data as intrusive or normal. Generally, neural networks consist of layers of interconnected nodes where each node producing a non-linear function of its input and input to a node may come from other nodes or directly from the input data. Also, some nodes are identified with the output of the network [39].

Naïve Bayes Algorithm

The Naive Bayes Classifier technique is based on the Bayesian theorem and is especially used once the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is conjointly possible to add new data at runtime and have a much better probabilistic classifier. A naive mathematician classifier considers that the presence (or absence) of a specific feature (attribute) of a class is unrelated to the presence (or absence) of the other feature when the class variable is given. For example, a fruit may be measured as an apple if it is red and round. Even if these features depend upon one another or upon the existence of other options of a class, a naive mathematician classifier considers all of those properties to independently contribute to the likelihood that this fruit is an apple.

The algorithm works as follows, Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c).

Naive Bayes categorified considers that the impact of the worth of a predictor (x) on a given class (c) is independent of the values of different predictors.

Among feature of Naive Bayes algorithm, it is Simple to implement, Great Computational efficiency and classification rate and it predicts accurate results for most of the classification and prediction problems [39].

$$\frac{p(c|x) = p(x|c)p(c)/p(x)}{p(x)}$$

$$p(c|X) = p(x_1|c) \times p(x_2|c) \times \ldots \times p(x_n|c) \times p(c)$$

Where:

- $p(c|X)$: Posterior probability
- $p(x|c)$: Likelihood

- $p(c)$: Class prior probability
- $p(x)$: Predictor prior probability

SVM Algorithm

SVM has attracted a great deal of attention in the last decade and actively applied to various domain applications. SVMs are usually used for learning classification, regression or ranking function. SVM is based on statistical learning theory and structural risk minimization principle and has the aim of determining the location of decision boundaries also known as hyper plane that produce the optimal separation of classes. Maximizing the margin and therefore thereby making the most important potential distance between the separating hyper plane and the instances on either aspect of it has been established to reduce an upperbound on the expected generalization error. The efficiency of SVM based classification does not directly depend on the dimension of classified entities. Though SVM is that the most robust and correct classification technique, there are several problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally expensive, as solving quadratic programming methods require large matrix operations as well as time-consuming numerical computations. Training time for SVM scales quadratically within the range of examples, thus researches try all the time for an additional efficient training algorithmic rule, leading to many variants based mostly algorithms. Among feature of Support vector machine algorithm, it has high accuracy and Works well even if data is not linearly separable in the base feature space[39].

ANN Algorithm

Artificial neural networks (ANNs) are types of computer architecture inspired by biological neural networks (Nervous systems of the brain) and are used to approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are presented as systems of interconnected “neurons” which may figure values from inputs and are capable of machine learning moreover as pattern recognition because of their adaptive nature. An artificial neural network operates by creating connections between many different processing elements each corresponding to a single neuron in a biological brain. These neurons could also be truly made or simulated by a computing device system. Each neuron takes many input signals then based on an internal weighting produces a single

output signal that is sent as input to another neuron. The neurons are strongly interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output. In general, one or more hidden layers are sandwiched in between the two. This structure makes it not possible to forecast or recognize the precise flow of knowledge. Among feature of Artificial Neural Network algorithm: It is easy to use, with few parameters to adjust, a neural network learns and reprogramming is not needed, Easy to implement and Applicable to a wide range of problems in real life[39].

2.3.2.2. Clustering Algorithm

Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy (a tree of cluster). Among the two of hierarchical clustering there are Agglomerative and Divisive techniques.

- Agglomerative (bottom-up) techniques starting with one point (singleton) clusters and recursively merging two or more most similar clusters to one parent cluster until the termination criterion is reached.
- Divisive (top-down) techniques starting with one cluster of all objects and recursively splitting each cluster until the termination criterion is reached[34].

Data Partitioning

Among popular data partitioning algorithm there area K-means algorithm and Density-Based Algorithm.

K-means algorithm

General characteristics of the approach: each of the *K* clusters $C_j$is represented by the mean (or weighted average) C_j of its objects, the centroid. And the clusters are iteratively recomputed to achieve stable centroids. An accepted and popular measure for the "intra-cluster variance" is the square-error criterion:

$$E = \sum_{i=1}^{k} \sum p \in c_i ||p - m_i||^2$$

(With m_i as the mean of cluster C_i)[34]

Algorithm2.1: Basic K-Means Algorithm:

1. Select K data points as the initial centroids.
2. (Re) Assign all points to their closest centroids.
3. Recomputed the centroid of each newly assembled cluster.
4. Repeat steps 2 and 3 until the centroids do not change[34].

Density-Based Partitioning algorithm

The basic idea is; an open set of data points in the Euclidean space can be divided into a set of components, with each component containing connected points (i.e., points which are close to each other)[34].

Algorithm 2.2: Basic density-based algorithm

1. Compute the E-neighborhoods for all objects in the data space.
2. Select a core object "*CO"*.
3. For all objects "*co"*E"*CO"*: add those objects "*y"* to "*CO"* which is "density connected" with "*co"*.
   Proceed until no further "ys" are encountered.
4. Repeat step 2 and 3 until either all core objects have been processed or any of the non-processed core objects is connected to another previously already processed object[34].

2.3.2.3. Association Rule Algorithm

AIS

The AIS algorithmic rule was the first revealed algorithm developed to get all massive itemsets in a transaction database. And this algorithm was proposed by Agrawal, Imielinski, and Swami for mining association rule. It focuses on improving the quality of databases together with the necessary functionality to process decision support queries. The AIS algorithmic rule makes multiple passes over the whole database. During each pass, it scans all transactions. In the 1st

pass, it counts the support of individual items and determines which of them are large or frequent in the database. Large itemsets of every pass are extended to get candidate item sets. After scanning dealing, the common item sets between large itemsets of the previous pass and items of this transaction are determined. To make this algorithm additional efficient, an estimation method was introduced to prune that item sets candidates that have no hope to be large, consequently, the unnecessary effort of counting those itemsets can be avoided. In addition to that, in the AIS algorithm, the frequent itemsets generate by scanning the databases several times. The drawback of this algorithm is, it requires more space and it wastes much effort. As well as this algorithm needs several passes over the total database [28][40].

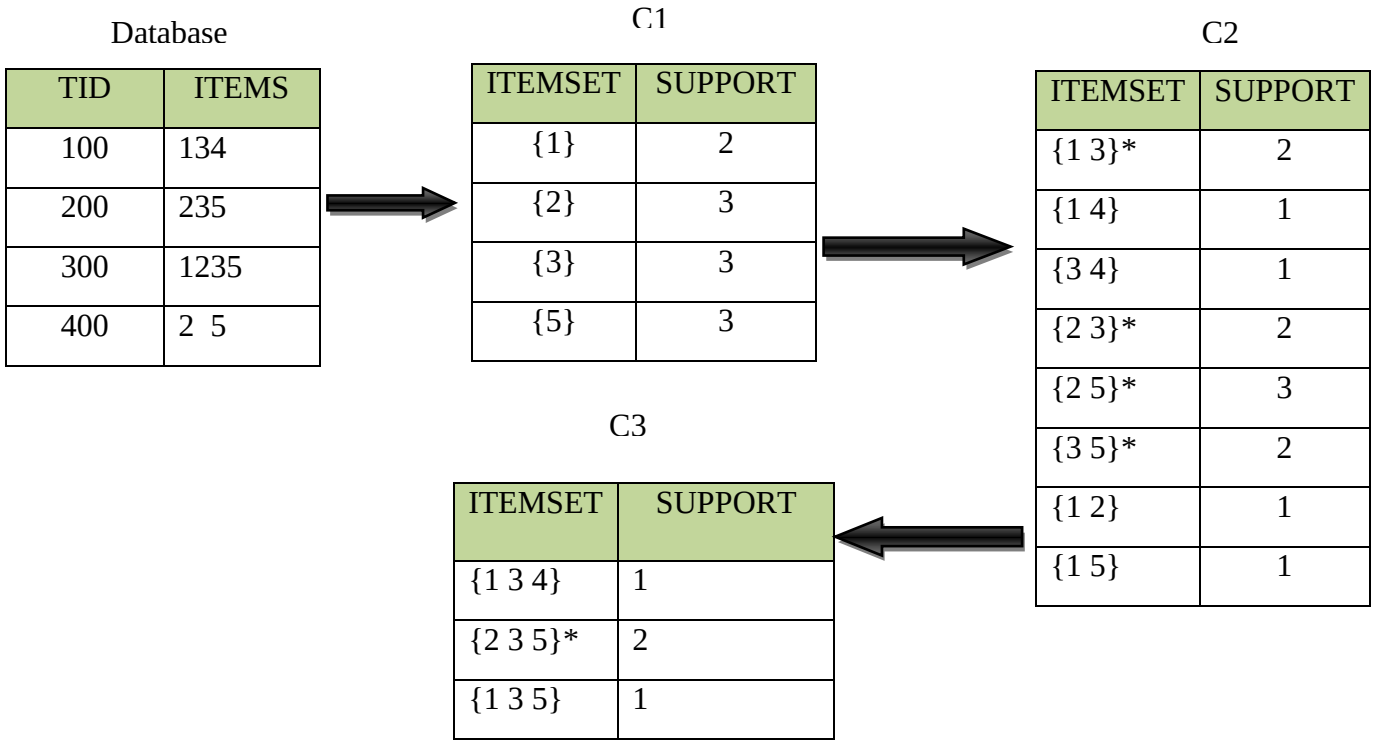


Figure2.4: Sample item set in AIS algorithm [28]

SETM

Similar to the AIS algorithmic rule, the SETM algorithmic rule makes multiple passes over the info. In addition to that, the SETM remembers the TIDs of the generating transactions with the candidate item sets.In the SETM algorithm, candidate itemsets are generated on-the-fly as the database is scanned, but counted at the end of the pass.

Then new candidate itemsets are generated the same way as in the AIS algorithm, but the transaction identifier TID of the generating transaction is saved with the candidate item set in a sequential structure. It separates the candidate generation process from counting. At the top of the pass, the support count of candidate itemsets is determined by aggregating the sequential structure. The SETM algorithm has the same disadvantage as the AIS algorithm. Another disadvantage is that for every candidate itemset, there are several entries as its support value[28][38].

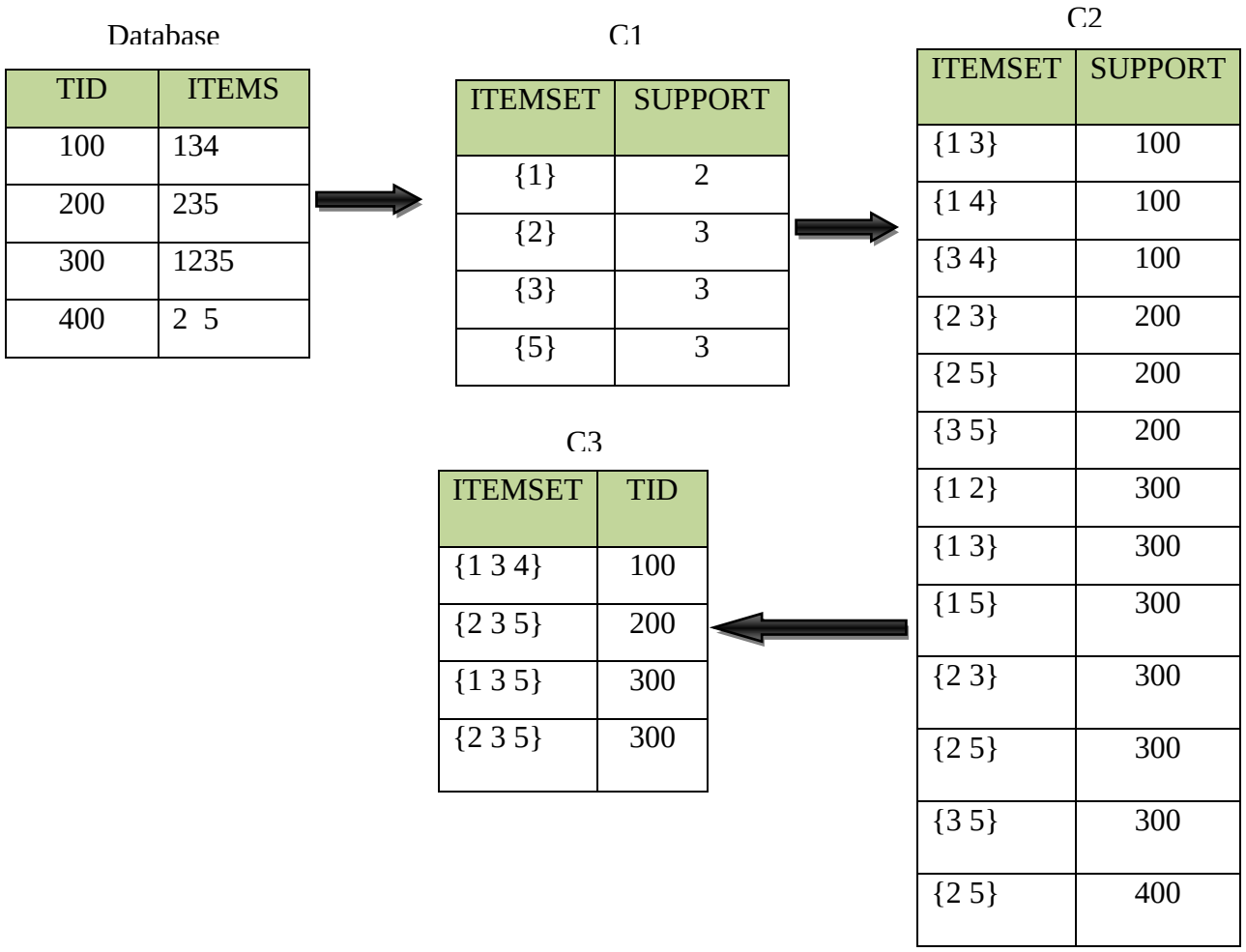


Figure2.5:Sample item set in SETM algorithm [28]

Apriori Algorithm

It is one of the most well-known association rule algorithms. The fundamental differences of this algorithm from the AIS and SETM algorithms are the way of generating candidate item-sets and the selection of candidate

itemsets for counting. The Apriori generates the candidate item-sets by connection the massive item-sets of the previous pass and deleting those subsets that square measure little within the previous pass while not considering the transactions in the database, frequent subsets are extended one item at a time and this step is known as the candidate generation process. Then teams of candidate's square measure tested against the data. To count candidate item sets with efficiency, Apriori uses breadth-first search method and a hash tree structure. It identifies the frequent individual things within the database and extends them to larger and larger item sets as long as those item sets seem sufficiently usually within the database.

Apriori algorithmic rule confirms frequent item-sets which will be wont to determine association rules that highlight general trends within the database. Apriori uses pruning whereas guaranteeing completeness. The Apriori algorithmic program is predicated on the Apriori principle. The algorithm for use a level-wise search, k-item sets exist used to explore (k+1)-item sets, to mine frequent item-sets from the transactional database for Boolean association rules. There are two drawbacks to the Apriori algorithm.

- First is the complex candidate generation process which uses most of the time, space and memory.
- The second one is, it requires multiple scans of the database[28][38][41].

Algorithm2.3: Procedure for Apriori algorithm[40]

1. CIk: Candidate itemset having size k
2. FIk: Frequent itemset having size k
3. FI1 = Frequent itemset;
4. For (k=1; FIk! = null; k++) do
5. BeginCIk+1 = candidates generated from FIk;
6. For each dealing t in info D do

7. Increment the count price of all candidates in
8. CIk+1 that area unit contained in t
9. FIk+1 = candidates in CIk+1 with min_support
10. End
11. Return FIk

a) C1

ITEMS	COUNT NUMBER
I1	7
I2	8
I3	6
I4	2
I5	3
I6	1

b) L1

LARGE 1 ITEMS
I1
I2
I3
I5

c) C2

ITEMS	COUNT NUMBER
I1, I2	5
I1, I3	4
I1, I5	3
I2, I3	4
I2, I5	3
I3, I5	1

d) L2

LARGE 2 ITEMS
I1, I2
I1, I5
I2, I5
I2, I3
I1, I3

e) C3

ITEMS	COUNT NUMBER
I1, I2, I5	3
I1, I2, I3	2

Figure 2.6: Sample item sets in Apriori algorithm [28]

Apriori _Tid Algorithm

In this algorithm, the database is not used for counting the support of candidate item-sets after the first pass. The process of candidate item set generation is the same as the Apriori algorithm.

Another set C’ is generated by which each member has the TID of each transaction and the large item-sets present in this transaction. The set generated i.e. C’ is used to count the support of each candidate item set [28].

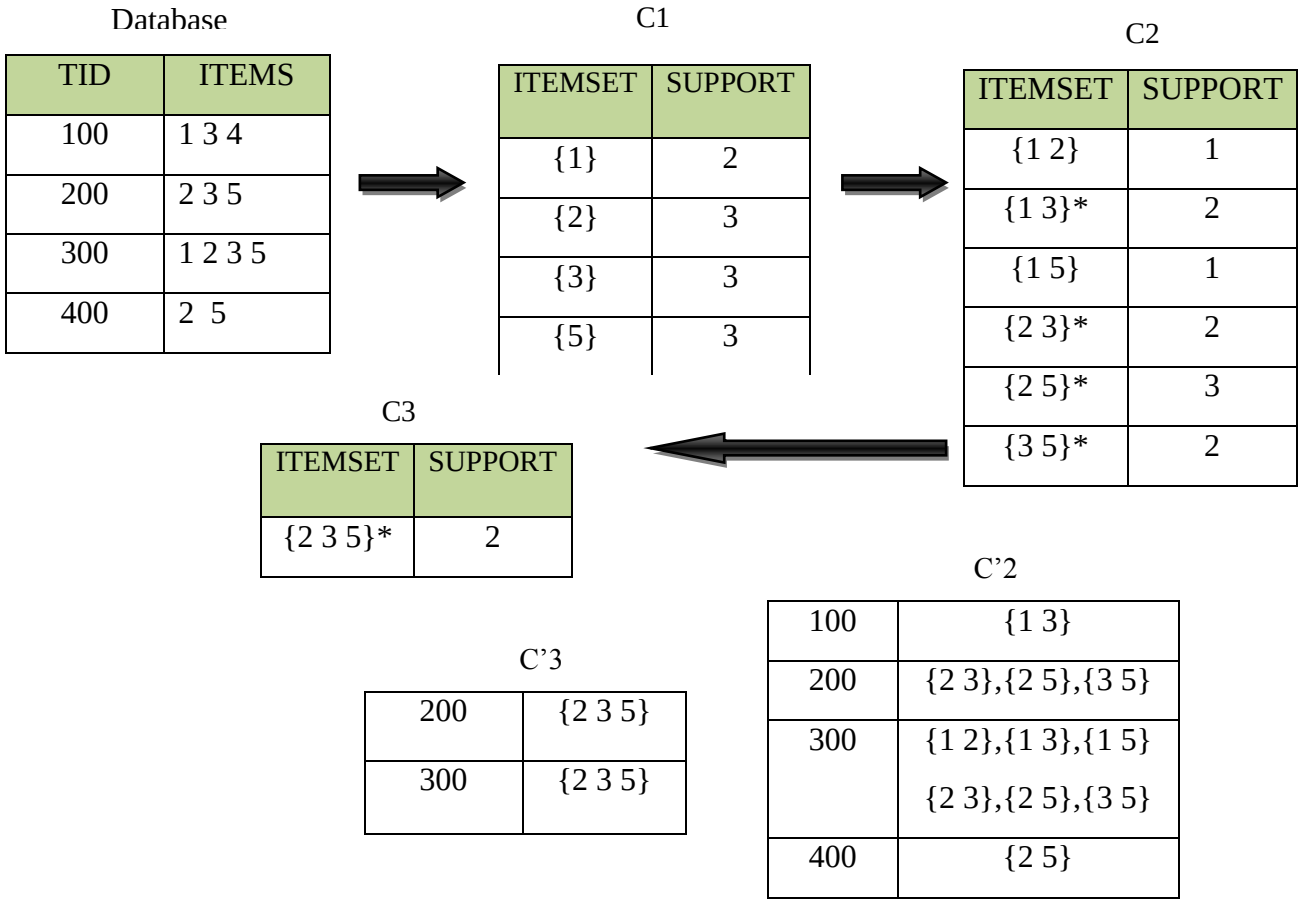


Figure 2.7: Sample item sets in Apriori _Tid Algorithm[28]

Apriori hybrid algorithm

Apriori hybrid uses features of both Apriori and Apriori-tid algorithms. It uses the Apriori algorithm in earlier passes and Apriori_Tid algorithm in later passes [28].

FP-Growth algorithm

FP-Growth algorithm has a capacity to mine accurately a frequent item sets, with better performance and scalability. It can be better to meet the requirements of big data mining and

efficiently mine frequent item-sets and association rules from the large datasets[28]. FP-growth requires constructing FP-tree. For that, it requires two passes and scans. It first computes a list of frequent items sorted by frequency in descending order (F-List) and during its first database scan. In the second scan, the database is compressed into an FP-tree. This algorithm performs mining on FP-tree recursively. There is a problem of finding frequent item-sets that are converted to searching and constructing trees recursively. The frequent item-sets are generatedwith solely two passes over the database and with none candidate generation method.

There are two sub-processes of frequent patterns generation method that includes: construction of the FP-tree, and generation of the frequent patterns from the FP-tree.

FP-tree is constructed over the data-set using 2 passes are as follows: Pass 1: 1) Scan the data and find support for each item. 2) Discard infrequent items.3) kind frequent items in descending order that relies on their support.

By using this order we will build FP-tree, in order that common prefixes will be shared[28].

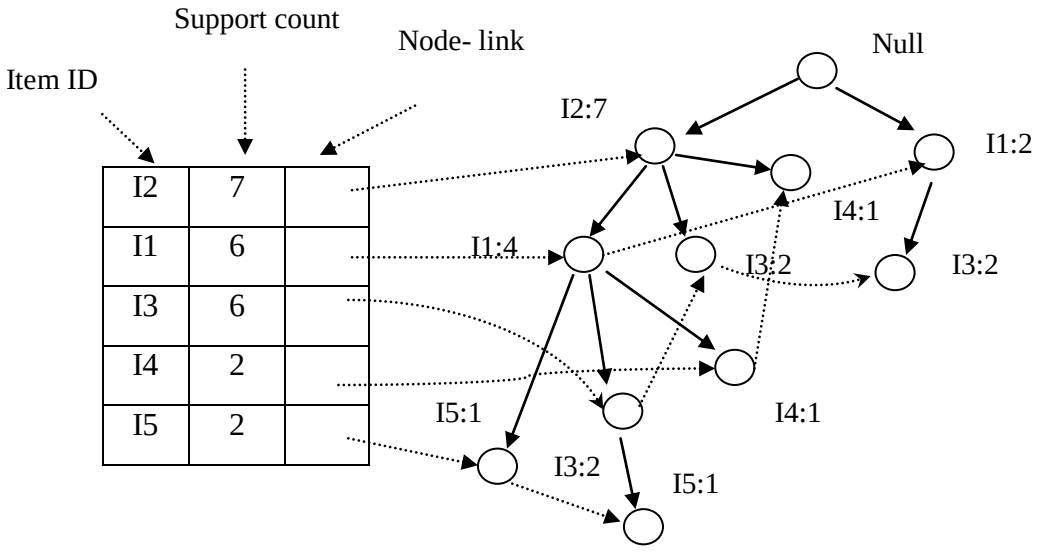


Figure 2.8: Sample for FP-Tree[40]

The advantages of the FP-Growth algorithm relative to the Apriori algorithm are scanning the transaction database two times; no candidate item set is generated in the mining process. But for massive data mining frequent patterns, the FP-Growth algorithm has many disadvantages, such

as excessive memory and low efficiency. The parallel algorithm of traditional MPI parallel programming model has shortcomings of node failure and network communication failure [42]. On the other hand, FP-growth uses a compressed representation of the database thus the irrelevant information is pruned. However, it cannot be used for interactive and incremental mining system as changes in threshold value or new insertions in the database may lead to a repetition of the whole process if we employ the FP-tree method [37].

FP-tree is made over the data-set exploitation pair of passes are, as follows: [41]

Pass 1

- Scan the info and realize support for every item.
- Discard rare things.
- Type frequent things in downward order that is based on their support.
- By exploitation this order we will build FP-tree, so common Prefixes will be shared.

Pass 2

- Here nodes correspond to things and it's a counter.
- FP-growth reads one dealing at a time then maps it to a path.
- Mounted order is employed, so methods will overlap once transactions share the things. In this case, counters are incremented. Some pointers are maintained between nodes that contain identical item, by creating on an individual basis coupled lists. The lot of methods that overlap, higher the compression. FP-tree could slot in memory. Finally, frequent item sets are extracted from the FP_Tree.

Algorithm 2.4: The Procedure FP-growth (Tree T, A)[41]

1. If Tree T contains a single path P,
2. Then for each combination of the nodes in the path P do Generate pattern B U A with Support = minimum support of nodes in B
3. Else for each Hi in the header of the
   tree T do
   {Generate pattern B=Hi U A with
   Support = Hi. Support;
   Construct B's conditional pattern base and B's conditional

FP_Tree that is B;

4. If Tree B ≠Ø
5. Then call FP-growth (Tree B, B)}

2.3.3. Data Mining Tools

According to[31][39]there are various open source tools available for data mining. Some of the tools work for clustering, some for classification, regression, association and some for all.

- **R**- This is an open source language used for statistical and data analysis. It can run in multiple platforms (e.g. Windows, Mac OS or Linux) and supports all kinds of techniques and algorithms'.
- **WEKA**- It stands for Waikato Environment for knowledge analysis. It is non-propriety, freely available, and application-neutral standard for data mining projects. It is widely adopted in academic and business and has an active community and it also contains tools for data preprocessing, classification, clustering, association rules, and visualization but it is not capable of multi-relational data mining.
- **Orange**- It is a free data mining tool that also supports data visualization and also analyzes data. Data mining is done through visual programming or Python scripting but not support time series analysis and big data processing.
- **Tanagra**- It is an open source environment for teaching and research and is the successor to the SPINA software but the basic drawback of other tools; Tanagra das do not support text analysis, big data processing, and time series. The following table shows tools that support data mining techniques.

Table 2.1: Data Mining Tools [39]

	R	WEKA	Orange	Tanagra
K-Means Clustering	Supports	Supports	Supports	Supports
Regression	Supports	Supports	Supports	Supports
Naïve Bayesian Classification	Supports	Supports	Supports	Supports
Decision Tree	Supports	Supports	Supports	Supports
Time Series Analysis	Supports	Supports	Does not Supports	Does not Supports
Big Data Processing	Supports	Supports	Does not Supports	Does not Supports
Text Analysis	Supports	Supports	Supports	Does not Supports

2.4. Educational Data Mining

Educational data mining (EDM) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to investigate due to the massive amount of data within which they exist. And it is concerned with developing methods for exploring the unique types of data that come from educational environments [43]. On the other hand, EDM refers to techniques, tools, and research designed to automatically extract meaning from large repositories of data generated by or related to people's learning activities in educational settings. And it also an evolving discipline concerned with developing approaches for discovering relationships in the unique and increasingly large-scale data that come from educational domains, and using such approaches to discover patterns and makes predictions that characterize learner behavior and achievement, domain content knowledge, assessment outcomes, educational functionalities, and applications [44].

The field of EDM provides new information that would be difficult to distinguish by simply looking at the raw data and the objective is to process meaningful information about learning for continual pedagogical improvement.

Accordingly, EDM has been broken into four phases. In the first phase, relationships between data are discovered by using statistical techniques such as classification, regression, clustering, factor analysis, social network analysis, association rule mining, and sequential pattern mining. In the second phase, the relationships are then theoretically validated. In the third phase, the validated relationships are used to make predictions about phenomena in future learning contexts. In the final phase, these predictions are used to support pedagogical and policy-level decisions for improved student outcomes [44].

EDM involved with developing strategies for exploring the distinctive kinds of information that return from academic environments and analyze information generated by any form of data system supporting learning or education (in schools, colleges, universities, and other academic or professional learning institutions [43].

2.5. Related Works

Data mining techniques are used in many applications. The effect and future trends have been stated. Many users have designed prediction systems using these techniques based on that several studies addressed the analysis of educational data to get useful information that affects learning quality [31][32]. According to that in [45] applied the Association rule mining technique using the Apriori algorithm to analyze student's performance and also improvising the subject teaching in that particular subject. Based on that Apriori Algorithm found necessary since a large number of Association rules or patterns or knowledge is generated from the large volume of a dataset. But most of the association rules have redundant information and thus all of them cannot be used directly for an application. So pruning or grouping rules by some means are necessary to get very important rules or knowledge. In the experiment, each field of specialization offers students five subjects during each semester within a period of four years and the overall marks obtained by students in the different subjects are utilized in their experiment in finding related subjects. The main objective is to determine the relationships that exist between different courses offered as this is required for optimizing the organization of courses in the syllabi. Instead of that this

experiment performed by collecting a student obtains similar marks for related subjects hypothesis and This assumption is justified by the fact that similar marks obtained by a student for different subjects imply that the student is meeting the expectations of the course similarly and applied association rule mining to determine the relationships of student marks with attendance.

In the same direction, authors in [46] discuss association rule discovery techniques improve the quality of education by analyzing the data and discover the factors that affect the academic results. By using Association rules discovery techniques, compare student's performance in the subject's common at Graduation and Post-Graduation level and to predict the factors which can explain their success or failure. Instead of that Apriori Algorithm used for mining frequent itemsets and then generate strong association rules from the frequent item sets. The required data for the study is taken from MCA course of a renowned university in Haryana which consists the results of the subjects which are common in their Graduation and Post-graduation degree programs and also data processing made since in data analysis which has a significant impact on the accuracy of the predicted results and as a data mining tool Tanagra were used to mine Association rules.

In [32], conducts two or more techniques by combining together to make an appropriate process to identify causes that concern academic performance using data mining techniques, based on that by using Association rule mining and SPSS as a data mining tool to analyze the collected data, hidden patterns find out from students and course records, then using these patterns or causes as early predictor to expected success rate and handling their weakness. The study was conducted a two years retrospect data and 150 samples of the population and "12" attributes were selected to identify the course and also "7" attributes from students record. Instead of that correlation is done between course and student to find out the features which have an obvious effect on success rate than a simple data mining based prediction model presented, such as both classification and clustering techniques to identify features affecting student's performance.

In [47] uses a classification method to predict students placement into the department. On the other hand, this study mainly concerned with how the department placement is held and how many of the students are placed to the departments based on their choice of preference. To

undertake the study the data was taken from the registrar office of Gondar University and the data set contains about 1496 instances of placed students. After that to make the data more suitable for the particular data mining software, preprocessing and cleansing is made. For that matter, MS Excel was used for preprocessing and to build the predictive model J48, Naïve Bayesian, and Random Forest classification algorithms were experimented by using Weka software. Based on that, the Random Forest algorithm more appropriate for the data given to build the predicting model for department placement with high accuracy, And supplied test data is better rather than using cross-validation since supplied test data has better accuracy when we are using these algorithms'.

In [48] the authors applied a classification model to predict students' dropouts through their university programs. In general 1290 records are selected from computer science students between 2005 and 2011 for the subject of the study, since the major goal of this study is classifying whether the student is dropped or graduated from the major. Having these To classify and predict dropout students, Decision Tree (DT) and Naive Bayes (NB) have experimented on the data set. Based on that to test the accuracy of each model, 10-fold cross-validation is used and selected Decision Tree (DT) classification algorithm based on the accuracy they found from the experiment. On the other hand, the study also includes discovering hidden relationships between student dropout status and enrolment persistence by mining frequent cases using the FP-growth algorithm.

In [49] the authors apply classification technique based on a decision tree to predict student's dropout rates and causes of dropout in the beginning stage of their study either before or after completion of their first-year study program. The data is collected from undergraduate ICT Courses at Residential University and the collected data set contained 220 instances. Having this association rule mining technique was used to discover the relationship between unrelated variables in a large database. On the other hand dropout risk factors are identified by using the data mining attributes such as frequent patterns, correlations, similarity measures, associations rule mining. Finally, by using the "R" programming language, the Classification model was implemented.

2.6. Summary

In this chapter, the general background and distinction regarding ML and DM have been discussed. In addition to this, to fulfill the mining process or knowledge extraction DM uses either predictive or descriptive nature. Theoretical backgrounds about knowledge extraction processing that are applied to EDM such as preprocessing steps are discussed. DM techniques, algorithms, and tools also widely expressed which are used for EDM to find out useful and easily understandable correlation and patterns in existing data. When we come to related work, several related works are discussed which have similar functionality with the study. The strength and weaknesses of the related works are discussed. Likewise, as indicated in chapter one there is a research gap, which is there is no research, documented related to occupational assessment quality and it factors in the use of data mining technique (DMT) and machine learning (ML). The summary of related work addressed as follows.

Table 2.2: Summary of related works

No	Index	Title and Year	Methodologies/ Algorithm	Findings	Gap/Strength
1	[45]	Finding Association Rule using Apriori Algorithm on Educational Domain.(April, 2015).	Used ARM technique to analyze student's performance, having that Apriori algorithm was used to associate student's dataset.	By Appling Apriori algorithm, the study found strong relations among students attendance and students mark.	As strength, the study indicates, pruning or grouping rules by some means are necessary to remove redundant rules.
2	[46]	Mining Association Rules in Student's Assessment Data (September 2012).	By using Association rules discovery techniques, compare student's performance in the subject's common at Graduation and Post-Graduation level.	Apriori Algorithm used for mining frequent item sets and then generate strong association rules from the frequent item sets to predict the factors .	As a weakness', the study used, Tanagra as a data mining tool to mine Association rules. Since Tanagra does not Supports big data analysis.
3	[32]	Predicting Student Academic Performance in KSA using Data Mining Techniques. (2017)	By using ARM, handled patterns which have a frequent correlation.	By combining classification and clustering techniques, the studycan be able	Strength conducts two or more techniques by combining together to make an appropriate process to identify

				to identify features affecting student's performance.	factors Weakness SPSS as a data mining tool to analyze the collected data, and also the study used a small sample of data for analysis.
4	[47]	Application of Data Mining Techniques to Predict Students Placement in to Departments, (*2016*).	The study uses a classification method to predict student's placement into the department.	As a findings, Random Forest algorithm more appropriate for the data given to build the predicting model for department placement with high accuracy	As strength, the study used J48, Naïve Bayesian, and Random Forest classification algorithms were experimented by using Weka softwareto build the predictive model.
5	[48]	Data mining in higher education: university student dropout case study, (January 2015).	Applied a classification model to predict students' dropouts through their university programs.	To discover a frequent item-sets, the study used FP-growth algorithm. And also the study selected Decision Tree (DT) classification	Strength For the experiment, the study compares different algorithms.

				algorithm based on the accuracy found from the experiment.	
6	[49]	Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue, (*January 2015)*	The study applied classification technique based on a decision tree to predict student's dropout rates and causes of dropout in the beginning stage of their study either before or after completion of their first-year study program.	Association rule mining technique were used to discover the relationship between variables. Having that, dropout risk factors are identified by using the data mining attributes such as frequent patterns, correlations, similarity measures.	<u>Strength</u> Maintaining prototype for the generated rule.

CHAPTER THREE: METHODOLOGY

To achieve the objective of the study, the following research methodologies were employed, including investigating whether there are similar researches proposed which is similar functionality. Besides that, the study made a deep study in the literature written on the EDM area to have a clear picture of the work before starting the actual work. Papers written on educational data mining reviewed to get an understanding of the various techniques, methods, and algorithms of educational data mining.

3.1. Study Design

In this study, the descriptive data model has been employed to analyze factors. Similarly, the predictive data model has been conducted. Since to evaluate the efficiency of the proposed algorithm, also to make a predictive model as a supportive for descriptive results as a justification. Having that, the proposed association rules mining algorithm which is the Apriori done first then the preferred classification algorithm approaches were tested and compute with the Apriori algorithm to evaluate the efficiency and valuable of apriori algorithm. Accordingly, the proposed architecture encompasses all processing components and adopted the following methodology from CRISP-DM, which most mining process needs to follow. The CRISP-DM (CRoss Industry Standard Process for Data Mining) process model aims to make large data mining projects, less costly, more reliable, more repeatable, more manageable, and faster[50].

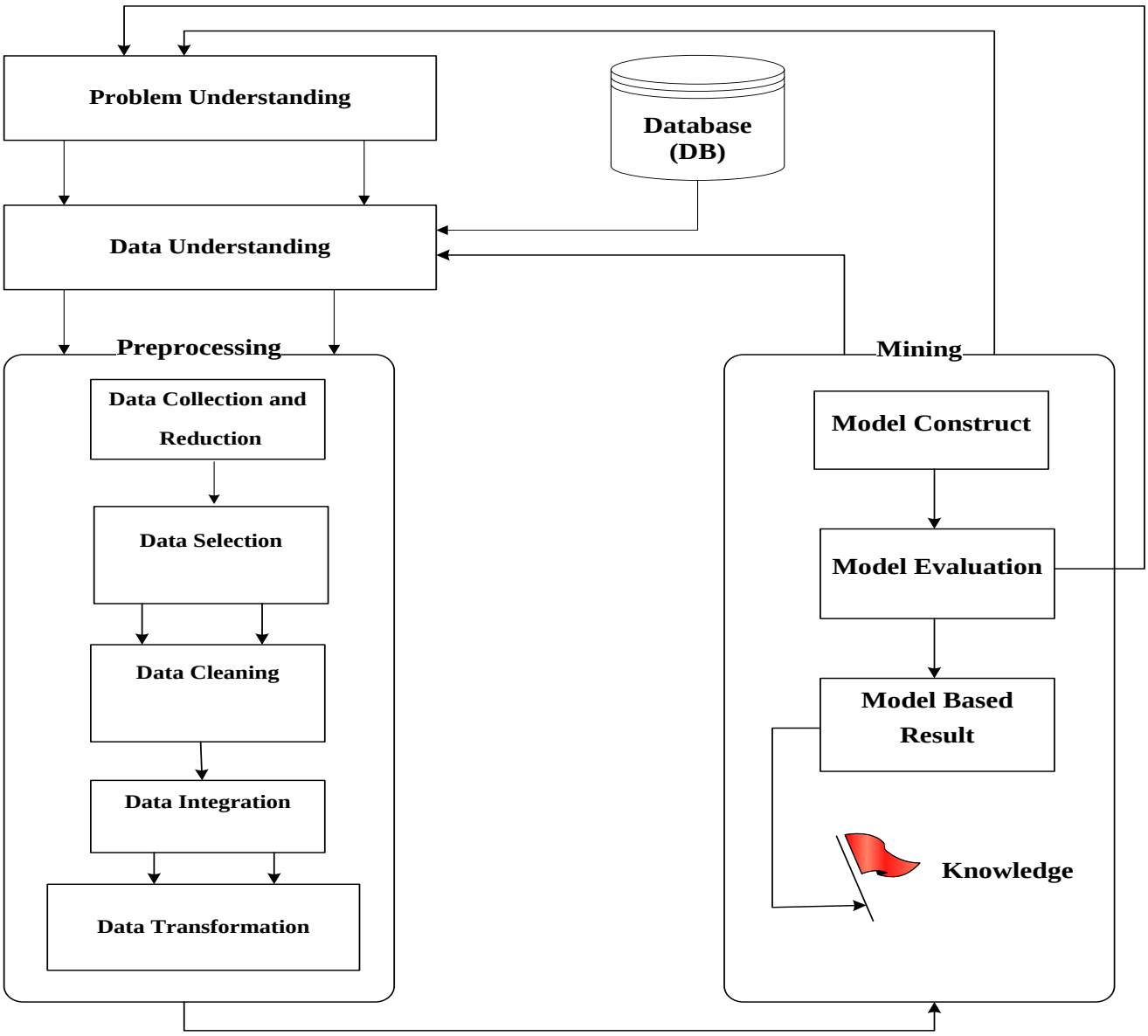


Figure 3.1: The proposed data mining process

3.2. Problem Understanding

The first step of the model is to understand the domain that needs to be addressed. In order to achieve the goal, different tasks are carried out such as discussion with domain experts to identify the domain problem. In the study, the researcher identifies OCACC center experts and Assessors to define the domain problem and determine the data mining goal.

3.3. Data Source

Addis Ababa city administration Occupational Competency Assessment and Certification Center (OCACC), have five branches. These branches also have their own database. According to that, all candidate demographic information is registered and stored OCACC database. For the experiment, this database is used as a data source. Quantitative data are required to obtain descriptive information to investigate the issue under consideration and the data source obtained from OCACC Smart Candidate Management System (SCMS) database and Access database, according to that, the collected data contains about 324,236 instances and 36 Attributes.

3.3.1. Source population

The target population of this study is all professionals male and female, who are registered for the assessment in Addis Ababa OCACC, not only TVET trainees.

3.3.2. Study population

The study conducts all candidates, male and female who are registered in OCACC for the assessment in the last five years (2014-2018).

3.3.3. Sample size and Sampling technique

As indicated before this study is knowledge discovery or data mining. Hence, for the experiment the study conducts convenience sampling technique. From the collected data 321,236 data were used.

3.3.4. Inclusion criteria

- Candidates those who are registered, scheduled, Assessed and submitted their results in the database in a given five years.

3.3.5. Exclusion criteria

- Candidates those who are registered but not scheduled, Assessed and submitted their results in the database in a given five years. According to that 3000 data were excluded from 324,236 data.

3.4. Data Understanding

Next to identifying the problem we proceed to the central item in the data mining process which is data understanding. This includes listing out attributes with their respective values and proceed visualize the actual data which is stored in the database. Data visualization technologies can display the data graphically which stored in the databases. Much research has been conducted on visualization and the field has advanced a great deal especially with the advent of multimedia computing[51].

The main goal of data visualization is to correspond information clearly in actual fact through graphical means. It doesn't mean that data visualization needs to look uninteresting to be functional or extremely sophisticated to look beautiful. Data presentation can be beautiful, elegant and descriptive. There is a variety of conventional ways to visualize data, such as histograms, pie charts, and bar graphs are being used every day, in every project and on every possible occasion [47].

To present the actual data stored in the database clearly, data visualization techniques were used.

The city government of Addis Ababa OCACC has 732,023 candidates record who took (OCA) in the various areas from 2000 EC up to 2010 EC in all branches. According to that 383,247 can be able to pass the assessment in the given year [18].

To find out the major factors, three branches were selected which provide a routine assessment. Based on that, (324,236) records were collected, among this "3000" records were discarded since the registered candidates were not assessed in various cases, such as not scheduled, in case of absence, etc. The following table and figure clearly show the status of the required data.

Table 3.1: The collected data set

Competent	Not yet competent	Not assessed	Total
132573	188663	3000	324,236

Description

- 58% of assessed candidates are not passed the assessment.
- Only 41% of assessed candidates are passed the assessment

- The remaining 0.9% of registered candidates is not assessed in several cases.

Table 3.2: Types of assessment with their status

Type of Assessment	Competent (C)	Not Yet Competent (NYC)
Practical	75864	106078
Both	28959	67153
Knowledge	726	1188
Unit of Competency (UNC)	24165	13654

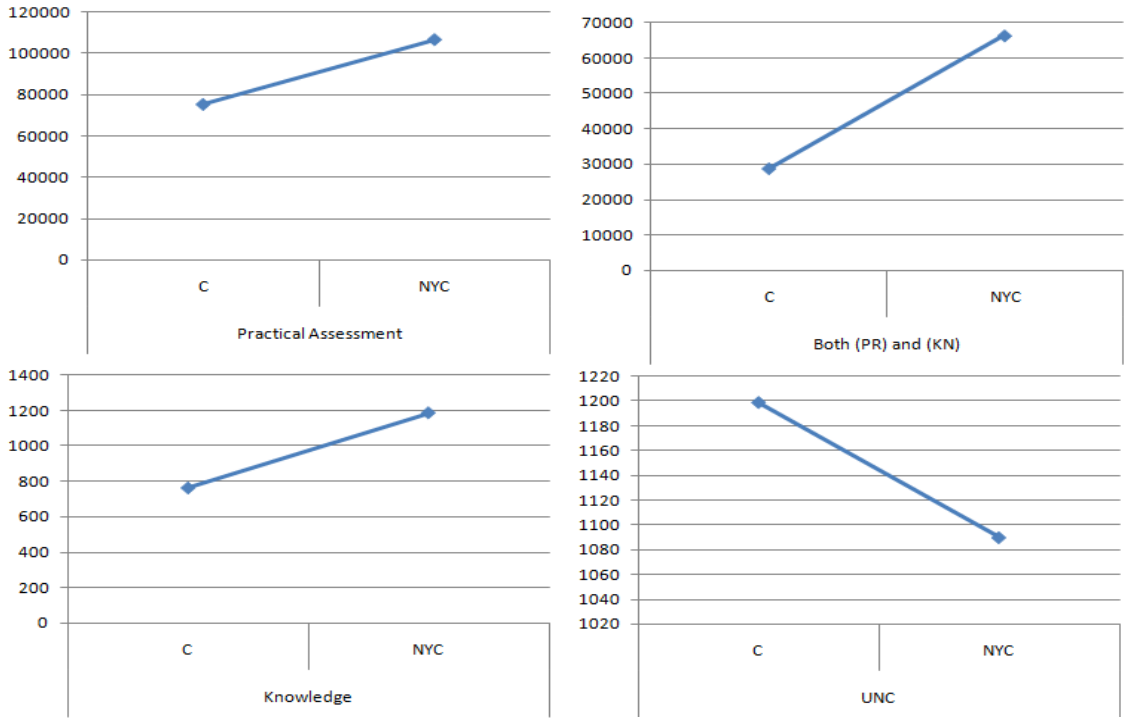


Figure 3.2: The degree of fall in each assessment type

Description

- It is known that Addis Ababa OCACC classifies the assessments into two main assessments, such as competency-based and qualification based assessment, in addition to that qualification based assessment classified into three types of assessments, namely Practical assessment, Knowledge assessment and Both practical and Knowledge assessment at a time. In the case of competency-based assessment, the candidates assessed only a unit of competency (UNC) from a given level. According to qualification based assessment, the candidate assessed all competencies from the given level and to continue the next level the candidate should have to pass the existing level. Based on that the containing table show performance of candidates and on which type of assessment they are well and vice versa.

Table 3.3: Qualification based assessment result

Qualification based assessment result	
C	105302
NYC	173851

Description

- From the above table, the degree of failure is higher in the case of qualification based assessment.
- All qualification based assessments are difficult for candidates.

Table 3.4: Competency based assessment result

Competency based assessment result	
C	24165
NYC	13654

Description

- From the above table, the degree of flirty is lower in the case of competency based assessment.
- Approximately competency-based assessments are not difficult for candidates.
- The following charts more describe the above figure 3.3 and 3.4.

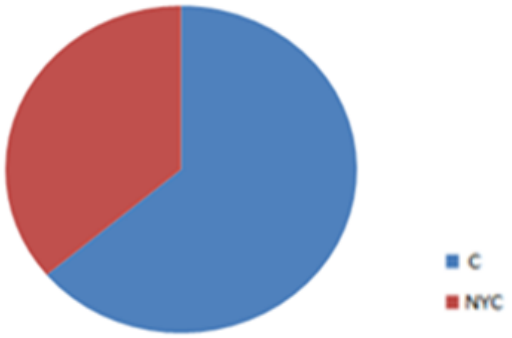


Figure 3.3: Competency Based Assessment

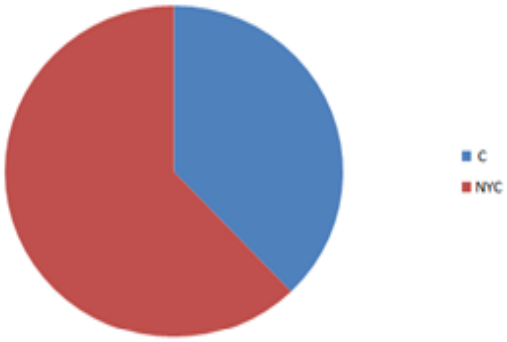


Figure 3.4: Qualification Based Assessment

3.5. Preprocessing

After the data has collected, tasks such as processing and cleaning are imposed to make the data more suitable for the particular data mining software, which is used in the study. This includes attribute selection which means attributes for the association, handling noisy data, accounting for missing data fields, and preparing the processed data in a file format acceptable to mining

software. On the other hand, the data selection process conducted to make the data mining process more valuable and to have a better evaluation, the selection of relevant data from the database is more important. This means, there are several tables are found in the database, based on that to analyze the candidate's performance, candidates' demographic table and the required parameter for the mining process selected. In this study there for data integration and transformation conducted since the data is gathered from different places and different database sources. In the case of gathering from different sources and places, the data analysis task involved data integration. Having that, the source data collected from three branches and all data are stored in SQL and Access database. Instead of that, all the three branches SQL data exported into Excel format and the Access data transferred into Excel data. Finally, all Excel data merged.

3.5.1. Data Collection and Reduction

Data collection is the most important and essential stage to acquire the fine and final data that can be taken as correct and suitable for further data mining tasks [52]. Based on section three on 3.3, the study is conducted in Addis Ababa city administration Occupational Competency Assessment and Certification Center (OCACC), to obtain the quantitative data; the data is obtained from OCACC SCMS and Access database. Having this, there are five branches allocated in different places in Addis Ababa city and the organization has no central database, according to that a five-year retrospective data were collected, which means (2014 to 2018) from three branches. The reason for time frame selection and branch selection is several occupational standards (OS), occupational competencies' (OC) and four branches were added in this given year, which means before five years ago there is only one branch the organization had. And also the selected branches have a routine assessment. Regarding this, 324,236 records were collected.

3.5.2. Description of the collected data

Description of the data is very important in the data mining process in order to clearly understand the data. Without understanding, useful analysis cannot perform. As indicated before, this study is done by collecting the data from the OCACC SCMS database and Access database.

In this section, some collected data sources described above, the attributes with their data types and descriptions are shown in the following table 3.5 below. The attributes have different data types like String, Nominal and Numeric data type.

Table 3.5: Partial attributes with their description

No	Attributes	Data type	Description
1	Registration Number	String	The candidate registration number
2	Name	String	Name of candidates
3	Sex	Nominal	Gender of candidates
4	Age	Numeric	Age of the candidates
5	Nationality	String	Nationality of the candidates
6	Sub City	String	Sub city of the candidates that comes from
7	Marital Status	Nominal	Marital status of the candidates
8	College/Institute Type	Nominal	Type of training institute that the candidates comes from
9	College /Institute Name	String	Name of institute that the candidate comes from
10	Training Start	date	Year of training start

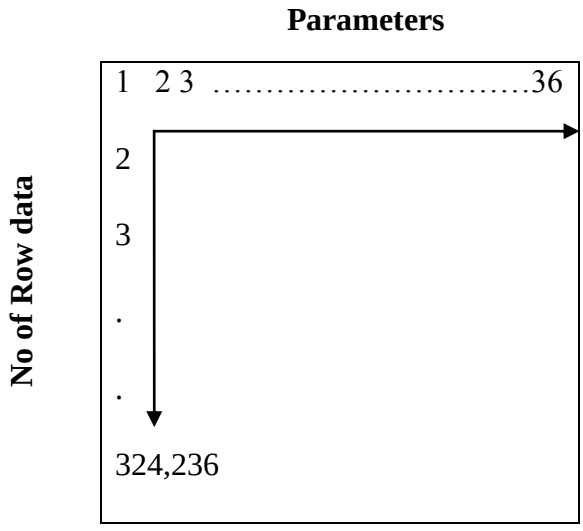


Figure 3.5: Number of Collected Rows and Parameters

3.5.3. Data Selection

Some of the attributes were ignored in the initial dataset that was not relevant to the data mining experiment goal. In this stage a target dataset were used. The whole target dataset may not be taken for the data mining task. Irrelevant or unnecessary data are eliminated before starting the actual data mining function. Having this, unrelated attributes that have no relevance to the data mining experiment were removed as shown in table 3.6.

Table 3.6: Partial reduced attributes

No	Attributes	Data type	Reason for Removal
1	Registration Number	String	Have no relation with candidates performance
2	Name	String	Have no relation with candidates performance
3	Nationality	String	Have no relation with candidates performance
4	College /Institute Name	String	Have no relation with candidates performance

Accordingly, as shown in the above table 3.6, attributes were removed which have no value to determine or analyze the issue under consideration. By applying the Data Selection preprocessing method in our data, 324,236 datasets and 10 attributes were selected for analyses which are important to determine the factors. The following table shows the remaining attributes.

Table 3.7: Selected attributes for the experiment

No.	Parameter Name	Description	Data Type
1	Gender	Gender of candidates from the registration record.	Nominal
2	Age	Age of the candidates.	Numeric
3	Marital status	Marital status of the candidates.	Nominal
4	Mode of training	Contains how the candidates learn their education.	Nominal
5	Education background	Contains candidate's level.	Nominal
6	Employment condition	Contains candidate's employment condition.	Nominal
7	Apply for UNC (Yes or No)	Describes candidate's application, either for unit of competency or qualification based.	Nominal
8	Apply for	Apply for practical, knowledge or both for practical and knowledge.	Nominal
9	Year gap	Indicates the year gap that candidates came to apply for the assessment after they accomplish their training.	Numeric
10	Status	Describes the final assessment result.(Competent or Not yet competent)	Nominal

In this research, based on the assessment of the existing situation and discussion with domain experts we derived one attribute (Year gap) from training end and date of application which is necessary to determine candidate performance.

3.5.4. Data Cleaning

This phase is used for making sure that the data is free from different errors. Otherwise, different operations like removing or reducing noise by applying smoothing techniques for that attribute are done [53].

According to the study, the collected data have some missing values and noisy data, having this to make available for mining applications; some missing values and noisy data handled by filling manually and removed some irrelevant data, such as null values. From this, remove missing values and smoothing noisy data from the data set performed.

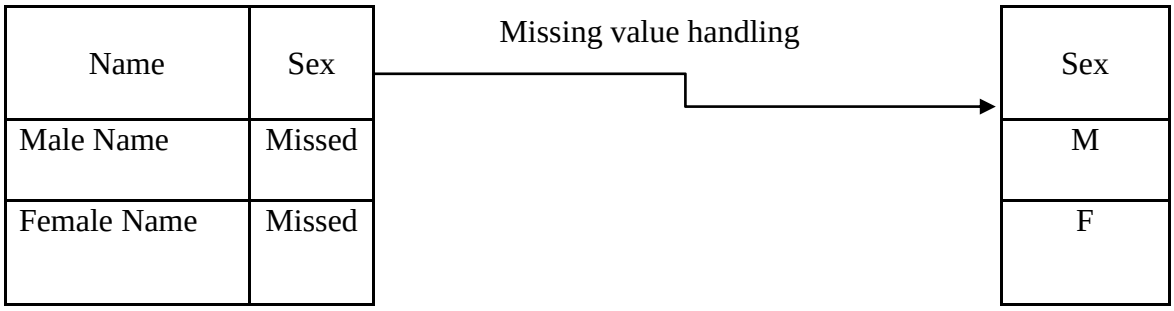


Figure 3.6: Highlighted Records of Partial Missing Value Handling

Description

From the above figure, some missing genders have been feeling by comparing the candidate's name. It is known that genders can be identified using the name. However, on the process of handling missing sex value, some common names were faced. Having that, these common names were removed. Since it is not possible to take either they are female or male to feel the missing value under attribute sex.

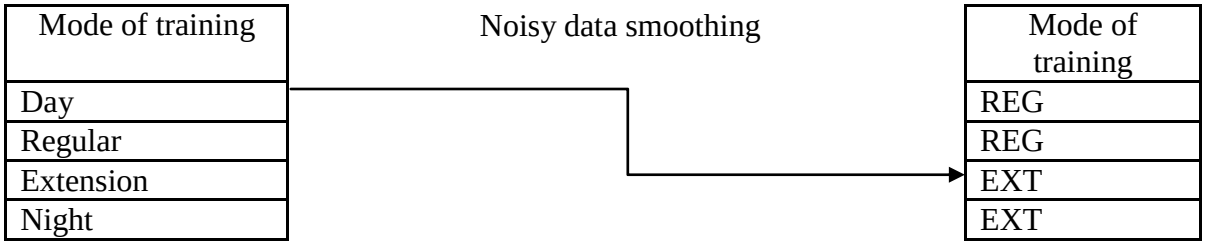


Figure 3.7: Highlighted Records of Partial Smoothing Noisy data

Description

From the above figure, noisy data have been handled by using the nature of the data. For example, from the above figure 3.10, mode of training has four different instances but these instances can be able to generalize using a common name. However, from the above figure (Day and Regular) handled by REG and (Extension and Night) handled by EXT. likewise, The selected data are cleaned further by removing the records that had incomplete (invalid) data and/or missing values under each column. Having that, from the selected instances which is 324,236 data 3000 instances are not available since these 3000 data or candidates are not scheduled for the assessment and have no results. According to that 321,236 instances are available for the data mining process. Data cleaning is done by using MS-Excel 2007. The selected data are presented below after cleaning preprocesses.

Table 3.8: Percentage of Missing values

No	Column name	# of Missing value	%
1	Gender	0	0
2	Age	862	0.27
3	Marital Status	10	0.003
4	Year Gap	33	0.01
5	Mode of Training	2228	0.69
6	Employment Condition	6351	1.97

7	Educational Background	13911	4.33
8	Apply for (Unit of competency)UNC	0	0
9	Applied for	0	0
10	Status	0	0

3.5.5. Data Integration

The data is gathered from different places and different database sources. According to that, the data analysis task involved data integration. For example: as figure 3.11, shows, the source data collected from three branches, these branches have not connected by the network infrastructure and there is no central database system. All data are stored in SQL and Access database. Instead of that, all SQL data exported, the exported and access data transfer into Excel data. Finally, all Excel data integrated with each other. The following diagram will describe the integration process.

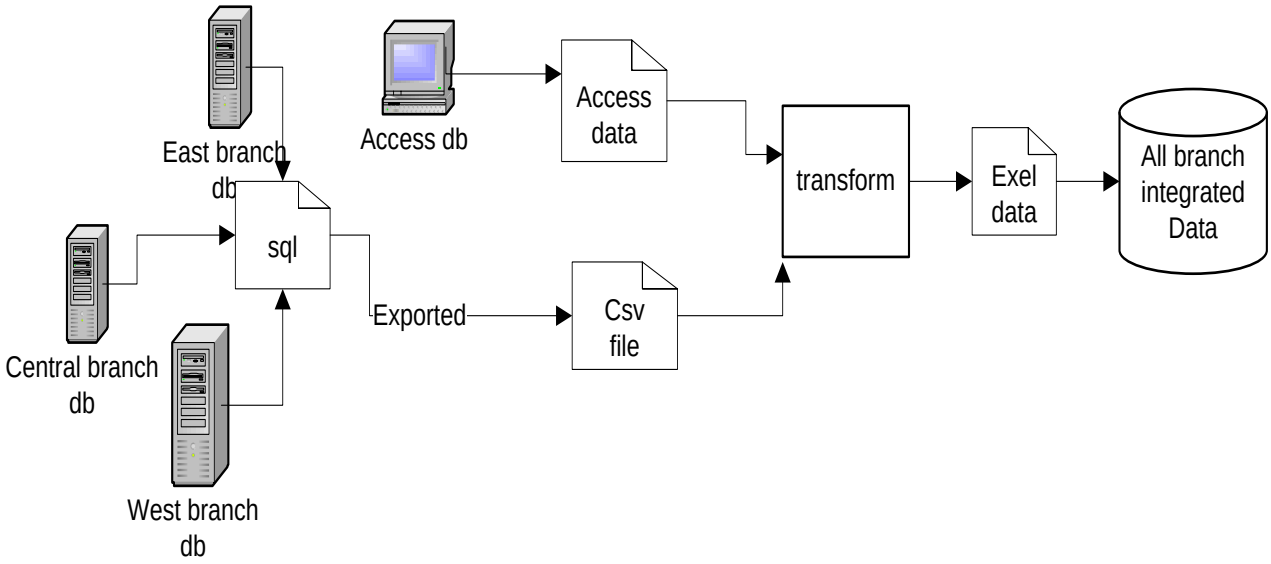


Figure 3.8: Data Integration Process

3.5.6. Data Transformation

According to[52]Data transformation is a process of converting source format data into a required format which is appropriate for mining. Several transformation techniques are applied to transforming source format to targeted format, they are,

- Smoothing: removes noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction [52].

Based on that, to make the mining process more suitable, the study used a data transformation process to make more efficient the data by converting it into an appropriate categorical format, and a categorical variable is constructed to make the attribute more valuable for the visualization process. The transformation process is shown in Table 3.9.

Table 3.9: Data transformation process

Attributes	Row data	Transformed data
Age	15-29	YNG (Young)
	30-49	ADT (Adult)
	=>50	OLD (Old)
Employment condition	Employed	EMP
	Unemployed	UNP
Marital status	Single	SNG
	Married	MRD
	Divorced	DIV
	Widow	WID

Status	Competent	C
	Not yet competent	NYC
Educational background	Formal	FRM
	Non formal	NFR
	Work experience	WEX
Gender	Male	M
	Female	F

Gender	Age	Marital Status	Year gap	Mode of training	Employment condition	Education background	Occ level	OCC Name	Apply for UNC	Status
M	YNG	SNG	3 YEAR	REG	UNEMP	LEVEL 2	LEVEL 2	AEES2	No	C
M	YNG	SNG	5 YEAR	REG	UNEMP	LEVEL 2	LEVEL 2	AEES2	No	NYC
M	YNG	SNG	5 YEAR	REG	UNEMP	LEVEL 2	LEVEL 2	AEES2	No	C
M	YNG	SNG	7 YEAR	REG	UNEMP	LEVEL 2	LEVEL 2	AEES2	No	C
F	YNG	SNG	1 YEAR	REG	EMP	LEVEL 2	LEVEL 2	AEES2	No	C
M	YNG	SNG	7 YEAR	REG	UNEMP	LEVEL 2	LEVEL 2	AEES2	No	C
M	YNG	SNG	5 YEAR	REG	EMP	LEVEL 2	LEVEL 2	AEES2	No	C
M	YNG	SNG	1 YEAR	EXT	UNEMP	DEGREE	LEVEL 2	AEES2	No	C
M	YNG	SNG	1 YEAR	EXT	UNEMP	LEVEL 3	LEVEL 2	AEES2	No	C
M	YNG	SNG	1 YEAR	EXT	UNEMP	LEVEL 3	LEVEL 2	AEES2	No	C
M	YNG	SNG	1 YEAR	EXT	UNEMP	LEVEL 3	LEVEL 2	AEES2	No	C
M	YNG	SNG	1 YEAR	EXT	UNEMP	LEVEL 3	LEVEL 2	AEES2	No	NYC

Figure 3.9: Highlight of Transformed records

3.6. Mining

Generally, data mining clarifies in section two as the extraction of remarkable non-trivial, concealed, previously unknown and potentially useful patterns or knowledge from massive data. On the other hand, it also expressed as Knowledge discovery from huge data. According to that, the study used association rule mining to identify valid, useful and easily understandable correlations and patterns present in the collected data among data mining techniques. Having that to generate strong association rules from the frequent item-sets the study found Apriori algorithm preferable to find combinations of item sets that occur frequently from a given database. Besides association rule mining, the study conducted a classification approach for farther analysis as described in section 3.1.

In these section different techniques, algorithms and tools applied for discovering interesting knowledge. To make data preprocessing the study used MS Excel, Visio 2007 was used for designing the proposed mining process. To mine hidden knowledge from the pre-processed dataset, WEKA software conducted to implement ARM and classification to analyze the collected data, WEKA (Waikato Environment for Knowledge Analysis) 3.8, were used. WEKA is chosen since it is it contains tools for data preprocessing, classification, clustering, regression, association rules, attribute selection and visualization. WEKA is written in Java programming language and contains a Graphic user interface (GUI) for interacting with data files and producing visual results. It also has a general Application Page Interface (API), so WEKA can be embedded like any other library in applications [55].

3.6.1. Model Construct

In the model construction phase, the study used association rule mining to find a strong correlation, and then a given rule constructed using the Apriori algorithm. On the other hand, the study constructs model using a classification algorithm to compute with the Apriori algorithm.

3.6.2. Model Evaluation

According to model construction, the generated rule evaluated using the proposed algorithms. Hence, the generated patterns are evaluated using their accuracy test criteria.

3.6.3. Model Based Result

Based on the model evaluation, the appropriate rules are selected, tested and result analyses are prepared.

3.6.4. Knowledge

Application of the discovered knowledge. It is the final discovered data from the database which is something interesting and non-trivial patterns.

Table 3.10: Description of the selected attributes

No.	Parameter Name	Description	Data Type
1	Gender	Gender of candidates from the registration record.	Nominal
2	Age	Age of the candidates.	Nominal
3	Marital status	Marital status of the candidates.	Nominal
4	Mode of training	Contains how the candidates learn their education.	Nominal
5	Education background	Contains candidate's level.	Nominal
6	Employment condition	Contains candidate's employment condition.	Nominal
7	Apply for UNC (Yes or No)	Describes candidate's application, either for unit of competency or qualification based.	Nominal
8	Year gap	Indicates the year gap that candidates came to apply for the assessment after they accomplish their training.	Nominal
9	Applied for	Describes candidate's application, either for practical assessment, Knowledge assessment or both practical and knowledge assessment.	Nominal
10	Status	Describes the final assessment result.(Competent or Not yet competent)	Nominal

All the attributes were assigned categorical values as it is shown below

1. **Gender** - M=MALE, F=FEMALE
2. **Age** - classified in to three categories such as, from 14-29=YNG which means Yang, 30-49= ADT, which means Adult and >=50=OLD

3. **Marital status** - classified in to four categories, such as MRD= married, SNG=single, DIV=divorced and WID=widow.
4. **Mode of training** - classified in to four categories, EXT= extension, REG= regular, DIST=distance and STT=short term training.
5. **Education background** - Degree, Diploma, Masters, Level 1-Level5, WEX=work experience.
6. **Employment condition** - EMP=employed and UNEMP= unemployed.
7. **Apply for UNC** - Yes or No
8. **Applied for** – PR, KN, BOTH
9. **Year gap**–NYG and YGP, Indicates the year gap that candidates to apply for the assessment after they accomplished their training.
10. **Status** - C = competent and NYC = not yet competent.

3.7. Summary

In this chapter, as a methodology, the study follows a descriptive model that is used to associate and summarizing the past data. In this study, more than one technique and lgorithmsalso used in order to compute and getting an accurate, preferable an acceptable result. The study also adopts the CRISP data model to design and describe the proposed data mining process. The design is composed of two components; preprocessing and mining. In these two components, there is another sub-process are presented. The relationships among these designed components and the responsibility of each component also presented. This section also described the collected data set and the whole mining process, which means the chapter described how the problem is understood next to how listing out the attributes which are needed for the mining process. Similarly, actual data presented graphically by using visualizing techniques to have a clear picture and understanding of the existing data. Thus in this study, the data was collected from different branches, according to that the integration, data transformation, and reduction have been conducted.

CHAPTER FOUR: EXPERIMENT AND RESULTS

4.1. Dataset

According to section three on 3.3, the set of data collected from OCACC databases. In this study, therefore, the collected data conducted the required data preparation procedures. In this case, to collect the exported data the study used Excel format and all preprocess events are accomplished by using Microsoft office Excel 2007. To run the experiments, WEKA Data Mining tool were used. In Section 2.3.3 and 3.6, a global description of WEKA is given.

4.2. Implementation

Experiments carried out using WEKA data mining tools; the analysis implemented using WEKA 3.8.3 for Association rule mining and Classification process, based on that, the exported data converted in to "CSV MS-DOS" format to manipulate the experiment. An example CSV files for the candidate's data is given in table4.1.

Table 4.1: Partial CSV data file

Gender	Age	Marital status	Year gap	Mode of training	Employment condition	Education background	Apply for UNC	Applied for	Status
F	YNG	MRD	YGP	EXT	UNEMP	LEVEL 2	No	PR	C
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 2	No	Both	NYC
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 3	No	PR	C
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 3	No	PR	NYC
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 3	No	Both	NYC
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 1	No	PR	NYC
M	YNG	MRD	YGP	REG	UNEMP	LEVEL 2	No	PR	NYC
M	YNG	MRD	YGP	EXT	UNEMP	LEVEL 1	No	Both	NYC
F	YNG	MRD	YGP	EXT	EMP	LEVEL 4	No	Both	NYC
F	YNG	MRD	YGP	EXT	UNEMP	LEVEL 1	No	Both	NYC
F	YNG	MRD	YGP	REG	UNEMP	LEVEL 3	No	Both	NYC
M	ADT	MRD	YGP	STT	UNEMP	LEVEL 3	No	Both	NYC
F	YNG	MRD	YGP	REG	UNEMP	LEVEL 1	No	Both	NYC

4.3. Experiment

To analyze the candidate's performance during the generation of association rule model, the study used Apriori Algorithm and for Classification Algorithm, Bayes Network, NaïveBays, Logistic Regression, Hoeffding tree, J48 Tree were conducted for the experiment. Instead of that to make the experiment 321,236 instances and 10 attributes were used for both classification and association model. The data has been experimented starting from (0.9 up to 0.7 confidence) to testify accuracy of the generated rule from association rule mining. Likewise, the data has been experimented by 10 fold cross-validation for 321,236 dataset (with 66% train and 34% test), were investigated during the generation of a classification model. Similarly, the data has been experimented with 80% train and 20% test using 321,236 dataset. In the experiments, variable "STATUS"was set as the independent or class variable and the remaining other attributes were set as dependent variables.

4.4. Result and Findings of the Study

The results obtained from various data mining algorithms i.e. for Association Rule Mining, Apriori Algorithm were conducted and for Classification Algorithm, Bayes Network, NaïveBays, Logistic Regression, Hoeffding tree, J48 Tree conducted. Having the generated set of rules the study selects the feasible rules which have a possible role to realize a candidate's performance. For apriori test, the study conducts 321,236 sets of data with 10 attributes. Using the Apriori Algorithm the study finds the association rules that have min Support=0.1(10%) and minimum confidence=0.9(90%).because this can generate more frequent item set. If we set minimum support more, then this can remove many attributes, but minimum number of attributes is not sufficient to give a proper decision[21].The summary of Apriori algorithm results as follows.

Table 4.2 : Apriori Algorithm results

Experment	*Results of Apriori Algorithm with 321,236 Instances and 10 Attributes*			
	Algorithm	Minimum metric (confidence)	Number of cycles performed	Generated sets of large Item sets
1	Apriori	0.9	15	5
2		0.8	10	4
3		0.7	9	3

For the selection of interesting rules, the study used minimum confidence 0.9. since as it has been indicated in the result of 321,236 instances from table 4.2, if the minimum metric or confidence is decline the number cycle and the set of large item sets although be decreases. Having that, to select rules for summarization, the study focused on the rules generated from the superset of frequent or large item set consisting of the highest size of large item set. The following are rules selected for summarization from experiment.

Interesting rules found:

(1) Mode_of_training=REG Apply_for_unc=No Status=NYC 91813 ==>Employment_condition=UNEMP 84582 <conf:(0.92)> lift:(1.06) lev:(0.02) [5101] conv:(1.71)

(2) Age=YNG Mode_of_training=REG Status=NYC 114182 ==>Employment_condition=UNEMP 105164 <conf:(0.92)> lift:(1.06) lev:(0.02) [6319] conv:(1.7)

(3) Apply_for_unc=No Status=NYC 135724 ==>Employment_condition=UNEMP 123419 <conf:(0.91)> lift:(1.05) lev:(0.02) [5925] conv:(1.48)

(4) 25. Age=YNG Mode_of_training=REG Applied_for=PR 133852 ==>Employment_condition=UNEMP 121655 <conf:(0.91)> lift:(1.05) lev:(0.02) [5782] conv:(1.47)

(5) Mode_of_training=REG Employment_condition=UNEMP Status=NYC 118345 ==>Year_gap=YGP 106661 <conf:(0.9)> lift:(1.22) lev:(0.06) [19466] conv:(2.67)

The meaning of those five large set of rules which are selected from 321,236 data set is, those candidates he/she have no year gap after accomplishing their training session, their status will be competent, which means the candidate can pass the assessment. And also those who have a year gap as well as no work experience it is difficult to pass the assessment. On the other hand candidates, those who are applied for qualification based assessment cannot pass the assessment. Likewise, from the generated rule if the candidate who are young and mode of training regular did not pass the assessment. And the result also indicated that, practical based assessment is difficult for candidates rather than qualification based assessment. The generated rules are presented on Annex I (A).

After association rule mining the study conducted classification algorithms to measure the efficiency of apriori algorithm and to take justified results. As we discussed before in section 4.4, the study conducted five classification algorithms, the experiment result also found on Annex II. in order to find out the accuracy of each classification algorithm and selecting the top one, the study conducted a statistical F-measure test accuracy.

In[56], statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is also, the average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. According to that to find out F-measure the study collected weighted Average of Precision and Recall from each algorithm. The summary of classification algorithms with F-measurer results as follows in table 4.3.

The general formula for **F-score** or **F-measure** is:F1 Score = 2 *Precision * recall/ Precision + recall.

Table 4.3: Summary of Classification algorithms

Experment	*F-measure with 10 fold cross validation with 34% supplied test set and 20% supplied test set from 321,236 instances and 10 attributes*							
	Algorithm	**Precision with supplied test**		**Recall with supplied test**		**F-measure with supplied test**		**Accuracy**	
		34%	**20%**	**34%**	**20%**	**34%**	**20%**	**34%**	**20%**
1	Bayes Network	0.73	0.73	0.73	0.73	0.73	0.73	73.6	73.6
2	NaïveBays	0.73	0.73	0.73	0.73	0.73	0.73	73.5	73.5
3	Logistic Regression	0.75	0.75	0.75	0.75	0.74	0.74	74.9	74.9
4	Hoeffding tree	0.78	0.77	0.78	0.77	0.77	0.77	77.7	77.7
5	J48 Tree	0.78	0.78	0.78	0.77	0.77	0.77	77.9	77.9

Looking at the results in Table 4.3 we can see the following. J48 Tree and Hoeffding tree has got the best overall performance from the existing classification models using 66% percentage split for training data and 34% percentage test data as well as 80% percentage split for training data and 20% percentage test data from 321,236 total numbers of data.

4.5. Measuring Efficiency of Algorithms to Compute Factors

In addition to Association rule mining, classification algorithms were performed and a test has been conducted to measure the efficiency of the proposed algorithm. This could be realized by testing the time it takes the algorithms to generate interesting rules and expected outcomes. The total time the algorithms took is computed as the sum of the time taken by each algorithm to accomplish its task. The experiment indicated that very short time and better accuracy was needed to generate rules, instead of that to analyze the association of attribute to get a frequent item, the study used the Apriori algorithm. As it has been indicated in experiment one on (Annex I) from the total 321,236 amounts of data and 10 attribute we can observe that Apriori algorithm

with Minimum support 0.1 and 0.9 confidence, generates five sets of large itemsets with fifteen (15) numbers of cycles.

Experiment two on Apriori algorithm with 10 attributes shows the Apriori algorithm with minimum support 0.1 and 0.8 confidence, generates three sets of large item sets with ten (10) numbers of cycles.

Finally, the third experiment on Apriori algorithm, with 10 attributes also shows the Apriori algorithm with minimum support 0.1 and 0.7 confidence, generates three sets of large item sets with nine (9) numbers of cycles.

On the other hand, a better accuracy selected by computing the f-measure of the conducted classification algorithms'. To find out F-measure the study collected weighted Average of Precision and Recall from each classification algorithms. Table: 4.3 shows the summary of classification algorithms from the total of 321,236 instances with F-measure results,based on results J48 Tree has got the best overall performance from the existing classification algorithms.

Based on classification experiment, in experiment one, from the above Table: 4.3 and Annex II (A) we can observe that Bayes Network with 10 attributes scored the accuracy of 73.6% and 0.73% F-measure with 10 fold cross-validations. In this Experiment from the total 321,236 amounts of data, 236506 (73.62%) of the records were correctly Classified, while the remaining 84730 (26.37%) of the records were incorrectly classified.

Experiment two on the Naïve Bayes algorithm with 10 attributes shows the accuracy of 73.5% and 0.73% F-measure with 10 fold cross-validation. According to Annex II (B) from the total 321,236 amounts of data, 236244 (73.54%) of the records were correctly Classified, while the remaining 84992 (26.45%) of the records were incorrectly classified.

The third experiment on Logistic Regression, 10 attributes shows the accuracy of 74.99% and 0.74% F-measure with cross-validation. In this case from the total data 240899 (74.99%) of the records were correctly Classified, while the remaining 80337 (25%) of the records were incorrectly classified see Annex II (C).

The fourth experiment on the Hoeffding tree, 10 attributes has shown an increase in accuracy of 77.76% and 0.77% F-measure with cross-validation. In this case from the total data, 249819

(77.76%) of the records were correctly Classified, while the remaining 71417 (22.23%) of the records were incorrectly classified see Annex II (D).

Experiment five on Annex II (E), J48 tree algorithm with 10 attributes shows the accuracy of 77.94% and 0.77% F-measure with 10 fold cross-validation. According to that from the total 321,236 amounts of data, 250377 (77.94%) of the records were correctly Classified, while the remaining 70859 (22%) of the records were incorrectly classified.

Therefore, the result gained from the experiments shows that the best prediction model using the J48 tree algorithm is more acceptable than the remaining three classification algorithms used in this study.

4.6. Comparison of the Resulted Association and Classification Algorithm

According to section 4.5, the efficiency of Apriori and classification algorithms were measured. Having that, from the experiment model, the results were compared with each other. The purposes of comparison were to measure the efficiency of the proposed algorithm and also it is appropriate for finding out the problem domain of the research. From the above association and classification models evaluation, the Apriori algorithm selected based on the resulted high-efficiency and performance criteria using minimum support 0.1 and 0.9 confidence.

4.7. Discussion

In analyzing educational data, it is interesting and has an important role in the educational environment. In this regard, the proposed solution promotes different algorithms and justified some interesting rules and factors that affect a candidate's performance using data mining techniques.

For the experiment, the data collected from Addis Ababa Occupational Competency Assessment and Certification Center (OCACC). According to that, the collected data contains about 324,236 instances and 36 Attributes. From this collected data 321,236 instances and 10 attributes were selected. Since 3000 candidates were not scheduled and assessed from the given data. According to that, the selected data applied to the selected classification and association rule mining

algorithms. Regarding this, all the selected algorithm results compute each other by their measurement criteria. In contrast to this, Association rule mining becomes more efficient to discover a hidden pattern in the existing data. Instead of the generated result from the Apriori algorithm experiment in general, the highest failure rate occurs on the age of young, mode of training regular candidates. Likewise, the candidates who are applied for qualification assessment cannot pass the assessment. On the other hand from the generated rule those candidates who have no year gap after accomplishing their training session, their status will be competent. The result also indicates that, the candidates who have a year gap as well as no work experience it is difficult to pass the assessment. On top of that, using the conveniencedata which is 321,236 instances, the study experimented with minimum support 0.1 and confidence 0.9 up to 0.7 respectively in association rule mining process. Likewise, in classification process the study used similar data which is used in association rule mining process. Instead of that to make the experiment the study used 10 fold cross validation with 66% of split and 34% supplied test, similarly the study used 80% of split and 20% supplied test for the data set as indicated in table 4.3.

According to section 4.5, the Apriori algorithm selected based on efficiency and performance criteria. Since as it has been indicated from the result tables such as 4.2, the hugeness of the data is not affect the efficiency of algorithm. However, if the confidence is decreases the efficiency of algorithm also decreases. In contrast with classification algorithm, predictive models are limited inlarge sets of data. According to that, this result is found to be good enough to take a considerable decision. Regarding this, the study can be able to realize the dominant factor using apriori algorithm through ARM.

Though some of the related works, [45] and [46], used association rule techniques, to analyze student's performance and although they mentioned Apriori Algorithm found necessary since a large number of Association rules or patterns or knowledge is generated from the large volume of a dataset. But most of the association rules have redundant information and thus all of them cannot be used directly for an application. So pruning or grouping rules by some means are necessary to get very important rules or knowledge. In general, through association rule discovery techniques, it can be able to improve the quality of education by analyzing the data and discover the factors that affect the academic results.

From the above findings, some domain experts were interviewed such as Assessors and Assessment experts those who are working and facilitates the assessment in the OCACC office. The experts indicated that the failure might emanate from the number of versions that the assessment has. Means these versions may circulate in each assessment and candidates have no awareness as well as full of knowledge about the assessment. Since some field of assessments starts from the common or general knowledge. For example, every business assessment candidates have to face Basic Clerical Works one (BCW1) to pass to their occupation. According to that the study responses to the research questions which are generated in section one in 1.4.

RQ1. What are the patterns that are frequently associate each other in the candidate database?

Having the research question from the selected 10 attributes six patterns with status NYC (Not Yet Competent) is frequently occurred. This means age with the value of YNG (Young), mode of training with the value of REG (Regular), employment condition with the value of UNEMP (Unemployed), applied for UNC(Unit of Competency) with the value of (NO), year gap with the value of YGP (Year gap) and applied for PR (Practical) assessment occur frequently together.

RQ2. Are the rules generated from frequent patterns are interesting and have influence on candidates' performance?

As it has been indicated in the result all frequent patterns are interesting and have influence on candidate's performance. According to the study result year gap has a major influence since the candidates do not assess immediately after accomplishing their education, Unemployment condition; which means candidates or professionals have no work environment exposure. Since the OS emerged from industries based on the marketing need however the study indicates that candidates have no awareness about occupational standards and occupational competencies (OC) which are derived from OS. and from the response of domain experts, high failures found on the business field especially on basic clerical work level one (BCW1) and according to OCACC assessment rule, business field assessment candidates have to face BCW1, having that the derived OS for level one assessment is not compatible with candidates. Since any business candidates begin their assessment from BCW level one. For example, the assessed candidate's profession might be Accounting, Marketing, Legal services, etc. having that the candidates

especially business professionals have no awareness or full of knowledge about the occupational standards.

RQ3. Has the discovered knowledge negative implication on the assessment process?

According to the study result, frequent failure occurs on candidates those who are young, regular student and those who have no work experience. Having that, the readiness of the candidates for the assessment is very low since, after accomplishing their study, they have no work environment exposure. Occupational competencies are prepared by industry experts; in addition to that candidates have no full chance to train into the industry. Likewise, Addis Ababa OCACC provides the assessment in different TVET colleges other than industries. Yet, the rule of the assessment is, candidates should have to assess in the industry. Having that, frequent failure occurs on candidates those who are young, unemployed, regular students, those who have year gap and those who are applied for practical based assessment, because of low work exposure and psychological readiness. Since the assessment provides in the teaching college. Hence candidates have no full of readiness for the assessment. However, the discovered knowledge has negative implication on the assessment process.

CHAPTER FIVE: CONCLUSION AND FUTURE WORKS

5.1. Conclusion

This work is an attempt to use Data Mining techniques to analyze factors that affect a candidate's performance. In this study five classification techniques and one association rule mining conducted on the candidate's dataset. From the analysis result of Association Rule Mining, the Apriori algorithm performs three types of experimentation by decreasing confidences from top to down. According to that, as it has been indicated in experiments respectively, experiment one with minimum support 0.1 and 0.9 confidence generates five large item sets with 15 number of cycle. According to that, interesting rules can retrieve from the generated set of large item set. Likewise, from the classification experiments, Bayes Network performs with the accuracy of 73.6%, Naïve Bays 73.5, Logistic Regression 74.9, Hoeffding Tree 77.7 and J48 Tree 77.9. From the experimental result, we notice that the Apriori algorithm is the most suitable method for this type of candidate dataset to realize candidate performance. This study aims to dig out the major patterns that are hidden and questionable in the organization that releases a candidate's performance. Though when we are seeing the generated rules it might look nontrivial, when we associate and make a deep study, the rules have a major influence on a candidate's performance. According to the generated rules which are frequently associated, as it has been indicated in the ARM analysis result, if a candidate are young, single, unemployed with no work experience are not competent. Additionally, if the candidates also young, single, mode of training regular and applied for qualification based assessment, he or she also be incompetent. Likewise, if the candidates those who are young, mode of training regularly and have a year gap to assess their competency after accomplish their study, she or he will not also be incompetent. As a researcher, from the above findings and domain experts, it is possible to conclude that occupational standard (OS) can make a higher influence on candidate's performance and also the candidates have no full of readiness for the assessment since candidates have no full of knowledge on OS which is presented for the assessment. This might derive from the candidates have no capacity for occupational competencies (OC) that are derived from OS for the assessment. On the other hand, the majority of candidates who are failed in the assessment are young, those who have year gap and regular students. On the other side extension, distance, short term training students, adult and old are not failed in contrast with regular students. Instead of that, it is possible to conclude,

Regular students have no work environment exposure and have no psychological readiness for the assessment because of low practice in the work environment. On the other hand, from the domain expert's side view, the assessments are not given in the industry; there isno cooperative training in the industry as much as required, candidates also complained there is a gap between occupational competency and training. Having that, if candidates have work experience and going to the assessment center for the assessment, the assessments not make them suitable. Hence the required assessment gives in the educational institutions, unlike industries. So these situations make candidates believe that, the assessment as a normal examination. Finally, on top of the results from the study, Discovering factors of candidates or any other process in the educational environment could be supported by the use of data mining technique. Though, as discussed in [30], the experience of different organizations are stated who used data mining techniques and presents the challenges of the organization before using the data mining technique and also identify the results that they are found by using a data mining technique. Overall, this study has to be able to implement for the assessment office of OCACC, TVET institutions, Colleges, Universities, and other academic regulatory institutions.

5.2. Recommendations

The study is done for academic purpose but it can be implemented for the OCACC office and other TVET institutions. The data used for the study is five-year data and three branches; therefore, if other researchers are conducted on this area by using different year and all branches data, it can be helpful for the organization to implement the application of data mining. Also, five algorithms were used from many classification algorithms, so other better algorithms can be selected by performing better than these algorithms used in this study to compute with association rule.

From the perspective of the organization the following recommendations are forwarded, to deal with the factors identified in this study.

1. It is recommended to broadcast online all Occupational Standards (OS) for the assessment to have a better awareness.

2. During training, training centers should work in collaboration with other centers or industries to make candidates a psychological preparation. Candidates should be assigned on the apparent ship or cooperative training. Therefore each candidate will have exposure in all types of practical activities.

4. Finally, OCACC has a responsibility to assess each candidate in the industry who is assigned for the apparent ship. Since if the candidates assessed in the industry their psychological setup will be maintained and also the candidates consider as they are working rather than examination.

5.3. Contribution

As a contribution, the study can be able to analyze huge amounts of data using data mining techniques which is difficult to discover traditionally.

The analysis ability of algorithms is tested using a large data size. Based on that, this study provides the following notable contributions.

- In a classification algorithm, the study justified that the training data size should be less to increase efficiency.
- In association rule mining, the study justified that, if the data size is as much as larger the efficiency increase. Likewise, confidence decreases, the efficiency also be decreased.

5.4. Future Works

In this section, we insight a key point that remains a challenge in candidate performance analysis, and of the challenge in this work. The data were collected from OCACC database only because of time limitation and resources. Based on that, to achieve a better resultin analyzing factors affecting candidates' or professionals' performance during the assessment process, further investigation can be done. This means we can find the candidate's interest and attitude from their experience, course selection, and their high school results by combining OCACC, TVET institutes, and another educational background.

REFERENCES

[1] R. Johan and J. Harlan, "Education nowadays," *Int. J. Educ. Sci. Res.*, vol. 4, no. 5, pp. 51–56, 2014.

[2] J. L. D. Espiritu, *Handbook of Comparative Studies on Community Colleges and Global Counterparts*, no. March. 2018.

[3] AU, "What is Technical and Vocational Education and Training (TVET)?," pp. 1–2, 2006.

[4] T. E. Journal, E. Vol, and 10. No, "The Ethiopian Journal of Education Vol. 26 No. 1 June 2006 31," vol. 26, no. 1, pp. 31–51, 2006.

[5] A. Hagos Baraki and E. van Kemenade, "Effectiveness of Technical and Vocational Education and Training (TVET)," *TQM J.*, vol. 25, no. 5, pp. 492–506, 2013.

[6] MOE, "Ethiopian education development roadmap: an integrated executive summary," no. July, pp. 1–101, 2018.

[7] A. P. Opoko *et al.*, "The Role of Technical and Vocational Education and Training (Tvet) in Nation Building: a Review of the Nigerian Case," *Int. J. Mech. Eng. Technol.*, vol. 09, no. 13, pp. 1564–1571, 2018.

[8] P. Krishnan and I. Shaorshadze, "Technical and vocational Education and Training in Ethiopia," *Int. Growth Centre, London Sch. Econ. Polit. Sci.*, no. February, 2013.

[9] E. T. Hailu, "Analysing the Labour Outcomes of TVET in Ethiopia : Implication of Challenges and Opportunities in Productive Self-employment of TVET Graduates," no. December, 2012.

[10] L. Geleto, "Technical Vocational Education Training Institute Curriculum Development in Ethiopia," *J. Educ. Vocat. Res.*, vol. 8, no. 3, p. 16, 2018.

[11] Ministry of Education, "Occupational Assessment and Certification," pp. 1–13, 2010.

[12] E. October, "USER ' S MANUAL IN DEVELOPING / FORMULATING ASSESSMENT ITEM / TOOL," 2008.

[13] "ANG," "Addis ababa city government OCACC Regulatory," No, 64/2014.

[14] M. of Education, "Manual for Assessment Center," pp. 1–5, 2010.

[15] J. Raven, L. Cunningham, C. Commentary, and J. Raven, "CONTENTS," pp. 1–2, 2001.

[16] R. Geometry and G. Analysis, "Competence Standards for Technical and Vocational Education and Training TVET".

[17] "MoE," "Ethiopian National TVET strategic document," MOE, Addis Ababa, 2008."

[18] "MoE", "National Technical and Vocational Education and Training" (TVET) Strategy, Addis Ababa, (2012)."

[19] "Addis Ababa Occupational Competency Assessment and Certification Center (OCACC) report 2018 " .

[20] R. Castro, "SWOT ANALYSIS: A THEORETICAL REVIEW," *Ekp*, vol. 13, no. 3, pp. 1576–1580, 2017.

[21] I. Technology, *Prediction and Analysis of Student Performance by Data Mining in WEKA Report of Project submitted for the partial fulfillment of the requirements for the*. .

[22] J. Fürnkranz, D. Gamberger, and N. Lavrač, "Foundations of Rule Learning," no. 2003, 2012.

[23] J. P. Jiawei Han, Micheline Kamber, *Data Mining – Concepts & Techniques*. 2011.

[24] K. B. Agyapong, D. J. . Hayfron-Acquah, and D. M. Asante., "An Overview of Data Mining Models (Descriptive and Predictive)," *Int. J. Softw. Hardw. Res. Eng.*, vol. 4, no. 5, pp. 53–60, 2016.

[25] J. Han, "Data Mining: Concepts and Techniques Why Data Mining? The Explosive Growth of Data: from terabytes to petabytes," 2007.

[26] M. Maurizio, "Data Mining Concepts and Techniques," 2011.

[27] M. Al-Saleem, N. Al-Kathiry, S. Al-Osimi, and G. Badr, "--ok--only on academic records--Mining Educational Data to Predict Students' Academic Performance," *Mach. Learn. Data Min. Pattern Recognition, Mldm 2015*, vol. 9166, no. 6, pp. 403–414, 2015.

[28] A. Rai, "An Overview of Association Rule Mining and its Applications," *Upgrad blog*, vol. 5, no. 1, pp. 1–7, 2018.

[29] B. B. Bezabeh, "the Application of Data Mining Techniques To Support Customer Relationship Management: the Case of Ethiopian Revenue and Customs Authority," *Int. J. Adv. Stud. Comput. Sci. Eng.*, vol. 6, no. 6, pp. 35–41, 2017.

[30] F. Baudier, "Preface: Creer les conditions favorables pour un depistage organise du cancer du col de l'uterus," *Sante Publique (Paris).*, vol. 12, no. SPEC. ISS., pp. 5–10, 2000.

[31] M. Gera and S. Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity," *Int. J. Comput. Appl.*, vol. 113, no. 18, pp. 22–29, 2015.

[32] N. A. Yassein, R. G. M Helali, and S. B. Mohomad, "Predicting Student Academic

Performance in KSA using Data Mining Techniques," *J. Inf. Technol. Softw. Eng.*, vol. 07, no. 05, 2017.

[33] A. Badr El Din Ahmed and I. Sayed Elaraby, "PER: A prediction for Student's Performance Using Decision Tree ID3 Method," *India - World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[34] L. Wanner, "Introduction to Clustering Techniques," *Iula*, 2004.

[35] B. Liu, W. Hsu, and Y. Ma, "KDD98-012.pdf," 1998.

[36] C. Paper and K. Mcgarry, "Advances in Computational Intelligence Systems," vol. 840, no. November 2017, 2019.

[37] Q. Zhao, "Association Rule Mining : A Survey Association Rule Mining : A Survey," vol. 5, no. March, pp. 2320–2324, 2016.

[38] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 7, pp. 58–63, 2017.

[39] T. Lindgren, K. Andersson, and D. Norbäck, "Analyzing Performance of Students by Using Data Mining Techniques," *Aviat. Sp. Environ. Med.*, vol. 77, no. 8, pp. 832–837, 2006.

[40] H. Bathla and M. K. Kathuria, "Association Rule Mining : Algorithms Used," *Int. J. Comput. Sci. Mob. Comput.*, vol. 4, no. 6, pp. 271–277, 2015.

[41] P. Prithiviraj and R. Porkodi, "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study," *Open J. Comput. Sci. Eng. Surv.*, vol. 3, no. 1, pp. 98–119, 2015.

[42] C. Fu, X. Wang, L. Zhang, and L. Qiao, "Mining algorithm for association rules in big data based on Hadoop," *AIP Conf. Proc.*, vol. 1955, no. April, 2018.

[43] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, 2013.

[44] B. M. M. Alom and M. Courtney, "Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia," *Int. J. Inf. Technol. Comput. Sci.*, vol. 10, no. 2, pp. 1–9, 2018.

[45] K. K. Jain and A. B. Raut, "Available Online at www.ijeecse.com Finding Association Rule using Apriori Algorithm on Educational Domain Available Online at www.ijeecse.com," vol. 2, no. 2, pp. 64–67, 2015.

[46] C. Bambrah, M. Bhandari, N. Maniar, and V. Munde, “Mining Association Rules in Student Assessment Data,” *Ijarcce.Com*, vol. 3, no. 3, pp. 5340–5342, 2014.

[47] G. B. Tarekegn, “Application of Data Mining Techniques to Predict Students Placement in to Departments,” *Int. J. Res. Stud. Comput. Sci. Eng.*, vol. 3, no. 2, pp. 10–14, 2016.

[48] G. S. Abu-Oda and A. M. El-Halees, “Data Mining in Higher Education : University Student Dropout Case Study,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 15–27, 2015.

[49] E. Balraj and D. Maalini, “A survey on Predicting Student Dropout Analysis Using Data Mining Algorithms,” *Int. J. Pure Appl. Math.*, vol. 118, no. 8, pp. 621–627, 2018.

[50] R. Wirth and J. Hipp, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.

[51] S. Sood, “Data Visualization A Tool of Data Mining,” *Data Vis. A Tool Data Min.*, vol. 4333, no. 3, pp. 197–198, 2011.

[52] M. H. Uma K, “Data Collection Methods and Data Pre- processing Techniques for Healthcare Data Using Data Mining,” *Int. J. Sci. Eng. Res.*, vol. 8, no. 6, pp. 1131–1136, 2017.

[53] A. Arora, P. K. Malhotra, S. Marwah, A. Bhardwaj, and S. Dahiya, “Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets,” 2012.

[54] C. B. Madsen *et al.*, “Two Popular Open-Source Programming Languages to Consider for Your Data Science Toolkit,” 2014.

[55] F. M. Nafie Ali and A. A. Mohamed Hamed, “Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents,” *J. Inf. Telecommun.*, vol. 2, no. 3, pp. 231–245, 2018.

[56] Y. Sasaki, “The truth of the F-measure,” " Research Fellow, University of Manchester, School of Computer Science"pp. 1–5, 2007.

,

ANNEXES

Annex (I): Sample Results Discovered from Association Rule Mining

A. Experiment One

Relation: 321,236 nominaldt

Instances: 321,236

Attributes: 10

Gender, age, marital_status, Year gap, mode_of_training

Employment_condition, education_background, apply_for_unc

Applied_for, status.

=== Associator model (full training set) ===

Minimum support: 0.25 (80309 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 15

Size of set of large itemsetsL(2): 53

Size of set of large itemsetsL(3): 97

Size of set of large itemsetsL(4): 67

Size of set of large itemsetsL(5): 10

Interesting rules found:

1. Marital_status=MRD Mode_of_training=REG Status=NYC 90348
==>Employment_condition=UNEMP 83818 <conf:(0.93)> lift:(1.07) lev:(0.02) [5605]
conv:(1.86)

2. Mode_of_training=REG Apply_for_unc=No Status=NYC 91813 ==>Employment_condition=UNEMP 84582 <conf:(0.92)> lift:(1.06) lev:(0.02) [5101] conv:(1.71)

3. Age=YNG Mode_of_training=REG Status=NYC 114182 ==>Employment_condition=UNEMP 105164 <conf:(0.92)> lift:(1.06) lev:(0.02) [6319] conv:(1.7)

4. Age=YNG Year_gap=YGP Mode_of_training=REG Status=NYC 102555 ==>Employment_condition=UNEMP 94371 <conf:(0.92)> lift:(1.06) lev:(0.02) [5591] conv:(1.68)

5. Age=YNG Marital_status=MRD Mode_of_training=REG Applied_for=PR 100202 ==>Employment_condition=UNEMP 92124 <conf:(0.92)> lift:(1.06) lev:(0.02) [5381] conv:(1.67)

6. Age=YNG Marital_status=MRD Year_gap=YGP Mode_of_training=REG 107142 ==>Employment_condition=UNEMP 98365 <conf:(0.92)> lift:(1.06) lev:(0.02) [5614] conv:(1.64)

7. Gender=F Marital_status=MRD Mode_of_training=REG 89788 ==>Employment_condition=UNEMP 82377 <conf:(0.92)> lift:(1.06) lev:(0.01) [4649] conv:(1.63)

8. Age=YNG Marital_status=MRD Mode_of_training=REG Apply_for_unc=No 91537 ==>Employment_condition=UNEMP 83899 <conf:(0.92)> lift:(1.06) lev:(0.01) [4657] conv:(1.61)

9. Marital_status=MRD Apply_for_unc=No Status=NYC 96488 ==>Employment_condition=UNEMP 88365 <conf:(0.92)> lift:(1.06) lev:(0.02) [4837] conv:(1.6)

10. Mode_of_training=REG Status=NYC 129229 ==>Employment_condition=UNEMP 118345 <conf:(0.92)> lift:(1.06) lev:(0.02) [6474] conv:(1.59)

11. Age=YNG Apply_for_unc=No Status=NYC 116062 ==>Employment_condition=UNEMP 106234 <conf:(0.92)> lift:(1.06) lev:(0.02) [5761] conv:(1.59)

12. Age=YNG Marital_status=MRD Mode_of_training=REG 144200 ==>Employment_condition=UNEMP 131987 <conf:(0.92)> lift:(1.06) lev:(0.02) [7156] conv:(1.59)

13. Age=YNG Year_gap=YGP Apply_for_unc=No Status=NYC 101884 ==>Employment_condition=UNEMP 93207 <conf:(0.91)> lift:(1.06) lev:(0.02) [5008] conv:(1.58)

14. Year_gap=YGP Mode_of_training=REG Status=NYC 116636 ==>Employment_condition=UNEMP 106661 <conf:(0.91)> lift:(1.06) lev:(0.02) [5691] conv:(1.57)

15. Age=YNG Marital_status=MRD Status=NYC 112809 ==>Employment_condition=UNEMP 103006 <conf:(0.91)> lift:(1.05) lev:(0.02) [5349] conv:(1.55)

16. Gender=F Year_gap=YGP Mode_of_training=REG 94969 ==>Employment_condition=UNEMP 86548 <conf:(0.91)> lift:(1.05) lev:(0.01) [4335] conv:(1.51)

17. Gender=F Age=YNG Mode_of_training=REG 113316 ==>Employment_condition=UNEMP 103233 <conf:(0.91)> lift:(1.05) lev:(0.02) [5137] conv:(1.51)

18. Age=YNG Year_gap=YGP Mode_of_training=REG Apply_for_unc=No 99127 ==>Employment_condition=UNEMP 90297 <conf:(0.91)> lift:(1.05) lev:(0.01) [4484] conv:(1.51)

19. Age=YNG Year_gap=YGP Mode_of_training=REG 143327 ==>Employment_condition=UNEMP 130531 <conf:(0.91)> lift:(1.05) lev:(0.02) [6455] conv:(1.5)

20. Age=YNG Marital_status=MRD Year_gap=YGP Status=NYC 100649 ==>Employment_condition=UNEMP 91660 <conf:(0.91)> lift:(1.05) lev:(0.01) [4530] conv:(1.5)

21. Marital_status=MRD Year_gap=YGP Mode_of_training=REG 122780 ==>Employment_condition=UNEMP 111797 <conf:(0.91)> lift:(1.05) lev:(0.02) [5509] conv:(1.5)

22. Age=YNG Year_gap=YGP Mode_of_training=REG Applied_for=PR 93479 ==>Employment_condition=UNEMP 85091 <conf:(0.91)> lift:(1.05) lev:(0.01) [4168] conv:(1.5)

23. Marital_status=MRD Mode_of_training=REG Applied_for=PR 116971 ==>Employment_condition=UNEMP 106408 <conf:(0.91)> lift:(1.05) lev:(0.02) [5148] conv:(1.49)

24. Apply_for_unc=No Status=NYC 135724 ==>Employment_condition=UNEMP 123419 <conf:(0.91)> lift:(1.05) lev:(0.02) [5925] conv:(1.48)

25. Age=YNG Mode_of_training=REG Applied_for=PR 133852 ==>Employment_condition=UNEMP 121655 <conf:(0.91)> lift:(1.05) lev:(0.02) [5782] conv:(1.47)

26. Age=YNG Mode_of_training=REG Apply_for_unc=No 126881 ==>Employment_condition=UNEMP 115315 <conf:(0.91)> lift:(1.05) lev:(0.02) [5476] conv:(1.47)

27. Year_gap=YGP Apply_for_unc=No Status=NYC 120017 ==>Employment_condition=UNEMP 109022 <conf:(0.91)> lift:(1.05) lev:(0.02) [5125] conv:(1.47)

28. Gender=F Age=YNG Status=NYC 96067 ==>Employment_condition=UNEMP 87209 <conf:(0.91)> lift:(1.05) lev:(0.01) [4045] conv:(1.46)

29. Gender=F Mode_of_training=REG 126140 ==>Employment_condition=UNEMP 114402 <conf:(0.91)> lift:(1.05) lev:(0.02) [5205] conv:(1.44)

30. Marital_status=MRD Mode_of_training=REG Apply_for_unc=No 102796 ==>Employment_condition=UNEMP 93215 <conf:(0.91)> lift:(1.05) lev:(0.01) [4226] conv:(1.44)

31. Year_gap=YGP Mode_of_training=REG Apply_for_unc=No 111428 ==>Employment_condition=UNEMP 100976 <conf:(0.91)> lift:(1.05) lev:(0.01) [4515] conv:(1.43)

32. Status=NYC 188663 ==>Year_gap=YGP 170964 <conf:(0.91)> lift:(1.23) lev:(0.1) [31961] conv:(2.81)

33. Gender=F Status=NYC 109596 ==>Year_gap=YGP 99284 <conf:(0.91)> lift:(1.23) lev:(0.06) [18535] conv:(2.8)

34. Age=YNG Mode_of_training=REG 196344 ==>Employment_condition=UNEMP 177831 <conf:(0.91)> lift:(1.05) lev:(0.02) [7860] conv:(1.42)

35. Year_gap=YGP Mode_of_training=REG 163716 ==>Employment_condition=UNEMP 148109 <conf:(0.9)> lift:(1.05) lev:(0.02) [6383] conv:(1.41)

36. Marital_status=MRD Mode_of_training=REG 166404 ==>Employment_condition=UNEMP 150524 <conf:(0.9)> lift:(1.04) lev:(0.02) [6471] conv:(1.41)

37. Year_gap=YGP Mode_of_training=REG Applied_for=PR 108038 ==>Employment_condition=UNEMP 97685 <conf:(0.9)> lift:(1.04) lev:(0.01) [4158] conv:(1.4)

38. Age=YNG Status=NYC 159355 ==>Employment_condition=UNEMP 144060 <conf:(0.9)> lift:(1.04) lev:(0.02) [6109] conv:(1.4)

39. Mode_of_training=REG Apply_for_unc=No Applied_for=PR 89501 ==>Employment_condition=UNEMP 80899 <conf:(0.9)> lift:(1.04) lev:(0.01) [3419] conv:(1.4)

40. Gender=F Employment_condition=UNEMP Status=NYC 98621 ==>Year_gap=YGP 89114 <conf:(0.9)> lift:(1.23) lev:(0.05) [16452] conv:(2.73)

41. Employment_condition=UNEMP Status=NYC 168797 ==>Year_gap=YGP 152506 <conf:(0.9)> lift:(1.23) lev:(0.09) [28139] conv:(2.73)

42. Gender=F Age=YNG Status=NYC 96067 ==>Year_gap=YGP 86713 <conf:(0.9)> lift:(1.23) lev:(0.05) [15932] conv:(2.7)

43. Mode_of_training=REG Status=NYC 129229 ==>Year_gap=YGP 116636 <conf:(0.9)> lift:(1.22) lev:(0.07) [21422] conv:(2.7)

44. Gender=F Mode_of_training=REG Employment_condition=UNEMP 114402 ==> Age=YNG 103233 <conf:(0.9)> lift:(1.08) lev:(0.02) [7934] conv:(1.71)

45. Mode_of_training=REG Apply_for_unc=No 142258 ==>Employment_condition=UNEMP 128326 <conf:(0.9)> lift:(1.04) lev:(0.02) [5176] conv:(1.37)

46. Age=YNG Year_gap=YGP Status=NYC 143709 ==>Employment_condition=UNEMP 129620 <conf:(0.9)> lift:(1.04) lev:(0.02) [5214] conv:(1.37)

47. Age=YNG Status=NYC 159355 ==>Year_gap=YGP 143709 <conf:(0.9)> lift:(1.22) lev:(0.08) [26299] conv:(2.68)

48. Mode_of_training=REG Employment_condition=UNEMP Status=NYC 118345 ==>Year_gap=YGP 106661 <conf:(0.9)> lift:(1.22) lev:(0.06) [19466] conv:(2.67)

49. Mode_of_training=REG Applied_for=PR 155295 ==>Employment_condition=UNEMP 139959 <conf:(0.9)> lift:(1.04) lev:(0.02) [5523] conv:(1.36)

50. Marital_status=MRD Status=NYC 134255 ==>Employment_condition=UNEMP 120964 <conf:(0.9)> lift:(1.04) lev:(0.01) [4742] conv:(1.36)

B. Experiment Two

Relation: 321,236 nominaldt

Instances: 321,236

Attributes: 10

Gender, age, marital_status, Year gap, mode_of_training

Employment_condition, education_background, apply_for_unc

Applied_for, status.

=== Associator model (full training set) ===

Minimum support: 0.5 (160618 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 9

Size of set of large itemsetsL(2): 18

Size of set of large itemsetsL(3): 4

Interesting rules found:

1. Status=NYC 188663 ==>Year_gap=YGP 170964 <conf:(0.91)> lift:(1.23) lev:(0.1) [31961] conv:(2.81)

2. Age=YNG Mode_of_training=REG 196344 ==>Employment_condition=UNEMP 177831 <conf:(0.91)> lift:(1.05) lev:(0.02) [7860] conv:(1.42)

3. Mode_of_training=REG 225310 ==>Employment_condition=UNEMP 202208 <conf:(0.9)> lift:(1.04) lev:(0.02) [7162] conv:(1.31)

4. Status=NYC 188663 ==>Employment_condition=UNEMP 168797 <conf:(0.89)> lift:(1.03) lev:(0.02) [5475] conv:(1.28)

5. Age=YNG Year_gap=YGP 197754 ==>Employment_condition=UNEMP 175216 <conf:(0.89)> lift:(1.02) lev:(0.01) [4024] conv:(1.18)

6. Age=YNG Marital_status=MRD 197946 ==>Employment_condition=UNEMP 175065 <conf:(0.88)> lift:(1.02) lev:(0.01) [3707] conv:(1.16)

7. Age=YNG Applied_for=PR 184035 ==>Employment_condition=UNEMP 161918 <conf:(0.88)> lift:(1.02) lev:(0.01) [2602] conv:(1.12)

8. Age=YNG 267594 ==>Employment_condition=UNEMP 235394 <conf:(0.88)> lift:(1.02) lev:(0.01) [3743] conv:(1.12)

9. Mode_of_training=REG Employment_condition=UNEMP 202208 ==> Age=YNG 177831 <conf:(0.88)> lift:(1.06) lev:(0.03) [9388] conv:(1.39)

10. Apply_for_unc=No 207908 ==>Employment_condition=UNEMP 182698 <conf:(0.88)> lift:(1.02) lev:(0.01) [2716] conv:(1.11)

11. Year_gap=YGP 236680 ==>Employment_condition=UNEMP 206791 <conf:(0.87)> lift:(1.01) lev:(0.01) [1902] conv:(1.06)

12. Mode_of_training=REG 225310 ==> Age=YNG 196344 <conf:(0.87)> lift:(1.05) lev:(0.03) [8657] conv:(1.3)

13. Marital_status=MRD 239886 ==>Employment_condition=UNEMP 207970 <conf:(0.87)> lift:(1) lev:(0) [305] conv:(1.01)

14. Applied_for=PR 223172 ==>Employment_condition=UNEMP 192997 <conf:(0.86)> lift:(1) lev:(-0) [-198] conv:(0.99)

15. Apply_for_unc=No 207908 ==> Age=YNG 176274 <conf:(0.85)> lift:(1.02) lev:(0.01) [3083] conv:(1.1)

16. Year_gap=YGP Employment_condition=UNEMP 206791 ==> Age=YNG 175216 <conf:(0.85)> lift:(1.02) lev:(0.01) [2956] conv:(1.09)

17. Employment_condition=UNEMP 278087 ==> Age=YNG 235394 <conf:(0.85)> lift:(1.02) lev:(0.01) [3743] conv:(1.09)

18. Marital_status=MRD Employment_condition=UNEMP 207970 ==> Age=YNG 175065 <conf:(0.84)> lift:(1.01) lev:(0.01) [1823] conv:(1.06)

19. Employment_condition=UNEMP Applied_for=PR 192997 ==> Age=YNG 161918 <conf:(0.84)> lift:(1.01) lev:(0) [1148] conv:(1.04)

20. Year_gap=YGP 236680 ==> Age=YNG 197754 <conf:(0.84)> lift:(1) lev:(0) [596] conv:(1.02)

C. Experiment Three

Relation: 321,236 nominaldt

Instances: 321,236

Attributes: 10

Gender, age, marital_status, Year gap, mode_of_training

Employment_condition, education_background, apply_for_unc

Applied_for, status.

=== Associator model (full training set) ===

Minimum support: 0.55 (176680 instances)

Minimum metric <confidence>: 0.7

Number of cycles performed: 9

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 8

Size of set of large itemsetsL(2): 11

Size of set of large itemsetsL(3): 1

Interesting rules found:

1. Age=YNG Mode_of_training=REG 196344 ==>Employment_condition=UNEMP 177831 <conf:(0.91)> lift:(1.05) lev:(0.02) [7860] conv:(1.42)

2. Mode_of_training=REG 225310 ==>Employment_condition=UNEMP 202208 <conf:(0.9)> lift:(1.04) lev:(0.02) [7162] conv:(1.31)

3. Age=YNG 267594 ==>Employment_condition=UNEMP 235394 <conf:(0.88)> lift:(1.02) lev:(0.01) [3743] conv:(1.12)

4. Mode_of_training=REG Employment_condition=UNEMP 202208 ==> Age=YNG 177831 <conf:(0.88)> lift:(1.06) lev:(0.03) [9388] conv:(1.39)

5. Apply_for_unc=No 207908 ==>Employment_condition=UNEMP 182698 <conf:(0.88)> lift:(1.02) lev:(0.01) [2716] conv:(1.11)

6. Year_gap=YGP 236680 ==>Employment_condition=UNEMP 206791 <conf:(0.87)> lift:(1.01) lev:(0.01) [1902] conv:(1.06)

7. Mode_of_training=REG 225310 ==> Age=YNG 196344 <conf:(0.87)> lift:(1.05) lev:(0.03) [8657] conv:(1.3)

8. Marital_status=MRD 239886 ==>Employment_condition=UNEMP 207970 <conf:(0.87)> lift:(1) lev:(0) [305] conv:(1.01)

9. Applied_for=PR 223172 ==>Employment_condition=UNEMP 192997 <conf:(0.86)> lift:(1) lev:(-0) [-198] conv:(0.99)

10. Employment_condition=UNEMP 278087 ==> Age=YNG 235394 <conf:(0.85)> lift:(1.02) lev:(0.01) [3743] conv:(1.09)

11. Year_gap=YGP 236680 ==> Age=YNG 197754 <conf:(0.84)> lift:(1) lev:(0) [596] conv:(1.02)

12. Marital_status=MRD 239886 ==> Age=YNG 197946 <conf:(0.83)> lift:(0.99) lev:(-0.01) [-1882] conv:(0.96)

13. Applied_for=PR 223172 ==> Age=YNG 184035 <conf:(0.82)> lift:(0.99) lev:(-0.01) [-1870] conv:(0.95)

14. Mode_of_training=REG 225310 ==> Age=YNG Employment_condition=UNEMP 177831 <conf:(0.79)> lift:(1.08) lev:(0.04) [12729] conv:(1.27)

15. Year_gap=YGP 236680 ==>Marital_status=MRD 179618 <conf:(0.76)> lift:(1.02) lev:(0.01) [2874] conv:(1.05)

16. Age=YNG Employment_condition=UNEMP 235394 ==>Mode_of_training=REG 177831 <conf:(0.76)> lift:(1.08) lev:(0.04) [12729] conv:(1.22)

17. Marital_status=MRD 239886 ==>Year_gap=YGP 179618 <conf:(0.75)> lift:(1.02) lev:(0.01) [2874] conv:(1.05)

18. Employment_condition=UNEMP 278087 ==>Marital_status=MRD 207970 <conf:(0.75)> lift:(1) lev:(0) [305] conv:(1)

19. Employment_condition=UNEMP 278087 ==>Year_gap=YGP 206791 <conf:(0.74)> lift:(1.01) lev:(0.01) [1902] conv:(1.03)

20. Age=YNG 267594 ==>Marital_status=MRD 197946 <conf:(0.74)> lift:(0.99) lev:(-0.01) [-1882] conv:(0.97)

Annex (II): Sample Results Discovered from Classification Models

A. Bayes Network Risult

```
Time taken to build model: 0.41 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      236506               73.6238 %
Incorrectly Classified Instances     84730               26.3762 %
Kappa statistic                          0.4426
Mean absolute error                      0.3475
Root mean squared error                  0.4266
Relative absolute error                 71.6947 %
Root relative squared error             86.6584 %
Total Number of Instances           321236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.168    0.715      0.600    0.653      0.447    0.782     0.741     C
                 0.832    0.400    0.748      0.832    0.787      0.447    0.782     0.810     NYC
Weighted Avg.    0.736    0.304    0.734      0.736    0.732      0.447    0.782     0.782

=== Confusion Matrix ===

     a      b   <-- classified as
 79580  52993 |      a = C
 31737 156926 |      b = NYC
```

B. NaïveBays

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      236244               73.5422 %
Incorrectly Classified Instances     84992               26.4578 %
Kappa statistic                          0.441
Mean absolute error                      0.3472
Root mean squared error                  0.4271
Relative absolute error                 71.6217 %
Root relative squared error             86.7579 %
Total Number of Instances           321236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.169    0.713      0.600    0.652      0.445    0.781     0.741     C
                 0.831    0.400    0.747      0.831    0.787      0.445    0.781     0.810     NYC
Weighted Avg.    0.735    0.305    0.733      0.735    0.731      0.445    0.781     0.781

=== Confusion Matrix ===

     a      b   <-- classified as
 79525  53048 |      a = C
 31944 156719 |      b = NYC
```

C. Logistic Regression

```
Time taken to build model: 73.93 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      240899               74.9913 %
Incorrectly Classified Instances     80337               25.0087 %
Kappa statistic                          0.4656
Mean absolute error                      0.3505
Root mean squared error                  0.4184
Relative absolute error                 72.2946 %
Root relative squared error             84.995  %
Total Number of Instances           321236

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.580    0.131    0.757      0.580    0.657      0.476    0.800     0.744     C
               0.869    0.420    0.747      0.869    0.803      0.476    0.800     0.829     NYC
Weighted Avg.  0.750    0.300    0.751      0.750    0.743      0.476    0.800     0.794

=== Confusion Matrix ===

     a      b   <-- classified as
 76940  55633 |      a = C
 24704 163959 |      b = NYC
```

D. Hoeffding Tree

```
Time taken to build model: 1.2 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      249819               77.7681 %
Incorrectly Classified Instances     71417               22.2319 %
Kappa statistic                          0.5275
Mean absolute error                      0.3059
Root mean squared error                  0.3942
Relative absolute error                 63.1023 %
Root relative squared error             80.0613 %
Total Number of Instances           321236

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.632    0.120    0.788      0.632    0.701      0.536    0.837     0.819     C
               0.880    0.368    0.773      0.880    0.823      0.536    0.837     0.862     NYC
Weighted Avg.  0.778    0.266    0.779      0.778    0.773      0.536    0.837     0.844

=== Confusion Matrix ===

     a      b   <-- classified as
 83747  48826 |      a = C
 22591 166072 |      b = NYC
```

E. J48 Tree

```
Time taken to build model: 2.75 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      250377               77.9418 %
Incorrectly Classified Instances     70859               22.0582 %
Kappa statistic                          0.5304
Mean absolute error                      0.3109
Root mean squared error                  0.3946
Relative absolute error                 64.1319 %
Root relative squared error             80.1449 %
Total Number of Instances           321236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.629    0.115    0.794      0.629    0.702      0.540    0.830     0.817     C
                 0.885    0.371    0.772      0.885    0.825      0.540    0.830     0.845     NYC
Weighted Avg.    0.779    0.265    0.781      0.779    0.774      0.540    0.830     0.833

=== Confusion Matrix ===

     a      b   <-- classified as
 83326  49247 |      a = C
 21612 167051 |      b = NYC
```